3 Major language families

The names in *italics* after the Arabic numerals indicate the language families, the names in square brackets are those of the major languages of each group.

EUROPE AND ASIA MINOR 1) *Indo-European* (see below). 2) *Finno-Ugric* [Finnish, Estonian, Hungarian and various smaller languages in Russia as well as beyond the Urals, e.g. the Samoyed languages]. 3) *Caucasian* [languages of the mountainous region between the Black Sea and the Caspian Sea, characterised by many highly differentiated languages in a small area, Geogrian (south Caucasian) is the best known of these]. 4) *Altaic* (Turk languages) [Turkish and various other languages, such as Azeri, Uzbek, Turkmen, etc., stretching eastward of Turkey as far as the border with China, includes Mongolian and Tungusic languages]. 5) *Independent* The only surviving independent language in Europe is Basque which has not been proven to be genetically related to any surrounding language. From history we have other examples of language isolates, e.g. Etruscan in ancient Italy.

NORTH-EAST ASIA (SIBERIA AND ALASKA) 1) *Paleo-Asiatic* [consists of a few small languages spread over a vast area of eastern Siberia]. 2) *Eskimo-Aleut* [few speakers spread over a large area stretching from Siberia through Alaska and Canada to Greenland].

NORTH AFRICA 1) *Afro-Asiatic* (Hamito-Semitic) [Branches into Semitic, which includes Arabic proper, Hebrew, Ethiopic and Aramaic, and Berber (in the Atlas mountains), along with Cushitic, Egyptian (Coptic) and Chadic (Hausa); it is the language family with the oldest linguistic records].

SUBSAHARAN AFRICA 1) *Niger-Congo* [a very large group, grouping into Western Sudanic, with the branches Mande, West Atlantic, Gur and Kwa, and Benue-Congo of which the main branch is Bantu with over 500 languages stretching down to South Africa, includes Xhosa, Zulu and Kiswahili]. 2) *Nilo-Saharan* [a diverse group stretching across the Sahara to Sudan].

SOUTH AFRICA 1) *Khoisan* [Bushman, Hottentot and other minor indigenous languages of the South African peninsula, noted for the presence of clicks].

SOUTH ASIA (Indian subcontinent, Pakistan) 1) *Dravidian* [Telugu, Tamil, Kannada; the second most important family in India]. 2) *Munda* [consists of a number of languages spoken on the east coast of India]. The remaining languages are Indo-European.
SOUTH-EAST, EAST ASIA 1) Sino-Tibetan [divides into at least three sub-groups: Sinitic, the chief representative of which are the dialects/languages of Chinese, Tibeto-Burman including Burmese and Tibetan, Tai which contains the two major languages Thai and Lao]. 2) Mon-Khmer [Khmer spoken in Cambodia]. 5) Independent. There are a number of independent languages in South and South-East Asia: Burushaski in Kashmir (northern India) is spoken by approximately 30,000. The language Ainu is spoken by even fewer speakers on various islands in northern Japan. Apart from these cases there are the three national languages Vietnamese, Japanese and Korean. A possible link exists between the latter two, but it is tenuous and contested.

AUSTRALIA AND OCEANIA 1) Austronesian [Indonesian, Polynesian; consists of hundreds of island languages spread throughout a large area in the West Pacific]. 2) Papuan [a group in a small area (that of the state of Papua, one half of the island state of Papua New Guinea); it contains very many different languages in a small area and is comparable in diversity with the Caucasus]. 3) Australian [the indigenous language family of Australia, consists of many languages spoken in all by not more than 50,000 aborigines].

THE AMERICAS Very many languages in many families are spoken by the native Americans of both continents. Among the major North American families (Canada, United States, Mexico) are: Algonkian, Wakashan, Salishan, Athapascan, Penutian, Yuman, Iroquoian, Siouan, Muskogean, Uto-Aztecan, Oto-Maguean, Zoquean, Mayan. The major families of Central-South America are: Macro-Chibchan, Ge-Pano-Carib and Andean Equatorial. Some of these languages, such as Quechua [Andean Equatorial], are spoken over a very large area (in Chile, Peru, Bolivia, Columbia and Ecuador) while Guaraní (in Paraguay) has official status alongside Spanish.

3.1 The Indo-European family

The proto-language of Indo-European probably originated in the area of present-day Ukraine / Southern Russia and was spoken up until about 3000 BC. From this area the speakers of this language spread out in various directions, eventually yielding separate dialects, the inputs to major subdivisions of the Indo-European family. Various features (most phonological) are used to distinguish these early divisions of the proto-language, such as the treatment of original /k/ which was either shifted to /s/ or preserved as /k/. Languages classified along this axis are said to be either centum languages (from the initial /k/ in the Latin word for 100) or satem (from the initial /s/ in the corresponding word in Avestan, an Indo-Iranian language).
3.2 Individual groups of Indo-European

INDO-IRANIAN This consists of the languages in and around Iran and of those groups who spread into north-west India and later throughout the whole country. Hindi and Urdu (the latter a close relative in Pakistan) are the main languages of the Indic branch whose classical form is Sanskrit.

ANATOLIAN An extinct group consisting in the main of Hittite, the language of the Hittite Empire (1700-1200 BC). Tablets containing remains of this language were discovered and identified in Turkey in the early twentieth century.

ARMENIAN A branch available from the ninth century AD in a Bible translation. It has continued as East Armenian (in the republic of Armenia) and West Armenian in Eastern Turkey.

TOKHARIAN Remnants of this language (an A and B version) were discovered by a German expedition at the beginning of the twentieth century in western China. It died out towards the end of the first millennium.

HELLENIC (GREEK) The set of dialects known as Classical Greek belong to this branch. Almost unbroken records are available covering over 2500 years. It continues as modern Greek.

ALBANIAN Despite its small numbers, Albanian represents a separate branch of the Indo-European family. First records are available from the 15th century.

ITALIC The term ‘Italic’ is used for those dialects of ancient Italy which include Latin but also Oscan and Umbrian (which strictly speaking form a separate branch). It continues as the set of Romance languages.

CELTIC Once spoken over a wide area in central Europe, the Celtic languages were pressed further west by rival Indo-European peoples which began to fill central and western Europe (Germanic tribes and Romans). It continues as the languages of the Celtic fringe of the British Isles and Breton in French Brittany.

GERMANIC This branch probably originated in southern Scandinavia and spread out from there to cover the area of present-day Germany, the regions to the south (Austria and Switzerland), the North Sea coast, England and the entire Scandinavian peninsula along with the Faroes and Iceland.
BALTIC A branch of its own with three representatives Lithuanian, Lettish and Old Prussian. The last language has been extinct since the 18th century. Present-day Lithuanian is particularly archaic and of special interest to Indo-Europeanists.

SLAVIC The oldest written form of Slavic is Old Church Slavonic. Nowadays there are three main branches: 1) Southern Slavic [Slovene, Croatian, Serbian, Macedonian, Bulgarian], 2) Western Slavic [Polish, Sorbian, Wendish, Czech, Slovak] and 3) Eastern Slavic [Russian, White Russian, Ukrainian].
4 Writing systems

An alphabet is a system for representing sounds in writing. It is based on the principle of sound-symbol equivalence, hence the letter *c* in Latin corresponded to the sound /k/. This principle may be disturbed by later developments, e.g. Latin /k/ later developed into /ts/ and then into /tʃ/ in Italian (before front vowels). One symbol can also stand for more than one sound, again Latin *c* remained /k/ before back vowels in Italian, similar to the principle in English: *call* with /k-/, but *cease* with /s-/.

But because of historical developments, not all symbols stand for the same sounds. The letter *a* stands for /a/ in European languages but for /ei/ in English due to a major vowel shift. The same is true of consonants, for example, *j* stands for /dʒ/ in English but for /j/ in Swedish.

The term *alphabet* comes from ‘alpha’ and ‘beta’, the first two letters of the Greek alphabet. The letters of an alphabet may have their own names as with Arabic or the Runic alphabet, an early Germanic system used in the first centuries AD, e.g. *þ* ‘thorn’, *ƿ* ‘wynn’ (= ‘joy’). Alphabet systems tend to be economical and can get by with about 30 symbols (26 letters in English, for instance). The forms of letters may vary with no effect on their sound values, e.g. letters may appear in italics or **bold** or UPPERCASE.

**Orthography**, from Greek *orthos* ‘straight’ and *graphō* ‘I write’, is the spelling system of a language.

A different principle is found in languages which use characters (such as Chinese). In these cases a symbol stands for an entire word or at least for a syllable. Such languages have a very large number of symbols — a few thousand — as in principle there is one per word, though by means of repetition and combination the number required can be reduced. In character systems the art of calligraphy, from *kallos* ‘beauty’ and *graphō* ‘I write’ in Greek, is often important, e.g. in China.

Alphabet systems are a development from older pictographic systems in which stylised abstractions were used in writing, e.g. a circle for the sun, a vertical stroke for a man, etc.

4.1 Features of alphabets

Not all alphabets have a one-to-one correspondence between sounds and letters. Because of this two or more letters may be used to indicate a single sound. In English *sh* indicates the voiceless sibilant [ʃ]. The reverse is less usually the case but can be seen in English with *x* which stands for [ks] and in German with *z* which word-initially stands for [ts]. There may also be variation in writing due
to position in a word. In English \([d\ddot{z}]\) is shown word-initially by \(j\), as in journey, but word-medially and word-finally as -dge, as in fudge. A single phoneme may also be indicated by different letters depending on such factors as its origin. The sound \([z]\) can be shown by \(s\) in inflections, stays, or by \(z\), as in zinc, or even by \(x\) as in xenophobia.

There may be ambiguity in writing due to diachronic developments. English \(w\) and \(wh\) both represent \([w]\) nowadays but which were distinguished formerly as \([w]\) versus \([\mathring{w}]\). An additional situation can be seen with English \(th\): here one has two letters which represent one of two sounds \([\theta]\) or \([\delta]\), both of which are distinctive (cf. the voiceless and voiced fricatives in think and that respectively). As Latin did not have this type of sound it had to be indicated by a combination of letters in English. The Latin practice of using two different letters for voiceless and voiced sounds was not upheld in English in this case.

The pronunciation of a letter does not necessarily correspond to its phonetic value. The letter \(u\) in English has the pronunciation \([ju:]\) but different phonetic values in words like pool \([pu:l]\), pull \([pu:l]\) and cut \([k\ddot{u}:t]\).

*The basis for modern alphabets* The languages of western Europe are based on forms of Latin. A variant of Greek, which arose in the tenth century in Bulgaria, supplied the input to Cyrillic, the orthography of Russian and other eastern and some southern Slavic languages (Ukrainian, White Russian; Bulgarian, Serbian). It has also been used in modified form to represent the languages of the former Soviet Union, many of which are non-Indo-European such as the Turk languages. Religion has not infrequently been the reason for a country employing the Latin alphabet as with the Slavic languages Polish, Czech, Slovak and Croatian, all effected by western Christianity as opposed to the Greek orthodox church which influenced the eastern Slavs.

It is true of most alphabets that phonetic detail is not represented in writing. Thus the different types of \(t\&h\) found in English, \([\ddot{t}r], [\ddot{t}], [\ddot{r}]\), etc. are not reflected in writing. Many languages go further by not representing sounds which are phonemic and predictable because of morphological information. This is the case with languages which do not show vowels, or not always. Semitic languages (Arabic, Hebrew), which are root-inflecting, usually have predictable vowel values depending on adjacent consonants so that strictly speaking it is not necessary to indicate them in writing (as was not done in Classical Hebrew).

*Diacritics and punctuation* It is common for diacritics (additions to letters) to be used for sounds which are not present in the original alphabet on which they are based. In European languages diacritics are used liberally, cf. the accents of French, \(\acute{e}, \grave{a}, \dot{e}\); the Umlaute of German, \(\ddot{u}, \ddot{o}, \ddot{a}\); the \(\ddot{a}\) of Swedish, the palatalised nasal of Spanish \(\ddot{n}\), etc. Other European languages with diacritics are: Polish, Czech, Hungarian, Rumanian, Croatian.

Punctuation has nothing to do with the representation of the segmental
phonology of a language in writing. It is very vaguely connected with intonation inasmuch as the comma is intended to indicate a pause (but does not always do so), the question mark the rise in intonation at the end of a question, etc. Other items of punctuation are supposed to be syntactic in nature. The full stop, the colon and the semi-colon are used to delimit syntactic groups. The use of uppercase initials (capitals) is also syntactic (i.e. they only occur at the beginning of a sentence in English). German is unique in using uppercase initials for all nouns (Danish abandoned this practice after the Second World War).

Spaces, italics and direction Spaces between words are morphologically motivated (the elements between spaces are those which can move freely around) and do not represent pauses in speech. Certain languages, such as Ancient Greek, do not use spaces at all, a sentence is then written as a continuous stream of letters.

Some languages, such as Arabic, are written in an Italic (slanted) or cursive form. European languages have a cursive form for handwriting but a upright, non-bound form for printing. A language like Arabic is printed with linked, curved letters. It is typical of such languages that they have different forms of letters depending on position. Thus nearly every Arabic letter has four different forms: initial, medial, final and independent (some of these may be similar or the same). Other languages, such as Sanskrit, do not have true cursive alphabets but have ligatures, i.e. single forms which contain two letters, Old English had a ligature æ which is a combination of a and e. German has a ligature in β which, while it looks like Greek beta, is in fact a ligature form of s and z as is clear from its name ‘sz’ [ɛstʃt].

There are also languages which do not have a distinction between lower and uppercase letters, e.g. Thai.

Writing can be done from left to right or from right to left. All European languages are written from left to right. Semitic languages (Arabic, Hebrew) are written from right to left as is Persian (an Indo-European language). Sanskrit was originally written from right to left, later from left to right as are all modern languages descended from it.

Stress It is not normal for a language to indicate stress in writing. It is not necessary as languages are acquired orally by natives so that even with relatively unpredictable stress systems, like that of Russian, the speaker will usually know where stress lies in a word. An exception to the practice of not marking stress is Czech which uses a system of right-slanting strokes over vowels to indicate that they are stressed. Some of the Romance languages indicate stress when it is different from the general pattern. In Italian the penultimate syllable of a word is stressed. However, should the final syllable end in a stressed vowel then this is indicated by an acute accent, e.g. umilità ‘humility’.
4.2 The development of alphabet systems

The very earliest records of speech are available in pictographic form. A pictography is a system of picture writing, e.g. the character for ‘man’ is a picture of a man, etc. The next step is ideography where the characters are stylised representations of what they denote, e.g. the character for ‘man’ is no longer a picture of a man but an abstract sketch of a man. What is important for the later development of alphabets is that the ideographs be conventionalised, i.e. if a man is represented by a vertical stroke with a horizontal bar, then this representation is then kept to and the convention is established.

The oldest records of alphabet systems come from the area of the Eastern Mediterranean (Canaan, the Palestine, Phoenicia, Mesopotamia). It is called North Semitic as the languages spoken in this area form part of the northern branch of the Semitic family of languages. The type of writing used in Mesopotamia is known as *cuneiform* (from Latin *cuneus* ‘wedge’) because of its appearance as wedge-like strokes. Since the end of the nineteenth century *cuneiform* has been deciphered and it is clear that it represents a number of languages chief of which is Sumerian, an unknown and unaffiliated language which disappeared towards 2,000 BC, and Akkadian (divided into Babylonian and Assyrian) which is Semitic. Akkadian served as the model for Hittite *cuneiform* (Hittite is an Indo-European language once spoken in central Turkey).

Sumerian is assumed to have been the input for Egyptian writing. This is known under the term for the characters used, namely *hieroglyphs* (from Greek *hieros* ‘sacred’ and *gluphe* ‘carving’) which are a form of Egyptian pictographs used at a very early stage of alphabet development. This type of inscription is referred to as a *petroglyph* if it is carved on stone (typically in a cave).

Another very early form stems from the area of ancient Elam and is known as Proto-Elamite. It is a pictography which dates back to about 3,000 BC and may well be connected to the various forms of Proto-Indic found in fragments in the Indus valley. This would make it an ancestor, albeit at several removes, of Sanskrit and its daughter languages. There is also an Aramaic (Semitic) element in Indic, this element being present in other languages of the eastern part of the Middle East, such as Avestan (Indo-European).

Traces of an alphabet are available from another quarter in the second millennium BC. Pre-Hellenic Greece was marked by the Minoan culture centred on Crete which, as legend would have it, ended abruptly with the volcanic explosion on the island of Santorini (Thera) around 1500 BC. Excavations at Knossos on Crete by Sir Arthur Evans led to the finds of ‘Linear A’ (representing an unidentified non-European language) and those of Michael Ventris led to the decipherment of ‘Linear B’ (a very early form of Greek which predates Homer (eighth or ninth century BC) by about 500 years.

For the later development of alphabets, the *cuneiform* syllabaries which existed in an area from the Sinai peninsula to northern Syria are important.
These developed from an early form of Proto-Sinaitic around 1600 BC to a linear form with the earliest biblical inscriptions around 1,000 BC from Byblos, an ancient city in Phoenicia (the coastal area of present-day Syria and the Lebanon). Phoenician was used on the mainland and in Cyprus from where a number of linear inscriptions are available.

The term ‘linear’ implies that the inscriptions are so strongly abstracted from the earlier pictographic and ideographic representation that one can speak of an alphabet in the modern sense. The linear forms of Phoenician look quite like ancient Greek.

The main Indo-European branch leading down from Phoenician is Greek. In Greek a gradual evolution in the form of letters took place, with these supplying a model for the early Latin alphabet.