Chapter 7

Language families

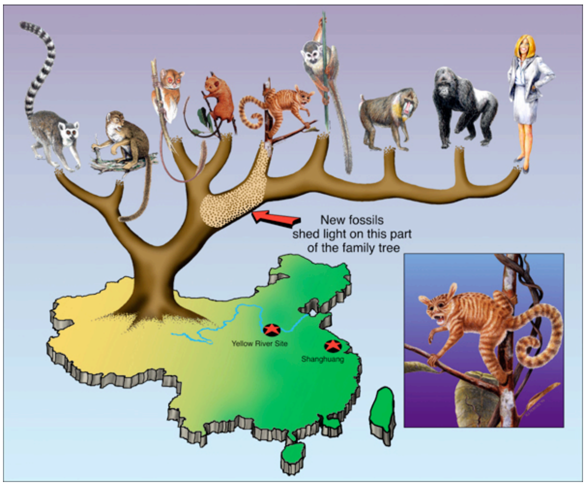


FIG. 1. *A family tree of mammals*

The **family tree** in FIG. 1 (growing in northwestern China, for reasons we shall not discuss here) shows the genealogical relationship between some mammals. Most of us are accustomed to «reading» such trees. When two mammals are standing at the end of two branches that part right below them, like the human being and the gorilla on the extreme right, we understand that these mammals are close relatives. We also understand that the point where two branches part represents the evolutionary stage of a mammal that was the common ancestor of the two branches. That is, humans and gorillas have a common ancestor, and this common ancestor and the baboon (the mammal to the left of the gorilla) had a common ancestor even further back in time, represented in the tree by the point where the baboon branch and the gorilla/human branch part. Of course, the word *tree* is used metaphorically when it designates an abstract «branching structure» like this.

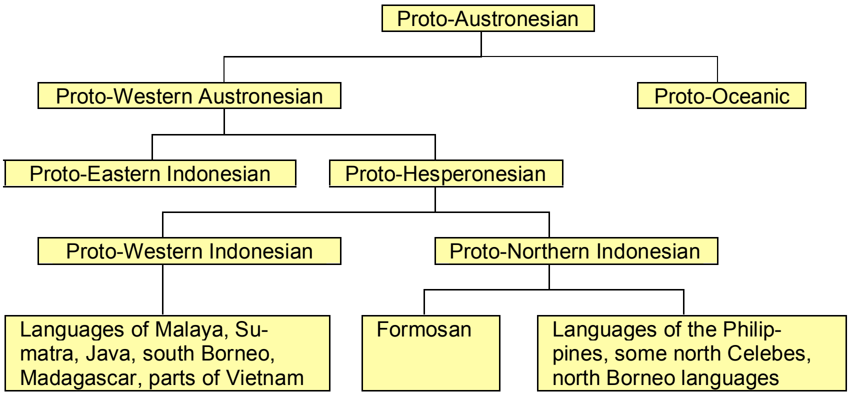

Proto-Austronesian
Proto-Western Austronesian
Proto-Oceanic
Proto-Eastern Indonesian
Proto-Hesperonesian
Proto-Western Indonesian
Proto-Northern Indonesian
Languages of Malaya, Sumatra, Java, south Borneo, Madagascar, parts of Vietnam
Formosan
Languages of the Philippines, some north Celebes, north Borneo languages

FIG. 2. *The Austronesian language family*

Family trees can also be drawn for languages, as illustrated in FIG. 2, which shows the main branches of the Austronesian language family according to one theory; we shall come back to these languages in § X. Like the «tree» in FIG. 2, family trees are often turned upside down, showing the oldest ancestor at the top and younger descendants below her or him. All Austronesian languages are descendants of Proto-Austronesian, a language that is not known from any written sources, but which is postulated to have existed in a distant past.

7.1 Protolanguages

It should be clear from the preceding lines that a **language family** is a group of languages with a common ancestor. This common ancestor is referred to as a **proto-language**. The proto-language split up into two or more dialects, which gradually became more and more different from each other—for example, because the speakers lived far from each other and had little or no mutual contact—until the speakers of one dialect could not understand the speakers of the other dialects any longer, and the different dialects had to be regarded as separate languages. When this scenario is repeated over and over again through centuries and millennia, large language families develop. Of course, the protolanguages of different families also had ancestors, which must have been members of older language families. Many of the branches of these older families may still exist, but they have separated so much that we are not able any longer to discover the family ties. In other cases most or all branches of an ancient language family may be extinct.

The American missionary organization SUMMER INSTITUTE OF LINGUISTICS (SIL) publishes regularly new editions of an interesting book called *Ethnologue: languages of the world*. Here you can find an overview of all languages of all countries in the world. There is an exciting web edition at
http://www.ethnologue.com/web.asp.
In *Ethnologue* you can also find a list of more than 100 hundred language families at
http://www.ethnologue.com/family_index.asp.
Many linguists will disagree about many of the details, but the list gives you an impression of the linguistic diversity of the earth.

Just like other families, most language families have several **branches**. This is illustrated in FIG. 2, according to which the Austronesian language family first split into two branches: *Proto-Western Austronesian* and *Proto-Oceanic*. The former branch then split into *Proto-Eastern Indonesian* and *Proto-Hesperonesian*, and so on. As can be observed from FIG. 2, the term **protolanguage** is not only used about the absolutely oldest stage; we also postulate a protolanguage as the ancestor language of each branch.

7.2 The comparative method

How do we prove that two or more languages are genetically related—that is, that they belong to the same language family? In the nineteenth century the **comparative method** was developed. This is the main tool of comparative linguistics, the branch of linguistics studying the historical and genetic relationship between languages. Before describing this method in detail, we shall take a look at **language change**. This will give us a better basis for understanding the comparative method.

7.2.1 Language change

It is a universal feature of human languages that they change through time, in different ways at different places and in different societal groups. *Why* this is so is not necessarily evident, but it is nevertheless true—always and everywhere. It may, however, be misleading to say that *languages change*, because what happens is rather that *human beings change their languages*.

They change them in many ways. The meanings of words are changed, the pronunciations of words are changed, new words are adopted, old words are discarded, inflections come and go, the structure of phrases, clauses, and sentences are changed.

They change them for many reasons. Children acquire the language or languages of their parents, but slightly modified. We change our language or languages through life, among other things to adjust it or them to new social conditions. Language is an important part of our personal identity, and we modify our language to signal group identity. Languages may be changed consciously or we may be unaware of the fact that we change them—or that other people change theirs.

We shall take a closer look at *sound change*, which to a surprisingly large degree turns out to be regular. The **regularity of sound change** implies that when a certain sound X changes to a slightly different sound X' in one word, the same change tends to take place in all words where sound X occurs, or in all words where sound X occurs in a particular context. *The regularity of sound change is the prerequisite for the comparative method.*

Compare the Latin and Italian words in Table 1.

Latin			Italian		
Orthography	Pronunciation	Meaning	Orthography	Pronunciation	Meaning
habere	/haˈbe:re/	'to have'	*avere*	/aˈve:re/	'to have'
herbam	/ˈherbã/	'grass'	*erba*	/ˈɛrba/	'grass'
hora	/ˈho:ra/	'time, hour'	*ora*	/ˈo:ra/'	'time, hour'
homo	/ˈhomo:/	'man'	*uomo*	/ˈwɔmo/	'man'

Table 1. *Four Latin and Italian words*

Italian is a direct descendant of Latin, and we can see that the words have changed in several ways. What interests us is the unexceptional fate of the Latin *h* /h/: it has disappeared in all Italian words. This is a completely regular sound change from Latin to Italian. The sound /h/ simply does not occur in Italian. We can formulate this regular sound change as in (1), where the symbol '>' means 'changed into' and the symbol 'Ø' means 'zero' or 'nothing'—that is '/h/ in Latin changed into noting in Italian':

(1) *A regular sound change from Latin to Italian: the disappearance of /h/*

Latin /h/ > Italian Ø

But Latin has more descendants than Italian, among others Portuguese, Castilian (Spanish), Catalan, French, and Romanian.[1] Let us take some Latin words and see what has happened to them in Portuguese, Castilian, Italian, and Romanian. Three words from these languages are presented in Table 2.

[1] The descendants of Latin are called *Romance languages*. Latin was one among several *Italic languages*, a branch of the *Indo-European language family*, which is described in § X.

MEANING	LATIN	PORTUGUESE[2]	CASTILIAN	ITALIAN	ROMANIAN
‘eight’	*octo* /ˈoktoː/□	*oito* /ˈojtu/□	*ocho* /ˈotʃo/□	*otto* /ˈɔtto/□	*opt* /ˈopt/□
‘milk’	*lactem* /ˈlaktẽ/□	*leite* /ˈlɐjtə/□	*leche* /ˈletʃe/□	*latte* /ˈlatte/□	*lapte* /ˈlapte/□
‘fact’	*factum* /ˈfaktũ/□	*feito* /ˈfɐjtu/□	*hecho* /ˈetʃo/□	*fatto* /ˈfatto/□	*fapt* /ˈfapt/□

TABLE 2. *Some words in Latin, Portuguese, Castilian, Italian, and Romanian*

The modern languages differ from Latin and from each other in several ways, and all of it is highly regular. We shall concentrate on one single change: the development of the Latin consonant cluster *ct* /kt/. It turns out that one regular sound change can be established from Latin to each of the four modern languages; cf. (2).

(2) *Four regular sound changes*

(a) Latin /kt/ > Portuguese /jt/
(b) Latin /kt/ > Castilian /tʃ/
(c) Latin /kt/ > Italian /tt/
(d) Latin /kt/ > Romanian /pt/

When a Latin word has the consonant cluster /kt/, the modern Portuguese version of the same word has the sound cluster /jt/, the modern Castilian version has /tʃ/, the Italian version has /tt/, and the Romanian version has /pt/. These generalizations are valid across the whole vocabularies of these languages, and they are just a few of all the regular sound changes of these languages.

It is also worth noticing that some sounds have not changed, or have only changed in some of the languages. For example, Latin *l* /l/ has survived word initially in all these languages, as observed in the word meaning ‘milk’ in TABLE 2, and supported by other words, like the one meaning ‘moon’: Latin *lunam* /ˈluːnã/, Portuguese *luna* /ˈlunɐ/, Castilian *luna* /ˈluna/, Italian *luna* /ˈluːna/, and Romanian *lună* /lunə/.

Latin *f* /f/ has remained /f/ in all the languages except Castilian, where it has disappeared (changed into nothing), as observed in the word meaning ‘fact’ in TABLE 2, and supported by other words, for example the one meaning ‘son’: Latin *filium* /ˈfiːljũ/, Portuguese *filho* /ˈfil^{j}ju/, Castilian *hijo* /ˈixo/, Italian *figlio* /ˈfil^{j}jo/, and Romanian *fiu* /ˈfiw/.

7.2.2 Regular sound correspondences

The regularity of sound changes can be observed without taking the ancestor language (the **protolanguage**) into account. Because the sound changes from the protolanguage to its descendants are regular, there are also **regular sound correspondences** between languages with a common ancestor. Continuing with our Romance examples, we observe that if a word starts with /fV/ in Portuguese, where V = any vowel, we also find /fV/ in Italian and Romanian, while Castilian has only /V/. On the basis of the words in TABLE 2 we can establish several sound correspondences between these four modern Romance languages. By taking more words into account, we can conclude that these sound correspondences are regular; cf. TABLE 3.

ENVIRONMENT	PORTUGUESE	CASTILIAN	ITALIAN	ROMANIAN

[2] The Portuguese pronunciation presented here is that of Portugal, not Brazil.

After a vowel	/jt/□	/tʃ/□	/tt/□	/pt/□
Word initially	/l/□	/l/□	/l/□	/l/□
Word initially	/f/□	Ø	/f/□	/f/□
In an accented syllable	/o/□	/o/□	/o/□	/o/□
Word finally	/u/□	/o/□	/o/□	Ø
Word finally	/ə/□	/e/□	/e/□	/e/

TABLE 3. *Some regular sound correspondences between Portuguese, Castilian, Italian, and Romanian*

The search for regular sound correspondences is an important part of the comparative method—because:

> If **regular sound correspondences** can be established between two or more languages, these languages are **genetically related**, that is, they belong to the same language family and are descendants of the same **protolanguage**.

It should be emphasized that what we are not only looking for *systematic resemblances* between languages, but also *systematic differences*. Of course, the regular sound correspondence Portuguese /l/ – Castilian /l/ – Italian /l/ – Romanian /l/ is a systematic resemblance, but the regular sound correspondence Portuguese /jt/ – Castilian /tʃ/ – Italian /tt/ – Romanian /pt/ is a systematic difference.

A study of the historical relationship between Latin and its descendants— the Romance languages—is particularly valuable because the proto-language is documented in texts. Several aspects of the comparative method can be tested. We can study regular sound correspondences between the modern languages and show that they are the result of regular sound changes from the protolanguage.

AUSTRONESIAN SOUND CORRESPONDENCES

When we compare languages whose history is totally or partially unknown, and we discover regular sound correspondences, we have a solid basis for the postulation of a genetic relationship and common ancestor. In TABLE 4, we have illustrated this with six words Malagasy, Indonesian, Samoan, and Maori—four Austronesian languages.

	MALAGASY	INDONESIAN	SAMOAN	MAORI
‘fire’	*afo* /afu/	*api* /api/	*afi* /afi/	*ahi* /ahi/
‘ten’	*folo* /fulu/	*se-puluh* /səpuluh/[3]	*se-fulu* /sefulu/[4]	*(tekau)*[5]
‘four’	*efatra* /efatʃa/	*empat* /əmpat/	*fa* /fa/□	*fa* /faː/
‘feather’	*volo* /vulu/	*bulu* /bulu/	*fulu* /fulu/	*huru* /huru/
‘fruit’	*voa* /vua/	*buah* /buah/	*fua* /fua/	*hua* /hua/
‘new’	*vao* /vau/	*baru* /baru/	*fou* /fou/	*hou* /hou/

TABLE 4. *Words from Malagasy, Indonesian, and Samoan*

[3] The element *se-* /□□/ in Indonesian *se-puluh* ‘ten’ means ‘one’. Cf. Indonesian *dua puluh* ‘twenty’, literally ‘to ten’.

[4] Samoan *se-* in *sefulu* ‘ten’ etymologically related to *se-* /□□/ ‘one’ in Indonesian. Cf. the preceding footnote.

[5] Maori *tekau* ‘ten’ is not related to the words meaning ‘ten’ in the other languages.

It does not take us long to discover lots of similarities between these four languages. Generally, the vowels are the same in all of them, even when the consonants vary considerably, like in the word meaning ‘fruit’—Malagasy *voa* /vua/, Indonesian *buah* /buah/, Samoan *fua* /fua/, and Maori *hua* /hua/. We shall concentrate on the two regular sound correspondences presented in Table 5. The first correspondence is based upon the words meaning ‘fire’, ‘ten’, and ‘four’, while the second correspondence is based upon the words meaning ‘feather’, ‘fruit’, and ‘new’.

Malagasy	Indonesian	Samoan	Maori
f /f/	*p* /p/	*f* /f/□	*h* /h/
v /v/	*b* /b/	*f* /f/□	*h* /h/

Table 5. *Two regular sound correspondences in Austronesian*

Malagasy *vao* /vau/, Indonesian *baru* /baru/, Samoan *fou* /fou/, and Maori *hou* /hou/ are quite different, and it is not immediately evident that these three words are **etymologically related**, that is, that they come from the same Proto-Austronesian word. It is only when we discover that the sound correspondence Malagasy /v/ / Indonesian /b/ / Samoan /f/ / Maori /h/ is *regular*—by recurring in many different words —that this etymological relatedness can be established beyond doubt.

But it is also important to emphasize that the etymological relatedness between the four words mentioned in the last paragraph is not based upon one regular sound correspondence only. *All* the sounds have to participate in regular sound correspondences, as illustrated in Table 6, where the words are written vertically and compared sound by sound.

Malagasy	Indonesian	Samoan	Maori
/v/	/b/	/f/	/h/
/a/	/a/	/o/	/o/
Ø	/r/	Ø	Ø
/u/	/u/	/u/	/u/

Table 6. *The sound correspondences between the individual sounds of the words meaning ‘new’ in four Austronesian languages*

7.2.3 Reconstructing the protolanguage

As mentioned earlier, a **protolanguage** is a language that is the ancestor of all the languages of a language family or a branch of a language family. There are very few—if any—cases where the protolanguage is known in all details. Even in the case of the Romance languages discussed in 7.2.1–7.2.2, the protolanguage is not fully known. The ancestor of the Romance languages was a variety of the Latin language, but not the literary Latin language of the classical literature. The protolanguage of the Romance languages was *Vulgar Latin*, the spoken Latin of ordinary people in the Roman Empire, which differed in several respects from the literary language. For most language families in the world, the protolanguage is completely unknown, and has to be **reconstructed** from the modern languages.

We shall illustrate the reconstruction of protolanguages on the basis of some data —presented in Table 7[6]—from three Chinese «dialects», those of Beijing, Guangzhou, and Fuzhou, which in reality are different Chinese languages. The protolanguage from which the modern Chinese dialects descend is called *Ancient Chinese*, which was spoken in the period 600–1200 after Christ. The reconstructed Ancient Chinese (AC) forms are preceded by an asterisk, which is the conventional way in historical and comparative linguistics to mark an older, reconstructed language stage that is not directly based upon written sources.[7]

Meaning	Ancient Chinese	Guangzhou	Beijing	Fuzhou
‘south’	*/nam/□	/naːm/□	/nan/□	/naŋ/□
‘three’	*/sam/□	/saːm/□	/san/□	/saŋ/□
‘year’	*/nien/□	/niːn/□	/niɛn/□	/n^{j}ieŋ/□
‘mountain’	*/ʃan/□	/ʃaːn/□	/ʃan/□	/ʃaŋ/□
‘can’	*/nəŋ/□	/naŋ/□	/nəŋ/□	/neŋ/□
‘east’	*/tuŋ/□	/tuŋ/□	/tuŋ/□	/təŋ/□

Table 7. *Some Chinese words ending in a nasal*

We shall concentrate on the nasals in the coda of these monosyllabic words. In the Guangzhou dialect, we find the three nasals /m n ŋ/ in the coda, in Beijing dialect we find the two nasals /n ŋ/ in the coda, while in the Fuzhou dialect we only find the nasal /ŋ/ in the coda. Furthermore, we can establish the three different regular sound correspondences of coda nasals in Table 8, where the AC reconstructions are also included.

Ancient Chinese	Guangzhou	Beijing	Fuzhou
*/m/□	/m/□	/n/□	/ŋ/□
*/n/□	/n/□	/n/□	/ŋ/□
*/ŋ/□	/ŋ/□	/ŋ/□	/ŋ/□

Table 8. *Chinese regular coda nasal correspondences*

When we reconstruct a protolanguage, two of the most important principles governing our work may be formulated as follows:

I. The number of phonemes in the protolanguage is identical to the number of regular sound correspondences between the modern languages.
II. Reconstruct the protolanguage in such a way that the simplest and most plausible sound changes from the protolanguage to the modern languages can be postulated.

Since the number of regular coda nasal correspondences between the Guangzhou, Beijing, and Fuzhou dialects is three, we conclude—on the basis of the first principle above—that the number of coda nasal phonemes in AC was also three. Furthermore, since one of the dialects has three different nasal coda phonemes, /m n ŋ/, we postulate the same three phonemes in AC, because then—in accordance with the

[6] Tones have been left out.
[7] Ancient Chinese was indeed a written language, but the writing system, i.e., the Chinese characters, does not give much information about the pronunciation.

second principle above—we can postulate some very simple sound changes from AC to the modern dialects—cf. (3).

(3) *Some regular Chinese sound changes*

(a) AC coda */m/ > Beijing /n/
(b) AC coda */m/ and */n/ > Fuzhou /ŋ/

No sound changes are required for Guangzhou in this case. In Beijing, AC coda */m/ has become /n/, while in Fuzhou, AC coda */m/ and */n/ have become /ŋ/.

7.2.4 Shared innovations and family trees

Sound changes can be used as data for the drawing family trees. Let us first formulate the sound changes in (3) in a slightly different way; cf. (4).

(4) *Some regular Chinese sound changes*

(a) AC coda */m/ > Proto-Beijing-Fuzhou /n/
(b) Proto-Beijing-Fuzhou coda */n/ > Fuzhou /ŋ/

In (4) we have assumed, like before, that the coda nasals of AC were like those of Guangzhou. But now we have also assumed that AC first split up into two branches: Proto-Beijing-Fuzhou and Guangzhou. What distinguishes Proto-Beijing-Fuzhou from Guangzhou is sound change (4a). Later, Proto-Beijing-Fuzhou split into two branches: Beijing and Fuzhou. What distinguishes Fuzhou from Beijing is sound change (4b), which describes how all occurrences of Proto-Beijing-Fuzhou coda */n/ changed into Fuzhou /ŋ/. This is described step by step in TABLE 8.

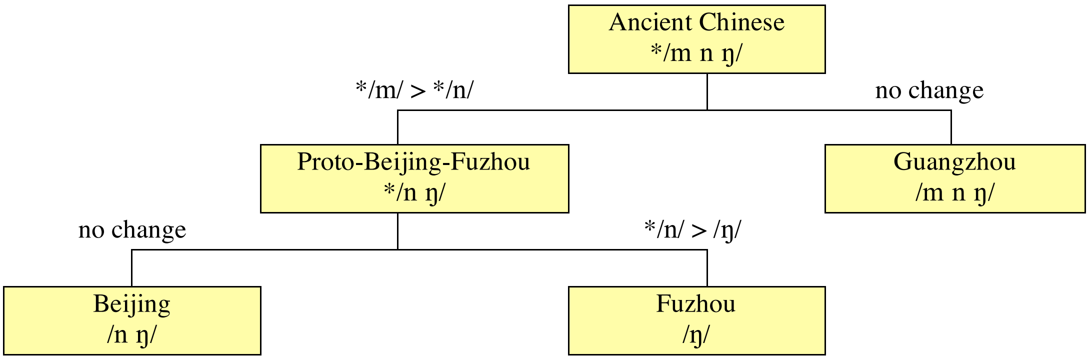


TABLE 8. *A hypothetical Chinese family tree*

The branching structure of this family tree is based upon **shared changes**. If a group of languages share a certain change, this is an indication that those languages were still one language at the time of that change. Those changes also distinguish this group from related languages.

7.3 Beyond the comparative method

7.3.1 Families of families

Linguists do not agree about the number of language families in the world. Some linguists may be of the opinion that some two language groups are related—that is, that they are branches of a single family—while others reject the evidence meant to prove the relatedness, concluding that the two groups constitute separate language families. We shall see many examples of this in the next section, where we shall describe a number of different language families in the world.

One reasonable question to ask is whether all languages of the world are related, if we only go far enough into the past. On the one hand we simply do not know. After some millennia, related languages seem to become so different from each other that our research methods simply do not work any more. On the other hand, this is a question about how many times in history spoken language has been invented from scratch. We may, if we like, distinguish a *monogenesis hypothesis* from a *polygenesis hypothesis*, where the former is a hypothesis about a common ancestry of all languages and the latter is a hypothesis of language being invented in may different parts of the world. Unfortunately, we can only speculate about this question, although, admittedly, there are linguists who try to reconstruct the «Proto-World» language, and who claim to have found some elements that recur in languages in all parts of the world.

7.3.2 Are there global cognates?

7.3.3 Genetic and linguistic classifications

7.4 Overview of language families

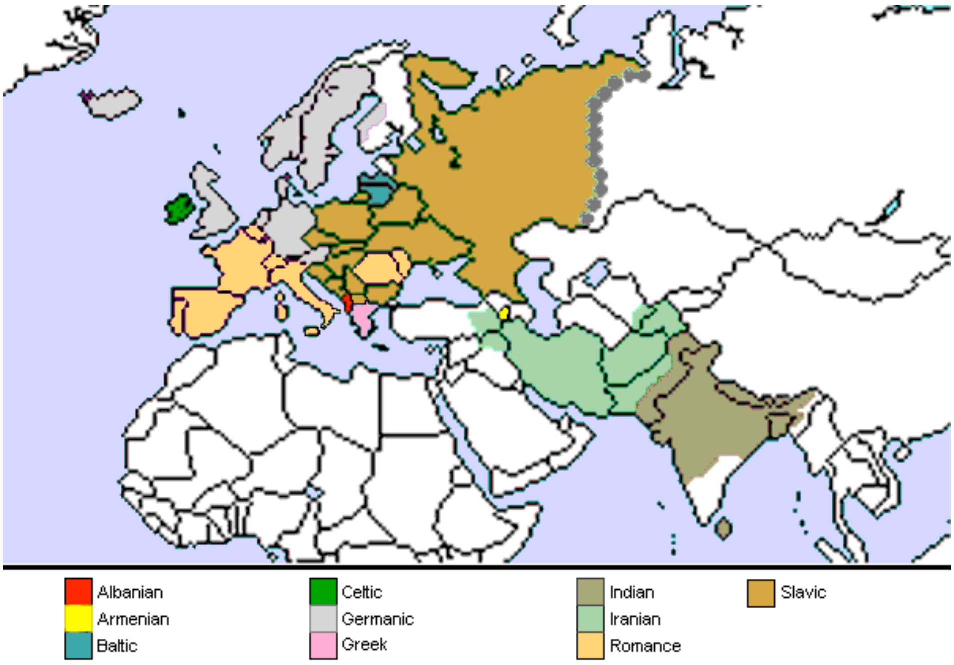
In this part of the chapter we shall give short descriptions of some of the most important language families in the world. There is no general agreement about the number of families or their internal structures. We shall, however, try to express traditional mainstream views.

We shall follow the common practice of using **–ic** as the suffix of words designating language families and major groups. For example, **Turkish** and **Tibetan** are individual languages, while **Turkic** and **Tibetic** are a language family and a major branch of the Sino-Tibetan language family, respectively.

7.4.1 Indo-European

With 443 languages, the **Indo-European** language family is one of the largest language families in the world. It has ten branches of living language, three of which are primarily spoken in Asia: **Armenian**, **Iranian**, and **Indo-Aryan** (also called **Indic**); cf. MAP 7. Here we shall concentrate on these three branches. Iranian and Indo-Aryan are usually treated as two sub-branches of an **Indo-Iranian** branch, on the basis of a large set of shared innovations; for example, Proto-Indo-European long and short */e a o/ all appear as long and short /a/ in Indo-Iranian.

Due to European colonization, **Germanic**, **Romance** languages have also spread to Asia and Africa, and **Slavic** languages to Asia.



MAP 7. *The Indo-European language family*

ARMENIAN

The **Armenian** language constitutes a separate branch of the Indo-European language family. There are approximately 6 millions first language speakers, a little over the half of which live in Armenia. The rest live in 29 other countries around the world, with the largest groups in Azerbaijan, Cyprus, Iran, Iraq, Israel, Jordan, Lebanon, Syria, and Turkey. There are many different dialects, but they are all inherently intelligible. There are two slightly different written varieties, *East Armenian*, based on the dialect of Yerevan, the capital of Armenia, and *West Armenian*, based on the dialect of Istanbul.

IRANIAN

The 84 **Iranian** languages are spoken in a vast continuous area from Pakistan and the Xinjiang province of China in the east and into Turkey in the west, with Iran in the middle. The six most important members of the group are:

Balochi (3 separate languages, mainly spoken in Pakistan, but also in neighboring countries: *Eastern Balochi*, 1.8 million speakers; *Western Balochi*, 3.4 million speakers; *Southern Balochi*, 1.8 million speakers),

Kurdish (2 separate languages: *Kurdi*, more than 6 million speakers in Iran and Iraq; *Kurmanji*, more 4 million speakers in Turkey, 1 million speakers in Syria, and smaller groups in Armenia, Azerbaijan, Iraq, and Iran),

Osetin (or *Ossete, Ossetic*; approximately 600 000 speakers, out of which 164 000 in Georgia, and the rest in neighboring countries),

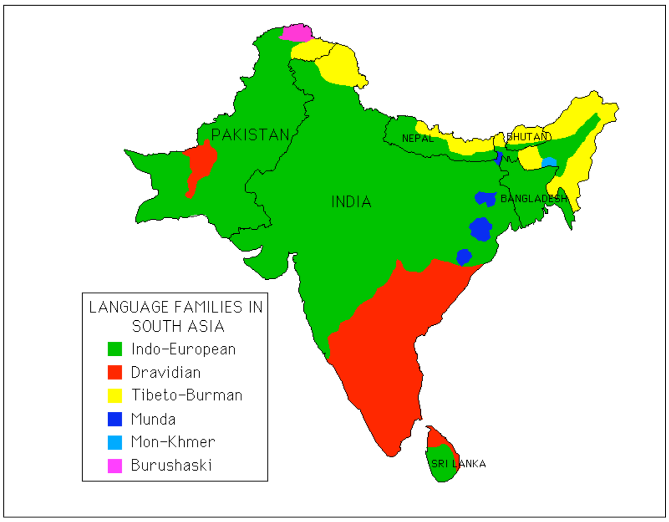
Pashtu (3 separate languages: *Northern Pashto*, 10 million speakers, mainly in Pakistan; *Southern Pashto*, 10 million speakers, mainly in Afghanistan; *Central Pashto*, in Pakistan, the number of speakers is not known),

Persian (or *Farsi*; two main varieties: *Western Persian*, 22 million speakers in Iran and more than 2 million speakers around the world; *Eastern Persian* or *Dari*, approaching 6 million speakers in Afghanistan and 1 million in Pakistan),

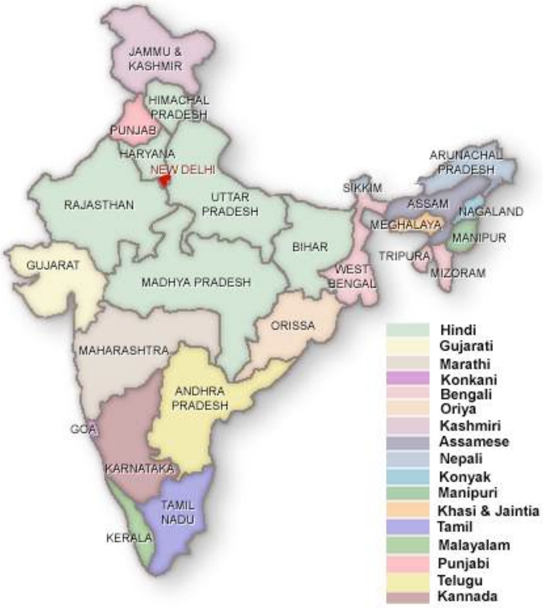
and **Tajiki** (3.5 million speakers in Tajikistan and 1 million in neighboring countries).

INDO-ARYAN

The 210 **Indo-Aryan** languages are primarily spoken in the countries of the Indian subcontinent, in practically the whole green area of MAP 8, with the exception of the Iranian-speaking areas of western Pakistan.



MAP 8. *The language families of India*



MAP 9. *Some major Indian languages*

Some major Indo-Aryan languages are shown on MAP 9,[8] among others:

Assamese (more than 15 million speakers, primarily in the Indian states of Assam, Meghalaya, and Arunachal Pradesh, but also in Bangladesh and Bhutan),

Bengali (207 million first language speakers, primarily in Bangladesh and in the Indian state of West Bengal),

Gujarathi (45.5 millions in the Indian states of Gujarat, Maharashtra, Rajasthan, Karnataka, Madhya Pradesh, and more than half a million speakers around the world),

Hindi (more than 180 million first language speakers throughout northern India: Delhi, Uttar Pradesh, Rajasthan, Punjab, Madhya Pradesh, northern Bihar, and Himachal Pradesh. The total number of first language speakers around the world is 366 millions and more than 120 million second language speakers.

Marathi (68 million first language speakers in the Indian state of Maharashtra and adjacent states and 3 million second language speakers),

Oriya (more than 32 million speakers in the Indian states of Orissa, Bihar, West Bengal, Assam, and Andhra Pradesh. Also spoken in Bangladesh),

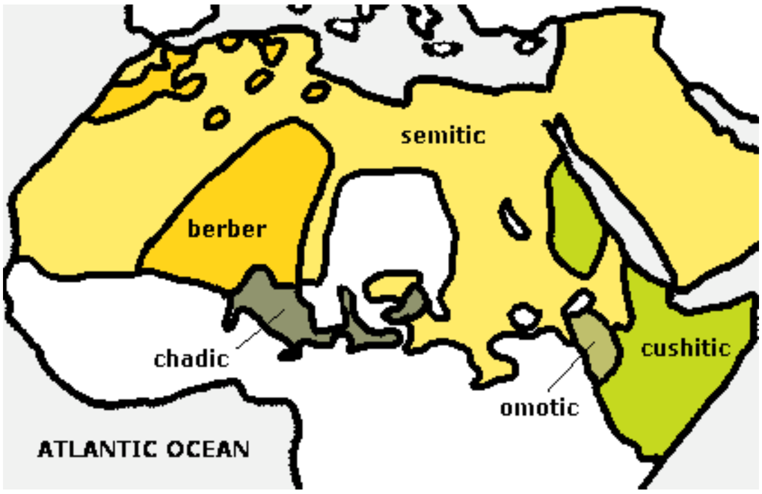
Punjabi (or *Panjabi*; 27 million speakers of *Eastern Punjabi* in the northwestern India, primarily in Punjab, but also in Rajasthan, Haryana, Delhi, and Jammu and Kashmir; probably around 45 million speakers of *Western Punjabi* in the Punjab area of Pakistan),

Urdu (very close to Hindi, but with a formal vocabulary borrowed from Arabic and Persian, while Hindi is de-Persianized and de-Arabicized, and borrows formal vocabulary from Sanskrit; more than 11 million first language speakers in Pakistan, 48 millions in India (Jammu and Kashmir and Muslims throughout the country), and a few millions around the world; close to 45 million second language speakers).

In addition, we should mention **Sinhalese** (more than 13 million speakers in Sri Lanka).

7.4.2 Afroasiatic

The **Afro-Asiatic** language family has the following five branches of living languages, all of which are shown on MAP 10: **Berber, Chadic, Cushitic, Omotic**, and **Semitic**. In addition, the extinct **Egyptian** language was a branch of this family. Earlier, this family was usually called *Hamito-Semitic*.



[8] The languages of southern India are **Dravidian**; cf. § 4.3.5.

Map 10. *The Afro-Asiatic language family*

Berber

The 26 **Berber** languages are distributed all over North Africa, from the Siwa Oasis in Egypt in the east to Senegal in the west, from Algeria in the north to Mali in the south; cf. Map 9. The most important Berber languages are:

Taqbaylit (or *Kabyle*; 2.5 million speakers in the Kabylia region of Algerie and half a million in other countries),

Tamasheq (may be four different languages: *Tamahaq*, 62 000 speakers in southern Algeria and adjacent areas; *Tamajaq*, 640 000 speakers in Niger and adjacent areas; *Tamajeq*, 250 000 speakers in Niger; *Tamasheq*, more than 250 000 in Mali),

Tamazight (3.5 million speakers in the Middle Atlas of Morocco and adjacent areas),

Tarifit (or *Rifi*; 2 million speakers, mainly in northern Morocco),

and **Tashelhiyt** (or *Shilha*; 3.5 million speakers, mainly in southwestern Morocco).

Chadic

There are altogether 195 **Chadic** languages, which are primarily spoken in Niger, Nigeria, Chad, Central African Republic, Sudan, Cameroon, and parts of Togo, Benin, and Ghana; cf. Map 9. More than 85% of all those who speak Chadic languages speak **Hausa**, which is spoken by 25 million first language speakers (20 million in Nigeria, 5 million in Niger, half million in Sudan) and 15 million second language speakers.

Cushitic

The 47 **Cushitic** languages are spoken in the Horn of Africa and along the Red Sea; cf. Map 9. The most important ones are:

Somali (more than 7 million speakers in Somalia, 3 millions in Ethiopia, and smaller groups in Djibouti and Kenya),

Oromo (more than 3.6 million speakers in the South Oromo Region in Ethiopia and smaller groups in Somalia and Kenya),

Sidamo (1.8 million speakers in south central Ethiopia),

and **Bedawi** (or *Beja*; probably around 1 million speakers in northwestern Sudan, along the Red Sea coast, and 120 000 in Eritrea).

Omotic

The 28 **Omotic** languages are spoken in southwestern Ethiopia. The most important ones are **Wolaytta** (1.2 million speakers), **Gamo-Gofa-Dawro** (1.2 million speakers), and **Kafa** (570 000 speakers).

Semitic

The living **Semitic** languages can be divided into two main branches, **North-west Semitic** and **South Semitic**, but there is no general agreement about all details in the classification.

North-west Semitic

The most important North-west Semitic languages are:

Arabic, with a large number of regional varieties, is spoken in Morocco, Algeria, Mauritania, Tunisia, Malta, Libya, Egypt, Sudan, Djibouti, Somalia, Saudi Arabia, Kuwait, Bahrain, Qatar, the United Arab Emirates, Oman, Yemen, Jordan, Syria, Iraq, Lebanon, and Israel/Palestine. The number of first language speakers exceeds 200 million. Modern Standard Arabic is the written form used in all these countries except Malta.

Aramaic has several varieties, the most important of which is **Assyrian Neo-Aramaic**, with 30 000 speakers in Iraq and 80 000 speakers in 25 countries around the world.

Hebrew, in its modern revived form, is spoken by around 5 million people, primarily in Israel.

South Semitic

South Semitic comprises *South Arabian* and *Ethio-Semitic*.

Among the *South Arabian* languages are **Soqotri** (70 000 speakers, mainly in Soqotra Island) and **Mehri** (58 000) in Yemen.

The *Ethio-Semitic* languages are spoken in Eritrea and Ethiopia. The most important ones are **Amharic** (more than 17 million first language speakers and 5 million second language speakers in Ethiopia) and **Tigrinya** (3.2 million speakers in Ethiopia and 2 million speakers in Eritrea).

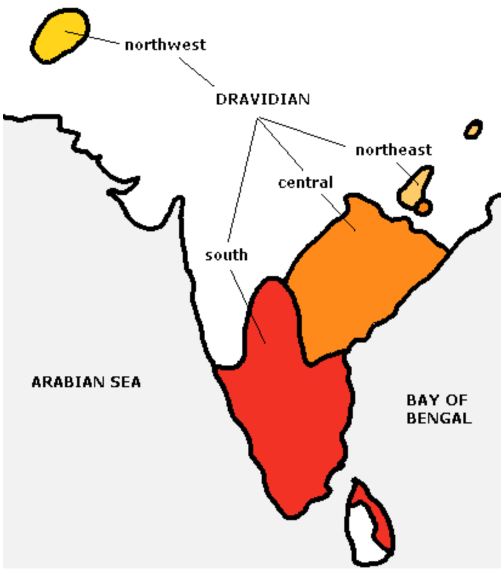
Egyptian

The old **Egyptian** language, with four and a half millennia of written records, was spoken in Egypt until the fourteenth century.

7.4.3 Dravidian

The 23 **Dravidian** languages are spoken in Afghanistan, Bangladesh, India, Nepal, Pakistan, and Sri Lanka. The family is divided into three branches: **South Dravidian**, **Central Dravidian**, and **North Dravidian**; cf. Map 6.

South Dravidian



Map 6. *The Dravidian language family*

The most important among the nine South Dravidian languages are **Malayalam** (35 million speakers in Kerala and Laccadive Islands), **Tamil** (66 million first language speakers and 8 million second language speakers in Tamil Nadu and neighboring states, including 3 millions in Sri Lanka and a few millions around the world), and **Kannada** (also called *Kanarese*; 35 million first language speakers and 9 million second language speakers in

Karnataka, Andhra Pradesh, Tamil Nadu, and Maharashtra).

Central Dravidian

There are twelve Central Dravidian languages, the most important of which is **Telugu** (70 million first language speakers and 5 million second language speakers in Andhra Pradesh and neighboring states).

Northern Dravidian

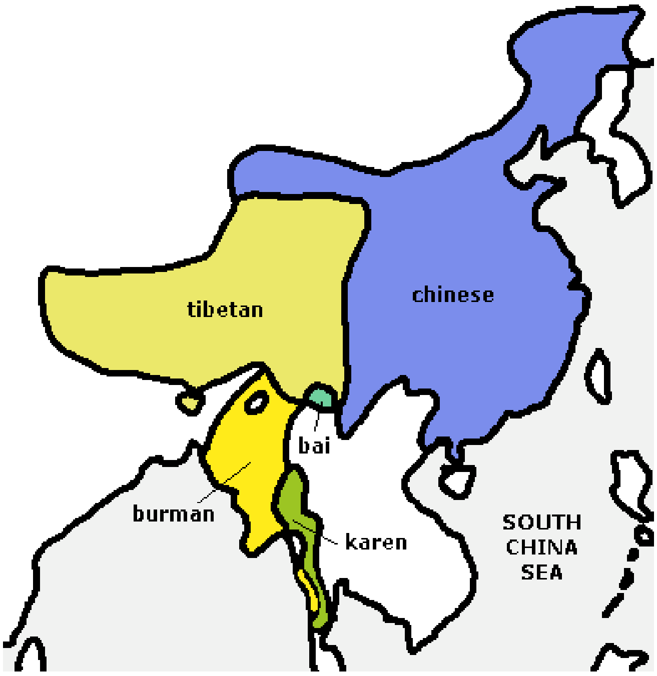
There are three Northern Dravidian languages, **Brahui** (2 million speakers in Pakistan and 200 000 in Afghanistan), **Kurukh** (2 million speakers eastern Indian and Bangladesh; there is another Northern Dravidian language called Kurukh in Nepal), and **Malto** (80 000 speakers in eastern India).

7.4.4 Sino-Tibetan

The **Sino-Tibetan** family has around 300 members, and the main branches are **Tibetic**, **Burmic**, **Bai**, **Karenic**, and **Sinitic**. The distribution of these branches is shown on Map 1. A short description of Tibetic, Burmic, and Sinitic follows.

Tibetic

Tibetic or **Bodic** (from *Bod*, the Tibetan name for Tibet), is divided into at least three different branches, the most important of which is **Bodish-Himalayish**, where we find **Tibetan**, the major Tibetic language, spoken by approximately 1,25 million people, primarily in Tibet, but also in Bhutan, Nepal, and India. Another Tibetic language, **Newari**, is spoken by around 700 000 people in Nepal and a few in India.



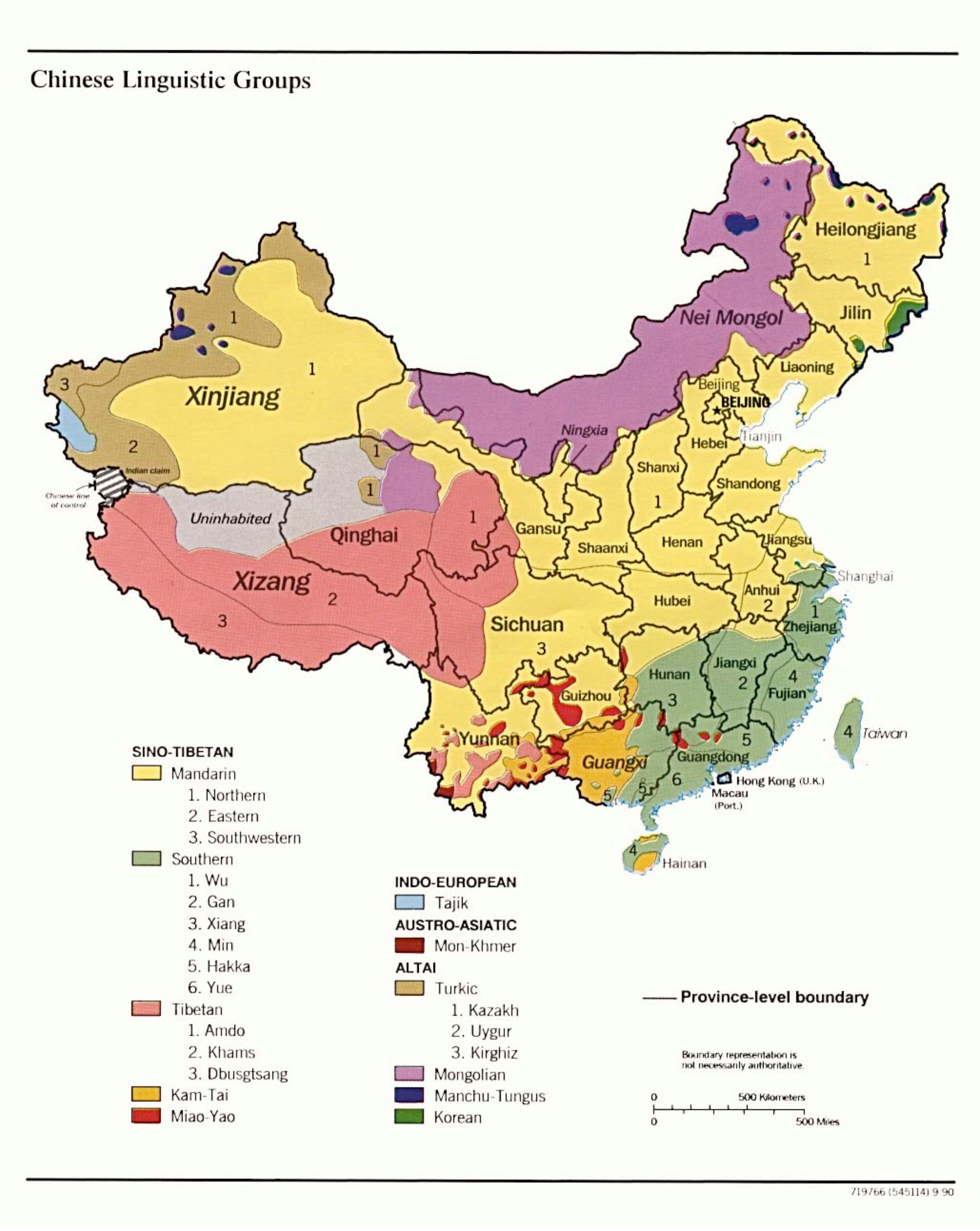
Map 1. *The Sino-Tibetan language family*

Burmic

Burmic has three branches, and the most important is **Burmish** or **Burmese-Lolo**, which includes **Burmese**, the most important Burmic language, spoken by approximately 32 million people, two thirds of which live in Myanmar (formerly Burma), and the rest primarily in Bangladesh and Thailand.

Sinitic

Sinitic is the same as Chinese, but the term Sinitic may be the preferable designation, since in emphasizes the fact that this is a group of languages, not a single language. The major divisions of Sinitic—cf. MAP 2—are **Mandarin** (Northern Chinese; more than 867 million speakers), **Wu** (including the Shanghai and Zhejiang dialects; around 80 million speakers), **Xiang** (Hunanese; around 40 million speakers), **Gan** (including the Jiangxi dialect; between 20 and 25 million speakers), **Kejia** (or Hakka, spoken by around 30 million people in large scattered areas of Guangxi and Guangdong), **Yue** (Cantonese; more than 71 million speakers in all countries, probably close to 52 million speakers in China), and **Min** (Fukienese; the number of speakers is probably close to 40 millions).

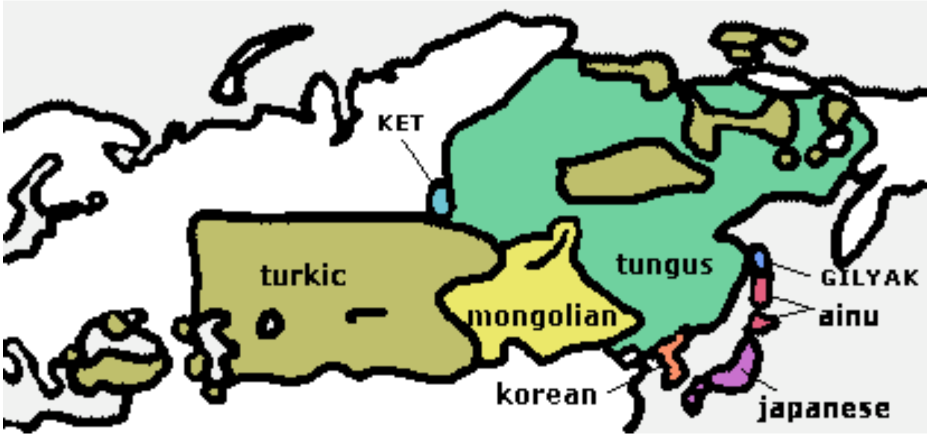


MAP 2. *The language groups of China*

7.4.5 Altaic

The **Altaic** family has three branches: **Turkic**, **Mongolian** and **Tungus**—cf. MAP 4. But the dominating view these days is that these are independent language families that only share some striking typological characteristics, like vowel harmony, an agglutinating morphological structure, and SOV sentence structure.

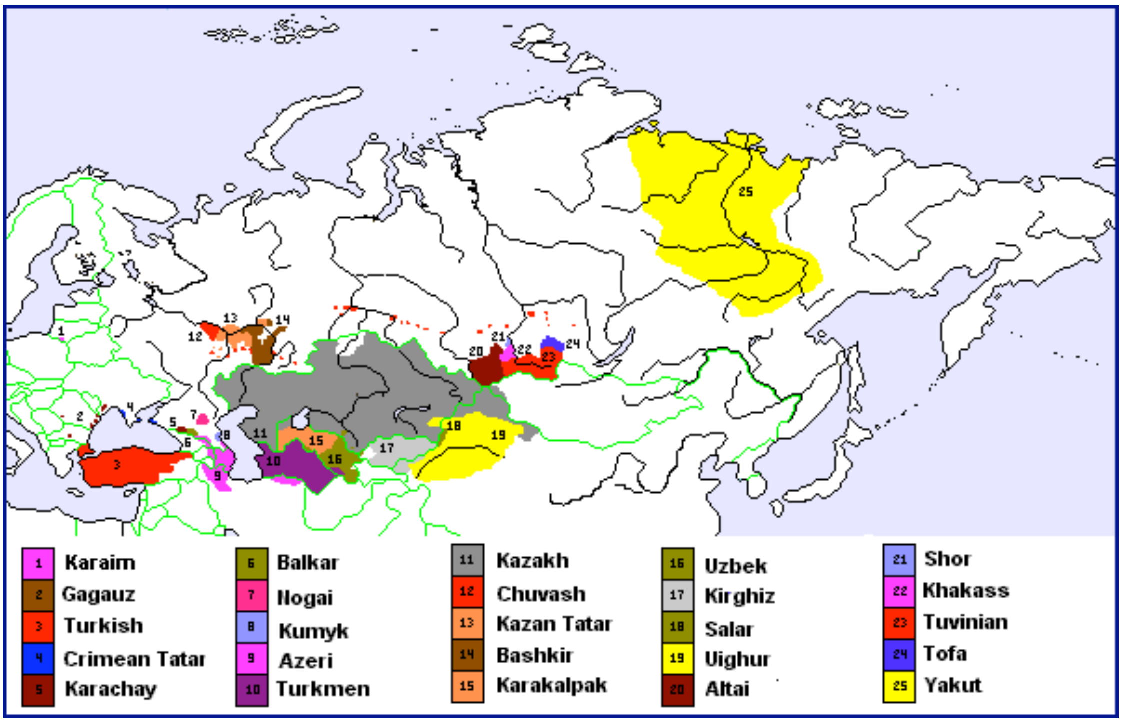
The two latter features are also shared by Japanese and Korean, which are therefore also sometimes claimed to belong to this hypothetical Altaic language family. There are even traces of vowel harmony in older stages of both languages.



MAP 4. *Altaic languages*

In the following we shall primarily be concerned with Turkic.

The **Turkic** language family has a wide distribution on the Eurasian continent, as can be seen from MAP 3.



MAP 3. *The Turkic language family*

The family comprises 25 languages, whose location and distribution is shown on the map. Turkic has two main branches, **r-Turkic** (with one language, **Chuvash**, number 12 on the map, spoken by close to 2 million people) and **z-Turkic** or **Common Turkic**, including all the other languages.

Z-Turkic is divided into four branches, **Southeast Turkic**, **Southwest Turkic**, **Northwest Turkic**, and **Northeast Turkic**.

The major Northeast Turkic language is **Tuvinian** (233.000 speakers, mainly in Tuvin AO in Russia). The other branches are described briefly below.

Southeast Turkic

There are two Southeast Turkic languages: **Uighur** (7.5 million speakers, mainly in China) and **Uzbek** (18.5 million speakers, mainly in Uzbekistan).

Southwest Turkic

There are four Southwest Turkic languages: **Turkish** (61 million speakers, primarily in Turkey), **Azerbaijani** (7 million speakers, primarily in Azerbaijan), **Turkmen** (6.4 million speakers, primarily in Turkmenistan), and **Gagauz** (200.000 speakers, primarily in Moldova).

Northwest turkic

Among the nine Northwest Turkish languages, the most important ones are: **Tatar** (7 million speakers, mainly in Tatarstan in Russia), **Kazakh** (8 million speakers, mainly in Kazakhstan), and **Kirgiz** (2.6 million speakers, mainly in Kyrgyzstan).

7.4.6 Japanese and Korean

On the basis of our present knowledge, the most sober conclusion to draw is that Japanese and Korean are language **isolates**, meaning that they both constitute a language family of their own. There is a widespread belief that Japanese and Korean are related.

The structural similarities between these two languages are striking—we have already mentioned, in 7.4.5, their «Altaic» structural features: they are both agglutinating languages with SOV syntax. Furthermore, both are «honorific languages»: **degrees of politeness** are expressed in grammar and lexicon to a larger extent than usual in the languages of the world.

Japanese

Japanese is the first language of 99,5% of the 127,4 million inhabitants of Japan[9] (2002) and several millions worldwide. There are two major dialect groups, *mainland Japanese* and *Ryukyuan*. Several Ryukyuan dialects differ so much from each other

[9] Japan has one ethnic minority, the 15 000 **Ainu** of Hokkaido. The Ainu language, another language isolate, is nearly extinct, being spoken by 15 persons in 1996. 0.5% of the Japanese population are Koreans.

and from mainland Japanese that they may be regarded as separate languages of the Japanese language family.

Korean

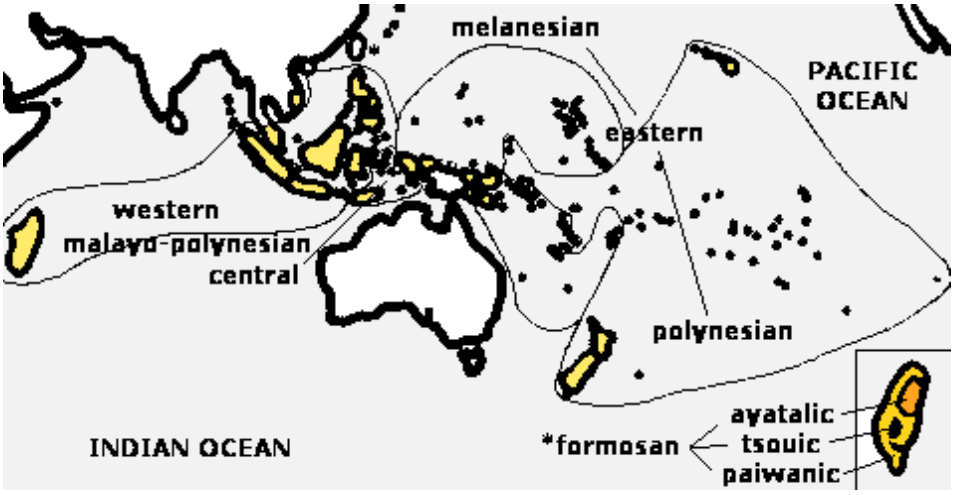
Korean is spoken by the whole populations of North and South Korea— around 70 million people—and by several millions around the world. The Koran peninsula may be divided into seven dialectal zones, but the language is relatively homogeneous, and all dialects are mutually intelligible.

7.4.7 Uralic

Uralic

7.4.8 Austronesian

There are more than 1 200 different Austronesian languages, spoken from Easter Island in the East to Madagascar in the west; cf. MAP 5. The family can be divided into major branches, **Western Austronesian** and **Eastern Austronesian**, also called **Oceanic**. One possible classification of this family was shown in FIG. 2 at the beginning of this chapter. There are two main points of disagreement concerning the classification. First, some scholars claim that the **Formosan** is not a branch of Western Austronesian, but constitutes a third branch of the family. Secondly, the languages of Melanesia differ considerably from other Austronesian languages, and there position within the family is uncertain.



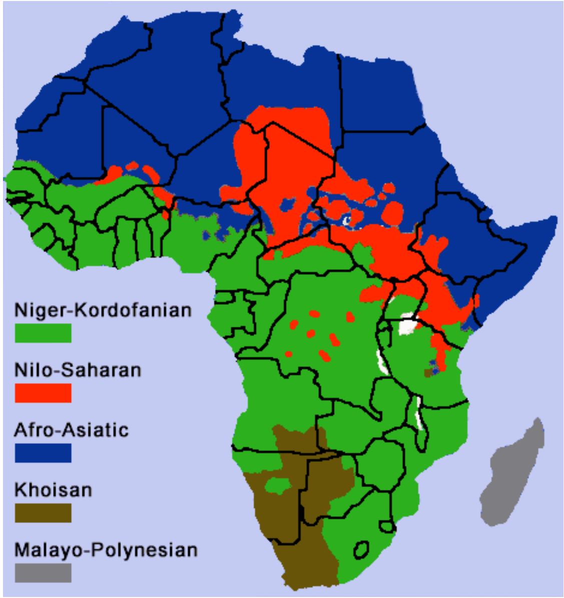
MAP 5. *The Austronesian language family*

Western Austronesian

Many of the 200 different **Western Austronesian** languages are spoken by millions of speakers. The most important ones are **Malay** (more than 18 million speakers), **Indonesian** (around 30 million first language speakers and more than 140 million second language speakers), **Javanese** (75 millions), **Sundanese** (27 millions), **Malagasy** (Madagascar; 15 millions), **Tagalog** (Philippines; 17 million first language speakers and 40 million second language speakers), and **Buginese** (Sulawesi; 4 millions). Indonesian or *Bahasa Indonésia* is really a variety of Malay that was selected as the national language of Indonesia.

Oceanic or Eastern Austronesian

While the Western Austronesian languages are spoken by more than 200 million people, the 300 **Oceanic** languages are spoken by less than 2 million people. Most of these languages are spoken outside Asia (and Africa), and will not be treated in any detail here. Let us only mention the well-known **Polynesian** sub-branch, including languages like **Samoan** (426 000 speakers), **Tongan** (125 000 speakers), **Tahitian** (125,000), **Maori** (50–70 000 speakers), and **Hawai'ian** (1 000 first language speakers, 8 000 second language speakers).



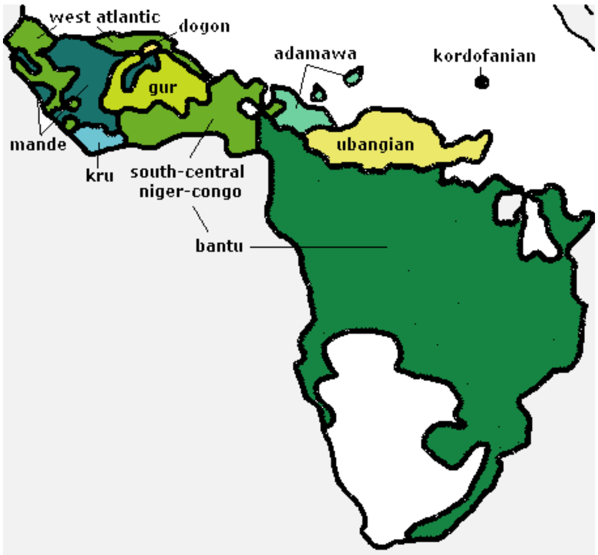
MAP 11. *African language families*

7.4.9 Niger-Congo

With its 1 436 languages—according to the most recent estimates—the **Niger-Congo** or **Niger-Kordofanian** language family is the largest in the world, and it occupies a greater part of the African continent than any other family—cf. MAP 11. Although the family contains several well-defined branches, there is no general agreement about the overall structure. We have chosen a classification with seven main branches: **Kordofanian**, **Mande**, **Atlantic**, **Ijoid**, **Dogon**, **North Volta-Congo**, and **South Volta-Congo**. The branches are described below, except the small Ijoid and Dogon branches.

Kordofanian

The 20 **Kordofanian** languages are spoken by small groups in the Nuba mountains in Sudan, for example **Koalib** (44 000 in 1984) and **Tegali** (35 000 in 1984). Cf. MAP 12.



MAP 12. *The Niger-Congo language family*

Mande

The **Mande** languages are spoken over a great part of the western half of West Africa, in Mali, Côte d'Ivoire, Guinea, Sierra Leone, and Liberia; they are also spoken in parts of the neighboring countries Burkina Faso, Senegal, Gambia, and Guinea Bissau. There are between 10 and 12 million speakers of around 35 different Mande languages, but over half speak forms of **Manding**, a widespread dialect cluster known under different names in different parts of West Africa: **Bambara** (or *Bamanankan*; 2.7 million first language speakers in Mali, but 80% of the population of 11 million speak it in varying degrees), **Jula** (or *Dyula*; 2.5 million first language speakers and 3 to 4 second language speakers, mainly in Burkina Faso and northern Côte d'Ivoire), and **Mandinka** (1.2 million speakers in Senegal, Gambia, and Guinea Bissau).

ATLANTIC

The **Atlantic** languages are a quite heterogeneous group of languages, most of which are spoken along the Atlantic coastline of West Africa, from the mouth of the Senegal River as far as Liberia. The most important languages are:

Fula or *Fulfulde, Pulaar, Pular, Fulani, Peul.* This language is the first language of at least 15 million people and at least 5 million second language speakers all over West Africa and neighboring areas, including Mauritania, Senegal, Gambia, Guinea, Mali, Burkina Faso, Benin, Togo, Niger, Nigeria, Cameroon, Chad, and Sudan.

Wolof is spoken by 3,2 million first language speakers and 4 million second language speakers in Senegal and adjacent areas (Gambia, Guinea, Guinea Bissau, Mali, and Mauritania).

Serer is spoken by 1 million people in Senegal and a few in Gambia.

Themne is spoken by 1.5 million people in Sierra Leone, 1.2 million of which are first language speakers.

NORTH VOLTA-CONGO

There are three sub-branches of the **North Volta-Congo** languages: **Kru, Gur,** and **Adamawa-Ubangi**.

The **Kru** languages are spoken in Liberia and south-west Côte d'Ivoire, by between 1 and 2 million people.

The **Gur** languages are spoken by between more than 6 million people in the greater part of Burkina Faso and into neighboring countries (Mali, Côte d'Ivoire, Ghana, Togo, Benin, and Nigeria). The most important Gur language is **Moore**, which is spoken by more than 5 million people in Burkina Faso and adjacent areas.

The **Adamawa-Ubangi** languages are spoken by almost 4 million people in an area that extends from north-west Nigeria, through northern Cameroon, southern Chad, Central African Republic, northern Gabon, Congo-Brazzaville, Congo-Kinshasa, and south-west Sudan. Among the more important languages is **Zande** (1.1 million speakers in Congo-Brazzaville, Central African Republic, and Sudan). The main lingua franca of the Central African Republic, the pidgin/creole **Sango**, with 400 000 first language speakers and 4.5 million second language speakers, contains elements primarily from French and Adamawa-Ubangi languages.

SOUTH VOLTA-CONGO

The **South Volta-Congo** languages are spoken in the areas called *South Central Niger-Congo* and *Bantu* on MAP 12. *South Central Niger-Congo* is an alternative name of the branch, while *Bantu*—as we shall come back to—is a major sub-group. South Volta-Congo is divided into two main branches: **Kwa** and **Benue-Congo**.

KWA

The **Kwa** languages are spoken by approximately 20 million people along the Atlantic coast of West Africa from Côte d'Ivoire to Nigeria. Among the better-known Kwa languages are **Akan** (7 million speakers in Ghana, or 44% of the population) and **Ewe** (1.6 million speakers in Ghana, or 13% of the population; close to 900 000 speakers in Togo, or 20% of the population).

BENUE-CONGO

The **Benue-Congo** languages occupy an area including most of Africa south of a line drawn from south-east Nigeria and across the continent to northern Kenya; cf. MAP 12. The internal structure of the branch is complicated, and scholars disagree about several details. According to *Ethnologue*, there are 938 different Benue-Congo languages, which are probably spoken by between 150 and 200 million people.

This language group may be divided into two branches, **West Benue-Congo** and **East Benue-Congo**.

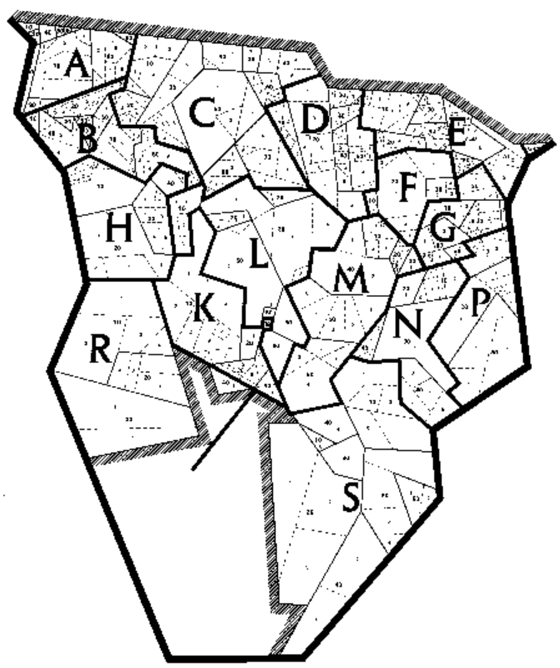
West Benue-Congo

The West Benue-Congo languages are spoken in most of southern Nigeria and into Benin. The most important languages are **Yoruba** (more than 20 million first language speakers in south-west Nigeria and 0.5 million in Benin; about 2 million second language speakers) and **Igbo** (18 million speakers in south-east Nigeria).

East Benue-Congo

The East Benue-Congo languages cover the remaining parts of the Benue-Congo area, and has three major branches: **Central Nigerian**, **Cross**, and **Bantoid**. Here, we shall concentrate on **Bantoid**.

The **Bantoid** languages are spoken in all the East Benue-Congo area except some small spots in northern Nigeria where *Central Nigerian* and *Cross* languages are found. The internal structure of *Bantoid* is very complicated, and cannot be discussed in any detail here. The disagreement among scholars is also conspicuous. Roughly, the Bantoid branch consists of the **Bantu** languages and several groups of small languages spoken in the Nigerian-Cameroon borderland. The largest non-Bantu Bantoid language is probably **Tiv** (2.2 million speakers in eastern Nigeria).



MAP 13. *Guthrie's Bantu language zones*

There are more than 400 different **Bantu** languages, spoken in a vast area shown on MAP 12 and MAP 13. The is no agreement among scholars about how to divide the Bantu group into sub-groups, and a geographically based classification constructed by the British bantuist Malcolm Guthrie is usually applied for practical purposes—cf. MAP 13. All Bantu languages have a code in this classification, consisting of a letter indicating the geographical area and a number within that area. For example, Swahili, has the code G.40. Here follows a list of some important Bantu languages (the Guthrie codes are included):

G.40 Swahili; the first language of around 5 million people, primarily in the coastal areas of Kenya and Tanzania. 30 million second language users all over East Africa.

H.10 Kongo; spoken by 1 million people in western Democratic Republic of Congo (Congo-Kinshasa), 1.2 million people in north-western Angola, and probably around 1 million in People's Republic of the Congo (Congo-Brazzaville).

H.20 Mbundu; spoken by 3 million people in Angola, or 25% of the population.

M.40 Bemba; spoken by 2 million people in Zambia and smaller groups in neighboring countries.

P.30 Makhuwa; four languages—**Makhuwa**, **Makhuwa-Marrevone**, **Makhuwa-Meetto**, **Makhuwa-Shirima**—spoken by around 5 million people in Mozambique.

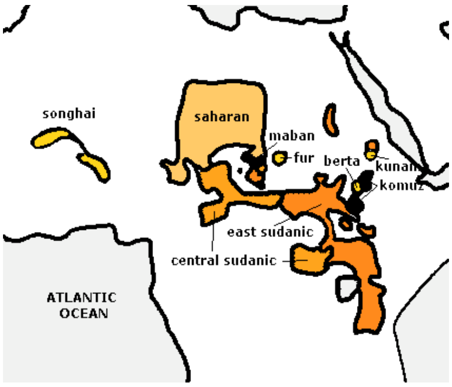
R.10 Umbundu; spoken by 4 million people in Angola, or 38% of the population.

S.10 Shona; spoken by 7 million people in Zimbabwe, 0.5 million in Mozambique, and smaller groups in Botswana, Malawi, and Zambia.

S.30 Sotho-Tswana; divided into three mutually intelligible dialects with their own written standards: **Southern Sotho** (2.7 million speakers in South Africa), **Northern Sotho** (3.8 million speakers in South Africa), and **Tswana** (2.8 million speakers in South Africa and 1 million in Botswana).

S.40 Nguni; divided into four mutually intelligible dialects with their own written standards: **Ndebele** (1.5 million speakers in Zimbabwe and 10 000 in Botswana), **Swati** (or *Swazi*; 1 million speakers in South Africa and 650 000 in Swaziland)**, Xhosa** (6.8 million speakers in south Africa and smaller groups in Botswana and Lesotho)**, Zulu** (close to 9 million speakers in South Africa and smaller groups in Botswana, Lesotho, Malawi, Mozambique, Swaziland).

7.4.10 Nilo-Saharan



MAP 14. *The Nilo-Saharan languages*

The **Nilo-Saharan** languages are spoken by more than 30 million people in fifteen African countries, from Tanzania in the east as far as Mali in the west; cf. MAP 14. According to M. Lionel Bender, the leading expert on Nilo-Saharan studies, the family has four branches: **Songay** (or *Songhai*), **Saharan**, **Kuliak**, and **Satellite-Core**. *Songay* and *Kuliak* will not be discussed here.

SAHARAN

The **Saharan** languages are spoken from Lake Chad and northwards into Sahara and eastwards into Sudan. The most important Saharan language is **Kanuri**, spoken by 3 million speakers in north-east Nigeria and 0.5 million speakers in Niger, Cameroon, Chad, and Sudan.

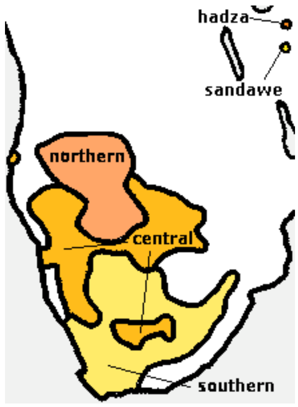
SATELLITE-CORE

The **Satellite-Core** group has six branches: **Maban**, **Fur**, **Central Sudanic**, **Berta**, **Kunama**, and **Core**. The five first branches contain lots of small languages that cannot be mentioned here.

On the other hand, **Core** contains four sub-branches: **East Sudanic**, **Koman**, **Gumuz**, and **Kadu**; only the first one, *East Sudanic*, contains languages of any size.

One sub-branch of a sub-branch of the **East Sudanic** group is a large group of 52 languages[10] known as **Nilotic**, spoken in northern Tanzania, western Kenya, Uganda, and southern Sudan. Here are some of the better-known Nilotic languages:

Dinka; more than 2 million speakers in southern Sudan.
Kalenjin; 2.5 million people in Kenya.
Luo; 3.2 million speakers in Kenya and 223 000 in Tanzania.
Maasai (or *Maa*); 430 000 speakers in Tanzania and 450 000 in Kenya.
Teso; 1 million speakers in Uganda and 217 000 in Kenya.



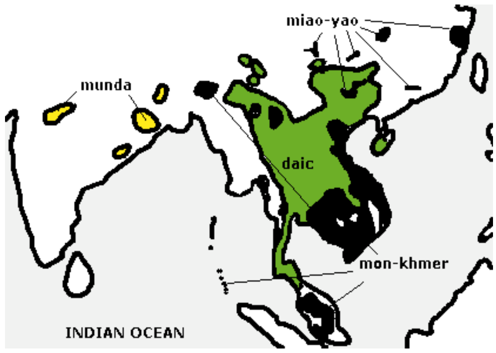
MAP 16. *Khoisan languages*

7.4.11 Khoi-San

The **Khoisan** family are the «click languages» of southern Africa, traditionally known as «Hottentot and Bushman languages». This hypothetical family may consist of not less than 5 unrelated languages families. The largest language is **Nama**, spoken by 176 000 people in Namibia and 56 000 in South Africa.

7.4.12 Austro-Asiatic

The **Vietnamese** language, spoken by 68 million people, belongs to the **Mon-Khmer** family, which is shown on MAP 15. Some scholars regard Mon-Khmer as a branch of a larger **Austro-Asiatic** super-family that also contains the **Munda** languages in India (including **Santali**, spoken by 6 million people in India and neighboring countries) and **Nicobarese**.



MAP 15. *Some Asian language families*

[10] 52 languages according to *Ethnologue*.

7.4.13 Australian

7.4.14 Indo-Pacific

7.4.15 Eskimo-Aleut

7.4.16 Na-Dené

7.4.17 Amerind