The World's Major Languages

Second Edition

Edited by
Bernard Comrie

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

Introduction

Bernard Comrie

1 Preliminary Notions

How many languages are there in the world? What language(s) do they speak in India? What languages have the most speakers? What languages were spoken in Australia, or in Mexico before European immigration? When did Latin stop being spoken, and when did French start being spoken? How did English become such an important world language? These and other similar questions are often asked by the interested layman. One aim of this volume – taking the Introduction and the individual chapters together – is to provide answers to these and related questions, or in certain cases to show why the questions cannot be answered as they stand. The chapters concentrate on an individual language or group of languages, and in this Introduction I want rather to present a linking essay which will provide a background against which the individual chapters can be appreciated.

After discussing some preliminary notions in Section 1, Section 2 of the Introduction provides a rapid survey of the languages spoken in the world today, concentrating on those not treated in the subsequent chapters, so that the reader can gain an overall impression of the extent of linguistic diversity that characterises the world in which we live. Since the notion of 'major language' is primarily a social notion – languages become major (such as English) or stop being major (such as Sumerian) not because of their grammatical structure, but because of social factors – Section 3 discusses some important sociolinguistic notions, in particular concerning the social interaction of languages.

1.1 How Many Languages?

Linguists are typically very hesitant to answer the first question posed above, namely: how many languages are spoken in the world today? Probably the best that one can say, with some hope of not being contradicted, is that at a very conservative estimate some 6,000 languages are spoken today. Laymen are often surprised that the figure should be so high, but I would emphasise that this is a conservative estimate. But why is it that linguists are not able to give a more accurate figure? There are several different reasons conspiring to prevent them from doing so, and these will be outlined below.

Even a couple of decades ago one could reasonably have observed that some parts of the world were simply insufficiently studied from a linguistic viewpoint, so that we did not know precisely what languages are spoken there. There are very few parts of the world where this still holds, since our knowledge of the linguistic situation in remote parts of the world has improved dramatically in recent years – New Guinea, for instance, has changed from being almost a blank linguistic map in the first half of the twentieth century to the stage where nearly all languages can be pinpointed with accuracy: since perhaps as many as one-fifth of the world's languages are spoken in New Guinea, this has radically changed any estimate of the total number of languages. Although new languages are still being discovered, this is no longer the major factor it would have been in the past.

A second and more important problem is that it is difficult or impossible in many cases to decide whether two related speech varieties should be considered different languages or merely different dialects of the same language. Native speakers of English are often surprised that there should be problems in delimiting languages from dialects, since present-day dialects of English are in general mutually intelligible (at least with some familiarisation), and even the language most closely related genetically to English, Frisian, is mutually unintelligible with English. With the languages of Europe more generally, there are in general established traditions of whether two speech varieties should be considered different languages or merely dialect variants, but these decisions have often been made more on political and social rather than on strictly linguistic grounds.

One criterion that is often advanced as a purely linguistic criterion is mutual intelligibility: if two speech varieties are mutually intelligible, they are different dialects of the same language, but if they are mutually unintelligible, they are different languages. But if applied to the languages of Europe, this criterion would radically alter our assessment of what the different languages of Europe are: the most northern dialects and the most southern dialects (in the traditional sense) of German are mutually unintelligible, while dialects of German spoken close to the Dutch border are mutually intelligible with dialects of Dutch spoken just across the border. In fact, our criterion for whether a dialect is Dutch or German relates in large measure to social factors – is the dialect spoken in an area where Dutch is the standard language or where German is the standard language? By the same criterion, the three nuclear Scandinavian languages (in the traditional sense), Danish, Norwegian and Swedish, would turn out to be dialects of one language, given their mutual intelligibility. While this criterion is often applied to non-European languages (so that nowadays linguists talk of the Chinese languages rather than the Chinese dialects, given the mutual unintelligibility of, for instance, Mandarin and Cantonese), it seems unfair that it should not be applied consistently to European languages as well.

In some cases, the intelligibility criterion actually leads to contradictory results, namely when we have a dialect chain, i.e. a string of dialects such that adjacent dialects are readily mutually intelligible, but dialects from the far ends of the chain are not mutually intelligible. A good illustration of this is the Dutch–German dialect complex. One could start from the far south of the German-speaking area and move to the far west of the Dutch-speaking area without encountering any sharp boundary across which mutual intelligibility is broken; but the two end points of this chain are speech varieties so different from one another that there is no mutual intelligibility possible. If one takes a simplified dialect chain A – B – C, where A and B are mutually intelligible,

as are B and C, but A and C are mutually unintelligible, then one arrives at the contradictory result that A and B are dialects of the same language, B and C are dialects of the same language, but A and C are different languages. There is in fact no way of resolving this contradiction if we maintain the traditional strict difference between language and dialects, and what such examples show is that this is not an all-or-nothing distinction, but rather a continuum. In this sense, it is not just difficult, but in principle impossible to answer the question how many languages are spoken in the world.

A further problem with the mutual intelligibility criterion is that mutual intelligibility itself is a matter of degree rather than a clear-cut opposition between intelligibility and unintelligibility. If mutual intelligibility were to mean 100 per cent mutual intelligibility of all utterances, then perhaps no two speech varieties would be classified as mere dialect variants; for instance, although speakers of British and American English can understand most of one another's speech, there are areas where intelligibility is likely to be minimum unless one speaker happens to have learned the linguistic forms used by the other, as with car (or auto) terms like British *boot*, *bonnet*, *mudguard* and their American equivalents *trunk*, *hood*, *fender*. Conversely, although speakers of different Slavonic languages are often unable to make full sense of a text in another Slavonic language, they can usually make good sense of parts of the text, because of the high percentage of shared vocabulary and forms.

Two further factors enter into the degree of mutual intelligibility between two speech varieties. One is that intelligibility can rise rapidly with increased familiarisation: when American films were first introduced into Britain they were initially considered difficult to understand, but increased exposure to American English has virtually removed this problem. Speakers of different dialects of Arabic often experience difficulty in understanding each other at first meeting, but soon adjust to the major differences between their respective dialects, and Egyptian Arabic, as the most widely diffused modern Arabic dialect, has rapidly gained in intelligibility throughout the Arab world. This can lead to 'one-way intelligibility', as when speakers of, say, Tunisian Arabic are more likely to understand Egyptian Arabic than vice versa, because Tunisian Arabic speakers are more often exposed to Egyptian Arabic than vice versa. The second factor is that intelligibility is to a certain extent a social and psychological phenomenon: it is easier to understand when you want to understand. A good example of this is the conflicting assessments different speakers of the same Slavonic language will often give about the intelligibility of some other Slavonic language, correlating in large measure with whether or not they feel well disposed to speakers of the other language.

The same problems as exist in delimiting dialects from languages arise, incidentally, on the historical plane too, where the question arises: at what point has a language changed sufficiently to be considered a different language? Again, traditional answers are often contradictory: Latin is considered a dead language, although its descendants, the Romance languages, live on, so at some time Latin must have changed sufficiently to be deemed no longer the same language, but a qualitatively different one. On the other hand, Greek is referred to as 'Greek' throughout its attested history (which is longer than that of Latin and the Romance languages combined), with merely the addition of different adjectives to identify different stages of its development (e.g. Ancient Greek, Byzantine Greek, Modern Greek). In the case of the history of the English language, there is even conflicting terminology: the oldest attested stages of English can be referred to either as Old English (which suggests an earlier stage of Modern English) or as Anglo-Saxon (which suggests a different language that is the

ancestor of English, perhaps justifiably so given the mutual unintelligibility of Old and Modern English).

A further reason why it is difficult to assess the number of languages spoken in the world today is that many languages are on the verge of extinction. While it has been the case throughout human history that languages have died out, recent social changes have considerably accelerated this process, as the languages of smaller speech communities are endangered by those with more speakers and more prestige. Economic factors often make it difficult for modern services (health, education, etc.) to be provided in the languages of smaller speech communities, and members of smaller speech communities interested in integrating into wider social networks often feel impelled to abandon their own language in favour of a language of wider currency or at least to encourage their children to do so. This is not just the ever increasing use of English as a result of globalisation. At a more local level, smaller languages are endangered in particular situations by French, Spanish, Indonesian, Swahili – even Tsamai (the last-mentioned a language of Ethiopia with around 10,000 speakers, to which the few remaining speakers of Birale are assimilating). Documentation and, where possible, preservation of endangered languages is one of the major tasks facing linguists in the twenty-first century.

1.2 Language Families and Genetic Classification

One of the basic organisational principles of this volume, both in Section 2 of the Introduction and in the arrangement of the individual chapters, is the classification of languages into language families. It is therefore important that some insight should be provided into what it means to say that two languages belong to the same language family (or equivalently: are genetically related).

It is probably intuitively clear to anyone who knows a few languages that some languages are closer to one another than are others. For instance, English and German are closer to one another than either is to Russian, while Russian and Polish are closer to one another than either is to English. This notion of similarity can be made more precise, as is done for instance in the chapter on the Indo-European languages, but for the moment the relatively informal notion will suffice. Starting in the late eighteenth century, a specific hypothesis was proposed to account for such similarities, a hypothesis which still forms the foundation of research into the history and relatedness of languages. This hypothesis is that where languages share some set of features in common, these features are to be attributed to their common ancestor. Let us take some examples from English and German.

In English and German we find a number of basic vocabulary items that have the same or almost the same form, e.g. English *man* and German *Mann*. Likewise, we find a number of bound morphemes (prefixes and suffixes) that have the same or almost the same form, such as the genitive suffix, as in English *man's* and German *Mann(e)s*. Although English and German are now clearly different languages, we may hypothesise that at an earlier period in history they had a common ancestor, in which the word for 'man' was something like *man* and the genitive suffix was something like *-s*. Thus English and German belong to the same language family, which is the same as saying that they share a common ancestor. We can readily add other languages to this family, since a word like *man* and a genitive suffix like *-s* are also found in Dutch, Frisian and the Scandinavian languages. The family to which these languages belong has been given

that the proto-language is not an attested language – although if written records had gone back far enough, we might well have had attestations of this language – but its postulation is the most plausible hypothesis explaining the remarkable similarities among the various Germanic languages.

Although not so obvious, similarities can be found among the Germanic languages and a number of other languages spoken in Europe and spreading across northern India as far as Bangladesh. These other languages share fewer similarities with the Germanic languages than individual Germanic languages do with one another, so that they are more remotely related. The overall language family to which all these languages belong is the Indo-European family, with its reconstructed ancestor language Proto-Indo-European. As is discussed in more detail in the chapter on Indo-European languages, the Indo-European family contains a number of branches (i.e. smaller language families, or subfamilies), such as Slavonic (including Russian and Polish), Iranian (including Persian and Pashto), and Celtic (including Irish and Welsh). The overall structure is therefore hierarchical: The most distant ancestor is Proto-Indo-European. At an intermediate point in the family tree, and therefore at a later period of history, we have such languages as Proto-Germanic and Proto-Celtic, which are descendants of Proto-Indo-European but ancestors of languages spoken today. Still later in history, we find the individual languages as they are spoken today or attested in recent history, such as English and German as descendants of Proto-Germanic and Irish and Welsh as descendants of Proto-Celtic. One typical property of language change that is represented accurately by this family-tree model is that, as time goes by, languages descending from a common ancestor tend to become less and less similar. For instance, Old English and Old High German (the ancestor of Modern German) were much closer to one another than are the modern languages – they may even have been mutually intelligible, at least to a large extent.

Although the family-tree model of language relatedness is an important foundation of all current work in historical and comparative linguistics, it is not without its problems, both in practice and in principle. Some of these will now be discussed.

We noted above that with the passage of time, genetically related languages will grow less and less similar. This follows from the fact that, once two languages have split off as separate languages from a common ancestor, each will innovate its own changes, different from changes that take place in the other language, so that the cumulative effect will be increasing divergence. With the passage of enough time, the divergence may come to be so great that it is no longer possible to tell, other than by directly examining the history, that the two languages do in fact come from a common ancestor. The best-established language families, such as Indo-European or Sino-Tibetan, are those where the passage of time has not been long enough to erase the obvious traces of genetic relatedness. (For language families that have a long written tradition, one can of course make use of earlier stages of the language, which provide more evidence of genetic relatedness.) In addition, there are many hypothesised language families for which the evidence is not sufficient to convince all, or even the majority, of scholars. For instance, the Turkic language family is a well-established language family, as is each of the Mongolic and Tungusic families. What is controversial, however, is whether or not these individual families are related as members of an even larger family. The possibility of an Altaic family, comprising Turkic, Mongolic, and Tungusic, is rather widely accepted, and some scholars would advocate increasing the size of this family by adding Korean and perhaps Japanese.

The attitudes of different linguists to problems of this kind have been characterised as an opposition between 'splitters' (who require the firmest evidence before they are prepared to acknowledge genetic relatedness) and 'clumpers' (who are ready to assign languages to the same family on the basis of quite restricted similarities). I should, incidentally, declare my own splitter bias, lest any of my own views that creep in be interpreted as generally accepted dogma. The most extreme clumper position would, of course, be to maintain that all languages of the world are genetically related, although there are less radical positions that would posit such 'macro-families' as Eurasiatic or Nostratic (including, inter alia, Indo-European, Uralic and Altaic), Dene-Caucasian (including, inter alia, Na-Dene, Sino-Tibetan, East Caucasian and West Caucasian), and Austric (including at least Austronesian and Austro-Asiatic). In the survey of the distribution of languages of the world in Section 2, I have basically retained my own splitter position, although for areas of great linguistic diversity and great controversy surrounding genetic relations (such as New Guinea and the Americas) I have simply refrained from detailed discussion.

While no linguist would doubt that some similarities among languages are due to genetic relatedness, there are several other possibilities for the explanation of any particular similarity, and before assuming genetic relatedness one must be able to exclude, at least with some degree of plausibility, these other possibilities. Unfortunately, in a great many cases it is not possible to reach a firm and convincing decision. Let us now examine some of the explanations other than genetic relatedness.

First, two languages may happen purely by chance to share some feature in common. For instance, the word for *dog* in Mbabaram, an Australian Aboriginal language, happens to be *dog*. This Mbabaram word is not, incidentally, a borrowing from English, but is the regular development in Mbabaram of an ancestral form something like **gudaga*, which is found in forms similar to this reconstruction in other related languages (it is usual to prefix reconstructed forms with an asterisk). If anyone were tempted to assume on this basis, however, that English and Mbabaram are genetically related, examination of the rest of Mbabaram vocabulary and grammar would soon quash the genetic relatedness hypothesis, since there is otherwise minimal similarity between the two languages. In comparing English and German, by contrast, there are many similarities at all levels of linguistic analysis. Even sticking to vocabulary, the correspondence *man*: *Mann* can be matched by *wife*: *Weib*, *father*: *Vater*, *mother*: *Mutter*, *son*: *Sohn*, *daughter*: *Tochter*, etc. Given that other languages have radically different words for these concepts (e.g. Japanese *titi* 'father', *haha* 'mother', *musuko* 'son', *musume* 'daughter'), it can clearly not be merely the result of chance that English and German have so many similar items. But if the number of similar items in two languages is small, it may be difficult or impossible to distinguish between chance similarity and distant genetic relatedness.

Certain features shared by two languages might turn out to be manifestations of language universals, i.e. of features that are common to all languages or are inherently likely to occur in any language. Most discussions of language universals require a fair amount of theoretical linguistic background, but for present purposes I will take a simple, if not particular profound, example. In many languages across the world, the syllable *ma* or its reduplicated form *mama* or some other similar form is the word for 'mother'. The initial syllable *ma* enters into the Proto-Indo-European word for 'mother' that has given English *mother*, Spanish *madre*, Russian *mat´*, Sanskrit *mātā*. In Mandarin Chinese, the equivalent word is *mā*. Once again, examination of other features of Indo-European

languages and Chinese would soon dispel any possibility of assigning Chinese to the Indo-European language family. Presumably the frequency across languages of the syllable *ma* in the word for 'mother' simply reflects the fact that this is typically one of the first syllables that babies articulate clearly, and is therefore interpreted by adults as the word for mother. (In the South Caucasian language Georgian, incidentally, *mama* means 'father' – and 'mother' is *deda* – so that there are other ways of interpreting baby's first utterance.)

Somewhat similar to universals are patterns whereby certain linguistic features frequently co-occur in the same language, i.e. where the presence of one feature seems to require or at least to foster the presence of some other feature. For instance, the study of word universals by Greenberg (1966) showed that if a language has verb-final word order (i.e. if 'the man saw the woman' is expressed literally as 'the man the woman saw'), then it is highly probable that it will also have postpositions rather than prepositions (i.e. 'in the house' will be expressed as 'the house in') and that it will have genitives before the noun (i.e. the pattern 'cat's house' rather than 'house of cat'). Thus, if we find two languages that happen to share the features: verb-final word order, postpositions, prenominal genitives, then the co-occurrence of these features is not evidence for genetic relatedness. Many earlier attempts at establishing wide-ranging genetic relationships suffer precisely from failure to take this property of typological patterns into account. Thus the fact that Turkic languages, Mongolic languages, Tungusic languages, Korean and Japanese share all of these features is not evidence for their genetic relatedness (although there may, of course, be other similarities, not connected with recurrent typological patterns, that do establish genetic relatedness). If one were to accept just these features as evidence for an Altaic language family, then the family would have to be extended to include a variety of other languages with the same word order properties, such as the Dravidian languages of southern India and Quechua, spoken in South America.

Finally, two languages might share some feature in common because one of them has borrowed it from the other (or because they have both borrowed it from some third language). English, for instance, borrowed a huge number of words from French during the Middle Ages, to such an extent that an uncritical examination of English vocabulary might well lead to the conclusion that English is a Romance language, rather than a Germanic language. The term 'borrow', as used here, is the accepted linguistic term, although the terminology is rather strange, since 'borrow' suggests a relatively superficial acquisition, one which is moreover temporary. Linguistic borrowings may run quite deep, and there is of course no implication that they will ever be repaid. Among English loans from French, for instance, there are many basic vocabulary items, such as *very* (replacing the native Germanic *sore*, as in the biblical *sore afraid*). Examples from other languages show even more deep-seated loans: the Semitic language Amharic, for instance, has lost the typical Semitic word order patterns, in which the verb precedes its object and adjectives and genitives follow their noun, in favour of the order where the verb follows its object and adjectives and genitives precede their noun; Amharic is in close contact with Cushitic languages, and Cushitic languages typically have the order object–verb, adjective/genitive–noun, so that Amharic has in fact borrowed these word orders from neighbouring Cushitic languages.

It seems that whenever two languages come into close contact, they will borrow features from one another. In some cases the contact can be so intense among the languages in a given area that they come to share a significant number of common features, setting

this area off from adjacent languages, even languages that may happen to be more closely related genetically to languages within the area. The languages in an area of this kind are often said to belong to a sprachbund (German for 'language league'), and perhaps the most famous example of a sprachbund is the Balkan sprachbund, whose members (Modern Greek, Albanian, Bulgarian, Macedonian, Rumanian) share a number of striking features not shared by closely related languages like Ancient Greek, other Slavonic languages (Bulgarian is Slavonic), or other Romance languages (Rumanian is Romance). The most striking of these features is loss of the infinitive, so that instead of 'give me to drink' one says 'give me that I-drink' (Modern Greek *ðos mu na pjo*, Albanian *a-më të pi*, Bulgarian *daj mi da pija*, Rumanian *dă-mi să beau*; in all four languages the subject of the subordinate clause is encoded in the ending of the verb).

Since we happen to know a lot about the history of the Balkan languages, linguists were not deceived by these similarities into assigning a closer genetic relatedness to the Balkan languages than in fact holds (all are ultimately members of the Indo-European family, though from different branches). In other parts of the world, however, there is the danger of mistaking areal phenomena for evidence of genetic relatedness. In South-East Asia, for instance, many languages share very similar phonological and morphological patterns: in Chinese, Thai and Vietnamese words are typically monosyllabic, there is effectively no morphology (i.e. words do not change after the manner of English *dog*, *dogs* or *love*, *loves*, *loved*), syllable structure is very simple (only a few single consonants are permitted word-finally, while syllable-initially consonant clusters are either disallowed or highly restricted), and there is phonemic tone (thus Mandarin Chinese *mā*, with a high level tone, means 'mother', while *mǎ* with a falling–rising tone, means 'horse'), and moreover there are a number of shared lexical items. For these reasons, it was for a long time believed that Thai and Vietnamese were related genetically to Chinese. More recently, however, it has been established that these similarities are not the result of common ancestry, and Thai and Vietnamese are now generally acknowledged not to be genetically related to Chinese. The similarities are the results of areal contact. The shared vocabulary items are primarily the result of intensive Chinese cultural influence, especially on Vietnamese. The tones and simple syllable structures can often be shown to be the result of relatively recent developments, and indeed in one language that is genetically related to Chinese, namely Classical Tibetan, one finds complex consonant clusters but no phonemic tone, i.e. the similarities noted above are neither necessary nor sufficient conditions for genetic relatedness.

In practice, the most difficult task in establishing genetic relatedness is to distinguish between genuine cognates (i.e. forms going back to a common ancestor) and those that are the result of borrowing. It would therefore be helpful if one could distinguish between those features of a language that are borrowable and those that are not. Unfortunately, it seems that there is no feature that can absolutely be excluded from borrowing. Basic vocabulary can be borrowed, so that for instance Japanese has borrowed the whole set of numerals from Chinese, and even English borrowed its current set of third person plural pronouns (*they*, *them*, *their*) from Scandinavian. Bound morphemes can be borrowed: a good example is the agent suffix *-er* in English, with close cognates in the other Germanic languages; this is ultimately a loan from the Latin agentive suffix *-arius*, which has however become so entrenched in English that it is a productive morphological device applicable in principle to any verb to derive a corresponding agentive noun.

At one period in the recent history of comparative linguistics, it was believed that a certain basic vocabulary list could be isolated, constant across languages and cultures,

such that the words on this list would be replaced at a constant rate. Thus, if one assumes that the retention rate is around 86 per cent per millennium, this means that if a single language splits into two descendant languages, then after 1,000 years each language would retain about 86 per cent of the words in the list from the ancestor language, i.e. the two descendants would then share just over 70 per cent of the words in the list with each other. In some parts of the world, groupings based on this 'glottochronological' method still form the basis of the only available detailed and comprehensive attempt at establishing genetic relations. It must be emphasised that the number of clear counter-examples to the glottochronological method, i.e. instances where independent evidence contradicts the predictions of this approach, is so great that no reliance can be placed on its results.

It is, however, true that there are significant differences in the ease with which different features of a language can be borrowed. The thing that seems most easily borrowable is cultural vocabulary, and indeed it is quite normal for a community borrowing some concept (or artefact) from another community to borrow the foreign name along with the object. Another set of features that seem rather easily borrowable are general typological features, such as word order: in addition to the Amharic example cited above, one might note the fact that many Austronesian languages spoken in New Guinea have adopted the word order where the object is placed before the verb, whereas almost all other Austronesian languages place the object after the verb; this happened under the influence of Papuan languages, almost all of which are verb-final. Basic vocabulary and bound morphology are hardest to borrow. But even though it is difficult to borrow bound morphology, it is not impossible, so in arguments over genetic relatedness one cannot exclude a priori the possibility that even affixes may have been borrowed.

2 Distribution of the World's Languages

In this section, I wish to give a general survey of the distribution of the languages of the world, in terms of their genetic affiliation. I will therefore be talking primarily about the distribution of language families, although reference will be made to individual languages where appropriate. The discussion will concentrate on languages and language families not covered in individual chapters, and at appropriate places I have digressed to give a brief discussion of some interesting structural or sociological point in the language being treated.

2.1 Europe

Europe, taken here in the traditional cultural sense rather than in the current geographical sense of 'the land mass west of the Urals', is the almost exclusive preserve of the Indo-European family. This family covers not only almost the whole of Europe, but also extends through Armenia (in the Caucasus), Iran and Afghanistan into Central Asia (Tajikistan), with the easternmost outpost of this strand the Iranian language Sarikoli, spoken just inside China. Another strand spreads from Afghanistan across Pakistan, northern India and southern Nepal, to end with Bengali in eastern India and Bangladesh; an off-shoot from northern India, Sinhalese, is spoken in Sri Lanka, and the language of the Maldives is the closely related Maldivian.

In addition, the great population shifts that resulted from the voyages of exploration starting at the end of the fifteenth century have carried Indo-European languages to

many distant lands. The dominant languages of the Americas are now Indo-European (English, Spanish, Portuguese, French), as is the dominant language of Australia and New Zealand (English). While in some countries these languages are spoken by populations descended primarily from European settlers, there are also instances where a variety of the European language is spoken by a population of a different origin, perhaps the best-known example being the creolised forms of European languages (especially English, French and Portuguese) spoken by the descendants of African slaves in the Caribbean. It should be noted that these population shifts have not led exclusively to the spread of European languages, since many languages of India, both Indo-European and Dravidian, have also extended as a by-product, being spoken now by communities in the Caribbean area, in East Africa, and in the South Pacific (especially Fiji).

Of the few European languages not belonging to the Indo-European family, mention may first be made of Basque, a language isolate, with no established genetic relations to any other language. It is spoken in the Pyrenees near the French–Spanish border. Basque is perhaps most noted for its ergative construction, whereby instead of having a single case (nominative) for both subjects of intransitive verbs and subjects (agents) of transitive verbs, with a different case (accusative) for objects (patients) of transitive verbs, Basque uses one case (absolutive) for both intransitive subjects and objects of transitive verbs, and a different case (ergative) for subjects of transitive verbs, as in the following sentences:

Jon etorri da.	'John came.'
Neska-k gizona ikusi du.	'The girl saw the man.'

In the first sentence, *Jon* is intransitive subject, and stands in the absolutive (no inflection); in the second sentence, *neska-k* 'girl' is transitive subject, and therefore stands in the ergative (suffix *-k*), while *gizona* 'man' is transitive object, and therefore stands in the absolutive, with no inflection.

Some other languages of Europe belong to the Uralic family. These include Hungarian, Finnish, Estonian, and Saami (formerly also called Lappish – properly a group of mutually unintelligible languages rather than a single language), to which can be added a number of smaller languages closely related to Finnish and Estonian. Other members of the Uralic family are spoken on the Volga and in northern Eurasia on both sides of the Urals, stretching as far as southern Siberia.

Turkish as spoken in the Balkans represents the Turkic family in Europe, but this is primarily an Asian family, and will be treated in the next section. The same is true of Afroasiatic, represented in Europe by Maltese.

2.2 Asia

Having just mentioned Turkish, we may now turn to the Turkic family, which is spoken in Turkey, parts of the Caucasus, some areas on the Volga, most of Central Asia (and stretching down into north-western Iran), and large parts of southern Siberia, with one off-shoot, Yakut, in north-eastern Siberia. Turkic is perhaps to be joined in a single language family (Altaic) with the Mongolic and Tungusic families. The Mongolic languages are spoken predominantly in Mongolia and northern China, though there are also isolated Mongolic languages in Afghanistan (Moghol) and just to the north of the Caucasus mountains (Kalmyk); the main member of the family is the language Mongolian

(sometimes called Khalkha, after its principal dialect), which is the official language of Mongolia. The Tungusic languages are spoken by numerically small population groups in Siberia, spreading over into Mongolia and especially north-eastern China. The Tungusic language best known to history is Manchu, the native language of the Qing dynasty that ruled China from 1644 to 1911; the Manchu language is, however, now almost extinct, having been replaced by Chinese. Whether Korean or Japanese can be assigned to the Altaic family is a question of current debate.

This is a convenient point at which to discuss a number of other languages spoken in northern Asia. All are the languages of small communities (a few hundred or a few thousand). They are sometimes referred to collectively as Paleosiberian (or Paleoasiatic), although this is not a genetic grouping. Three of them are language isolates: Ket, spoken on the Yenisey river, and the sole survivor of the small Yeniseic family; Yukaghir, spoken on the Kolyma river; and Nivkh (Gilyak), spoken at the mouth of the Amur river and on Sakhalin island. The small Chukotko-Kamchatkan family comprises the indigenous languages of the Chukotka and Kamchatka peninsulas: Chukchi, Koryak, Kamchadal (Itelmen); it has been suggested that they may be related to Eskimo-Aleut, which we treat in Section 2.5 on the Americas. Finally, we may mention here the recently extinct Ainu, apparently a language isolate, whose last native speakers lived in Hokkaido, the most northerly Japanese island.

One of the geographic links between Europe and Asia, the Caucasus, has since antiquity been noted for the large number of clearly distinct languages spoken there; indeed it was referred to by the Arabs as the 'mountain of tongues'. Some of the languages spoken in the Caucasus belong to other families (e.g. Armenian and Ossetian to Indo-European, Azerbaijani to Turkic), but there are in addition a number of languages with no known affiliations to languages outside the Caucasus: these are the Caucasian languages. Even the internal genetic relations of the Caucasian languages are the subject of debate. Few scholars now accept the genetic relatedness of all Caucasian languages, but there is ongoing debate over whether West Caucasian and East Caucasian together form a single North Caucasian family. The Kartvelian or South Caucasian family includes Georgian, the Caucasian language with the largest number of speakers (over four million) and the only Caucasian language to have a long-standing literary tradition (dating back to the fifth century). The West (North-West) Caucasian languages are on and close to the Black Sea coast, though also in Turkey as a result of emigration since the mid-nineteenth century; one Caucasian language, Ubykh, which died out in Turkey towards the end of the twentieth century, is noteworthy for the large number of its consonant phonemes – at one time it was considered the world record-holder. The East (North-East) Caucasian (or Nakh-Daghestanian) languages are spoken mainly in Daghestan, Chechnya and Ingushetia in the Russian Federation; the best-known language is Chechen with about a million speakers, though the family also includes languages like Hinuq, spoken by about 500 people in a single village.

Turning now to south-western Asia, we may consider the Afroasiatic family, which, as its name suggests, is spoken in both Asia and Africa. In Asia its main focus is the Arab countries of the Middle East, although Hebrew and Aramaic are also Afroasiatic languages of Asia, belonging to the Semitic branch of Afroasiatic. In addition Arabic is, of course, the dominant language of North Africa, where Afroasiatic is represented not only by a number of other Semitic languages (those of Ethiopia, the major one being Amharic), but also by Berber, the Cushitic languages of the Horn of Africa (including Somali, the official language of Somalia), and the Chadic languages of

northern Nigeria and adjacent areas (including Hausa). One branch of Afroasiatic formerly spoken in Africa, Egyptian (by which is meant the language of ancient Egypt, not the dialect of Arabic currently spoken in Egypt), is now extinct.

In South Asia (the traditional 'Indian subcontinent'), four language families meet. Indo-European languages, more specifically languages of the Indo-Aryan branch of Indo-European, dominate in the north, while the south is the domain of the Dravidian languages (although some Dravidian languages are spoken further north, in particular Brahui, spoken in Pakistan). The northern fringe of the subcontinent is occupied by Sino-Tibetan languages, to which we return below. The fourth family is Austro-Asiatic. The languages in this family with most speakers are actually spoken in South-East Asia: Vietnamese in Vietnam and Khmer (Cambodian) in Cambodia, and they are the only languages of the family to have the status of national languages. Languages of the family are scattered from central India eastwards into Vietnam, Western Malaysia and the Nicobar Islands. In India itself, the Austro-Asiatic language with most speakers is Santali. It is only relatively recently that the assignment of Vietnamese to this family has gained widespread acceptance. In addition, there is one language isolate, Burushaski, spoken in northern Pakistan, while the genetic affiliations of the languages of the Andaman islands remain unclear.

We have already introduced a number of South-East Asian languages, and may now turn to the other two families represented in this area: Tai-Kadai (also called Kadai and Kam-Tai) and Sino-Tibetan. While the Tai-Kadai group of languages, which includes Thai (Siamese) and Lao, was earlier often considered a branch of Sino-Tibetan, this view has now been abandoned; Tai-Kadai languages are spoken in Thailand, Laos, southern China, and also in parts of Burma (Myanmar) and Vietnam. Sino-Tibetan contains the language with the largest number of native speakers in the world today, Chinese (and this remains true even if one divides Chinese into several different languages, in which case Mandarin occupies first position). The other Sino-Tibetan languages traditionally form the Tibeto-Burman branch, which includes Tibetan and Burmese, in addition to a vast number of languages spoken predominantly in southern China, Burma (Myanmar), northern India and Nepal. Finally, the languages of the Hmong-Mien (Miao-Yao) family are spoken in southern China and adjacent areas.

In East Asia we find Korean and Japanese (the latter together with the closely related Ryukyuan varieties), whose genetic affiliations to each other or to other languages (such as Altaic) remain the subject of at times heated debate.

The Austronesian family, though including some languages spoken on the Asian mainland, such as Malay of Western Malaysia and Cham spoken in Cambodia and Vietnam, is predominantly a language of the islands stretching eastwards from the South-East Asian mainland: even Malay–Indonesian has more speakers in insular South-East Asia than on the Malay peninsula. Austronesian languages are dominant on most of the islands from Sumatra in the west to Easter Island in the east, including the Philippines, but excluding New Guinea (where Austronesian languages are, however, spoken in many coastal areas); Malagasy, the language of Madagascar, is a western outlier of the family; Austronesian languages are also indigenous to Taiwan, though now very much in the minority relative to Chinese.

2.3 New Guinea and Australia

The island of New Guinea, which can be taken linguistically together with some of the smaller surrounding islands, is the most differentiated area linguistically in the whole

world. Papua New Guinea, which occupies the eastern half of the island, contains some 800 languages for a total population of about five and a half million, meaning that the average language has just under 7,000 speakers. In many of the coastal areas of New Guinea, Austronesian languages are spoken, but the other languages are radically different from these Austronesian languages. These other languages are referred to collectively as either 'non-Austronesian languages of New Guinea' or as 'Papuan languages', though it should be realised that this is a negatively characterised term, rather than a claim about genetic relatedness. Though much work remains to be done, there has been considerable recent progress in classifying the Papuan languages genetically; in particular, there is growing evidence for a Trans-New Guinea family containing a large number of languages running east–west across the middle of the main island and on to some of the smaller islands to the west.

One syntactic property that is widespread among the Highland Papuan languages is worthy of note, namely switch reference. In a language with a canonical switch reference system, a sentence may (and typically does) consist of several clauses, of which only one is an independent clause (i.e. could occur on its own as a free-standing sentence), all the others being dependent; each dependent clause is marked according to whether or not its subject is the same as or different from the subject of the clause on which it is dependent. The examples below are from Usan:

Ye nam su-ab, isomei. 'I cut the tree and went down.'
Ye nam su-ine, isorei. 'I cut the tree and it fell down.'

The independent verbs, *isomei* and *isorei*, are respectively first person singular and third person singular. The dependent verbs, *su-ab* and *su-ine*, have respectively the suffix for same subject and the suffix for different subject. In the first example, therefore, the subjects of the two clauses are the same (i.e. I cut the tree and I went/fell down), while in the second sentence they are different (i.e. I cut the tree and some other entity – from the context only the tree is available – went/fell down). The words *ye* and *nam* mean respectively 'I' and 'tree'. One effect of switch reference is that the speaker of a language with switch reference must plan a discourse ahead to a much greater extent than is required by languages lacking switch reference, since in switch reference languages it is nearly always the case that the dependent clause precedes the independent clause, i.e. in clause *n* one has to mark the co-reference relation that holds between the subject of clause *n* and the subject of clause $n + 1$. This should, incidentally, serve to dispel any lingering notions concerning the primitiveness or lack of grammar in the languages of other societies. Although switch reference is found in many other parts of the world (e.g. in many indigenous languages of the Americas), it is particularly characteristic of the languages of the New Guinea Highlands.

Although the genetic classification of the indigenous languages of Australia, which numbered over 250 at the time of contact with Europeans, is the subject of at times acrimonious debate, there is a general consensus that a large Pama-Nyungan family can be identified, comprising most languages spoken in the south and centre and some in the north, while the other languages of the north form a number of small families and language isolates.

The Australian languages overall are characterised by an unusual consonant system, from the viewpoint of the kinds of consonant systems that are found most frequently across the languages of the world. Most Australian languages have no fricatives, and no

voice opposition among their stops. However, they distinguish a large number of places of articulation, especially in terms of lingual articulations: thus most languages have, in addition to labial and velar stops, all of palatal, alveolar, and retroflex stops, while many languages add a further series of phonemically distinct dentals. The same number of distinctions is usually found with the nasals, and some languages extend this number of contrasts in the lingual stops to the laterals as well. One result of this is that Europeans usually fail to perceive (or produce, should they try to do so) phonemic oppositions that are crucial in Aboriginal languages, while conversely speakers of Australian languages fail to perceive or produce phonemic oppositions that are crucial in English (such as the distinction among *pit*, *bit*, *bid*).

One Australian language, Dyirbal, spoken in the Cairns rainforest in northern Queensland, has played an important role in recent discussions of general linguistic typology, and it will be useful to make a short digression to look at the relevant unique, or at least unusual, features of Dyirbal – though it should be emphasised that these features are not particularly typical of Australian languages overall.

In English, one of the pieces of evidence for saying that intransitive and transitive subjects are just subtypes of the overall notion 'subject' is that they behave alike with respect to a number of different syntactic processes. For instance, a rule of English syntax allows one to omit the subject of the second conjunct of a coordinate sentence if it is co-referential with the subject of the first conjunct, i.e. one can abbreviate the first sentence below to the second one:

I hit you and I came here.
I hit you and came here.

It is not possible to carry out a similar abbreviation of the next sentence below, since its subjects are not co-referential, even though the object of the first conjunct is co-referential with the subject of the second conjunct:

I hit you and you came here.

In the above examples, the first clause is transitive and the second clause intransitive, but the notion of subject applies equally to both clauses. If we think not so much of grammatical labels like subject and object, but rather of semantic labels like agent and patient, then we can say that in English it is the agent of a transitive clause that behaves as subject. In the corresponding Dyirbal sentences, however, it is the patient that behaves as subject, as can be seen in the following sentences:

Ngaja nginuna balgan, ngaja baninyu.	'I hit you and I came here.'
Ngaja nginuna balgan, nginda baninyu.	'I hit you and you came here.'
Ngaja nginuna balgan, baninyu.	'I hit you and you came here.'

In these sentences, *ngaja* is the nominative form for 'I', while *nginuna* is the accusative form for 'you'; the verbs are *balgan* 'hit' (transitive) and *baninyu* 'come here' (intransitive). In the third sentence, where the intransitive subject is omitted, it must be interpreted as co-referential with the patient, not the agent, of the first clause. In Section 2.1 we mentioned ergativity in connection with Basque case marking. These Dyirbal examples show that Dyirbal has ergativity in its syntactic system: patients of transitive verbs,

rather than agents of transitive verbs, are treated as subjects, i.e. are treated in the same way as intransitive subjects. Note that in this sense Dyirbal grammar is certainly different from English grammar, but it is no less well defined.

Another unusual feature of Dyirbal is sociolinguistic. In many, if not all languages there are different choices of lexical item depending on differences in social situation, such as the difference between English *father* and *dad(dy)*. What is unusual about Dyirbal is that a difference of this kind exists for every single lexical item in the language. Under certain circumstances, in particular in the presence of a taboo relative (e.g. a parent-in-law), every lexical item of ordinary language (Guwal) must be replaced by the corresponding lexical item from avoidance style (Jalnguy). No doubt in part for functional reasons, to ease the memory load, it is usual for several semantically related words of Guwal to correspond to a single Jalnguy word, as when the various Guwal names for different species of lizard are all subsumed by the one Jalnguy word *jijan*.

The surviving textual materials in the Tasmanian languages, extinct since the end of the nineteenth century, are insufficient in scope or reliability to allow any accurate assessment of the genetic affiliations of these languages – certainly none is immediately apparent.

2.4 Africa

Africa north of the Sahara is the preserve of Afroasiatic languages, which have already been treated in Section 2.2. This section will therefore concentrate on the sub-Saharan languages, though excluding languages introduced into Africa by external colonisation (though one such language, Afrikaans, a descendant of colonial Dutch, is a language of Africa by virtue of its geographic distribution), and also Malagasy, the Austronesian language of Madagascar. It is useful to take as a starting point the classification of the sub-Saharan African languages into three groups as proposed by J.H. Greenberg in the mid-1960s – Niger-Congo, Nilo-Saharan and Khoisan – while keeping in mind that some of these groupings remain controversial, either in general or in particular details.

The Niger-Congo (Niger-Kordofanian) family covers most of sub-Saharan Africa, and includes not only major languages of West Africa such as Yoruba but also, as a low-level node on the family tree, the Bantu languages, dominant in most of East, Central and southern Africa. The assignment of some groups of languages to Niger-Congo, such as the Kordofanian languages spoken in the Kordofan mountains of Sudan and the Mande languages of western West Africa, remains controversial.

More controversial is the proposed Nilo-Saharan family, which would include languages spoken in a number of geographically discontinuous areas of northern sub-Saharan Africa including parts of southern Sudan running through northern Uganda and western Kenya to northern Tanzania, northern Chad and neighbouring areas and the bend of the Niger river in West Africa. While the languages of the first group form a well-defined language family, Nilotic, at the opposite extreme inclusion of Songhay on the bend of the Niger river is widely rejected.

Finally, the three main groups within Khoisan as proposed by Greenberg, namely Northern, Central (Khoe) and Southern, all spoken in southern, mainly south-western Africa, plus two geographically isolated languages of Tanzania, Hadza and Sandawe, are individually well-defined language families (or isolates). However, grouping them together as a single larger family is generally considered at least premature. Typologically, Khoisan languages are most noted for having click sounds as part of their regular

phoneme inventory, a sound type that has also been borrowed into some neighbouring Bantu languages such as Zulu and Xhosa.

2.5 The Americas

The genetic classification of the indigenous languages of the Americas is overall the most contentious, with proposals ranging from a single family covering nearly all these languages, associated especially with the name of J.H. Greenberg, to around 200 distinct families and isolates. Even widely cited intermediate proposals, such as the Hokan and Penutian families, remain controversial. In what follows, I have concentrated on some of the more widespread established families and on some of the other languages with relatively large numbers of speakers.

Two population groups of North America are distinct ethnically from the remainder, namely the Eskimos (Inuit, although this latter term properly only refers to part of the Eskimos overall) and Aleuts. The Eskimo-Aleut family contains two branches, Aleut and Eskimo. Eskimo is properly a number of different languages rather than a single language, and is spoken from the eastern tip of Siberia through Alaska and northern Canada to Greenland; in Greenland it is, under the name Greenlandic, an official language.

Another language family centred in Alaska is the Athapaskan family (more properly: Athapaskan-Eyak, with inclusion of the Athapaskan or Athabaskan languages and the single language Eyak as the two branches of the family). Most of the Athapaskan languages are spoken in Alaska and north-western Canada, though the Athapaskan language with most speakers, Navajo, is spoken in Arizona and adjacent areas. Navajo is the indigenous language of North America (Canada and the USA) with the largest number of speakers, about 150,000. Athapaskan-Aleut is related to Tlingit, together forming a grouping often referred to as Na-Dene, although this term is also used to include Haida, which may well rather be a language isolate.

Among the other major families of North America are Iroquoian (around Lakes Ontario and Erie), Siouan (the Great Plains), and Algonquian (much of the north-eastern USA and eastern and central Canada, though also extending into the Great Plains with Arapaho and Cheyenne). One interesting feature of the Algonquian languages to which it is worth devoting a short digression is obviation. In Algonquian languages, a distinction is made between two kinds of third person, namely proximate and obviative, so that where English just has one set of third person pronouns (e.g. *he*, *she*, *it*, *they*) and morphology (e.g. the third person singular present tense ending *-s*), Algonquian languages distinguish two sets. In a given text span (which must be at least a clause, but may be longer), one of the third person noun phrases is selected as proximate (the one which is in some sense the most salient at that part of the text), all other third person participants are obviative. In the remainder of the text span, the proximate participant is always referred to by proximate morphology, while other participants are referred to by obviative morphology. In this way, the ambiguity of an English sentence like *John saw Bill as he was leaving* (was it John that was leaving, or Bill?) is avoided. The following examples are from Cree:

Nāpēw atim-wa wāpam-ē-w, ē-sipwēhtē-t.
'The man saw the dog as he (the man) left.'

Nāpēw atim-wa wāpam-ē-w, ē-sipwēhtē-ýit.
'The man saw the dog as it (the dog) was leaving.'

In both sentences, 'the man' is proximate (indicated by the absence of any affix on *nā pēw* 'man'), and 'the dog' is obviative (indicated by the suffix *-wa* on *atim-wa* 'dog'). The morphology of the verb *wāpam-ē-w* 'he sees him' indicates that the agent is proximate and the patient obviative (this is important, since the word order can be varied). The prefix *ē-* on the second verb indicates that it is subordinate ('conjunct', in Algonquianist terminology). In the first sentence, the suffix *-t* on this second verb indicates a proximate subject, i.e. the subject must be the proximate participant of the preceding clause, namely *the man*. In the second sentence, the suffix *-ýit* indicates an obviative subject, i.e. the subject of this verb must be an obviative participant of the preceding clause, in this sentence the only candidate being *the dog*.

Another important family, Uto-Aztecan, includes languages spoken in both North America (the South-West) and Central America. Its Aztecan branch includes Nahuatl, whose varieties have in total over a million speakers. The ancestor of the modern dialects, Classical Nahuatl, was the language of the Aztec civilisation which flourished in Central Mexico before the arrival of the Spanish. Spoken to the south of Nahuatl entirely within Central America, the Mayan family has an equally glorious past, because of its association with the ancient Mayan civilisation. Mayan languages are spoken in southern Mexico and Guatemala, with some overspill into neighbouring Central American countries; the Mayan language with the largest number of speakers is Yucatec, with about 700,000, although several others have speaker numbers in the hundreds of thousands.

The major families of South America include Carib, Arawakan and Tupi. These language families do not occupy geographically continuous areas: Carib languages are spoken to the north of the Amazon, and predominate in the eastern part of this region; Arawakan languages, once also spoken in the West Indies, dominate further west and are also found well south of the Amazon; while Tupi languages are spoken over much of Brazil south of the Amazon and in Paraguay. One Tupi language, (Paraguayan) Guarani, with about five million speakers, is a co-official language of Paraguay and is unique among indigenous languages of the Americas in that most of its speakers are non-Indians. Hixkaryana, a Carib language spoken by about 600 people on the Nhamundá river, a tributary of the Amazon, has become famous in the linguistic literature as the first clear attestation of a language in which the word order is object–verb–subject, as in the following sentence:

Toto yahosɨye kamara. 'The jaguar grabbed the man.'

In Hixkaryana, *toto* means 'man', *kamara* means 'jaguar', while the verb *yahosɨye* has the lexical meaning 'grab' and specifies that both subject and object are third person singular. Since there is no case marking on the nouns, and since the verb morphology is compatible with either noun as subject or object, the word order is crucial to understanding of this Hixkaryana sentence (which cannot mean 'the man grabbed the jaguar'), just as the different subject–verb–object word order is crucial in English.

Quechua – properly a family of often mutually unintelligible languages rather than a single language – has about ten million speakers, primarily in Peru and Bolivia, though with offshoots north into Ecuador and Colombia and south into Chile and Argentina. It is of uncertain genetic affiliation, though often claimed to be related to the neighbouring Aymara language. Quechua was the language of the Inca civilisation, centred on Cuzco in what is now Peru.

3 The Social Interaction of Languages

As was indicated in the Preface, the notion of 'major language' is defined in social terms, so it is now time to look somewhat more consistently at some notions relating to the social side of language, in particular the social interaction of languages. Whether a language is a major language or not has nothing to do with its structure or with its genetic affiliation, and the fact that so many of the world's major languages are Indo-European is a mere accident of history.

First, we may look in more detail at the criteria that serve to define a language as being major. One of the most obvious criteria is the number of speakers, and certainly in making my choice of languages to be given individual chapters in this volume number of speakers was one of my main criteria. However, number of speakers is equally clearly not the sole criterion.

An interesting comparison to make here is between Chinese (or even more specifically, Mandarin) and English. Mandarin has far more native speakers than English, yet still English is generally considered a more useful language in the world at large than is Mandarin, as seen in the much larger number of people studying English as a second language than studying Mandarin as a second language. One of the reasons for this is that English is an international language, understood by a large number of people in many different parts of the world; Mandarin, by contrast, is by and large confined to China, and even taking all Chinese dialects (or languages) together, the extension of Chinese goes little beyond China and overseas Chinese communities. English is not only the native language of sizable populations in different parts of the world (especially the British Isles, North America and Australia and New Zealand) but is also spoken as a second language in even more countries, as is discussed in more detail in the chapter on English. English happens also to be the language of some of the technologically most advanced countries (in particular of the USA), so that English is the basic medium for access to current technological developments. Thus factors other than mere number of speakers are relevant in determining the social importance of a language.

Indeed, some of the languages given individual chapters in this volume have relatively few native speakers. Some of them are important not so much by virtue of the number of native speakers but rather because of the extent to which they are used as a lingua franca, as a second language among people who do not share a common first language. Good examples here are Swahili and Indonesian. Swahili is the native language of a relatively small population, perhaps a couple of million, primarily on the coast of East Africa, but its use as a lingua franca has spread through much of East Africa (especially Kenya and Tanzania) and beyond, so that the language is used by a total of perhaps around 50 million people. The Indonesian variety of Malay–Indonesian is the native language of perhaps 23 million, but is used as a second language by about 140,000,000 in Indonesia. In many instances, in my choice of languages I have been guided by this factor rather than by raw statistics. Among the Philippine languages, for instance, Tagalog does not have the largest number of native speakers, but I selected it because it is both the national language of the Philippines and used as a lingua franca across much of the country. A number of Indo-Aryan languages would surely have qualified for inclusion in terms of number of speakers, but they have not been assigned individual chapters because in social terms the major languages of the northern part of South Asia are clearly Hindi–Urdu and Bengali.

Another important criterion is the cultural importance of a language, in terms of the age and influence of its cultural heritage. An example in point is provided by the Dravidian

languages. Tamil does not have the largest number of native speakers; it is, however, the oldest Dravidian literary language, and for this reason my choice rested with Tamil. I am aware that many of these decisions are in part subjective, and in part contentious. As I emphasise in the Preface, the thing furthest from my mind is to intend any slight to speakers of languages that are not considered major in the contents of this volume; much of our knowledge of Language as a general characteristic of the human species comes precisely from the study of smaller, often endangered languages.

Certain languages are major even despite the absence of native speakers, as with Latin and Sanskrit. Latin has provided a major contribution to all European languages, as can be seen most superficially in the extent to which words of Latin origin are used in European languages. But even those languages that have tried to avoid the appearance of Latinity by creating their own vocabulary have often fallen back on Latin models: German *Gewissen* 'conscience', for instance, contains the prefix *ge-*, meaning 'with', the stem *wiss-*, meaning 'know', and the suffix *-en* to form an abstract noun – an exact copy of the Latin *con-sci-entia*; borrowings that follow the structure rather than the form in this way are known as calques or loan translations. Sanskrit has played a similar role in relation to the languages of India, including Hindi. Hebrew is included not because of the number of its speakers – as noted in the chapter on Hebrew, this has never been large – but because of the contribution of Hebrew and its culture to European and Middle Eastern society.

A language can thus have influence beyond the areas where it is the native or second language. A good example to illustrate this is Arabic. Arabic loans form a large part of the vocabulary of many languages spoken by Islamic peoples, even of languages that are genetically only distantly related to Arabic (e.g. Hausa) or that are genetically totally unrelated (e.g. Turkish, Persian and Urdu). The influence of Arabic can also be seen in the adoption of the Arabic writing system by many Islamic peoples. Similarly, Chinese loan words form an important part of the vocabulary of some East Asian languages, in particular Vietnamese, Japanese and Korean; the use of written Chinese characters has also spread to Japan and Korea, and in earlier times also to Vietnam.

It is important to note also that the status of a language as a major language is far from immutable. Indeed, as we go back into history we find many significant changes. For instance, the possibility of characterising English as the major language of the world is an innovation of the twentieth century. One of the most important shifts in the distribution of major languages resulted from the expansion of European languages, especially English, Spanish, Portuguese, and to a lesser extent French as a result of the colonisation of the Americas: English, Spanish and Portuguese all now have far more native speakers in the New World than in Britain, Spain or Portugal. Indeed, in the Middle Ages one would hardly have imagined that English, confined to an island off the coast of Europe, would have become a major international language.

In medieval Europe, Latin was clearly the major language, since, despite the lack of native speakers, it was the lingua franca of those who needed to communicate across linguistic boundaries. Yet the rise of Latin to such pre-eminence – which includes the fact that Latin and its descendants have ousted virtually all other languages from south-western Europe – could hardly have been foreseen from its inauspicious beginnings confined to the area around Rome. Equally spectacular has been the spread of Arabic, in the wake of the spread of Islam, from being confined to the Arabian peninsula to being the dominant language of the Middle East and North Africa.

In addition to languages that have become major languages, there are equally languages that have lost this status. The earliest records from Mesopotamia, often considered the

cradle of civilisation, are in two languages: Sumerian and Akkadian (the latter the language of the Assyrian and Babylonian empires); Akkadian belongs to the Semitic branch of Afroasiatic, while Sumerian is as far as we can tell unrelated to any other known language. Even at the time of attested Sumerian inscriptions, the language was probably already approaching extinction, and continued to be used in deference to tradition (as with Latin in medieval Europe). The dominant language of the area was to become Akkadian, but in the intervening period this too has died out, leaving no direct descendants. Gone too is Ancient Egyptian, the language of the Pharaohs and whose earliest texts are roughly contemporaneous with those of Sumerian. The linguistic picture of the Mediterranean and Near East in the year nought was very different from that which we observe today.

Social factors and social attitudes can even bring about apparent reversals in the family-tree model of language relatedness. At the time of the earliest texts from Germany, two distinct Germanic languages are recognised: Old Saxon and Old High German. Old Saxon is the ancestor of the modern Low German (Plattdeutsch) dialects, while Old High German is the ancestor of the modern High German dialects and of the standard language. Because of social changes – such as the decline of the Hanseatic League, the economic mainstay of northern Germany – High German gained social ascendancy over Low German. Since the standard language, based on High German, is now recognised as the standard in both northern and southern Germany, both Low and High German dialects are now considered dialects of a single German language, and the social relations between a given Low German dialect and standard German are in practice no different from those between a High German dialect and standard German.

One of the most interesting developments to have arisen from language contact is the development of pidgin and creole languages. A pidgin language arises from a very practical situation: speakers of different languages need to communicate with one another to carry out some practical task, but do not speak any language in common and moreover do not have the opportunity to learn each other's language properly. What arises in such a situation is, initially, an unstable pidgin, or jargon, with highly variable structure – considerably simplified relative to the native languages of the people involved in its creation – and just enough vocabulary to permit practical tasks to be carried out reasonably successfully. The clearest examples of the development of such pidgins arose from European colonisation, in particular from the Atlantic slave trade and from indenturing labourers in the South Pacific. These pidgins take most of their vocabulary from the colonising language, although their structure is often very different from those of the colonising language.

At a later stage, the jargon may expand, particularly when its usefulness as a lingua franca is recognised among the speakers of non-European origin, leading to a stabilised pidgin, such as Tok Pisin, the major lingua franca of Papua New Guinea. This expansion is on several planes: the range of functions is expanded, since the pidgin is no longer restricted to uses of language essential to practical tasks; the vocabulary is expanded as a function of this greater range of functions, new words often being created internally to the pidgin rather than borrowed from some other language (as with Tok Pisin *maus gras* 'moustache', literally 'mouth grass'); the structure becomes stabilised, i.e. the language has a well-defined grammar.

Probably at any stage in this development, from inception to post-stabilisation, the pidgin can 'acquire native speakers', i.e. become the native language of part or all of the community. For instance, if native speakers of different languages marry and have

the pidgin as their only common language, then this will be the language of their household and will become the first language of their children. Once a pidgin has acquired native speakers, it is referred to as a creole. The native language of many inhabitants of the Caribbean islands is a creole, for instance the English-based creole of Jamaica, the French-based creole of Haiti, and the Spanish- and/or Portuguese-based creole Papiamento (Papiamentu) of the Netherlands Antilles and Aruba. At an even later stage, social improvements and education may bring the creole back into close contact with the European language that originally contributed much of its vocabulary. In this situation, the two languages may interact and the creole, or some of its varieties, may start approaching the standard language. This gives rise to the so-called post-creole continuum, in which one finds a continuous scale of varieties of speech from forms close to the original creole (basilect) through intermediate forms (mesolect) up to a slightly regionally coloured version of the standard language (acrolect). Jamaican English is a good example of a post-creole continuum.

Even with hindsight, as we saw above, it would have been difficult to predict the present-day distribution of major languages in the world. It is equally impossible to predict the future. In terms of number of native speakers, it is clear that a major shift is underway in favour of non-European languages: the rate of population increase is much higher outside Europe than in Europe, and while some European languages draw some benefit from this (such as Spanish and Portuguese in Latin America), the main beneficiaries are the indigenous languages of southern Asia and Africa. It might well be that a later version of this volume would include fewer of the European languages that are restricted to a single country, and devote more space to non-European languages. Another factor is the increase in the range of functions of many non-European languages: during the colonial period European languages (primarily English and French) were used for most official purposes and also for education in much of Asia and Africa, but the winning of independence has meant that many countries have turned more to their own languages, using these as official language and medium of education. The extent to which this will lead to increase in their status as major languages is difficult to predict – at present, access to the frontiers of scholarship and technology is still primarily through European languages, especially English; but one should not forget that the use of English, French and German as vehicles for science was gained only through a prolonged struggle against what then seemed the obvious language for such writing: Latin. (The process may go back indefinitely: Cicero was criticised for writing philosophical treatises in Latin by those who thought he should have used Greek.) But at least I hope to have shown the reader that the social interaction of languages is a dynamic process, one that is moreover exciting to follow.

Bibliography

The most comprehensive and up-to-date index of the world's languages, with genetic classification, is Gordon (2005); while some data are no doubt questionable, this is certainly the most reliable such index available. For a more splitter-oriented classification, see Dryer (2005); for a more clumper-oriented one, Ruhlen (1991). Two series dealing with particular language families are the Cambridge Language Surveys (http://www.cup.cam.ac.uk/series/sSeries.asp?code=CLS) and the Routledge Language Family Series (http://www.routledge.com/books/series/Routledge_Language_Family_Series); the former concentrates on general properties of the language group in question, while the latter, initially inspired

by the first edition of the present work, provides sketches of individual languages. Though outside these series, Heine and Nurse (2000) provides a good overview of languages of Africa. Among several recent publications on endangered languages are Abley (2003) and Harrison (2007); see also http://www.ethnologue.com/nearly_extinct.asp for a list of 'nearly extinct' languages, i.e. for which 'only a few elderly speakers are still living'.

Readers wanting to delve deeper into problems of genetic classification should consult a good introduction to historical and comparative linguistics, such as Campbell (2004). For discussions of language universals and typology, reference may be made to Comrie (1989) or Croft (2002). A good introduction to language contact is Thomason (2001).

References

Abley, Mark. 2003. *Spoken Here: Travels among Threatened Languages* (Heinemann, London; Houghton Mifflin, Boston).

Campbell, L. 2004. *Historical Linguistics*, 2nd edn (Edinburgh University Press, Edinburgh).

Comrie, B. 1989. *Language Universals and Linguistic Typology*, 2nd edn (Basil Blackwell, Oxford; University of Chicago Press, Chicago).

Croft, W. 2002. *Typology and Universals*, 2nd edn (Cambridge University Press, Cambridge).

Dryer, M.S. 2005. 'Genealogical Language List', in M. Haspelmath, M.S. Dryer, D. Gil and B. Comrie (eds) *The World Atlas of Language Structures* (Oxford University Press, Oxford), pp. 584–644.

Gordon, R.G., Jr (ed.) 2005. *Ethnologue: Languages of the World*, 15th edn (SIL International, Dallas). Online version http://www.ethnologue.com/

Greenberg, J.H. 1966. 'Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements', in J. H. Greenberg (ed.) *Universals of Language*, rev. edn (MIT Press, Cambridge, MA), pp. 73–112.

Harrison, K. David. 2007. *When Languages Die: The Extinction of the World's Languages and the Erosion of Human Knowledge* (Oxford University Press, Oxford).

Heine, B. and Nurse, D. (eds) 2000. *African Languages: An Introduction* (Cambridge University Press, Cambridge).

Ruhlen, M. 1991. *A Guide to the World's Languages, Vol. I: Classification*, rev. edn (Stanford University Press, Stanford).

Thomason, S.G. 2001. *Language Contact: An Introduction* (Edinburgh University Press: Edinburgh).

Sources

I owe the Mbabaram example to R.M.W. Dixon. The Basque examples are from R. Etxepare (2003), 'Valency and Argument Structure in the Basque Verb', p. 364, in J.I. Hualde and J. Ortiz de Urbina (eds) *A Grammar of Basque* (Mouton de Gruyter, Berlin), pp. 363–426; the system is somewhat more complex than indicated in my text. The Usan examples are to be found in Ger P. Reesink (1983), 'Switch Reference and Topicality Hierarchies', *Studies in Language*, vol. 7, pp. 215–46. The discussion of Dyirbal is based on R.M.W. Dixon (1972), *The Dyirbal Language of North Queensland* (Cambridge University Press, Cambridge), especially Section 5.2.2 and Chapter 8. The Cree examples are from H. C. Wolfart and J.F. Carroll (1981), *Meet Cree*, 2nd edn (University of Alberta Press, Edmonton), p. 26. The Hixkaryana example is taken from D.C. Derbyshire (1985), *Hixkaryana and Linguistic Typology* (Summer Institute of Linguistics, Dallas; University of Texas at Arlington, Arlington) p. 32.