2 Languages of the World

BERNARD COMRIE

1 Introduction

The aim of this chapter is to provide readers with an overview of current views on the distribution of the languages of the world and on the genetic relations among those languages. Needless to say, the mention of individual languages will be on a selective basis, with emphasis on those languages that are most widely spoken or that have played an important role in history, although some departure from this principle will necessarily be made for parts of the world, like the Americas, Australia, and New Guinea, where there are few languages with large numbers of speakers.

The best currently available detailed account of the distribution of the world's languages, with information on geographic location, number of speakers, and genetic affiliation, is Grimes (1996a), which is accompanied by Grimes (1996b) and Grimes and Grimes (1996). This work lists over 6,700 languages spoken in the world today or having recently become extinct. While this figure is towards the high end of estimates that would be given by linguists, it is nonetheless a reasonable estimate, based where possible on a linguists' definition of "language" (as opposed to "dialect") as a speech variety that is not mutually intelligible with other speech varieties. This definition brings with it a number of ancillary problems. For instance, testing mutual intelligibility is far from straightforward (Casad 1974). There are, moreover, complicated cases, like intelligibility that is greater in one direction than the other, i.e. speakers of A understand B better than speakers of B understand A, and dialect chains, i.e. a geographic chain of dialects A—B— . . . —N such that each dialect is mutually intelligible with its neighbor(s), but the extremes of the chain, A and N, are not mutually intelligible. Added to this is the fact that for many speech varieties serious tests of mutual intelligibility have simply not been carried out.

The question of the genetic affiliation among the languages of the world is one that is currently fraught with controversy, in particular between those

who adopt a cautious stance, accepting that languages are genetically related only in the face of overwhelming evidence, and those who are more willing to accept genetic relatedness among languages on the basis of less compelling evidence. In this survey, I have in general included only language families that are universally or almost universally recognized by linguists, and I have specifically added notes of caution where I use terms that cover larger potential genetic groupings of languages. At the same time, this survey does have a duty to inform the reader about more speculative hypotheses that have gained the support of some reasonable set of linguists. I have therefore included notes on possible more widespread groupings, largely following Ruhlen (1994: 15–34). The other book dealing with the classification of the world's languages, Ruhlen (1987), likewise adopts in general an approach that includes both traditionally accepted and currently debated genetic groupings, though with an equally undisguised bias towards the latter; but the critical approach to less widely accepted groupings does not extend to its treatment of languages of the Americas.

In the space available, I have limited myself to geographic distribution and genetic affiliation, although there are a number of other questions that might have been touched on, such as the ways in which languages influence one another by contact, and more generally the historical processes that have given rise to the present-day distribution of the world's languages. Recent literature dealing with this latter problem includes Nichols (1992), Dixon (1997), and Nettle (1999).

Many of the references for individual language families are to volumes in the following series: Cambridge Language Surveys (Cambridge University Press, ongoing), Routledge Language Family Descriptions (Routledge, discontinued and effectively replaced by the following), Curzon Language Family Descriptions (Curzon, ongoing).

2 Languages of Europe and Northern Asia

2.1 Indo-European languages

The Indo-European language family (Ramat and Ramat 1998) covers most of Europe and spreads, with some breaks, across Iran and Central Asia down into South Asia. As a result of colonial expansion, it is now also dominant in the Americas and in Australia and New Zealand. In Europe itself, only a few peripheral areas are occupied by non-Indo-European languages, in particular areas where Basque and some Uralic languages are spoken and parts of the Caucasus. The Indo-European family subdivides into a number of well established branches.

The Germanic languages (König and van der Auwera 1994) are the dominant languages of northwestern Europe, extending into central Europe. This is

the language family that includes English, and also Dutch, German, and the Scandinavian languages (including Danish, Norwegian, Swedish, and Icelandic); an offshoot of German with considerable admixture from Hebrew-Aramaic and Slavic is Yiddish, the traditional language of Ashkenazi Jews and a widely spoken language of eastern Europe before the Holocaust. The Scandinavian languages form North Germanic, while the other languages cited are West Germanic; a third subbranch of the family, East Germanic, is now extinct, the only substantially attested language being Gothic.

The Celtic languages (Ball 1993, MacAulay 1993) were once also dominant languages of western and central Europe, but with the expansion of Germanic and Romance languages in particular they have retreated to the western fringes of Europe, the living languages being Welsh in Wales, Irish on the west coast of Ireland, Breton in Brittany (France), and Scots Gaelic in northwestern Scotland.

The Romance languages (Harris and Vincent 1988, Posner 1996) occupy most of southwestern Europe, and are the descendants of Latin, the language of the Roman Empire. Strictly speaking, the branch of Indo-European is Italic, since it includes a number of languages other than Latin that died out by the early centuries of the Common Era as a result of Roman and Latin expansion, so that all living Italic languages are in fact Romance languages. The major living languages are French, Catalan, Spanish, Portuguese, Italian, and Romanian.

Turning to eastern Europe, the northernmost Indo-European branch is Baltic, now consisting of the two languages, Lithuanian and Latvian. The Baltic languages have a particularly close relation to the Slavic (Slavonic) languages (Comrie and Corbett 1993), now dominant in much of eastern and central Europe and including three subbranches. The East Slavic languages are Russian, Belarusian (Belorussian), and Ukrainian. The West Slavic languages include Polish, Czech, and Slovak. The South Slavic languages are Slovenian, Serbo-Croatian, Bulgarian, and Macedonian. As a result of ethnic differences, what linguists would, on grounds of mutual intelligibility, consider a single Serbo-Croatian language is now often divided into Serbian and Croatian, with Bosnian sometimes added as a third ethnic variety.

Two further branches of Indo-European, each consisting of a single language, are found in the Balkans. Albanian consists of two dialect groups, Gheg in the north and Tosk in the south, which might well be considered distinct languages on the basis of the mutual intelligibility test, although there is a standard language based on Tosk. Hellenic includes only Greek, although it is customary to give a different name to the branch, in part because it includes varieties of Greek over more than three millennia, from Mycenean through Classical Greek and Byzantine Greek to the modern language. Armenian, spoken primarily in Armenia though also in the Armenian diaspora originating in eastern Turkey, is another branch of Indo-European consisting of a single language, although the differences between Eastern Armenian (spoken mainly in Armenia) and Western Armenian (spoken originally mainly in Turkey) are considerable, and there are two written languages.

Finally, with respect to the living languages, the Indo-Iranian languages are spoken from the Caucasus to Bangladesh. Indo-Iranian divides into two subbranches, Iranian and Indo-Aryan (Indic), the latter occupying an almost continuous area covering most of Pakistan, northern India, Nepal, and Bangladesh. The most widely spoken Iranian languages are Persian (Iran), with national variants Tajik (in Tajikistan) and Dari (in Afghanistan), Kurdish (mainly in the border area of Turkey, Iran, and Iraq), Pashto (in Afghanistan and Pakistan), and Balochi (in Pakistan).

The Indo-Aryan subbranch of Indo-Iranian (Masica 1991) includes Sanskrit, the classical language of Indian civilization; Pali, the sacred language of Buddhism; and a large number of modern languages, of which the most widely spoken are Hindi and Urdu, essentially different national forms of the same language, in India and Pakistan respectively; Sindhi and Western Panjabi (Lahnda) in Pakistan; Nepali in Nepal; and Kashmiri, Eastern Panjabi, Gujarati, Rajasthani, Marathi, Bhojpuri, Maithili, Assamese, and Oriya in India; Bengali in India and Bangladesh; and Sinhala, geographically separated from the other Indo-Aryan languages in Sri Lanka. It should also be noted that the various Romani languages, spoken by Rom (Gypsies), belong to the Indo-Aryan group of languages.

In addition, two branches of Indo-European consist of extinct but well attested languages. The best known of the Anatolian languages, spoken in what is now Turkey, is Hittite, language of a major ancient empire (seventeenth–twelfth centuries BCE). Tocharian is a family of two closely related languages, attested in texts from the latter half of the first millennium CE in what is now the Xinjiang region in northwestern China.

2.2 *Uralic languages*

The Uralic language family (Abondolo 1998) must once have been spoken over a continuous part of northeastern Europe and northwestern Asia, but inroads by other languages, primarily Indo-European and Turkic, have isolated many of the Uralic branches and languages from one another geographically. The family falls into two clear subgroups, Finno-Ugric and Samoyedic. The Samoyedic languages, all with small numbers of speakers, are spoken along the northern fringe of Eurasia, roughly from the Kanin peninsula to the Taymyr peninsula.

Finno-Ugric divides in turn into a number of branches: Balto-Finnic (around the Baltic Sea), Saamic (Lappish) (northern Scandinavia to the Kola peninsula), Volgaic (on the Volga, although the unity of this branch is now questioned), Permic (northeastern European Russia), and Ugric (western Siberia and Hungary, though the unity of Ugric is also questioned). The most widely spoken languages are two Balto-Finnic languages, Finnish and Estonian, and one of the Ugric languages, Hungarian. It should be noted that the present location of Hungarian is the result of a long series of migrations, so that Hungarian is now far distant in location from its closest relatives within Finno-Ugric.

2.3 Altaic families

Altaic is a proposed genetic grouping that would include minimally the Turkic, Tungusic, and Mongolic families, perhaps also Korean and Japanese. Each of these components is a well established language family, and Altaic lies perhaps at the dividing line that separates proponents of wide-ranging genetic groupings of languages from those that remain skeptical. Here the various families and the languages they contain will be noted without any commitment to the unity of the overall grouping.

The Turkic languages (Johanson and Csató 1998) are spoken, with interruptions, in a broad belt stretching from the Balkans in the west through the Caucasus and Central Asia and into Siberia. Classification of the Turkic languages has always been problematic, in part because most of the languages are very close to one another linguistically, in part because population movements and even, in recent times, language politics have tended to overlay new distinctions on old ones. It is recognized that two languages form separate branches of the family: Chuvash, spoken in the Chuvash Republic (Russia) on the Volga, and Khalaj, spoken by a small and dwindling population in the Central Province of Iran. Johanson and Csató (1998: 82–3) propose four other branches, listed here with representative languages. Southwestern (Oghuz) Turkic includes Turkish (Turkey), Azeri (Azerbaijani) (Azerbaijan, northwestern Iran), and Turkmen (Turkmenistan, also Iran and Afghanistan). Northwestern (Kipchak) Turkic includes Kumyk and Karachay-Balkar (both spoken in the Caucasus), Tatar and Bashkir (both spoken on the Volga), Kazakh (Kazakhstan and northwestern China), and Kirghiz (Kyrgyzstan). Southeastern (Uyghur) Turkic includes Uzbek (Uzbekistan) and Uyghur (mainly in northwestern China). Finally, Northeastern (Siberian) Turkic includes Tuvan and Altai (Oyrot) in southern Siberia and Yakut (Sakha) in the huge Sakha Republic in Russia.

The Tungusic languages have few speakers, scattered across the sparsely populated areas of central and eastern Siberia, including Sakhalin Island, and adjacent parts of northeastern China and Mongolia. One Tungusic language, Manchu, is well known in history as the language of the Manchu conquerors who established the Qing dynasty in China (1644–1911), but all but a few ethnic Manchu now speak Mandarin.

The Mongolic languages are spoken primarily in Mongolia and adjacent parts of Russia and China, although there is also one Mongolic language in Afghanistan while Kalmyk is spoken in Kalmykia (Russia) on the lower Volga. The most widely spoken Mongolic language is Mongolian (Mongolia, northern China), although both Buriat (to the south and east of Lake Baikal) and Kalmyk are languages of constituent republics of the Russian Federation.

The other two potential members of the Altaic family are Korean and Japanese. Korean (Sohn 1999) is a single language. Japanese (Shibatani 1990) is strictly speaking a small family, including not only Japanese but also the Ryukyuan

languages, which are not mutually intelligible with Japanese or with each other; the family is sometimes called Japanese-Ryukyuan.

2.4 Chukotko-Kamchatkan languages

Chukotko-Kamchatkan is a small language family spoken on the Chukotka and Kamchatka peninsulas in the far northeast of Russia. All of the languages, which include Chukchi, are endangered.

2.5 Caucasian families

Some of the languages spoken in the Caucasus belong to language families already mentioned, in particular Indo-European (Armenian, Iranian) and Turkic. But there remain a large number of languages that do not belong to any of these families. These languages are referred to as Caucasian, but it is important to note that this is essentially a negative characterization. Indeed, it is currently believed that there are two or three families represented among the "Caucasian" languages.

The Kartvelian (South Caucasian) family is spoken in Georgia with some extension into Turkey, and the main language, the only one to be used as a written language, is Georgian, the official language of the Republic of Georgia.

The other two Caucasian families are Northwest Caucasian (West Caucasian, Abkhaz-Adyghe) and Northeast Caucasian (East Caucasian, Nakh-Daghestanian), although Nikolayev and Starostin (1994) present a detailed argument for considering them to constitute a single North Caucasian family; I will treat them separately here.

The Northwest Caucasian languages are spoken in Abkhazia, the northwestern part of the geographic territory of the Republic of Georgia, and in parts of Russia to the north of this. The main languages are Abkhaz (in Abkhazia) and the varieties of Circassian (Kabardian and Adyghe) spoken in Russia and by a sizeable diaspora in the Middle East.

The Northeast Caucasian languages are spoken primarily in the constituent republics of the Russian Federation of Chechnya, Ingushetia, and Daghestan, with some spillover into Azerbaijan. The languages with the largest numbers of speakers are Chechen (Chechnya) and Avar (Dagestan).

2.6 Other languages of Europe and northern Eurasia

A number of other languages or small language families are or were spoken in Europe or northern Asia but do not, at least unequivocally, belong to any of the above families. Basque is a language isolate spoken in the Pyrenees,

divided by the Spain–France border. Etruscan was the language of Etruria in northern Italy before the spread of Latin; it is now known to be related to two less well attested languages, Rhaetian in the Alps and Lemnian on the island of Lemnos (Limnos) in the Aegean. Hurrian (sixteenth century BCE) and Urartean (ninth to seventh centuries BCE) are two related extinct languages once spoken in eastern Anatolia.

The Yeniseian family of languages has only one survivor, Ket, spoken on the Yenisei River in western Siberia, although other languages are known from historical records that became extinct from the eighteenth to the twentieth centuries. Yukaghir, spoken in the area of the Kolyma and Indigirka rivers in northeastern Russia, is sometimes treated as a language isolate, although many linguists believe that it is distantly related to Uralic. Nivkh (Gilyak) is a language isolate spoken at the mouth of the Amur River and on Sakhalin Island. Ainu (Shibatani 1990) is a virtually extinct language isolate spoken in northern Japan (Hokkaido Island). Some or all of the languages mentioned in this paragraph are often referred to collectively as Paleosiberian or Paleoasiatic, but this is essentially a negative characterization (they do not belong to any of the established language families), with no implication that they are related to one another.

2.7 Proposals for larger groupings

Two similar, but not identical, proposals have been made for grouping together a large number of the language families found in Europe and northern Asia. The Nostratic proposal, first worked out in detail by Illič-Svityč (1971–84), would include at least Indo-European, Uralic, Altaic, Afroasiatic (see section 4.1), Kartvelian, and Dravidian (see section 3.1). Eurasiatic, the subject of ongoing work by Joseph H. Greenberg, would include at least Indo-European, Uralic, Altaic, Chukotko-Kamchatkan, Eskimo-Aleut (see section 5.1), and possibly also Nivkh. For possibilities including some of the other languages, see section 3.10.

3 Languages of Southern, Eastern, and Southeastern Asia and Oceania

This section deals primarily with languages of southeast Asia and its island extensions into Oceania. There is unfortunately no up-to-date general survey of southeast Asia, or indeed of the individual language families, although James Matisoff is working on one for Cambridge University Press. Things are somewhat better for the islands, although this is an area where there is rapid ongoing work leading to frequent changes in accepted genetic classification.

3.1 Dravidian languages

The Dravidian languages (Steever 1998) are the dominant languages of southern India, with Tamil also spoken in northern Sri Lanka. The Dravidian family is divided into four branches, Northern, Central, South-Central, and Southern, although the four main, literary languages belong to the last two branches. Telugu, the language of the Indian state of Andhra Pradesh, is a South-Central Dravidian language, while the following are South Dravidian: Tamil (Tamil Nadu state in India, northern Sri Lanka), Malayalam (Kerala state in India), and Kannada (Karnataka state in India).

3.2 Austro-Asiatic languages

Austro-Asiatic languages are spoken from eastern India across to Vietnam and down to the Nicobar Islands and peninsular Malaysia, although in most of this region they are interspersed among other, more widely spoken, languages. The family has two branches, Munda and Mon-Khmer. Munda languages are spoken in eastern India and some neighboring regions. Most of the languages have small numbers of speakers, the main exceptions being Santali and Mundari. Mon-Khmer languages start in eastern India, but their largest numbers are in Myanmar, Thailand, Malaysia, Cambodia, Laos, and Vietnam. While most have few speakers, there are two notable exceptions. Khmer (Cambodian) is the dominant language of Cambodia, while Vietnamese is the dominant language of Vietnam. Another historically important Mon-Khmer language is Mon, still spoken in the delta area to the east of Yankon (Rangoon), as the Mon played an important role in the development of Burmese and Thai culture. Vietnamese is typologically quite unlike the other Mon-Khmer languages and has undergone considerable influence from Chinese, with the result that its membership in Mon-Khmer was for a long time not recognized.

3.3 Sino-Tibetan

Sino-Tibetan is one of the world's largest language families in terms of numbers of speakers, and includes the language most widely spoken as a native language, namely Mandarin Chinese. Sino-Tibetan languages are spoken primarily in China, the Himalayan region of India and Nepal, and Myanmar, with excursions into some neighboring countries, in addition to a large Chinese diaspora. (Ethnic Chinese make up, for instance, some three-quarters of the population of Singapore.) Our understanding of Sino-Tibetan has been increased considerably in recent years by the availability of descriptions of the less widely spoken languages; a major impetus here has been the Sino-Tibetan Etymological Dictionary and Thesaurus project (see http://www.linguistics.berkeley.edu/lingdept/research/stedt/) at the University of California at Berkeley.

The usual classification splits the family into two branches, Sinitic (consisting essentially of the Chinese languages; Norman 1988) and Tibeto-Burman. Recently, van Driem (1997) has proposed, on the basis of the most recent reconstructions of the phonology of Old Chinese, that Sinitic may actually be a subbranch of Tibeto-Burman, grouped most closely with the Bodic languages – the family as a whole would thus more properly be called Tibeto-Burman. In what follows, I will retain the traditional classification, though emphasizing that this is more for convenience than through conviction.

Sinitic consists primarily of the various Chinese languages, which in terms of mutual intelligibility are clearly sufficiently different from one another to be considered distinct languages, even if all stand under the umbrella of a reasonably homogeneous written language. The major varieties are Mandarin, Wu (including Shanghai), Gan, Hakka, Xiang, Yue (Cantonese), Northern Min, and Southern Min (including Taiwanese).

The main groupings within (traditional) Tibeto-Burman are Baric, Bodic, Burmese-Lolo, Karen, Nung (Rung), and Qiang; proposals for subgrouping vary. The Baric languages include Meithei (Manipuri) in Manipur State, India. Bodic includes a number of languages spoken in the Himalayas, the most widely spoken and culturally important being Tibetan. The Burmese-Lolo languages are spoken mainly in Myanmar and southern China and include Burmese. The Karen languages are spoken in Myanmar and adjacent parts of Thailand, the most widely spoken being S'gaw Karen (White Karen). The Nung and Qiang languages are spoken in Myanmar and southern China.

3.4 Daic languages

Daic is one of a number of names (others including Tai-Kadai and Kam-Tai) for a family of languages with three branches, Kadai, Kam-Sui, and Tai. Kadai and Kam-Sui contain languages with small numbers of speakers spoken in southern China and parts of Vietnam. Tai, by contrast, includes two of the dominant languages of southeast Asia, namely Thai (Thailand) and the closely related Lao (Laos). Other Tai languages are spoken in these countries and in southern China, though with some excursions into Vietnam and Myanmar. The most widely spoken Tai language of China is Zhuang. It is now conventional to use the spelling Thai for the language, Tai for the branch, and Daic for the family.

3.5 Hmong-Mien (Miao-Yao) languages

The Hmong-Mien or Miao-Yao languages are spoken in parts of southern China and stretching into southeast Asia, especially Vietnam. Hmong and Mien are the indigenous ethnic names, while Miao and Yao are the Chinese equivalents. Hmong and Mien are the two branches of the family, and each consists of

several languages. The most widely spoken variety is Hmong Njua (Western Hmong) in China and Vietnam.

3.6 Austronesian languages

Austronesian is one of the most extensive families, covering almost all the islands bounded by an area from Madagascar in the west via Taiwan and Hawaii to Easter Island in the east and down to New Zealand in the south, with the exception of most of New Guinea and all of Australia. Although predominantly an island language family, Austronesian languages are also dominant in peninsular Malaysia, while the Chamic languages are spoken in coastal areas of Vietnam and Cambodia as well as on Hainan Island, China. An overview of the Austronesian languages by Robert Blust is in preparation for Cambridge University Press.

Although the Austronesian languages of Taiwan are very much minority languages on an island where varieties of Chinese have become dominant, the internal diversity among the Austronesian languages of Taiwan, the so-called Formosan languages, is greater than that in all the rest of Austronesian put together, so there is a major genetic split within Austronesian between Formosan and the rest, the latter now usually called Malayo-Polynesian (although in some earlier work this term was used for the family as a whole). Indeed, the genetic diversity within Formosan is so great that it may well consist of several primary branches of the overall Austronesian family.

The basic internal classification of Malayo-Polynesian is reasonably well established. The primary branchings are into Western Malayo-Polynesian and Central-Eastern Malayo-Polynesian, with the dividing line running to the east of Sulawesi and through the Lesser Sunda islands. Western Malayo-Polynesian thus includes all the languages of the Philippines, the Asian mainland, western Indonesia, and Madagascar. It also includes all the Austronesian languages with large numbers of speakers, including Malay–Indonesian, the different national varieties of what is essentially the same standard written language, though with radically different local spoken varieties. Other widely spoken languages of Indonesia are Acehnese, Toba Batak, Lampung, and Minangkabau (all on Sumatra), Javanese, Madurese, and Sundanese (all on Java), Balinese (on Bali), and Buginese and Makassarese (on Sulawesi). Widely spoken languages of the Philippines include, in addition to the national language Tagalog, the following: Bikol, Hiligaynon, Ilocano, Pampangan, Pangasinan, and Waray-Waray. The other major Western Malayo-Polynesian language is Malagasy (Madagascar).

Central-Eastern Malayo-Polynesian further divides into Central Malayo-Polynesian and Eastern Malayo-Polynesian, the former comprising a number of languages spoken in parts of the Lesser Sunda islands and of southern and central Maluku. Eastern Malayo-Polynesian divides in turn into South Halmahera-West New Guinea and Oceanic, with the former including Austro-

nesian languages of southern Halmahera and parts of northwest Irian Jaya. Oceanic includes all other Austronesian languages of Melanesia, Micronesia (except that Palauan and Chamorro are Western Malayo-Polynesian), and Polynesia. Oceanic thus includes the Polynesian languages, spoken in the triangle whose points are Hawaii in the north, Easter Island in the east, and New Zealand in the south. Polynesian languages include Hawaiian, Tahitian, Maori, Samoan, Tuvaluan, and Tongan. Genetically just outside Polynesian within Oceanic is Fijian. Kiribati (Gilbertese) is a Micronesian language, also within Oceanic but outside Polynesian.

3.7 Papuan families

The island of New Guinea and immediately surrounding areas form the linguistically most diverse area on earth, with over 1,000 languages spoken by a population of between six and seven million. While some of the coasts of New Guinea itself and most of the smaller islands of the New Guinea area are occupied by Austronesian languages, most of the interior, together with some coastal and island areas, are occupied by so-called Papuan languages. The term "Papuan" is basically defined negatively as those languages of the New Guinea area that are not Austronesian. Until recently, two radically different approaches to the internal classification of Papuan languages prevailed among specialists. On the one hand, Wurm (1982) divided the languages into five major "phyla" (i.e. large-scale families) and six minor phyla, plus seven or more language isolates. The most widespread of these large families is the Trans New Guinea phylum, containing most of the languages spoken across the highland backbone of the island but also extending southwest as far as Timor and neighboring islands. The other major phyla in this classification are: West Papuan (northern Halmahera and parts of the Bird's Head in Irian Jaya), Geelvink Bay (part of the north coast of Irian Jaya, to the east of the Bird's Head), Torricelli (western parts of the north coast of Papua New Guinea), Sepik-Ramu (large parts of northwestern Papua New Guinea), and East Papuan (on islands from New Britain eastwards to the Solomons). (Note that Geelvink Bay is now called Cenderawasih Bay; the Bird's Head was formerly called the Vogelkop.) Foley (1986), by contrast, maintains that work to date allows only the identification of about sixty genetic units, with internal diversification about as for Romance, among the Papuan languages, with higher-level relations among them remaining a task for future research.

Ongoing work, some of it published in Pawley (1999) and including contributions by Foley among others, suggests that there may well be a firm basis for using traditional comparative methods for a stripped-down version of the Trans New Guinea family, which would still include a substantial number of the smaller genetic units found along the backbone of the main island, although by no means all of Wurm's Trans New Guinea phylum finds justification in the ongoing work. But this does indicate that the time may be ripe

or nearly ripe for a more systematic look at genetic relations among the Papuan languages.

As can be imagined from the low average ratio of speakers to languages, most Papuan languages have few speakers. The languages listed by Grimes (1996a) as having more than 100,000 speakers are Enga, Chimbu, and Medlpa in the highlands of Papua New Guinea, and Western Dani, Grand Valley Dani, and Ekari in the highlands of Irian Jaya. It is a general pattern that languages with more speakers tend to be found in the highlands, whose valleys are also the area of greatest population density.

3.8 Australian families

The classification of Australian languages is in something of a turmoil at present. Dixon (1980) proposed that all Australian languages form a single family, with the exception of Tiwi, spoken on islands off the north coast, and Djingili, in the Barkly Tableland. In a more recent work, Dixon (1997) takes a different stand, suggesting that the peculiar social history of Aboriginal Australia, with the absence of major power centers and continual contact among languages, may make the traditional comparative method unworkable for Australia. Many Australianists nonetheless retain the concept of language family, with about twenty language families in Australia, perhaps all or most being related as a single Australian language family. In particular, there is widespread acceptance of a Pama-Nyungan family that would include the languages spoken in most of the island-continent except some of those in the far north, although Dixon (1997) explicitly rejects the genetic unity of Pama-Nyungan. A new synthesis of Australian languages, to replace Dixon (1980), is currently in preparation by Dixon and others. No Australian language has a large number of speakers, the most viable languages having at most a few thousand.

The records of the extinct Tasmanian languages are sparse, and Dixon (1980) concludes that they are insufficient to exclude the possibility that they may have been related to Australian languages, though equally they are insufficient to establish such a relationship (or any other). Speakers of the Tasmanian languages must have been separated from the rest of humanity for about 12,000 years, from the time rising waters created the Bass Strait to the first visits by Europeans, making them the most isolated human group known to history; the genocide visited upon the Tasmanians in the nineteenth century is thus also a scientific tragedy of the first order.

3.9 Other languages of southern, eastern, and southeastern Asia

A number of living languages spoken in this region have so far eluded genetic classification, in particular Burushaski spoken in northern Pakistan, and Nahali

(Nihali) in central India. Burushaski is reasonably well described, while Nahali is in urgent need of a detailed description. The Andamanese languages, spoken on the Andaman islands (politically part of India) also lack any widely accepted broader genetic affiliation.

In addition, reference may be made to two extinct languages. Elamite was the language of Elam, an important empire in what is now southwestern Iran around 1000 BCE; it is possible that it may be related to Dravidian (McAlpin 1981). Sumerian was the language of ancient Sumer, and is noteworthy as being probably the first language to have had a writing system; it was still used as a literary language in the Old Babylonian period, although before or during this period it was replaced as a spoken language by Akkadian (see section 4.1).

3.10 Proposals for larger groupings

For the suggestion that Dravidian might belong to the proposed Nostratic macro-family, see section 2.7.

Benedict (1975), building largely on his own earlier work, proposes an Austro-Tai macro-family that would include Austro-Asiatic, Daic, Hmong-Mien, and Austronesian. Ruhlen (1994: 24–8) reports on attempts to set up a Dene-Caucasian grouping that would include Na-Dene (see section 5.1), Yeniseian, Sino-Tibetan, Nahali, Sumerian, Burushaski, North Caucasian, and Basque (for some of these languages, see section 2.6).

Greenberg (1971) proposed an Indo-Pacific grouping that would include all Papuan languages plus the Andamanese and Tasmanian languages, but this proposal does not seem to have been taken up in detail by other linguists.

The possibility of a link between (some) Australian and (some) Papuan languages is mooted by Foley (1986).

4 Languages of Africa and Southwestern Asia

The starting point for recent discussions of the classification of African languages is Greenberg (1963), who proposes a fourway division into Afroasiatic, Niger-Congo (Niger-Kordofanian), Nilo-Saharan, and Khoisan families. Afroasiatic and Niger-Congo are now generally accepted, while more controversy has surrounded Nilo-Saharan and Khoisan.

4.1 Afroasiatic languages

The Afroasiatic (formerly Hamito-Semitic) family is the dominant language family of most of north Africa and large parts of southwestern Asia, and

although individual languages have contracted or extended their geographical distribution, this distribution of the family as a whole goes back to antiquity. The family is generally considered to have six branches: Semitic in southwestern Asia, Eritrea, and much of Ethiopia, also of course now in most of North Africa as a result of the spread of Arabic; Egyptian in older times in Egypt; Berber across most of the rest of north Africa (though now in retreat before Arabic in most of this area); Chadic, in a belt centered on northern Nigeria and southern Niger; Cushitic in the Horn of Africa (Somalia, Djibouti, much of southern Ethiopia, and extending into Kenya and Tanzania to the east of Lake Victoria); and Omotic along the Omo River in southeastern Ethiopia. Omotic languages were formerly, and are still sometimes, considered a subbranch of Cushitic. There is need for an up-to-date survey of the family as a whole; in the meantime, reference may be made to Diakonoff (1988).

The Semitic languages are the best studied of the Afroasiatic branches, and Semitic languages can be traced back almost to the beginning of written history. The most recent survey is Hetzron (1997). The Semitic branch is divided into two subbranches, East Semitic and West Semitic. The East Semitic branch is extinct, although it contains Akkadian, the language of the Babylonian and Assyrian civilizations. West Semitic contains all the living Semitic languages as well as several historically important dead languages. The subdivision of West Semitic is more controversial, especially as regards the position of Arabic. The widely accepted current classification as given in Hetzron (1997) divides West Semitic into Central Semitic and South Semitic. Central Semitic subdivides into Arabic and Northwest Semitic. The older classification would put Arabic in South Semitic, and thus use Northwest Semitic for the other subbranch of West Semitic. The classification of Hetzron will be followed in the presentation here.

Arabic was, until the spread of Islam, the language of part of the Arabian peninsula, but as the language of Islam it has spread through much of southwestern Asia and north Africa, replacing the languages previously spoken across most of this area and becoming one of the modern world's major languages. The standard written language is still firmly rooted in the language of the Koran and medieval Arabic literature, but spoken varieties of Arabic are sufficiently different from one another that mutual intelligibility is not possible between extreme varieties. However, only one variety of Arabic has developed as a separate written language, namely Maltese.

Northwest Semitic includes the Canaanite languages and Aramaic. The best known of the Canaanite languages is Hebrew, used as the spoken and written language of the Jews until the early centuries CE, then as a written and liturgical language by Jews throughout the middle ages, to be revived as a spoken language starting in the late nineteenth century and reaching its culmination as an official language and the dominant spoken language of Israel. The other Canaanite languages are all extinct, the best known being Phoenician. Aramaic was a major lingua franca of the Near East from the eighth century BCE, but at present varieties of Aramaic are spoken in enclaves in Syria, Iraq, and Iran.

South Semitic includes the South Arabian languages spoken on the southern fringe of the Arabian peninsula. Most living South Semitic languages belong to the Ethiopian Semitic subgroup, and include Amharic, the dominant language of Ethiopia; Tigrinya, an important regional language of Ethiopia and Eritrea; and Tigré, another regional language of Eritrea. In addition, Ethiopian Semitic includes Ge'ez, the extinct language still used liturgically by the Ethiopian Church.

Egyptian, by which is meant here Ancient Egyptian, is a single language attested in various historical stages from the earliest writing in Egypt. The hieroglyphic writing system and its offshoots were used into the Common Era, but were soon replaced after Christianization by a Greek-based script, and this later variety of the language is called Coptic. Coptic survived as a spoken language to the late middle ages, when it was finally replaced completely by Arabic, although it continues in use as the liturgical language of the Coptic Church. A recent survey is Loprieno (1995).

The Berber languages are spoken in a scattered pattern across north Africa from just east of the Egypt–Libya border, though they are strongest in mountainous parts of Algeria and especially Morocco, and in the desert parts of Mali and Niger. Among the most widely spoken varieties are Kabyle (Algeria), Chaouia (Algeria), Tarifit (Northern Shilha) (Morocco, Algeria), Tachelhit (Central Shilha) (Morocco, Algeria), Tamazight (Southern Shilha) (Morocco, Algeria), Tamashek (the language of the Tuaregs, mainly in Mali and Niger).

Most of the Chadic languages have few speakers, but there is one significant exception, namely Hausa, the dominant indigenous language of northern Nigeria and southern Niger. Hausa is widely used as a lingua franca by speakers of other neighboring Chadic and non-Chadic languages.

The most widely spoken Cushitic languages are Somali (mainly in Somalia and Ethiopia), Sidamo (Ethiopia), Oromo (Galla) (Ethiopia), Afar (Ethiopia, Eritrea, Djibouti), and Bedawi (Beja) (Sudan). The most widely spoken Omotic language is Wolaytta (Ethiopia).

4.2 Niger-Congo languages

Niger-Congo languages cover most of Africa south of a line drawn from the mouth of the Senegal River in the west to where the equator cuts the coast of Africa in the east, with the major exception of the area in southwestern Africa occupied by the Khoisan languages. There are also considerable excursions of Niger-Congo to the north of this line, and less significant excursions of non-Niger-Congo languages to the south of this line, e.g. Cushitic and Nilotic languages spoken to the east of Lake Victoria. The internal structure of the Niger-Congo family was first worked out in detail in Greenberg (1963), although a number of changes have been proposed in more recent work, several of which are still the subject of debate. The most recent overview is Bendor-Samuel (1989), and the classification given there will be followed here.

One branch of Niger-Congo is spoken outside the area delimited above, namely Kordofanian, spoken in the Nuba mountains of Sudan, to the south of El-Obeid. While Greenberg considered Kordofanian genetically the most distant of the languages in the overall family, thus naming the family as a whole Niger-Kordofanian with two coordinate branches Kordofanian and Niger-Congo, the current view is rather that Kordofanian is at least no more distant genetically from the core of the family than are the Mande languages, and the name Niger-Congo is current for the family as a whole. It should be noted that one group of languages assigned tentatively by Greenberg to Kordofanian on the basis of fragmentary material, namely Kado or Kadu (formerly called Kadugli-Krongo), is now believed not to be Kordofanian or Niger-Congo, and perhaps Nilo-Saharan (Bender 1997: 25).

The Mande languages are spoken over most of west Africa to the west of 5°W and to the south of 15°N, although considerable parts of this territory, especially near the coasts, are occupied by other branches of Niger-Congo (Atlantic and Kru). Mende languages include Bambara, the major indigenous language of Mali, and some closely related languages such as Maninka; Jula, spoken in Côte d'Ivoire and Burkina Faso; Kpelle, the major indigenous language of Liberia; and Mende, the major indigenous language of Sierra Leone.

In Bendor-Samuel (1989: 21) the rest of Niger-Congo, once Kordofanian and Mende have been removed, is referred to as Atlantic-Congo, with Atlantic and Ijoid as the genetically next most divergent groups, the remainder being referred to as Volta-Congo. The remaining Niger-Congo groups, i.e. Bendor-Samuel's Atlantic-Congo, will be treated together in what follows.

Atlantic languages are spoken, predominantly in coastal areas, from the Senegal River in the north down into Liberia, although the most widely spoken Atlantic language, Fula (Fulfulde, Peul) has a different distribution. The Fulani, as the speakers of Fula are called, are pastoralists whose range is between the rain forest to the south and the desert to the north, with traditional seasonal moves along a north–south axis; the language is spoken in pockets from the Atlantic coast into Sudan and even Ethiopia, with concentrations in northern Nigeria and northern Cameroon. Another widely spoken Atlantic language is Wolof, the major indigenous language of Senegal.

The Kru languages are spoken in Liberia and southwestern Côte d'Ivoire, with relatively small numbers of speakers. Kru was included in Kwa (see below) by Greenberg (1963).

The Gur (Voltaic) group cover most of Burkina Faso, spreading also into northern parts of countries to the south. The Gur language with by far the largest number of speakers is Moore, the dominant indigenous language of Burkina Faso. One language sometimes considered to be Gur is Dogon, spoken around Bandiagara in Mali and adjacent parts of Burkina Faso, but current opinion questions this assignment and in Bendor-Samuel (1989) Dogon is considered at least provisionally a separate branch within Volta-Congo.

To the south of the Gur languages and continuing to the coast are the Kwa languages, stretching roughly from the Bandama River in the west to the

Benin–Nigeria border in the east. The precise extent of Kwa has shifted considerably since Greenberg (1963), and not all the innovations have been generally accepted. In Bendor-Samuel (1989), the term Kwa covers essentially Greenberg's Western Kwa, with his Eastern Kwa being mostly reassigned to Benue-Congo (see below). The least controversial part of these changes is the exclusion of Kru (see above) and Ijo (see below) from Kwa. In what follows, as in the geographical description given above, the restricted sense of Kwa as in Bendor-Samuel (1989) will be followed. Kwa languages, in this narrow sense, include Baule, an important regional language of southern Côte d'Ivoire; the Akan dialect cluster (Twi-Fante), the major indigenous language of Ghana; the Ga-Dangme dialect cluster, including Ga, the major indigenous language of the Ghanaian capital Accra; and the Gbe dialect cluster, including Ewe, a widely spoken indigenous language in Ghana and Togo, and Fongbe, the most widely spoken indigenous language of Benin.

Ijo, now usually considered a distinct branch of Niger-Congo, is spoken around the delta of the Niger River in Nigeria, and is the major indigenous language of Nigeria's Rivers State. Different varieties of Ijo are not all mutually intelligible, the most prestigious varieties being Kolokuma and Kalabari.

The Adamawa-Ubangi languages are spoken in a belt from eastern Nigeria into Sudan, with the main concentration in the Central African Republic. The languages of the Adamawa subgroup are spoken to the west, those of the Ubangi subgroup to the east. The most widely used Adamawa-Ubangi language is Sango, which is the national language of the Central African Republic; historically, it is a creole derived primarily from the Ubangi language Ngbandi.

The remaining branch of Niger-Congo, Benue-Congo, covers most of sub-Saharan Africa from the western border of Nigeria eastwards to the Indian Ocean and southwards to the Cape. Most of this area and population falls under Bantu, but from a historical linguistic viewpoint Bantu is a rather low-level subgroup within Benue-Congo and the present geographical distribution of Bantu is the result of an expansion from the Nigeria–Cameroon border area that took place for the most part within the last two millennia. The most widely spoken Benue-Congo languages outside Bantu are Yoruba, an official language in southwestern Nigeria; Edo, to the southeast of Yoruba; Nupe, to the northeast of Yoruba; Igbo, an official language in central southern Nigeria; Ibibio-Efik to the east of the Niger delta in Nigeria; and Tiv, a regionally important language of eastern Nigeria.

As already implied, the Bantu languages occupy most of Africa from the Nigeria–Cameroon border to the east and south, including several major indigenous languages. The most widely spoken Bantu language is Swahili, originally the language of Zanzibar and the adjacent coast, although it has now spread as a lingua franca and also, especially in Tanzania, as a first language across large parts of east Africa; it is the official language of Tanzania and an official language in Kenya. Comorian, the indigenous language of the Comoros, is closely related to Swahili. Several other widely spoken Bantu languages are here listed primarily by country: Fang (Equatorial Guinea, Gabon), Bangala

(Congo–Kinshasa), Kituba (Congo–Kinshasa), Lingala (Congo–Kinshasa), Kikongo (Congo–Kinshasa, Angola), Luba-Kasai (Congo–Kinshasa), Luba-Shaba (Congo–Kinshasa), Zande (Congo–Kinshasa and neighboring countries), Northern Mbundu (Angola), Southern Mbundu (Angola), Gikuyu (Kenya), Kamba (Kenya), Luyia (Kenya), Luganda (Uganda), Nyankore (Uganda), Soga (Uganda), Kirundi (Burundi), Kinyarwanda (Rwanda) – Kirundi and Kinyarwanda are essentially different national variants of the same language – Chagga (Tanzania), Haya (Tanzania), Makonde (Tanzania, Mozambique), Nyamwezi (Tanzania), Sukuma (Tanzania), Lomwe (Mozambique), Makua (Mozambique), Sena (Mozambique), Tsonga (Mozambique, South Africa), Nyanja (Malawi, Mozambique, Zambia), Tumbuka (Malawi, Zambia), Yao (Malawi, Tanzania), Nyakyusa-Ngonde (Malawi, Tanzania), Bemba (Zambia), Luvale (Zambia), Tonga (Zambia), Northern Ndebele (Zimbabwe), Shona (Zimbabwe, Zambia, Mozambique), Tswana (Botswana, South Africa), Southern Sotho (Lesotho, South Africa), Swati (Swaziland, South Africa), Northern Sotho (Pedi) (South Africa), Tsonga (South Africa), Venda (South Africa), Xhosa (South Africa), Zulu (South Africa).

4.3 Nilo-Saharan families

Nilo-Saharan, as proposed by Greenberg (1963), has proven to be more controversial than either Afroasiatic or Niger-Congo, although the most recent survey of the Nilo-Saharan languages (Bender 1997) is positive. The internal structure of Nilo-Saharan is also more controversial. In what follows I have therefore limited myself to citing some of the more widely spoken Nilo-Saharan languages and the branches of the family to which they belong.

Nilo-Saharan languages are not spoken in a continuous geographical area, and even in the areas mentioned below they are often interspersed with Afroasiatic (Chadic, Cushitic, also Arabic) and Niger-Congo languages. One Nilo-Saharan area is the middle course of the Niger River; another is Chad; a third is the Nile around the Egypt–Sudan border; while a fourth includes parts of southern Sudan, westernmost Ethiopia and Eritrea, northeastern Congo–Kinshasa, and parts of Kenya and Uganda to the north and east of Lake Victoria.

The westernmost language, or rather cluster of closely related languages, assigned to Nilo-Saharan is Songay, spoken along the Niger river in an area including the town of Timbuktu, although it is also the living language whose inclusion in Nilo-Saharan has proven most controversial (Bender 1997: 59). Another major western Nilo-Saharan language, assigned to the Saharan branch of the family, is Kanuri, the dominant indigenous language of Bornu State in northeastern Nigeria. Within the For(an) branch, mention should be made of For (Fur), spoken in the Darfur region in west-central Sudan.

Most of the more widely spoken Nilo-Saharan languages belong to the East Sudanic and Central Sudanic branches. East Sudanic includes the Nubian

languages of the Egypt–Sudan border area, of which the most widely spoken is Nobiin. It also includes the Nilotic languages, a grouping which includes the Luo (Lwo) languages Acholi (Uganda), Lango (Uganda), Alur (Uganda, Congo–Kinshasa), and Luo (Dholuo) (Kenya); the Dinka-Nuer languages Jieng (Dinka) (Sudan) and Naadh (Nuer) (Sudan); the Eastern Nilotic languages Maasai (Kenya, Tanzania), Turkana (Kenya), Karamojong (Uganda), and Teso (Uganda, Kenya), and the Southern Nilotic language Kalenjin (Kenya). Central Sudanic includes Ngambay (Sara-Ngambay) (Chad), Lugbara (High Lugbara) (Congo–Kinshasa, Uganda), Mangbetu (Congo–Kinshasa), Ndo (Congo–Kinshasa), and Badha (Lendu) (Congo–Kinshasa).

It has been suggested that Meroitic, the extinct language of the Meroë civilization (ca. 2300–1600 BP), might be a Nilo-Saharan language, but Bender (1997: 32) considers the available data insufficient to resolve the issue. Finally, it should be noted that there are some as yet virtually undescribed languages spoken in the general Nilo-Saharan area that are as yet insufficiently known to establish whether or not they might be Nilo-Saharan. Many Nilo-Saharan and possible Nilo-Saharan languages are spoken in regions of current unrest (southern Sudan) or recent unrest (Ethiopia), which accounts in part for the rather poor state of our knowledge of such languages.

4.4 Khoisan families

The Khoisan languages are spoken predominantly in southwestern Africa. The area occupied by Khoisan languages has certainly contracted as a result of the spread of Bantu and, more recently, Indo-European languages, and all Khoisan languages have small numbers of speakers, with the largest, Nama (Khoekhoe), spoken in Namibia and South Africa, having an estimated 146,000 (Grimes 1996a: 323). Two otherwise unclassified languages of East Africa, namely Hadza and Sandawe of Tanzania, were proposed for inclusion in Khoisan by Greenberg (1963). However, even the genetic unity of Khoisan with the exclusion of Sandawe and Hadza is not accepted by all specialists, some of whom prefer to treat Northern Khoisan, Central Khoisan, and Southern Khoisan as distinct families. Sands (1998) is a recent treatment, concluding that there are striking parallels among the three nuclear branches of Khoisan plus Hadza and Sandawe, but that it is not clear to what extent this reflects common genetic origin versus contact.

4.5 Proposals for larger groupings

The proposal that Afroasiatic might form part of a larger Nostratic macrofamily was discussed in section 2.7. Proposals that Niger-Congo and Nilo-Saharan might be distantly related are considered worth following up both in Bendor-Samuel (1989: 8–9) and by Bender (1997: 9).

5 Languages of the Americas

The internal and external genetic affiliations of the indigenous languages of the Americas have given rise to considerable debate in recent years, with proposals ranging from a total of three families (Greenberg 1987) to almost 200 (Campbell 1997). Since Campbell (1997) lists securely assured genetic units, but then also discusses proposals for broader genetic groupings, his account can serve as a survey that covers the range of proposals. In the space available, it would not make sense to list and discuss up to 200 genetic units, so in what follows a very selective choice will be made, concentration on larger families and languages with larger numbers of speakers.

5.1 Languages of North America

The languages of North America are surveyed in Mithun (1999).

The northern fringe of North America is home to the Eskimo-Aleut family. The family has two branches, Aleut, spoken on the Aleutian islands, and Eskimo. The latter starts in eastern Siberia and then stretches from Alaska to Greenland. Eskimo is properly a family of languages, with a major division between Yupik (Siberian and southern Alaskan varieties) and Inuit-Inupiak in northern Alaska, Canada, and Greenland. Greenlandic is the variety with most speakers, and is the national language of Greenland.

Another major family of North America is Na-Dene, although the precise extent of the family is controversial. Its core is Athabaskan, comprising most of the languages of the interior of Alaska, northwest Canada, with some languages (all extinct or moribund) in Oregon and northern California, and then a flowering in the geographically remote Apachean languages of the southwestern USA, including Navajo. It is established that the recently extinct Eyak language, spoken at the mouth of the Copper River in Alaska, is genetically related to Athabaskan, to give Athabaskan-Eyak. Less certain is whether Tlingit, (spoken on the Alaska panhandle) is related to these, which would justify the more inclusive term Na-Dene, and even more questionable whether Haida (spoken on Queen Charlotte Island) should be added.

Other language families of the Pacific Northwest include Wakashan (British Columbia and adjacent Washington state; the family includes Nootka) and Salishan (British Columbia, Washington state, with some excursion into Idaho and Montana). Other language families of California, sometimes extending to adjacent areas, are Miwok-Costanoan, Chumashan, and Yuman. The small Keresan family and the language isolate Zuni of New Mexico, though small in number of speakers (each in the thousands) are among the most vigorous indigenous languages of the USA, with high rates of acquisition by children.

The Siouan languages are a major language family of the North American Plains, stretching from north of the US–Canada border through the Dakotas,

Minnesota, and Wisconsin down to Arkansas, with outliers historically almost as far south as the Gulf and in Virginia. The Muskogean family, formerly concentrated in the southeastern USA, includes Choctaw, Chickasaw (these two arguably dialects of a single language), Alabama, and Seminole.

The Iroquoian languages are spoken around the Great Lakes, apart from Cherokee, originally spoken in Georgia; the family also includes Tuscarora, Huron (extinct), Seneca, and Mohawk. The Algic (Algonquian-Ritwan) family covers much of the northeast of North America, though also extending into western Canada and with two outliers on the Great Plains (Cheyenne and Arapaho). The family includes Blackfoot, the various forms of Cree spoken in Canada, and Ojibwa in Canada and the USA. These are all Algonquian languages. The two Ritwan languages, Wiyot (extinct) and Yurok (moribund), though indubitably related to Algonquian, are spoken in California.

Uto-Aztecan is one of the major language families of North America, spreading also into Meso-America. The Northern Uto-Aztecan languages include Shoshone, Comanche, Ute, and Hopi; while the Southern Uto-Aztecan languages include Pima-Papago (O'odham) in Arizona and Sonora; Cora and Huichol in Nayarit and Jalisco; and Nahuatl in central Mexico. Nahuatl was the language of the Aztec empire.

5.2 *Languages of Meso-America*

The languages of Meso-America are surveyed in Suárez (1983). The Uto-Aztecan family was discussed in section 5.1.

Other major language families of Meso-America are Otomanguean, Mixe-Zoquean, and Mayan. The Otomanguean languages are spoken mainly across the isthmus of Mexico, especially its southern part (Guerrero, Oaxaca, Puebla), including the Zapotecan languages and Mixtec, although some Otomanguean languages, such as Otomí, are spoken further north and separated from the mass of Otomanguean languages. Mixe-Zoquean languages are spoken in a number of geographically separated groups in the isthmus of Mexico; the Olmecs, the first of the great Meso-American civilizations, seem to have spoken a Mixe-Zoquean language. The Mayan languages cover or covered most of Mexico east of the isthmus and also Guatemala and Belize; individual languages include Yucatec, Chol, Kekchi, and K'iche' (Quiché); although Chol does not have one of the highest numbers of speakers among Mayan languages, it is important historically as the most direct descendant of the language of the Mayan hieroglyphic inscriptions.

For Chibchan languages of Meso-America, see section 5.3.

5.3 *Languages of South America*

For Amazonian languages, Dixon and Aikhenvald (1999) provide a survey; a comparable survey for Andean languages is in preparation by Willem Adelaar

and Pieter Muysken for Cambridge University Press. Campbell's (1997) discussion of South American languages is based on Kaufman (1990).

The Chibchan language family includes a number of languages scattered from Costa Rica in the west through Panama to Colombia in the east. Cariban languages are scattered across northeastern South America, mostly to the north of the Amazon, although some languages are spoken as far west as Colombia and there is a geographically isolated group well to the south along the upper course of the Xingu River.

The precise extent of the Arawakan family is a matter of ongoing debate, although the group of languages that are clearly genetically related are sometimes referred to as Maipurean. They are scattered at great distances from one another across much of northern South America, from the Caribbean coast as far south as Paraguay, with Garífuna (also misleadingly known as Black Carib) spoken in Central America.

Tucanoan languages are spoken in northwestern South America (Colombia, Ecuador, Peru, adjacent parts of Brazil). Panoan languages are spoken in the Peru–Brazil border area, with some spillover into Bolivia, while the Tacanan languages, now believed to be related to Panoan in a Panoan-Tacanan family, are spoken in Bolivia with some spread into Peru. The Gê (Je) family is spoken in Brazil. Tupian languages are spoken both on the Amazon River and its tributaries and in an area that includes Paraguay and adjacent parts of Brazil, Bolivia, and Argentina. The most widely spoken Tupian language is Paraguayan Guaraní, spoken by 95 percent of the population of Paraguay and a national language of the country.

Quechumaran includes the Quechua and Aymara branches, although the nature of the relationship between Quechua and Aymara – genetic or contact – continues to be debated. Quechua is strictly speaking a language family rather than a single language, since different varieties are not mutually intelligible; in terms of numbers of speakers, it is the largest indigenous language family of the Americas. Quechua was the language of the Inca empire, and partly as a result of this empire and later use as a lingua franca by the Spanish administration it achieved a spread from Colombia in the north to Argentina in the south, although most speakers are in Peru. The most widely spoken Quechua languages are South Bolivian Quechua in Bolivia, Cuzco Quechua in Peru, and Chimborazo Quichua in Ecuador. Aymara is spoken predominantly in western Bolivia.

5.4 Proposals for larger groupings

For proposals that would group Eskimo-Aleut as part of Eurasiatic, see section 2.7. For proposals that would group Na-Dene with Sino-Tibetan and possibly other families, see section 3.10. Otherwise, the main proposal is that of Greenberg (1987) to group all the remaining indigenous languages of the Americas into a single Amerind family.

There are also more modest proposals for larger genetic units within the indigenous languages of the Americas, excluding Eskimo-Aleut and Na-Dene, of which Penutian and Hokan are perhaps the most engrained in the literature. Penutian would group together a number of languages and language families of the northern part of western North America, including at least: Maidu and Miwok-Costanoan (together California Penutian); Chinookan; Coos, Kalapuyan, and Yakonan (together Oregon Penutian); Klamath-Modoc and Sahaptin (together Plateau Penutian); and Tsimshian. Hokan would group together a number of languages and language families of the southern part of western North America and extending into Meso-America, including at least: Yuman; Karok-Shasta, Pomo, and Yana (Northern Hokan); Chumash, Salinan, and Seri; Tequistlatecan (Chontal of Oaxaca); and Washo.

6 Pidgin and Creole Languages

Since the main concern of this survey is the geographic distribution of languages as spoken by native speakers, pidgin languages (Holm 1989) will only be considered to the extent to which they are being creolized. Pidgin languages that are relevant in this way include the closely related Krio of Sierra Leone, Pidgin of Cameroon, and Pidgin of Nigeria, all of which are English-based pidgins undergoing creolization and widely used as lingua francas in the relevant countries. In addition, mention must be made of the closely related Tok Pisin of Papua New Guinea, Pijin of the Solomon Islands, and Bislama of Vanuatu, all likewise English-based pidgins undergoing creolization and widely used as lingua francas.

Otherwise, creole languages are particularly prevalent in the Caribbean and the islands of the Indian Ocean. They include English-based Sranan, the lingua franca of Suriname, and the French-based creoles of Haiti in the Caribbean and Mauritius, Réunion, and the Seychelles in the Indian Ocean. The case of Sango, in origin a creolized form of Ngbandi, was discussed in section 4.2.

7 Deaf Sign Languages

Most works on languages of the world deal exclusively with spoken languages, and this is certainly the emphasis of this survey. Recent work on deaf sign languages has shown, however, that deaf sign languages are languages in their own right, differing considerably in structure from the spoken languages used in the same territory (see chapter 22). Indeed, genetic relations among deaf sign languages often do not match those of the "corresponding" spoken languages, e.g. American Sign Language (ASL) is more closely related to French

Sign Language than it is to British Sign Language. Grimes and Grimes (1996) list 104 deaf sign languages, though without giving any internal genetic classification, and it is unfortunately true to say that our knowledge of all but a handful of deaf sign languages (such as ASL) is so poor that it is not at present possible to undertake such a task. This is clearly an area that merits further investigation.