Chapter 16
Comparative method: Establishing language relationship

> The Sanscrit language, whatever be its antiquity, is of a wonderful structure; more perfect than the Greek, more copious than the Latin, and more exquisitely refined than either, yet bearing to both of them a stronger affinity, both in the roots of verbs and in the forms of grammar, than could possibly have been produced by accident; so strong indeed, that no philologer could examine them all three, without believing them to have sprung from some common source, which, perhaps, no longer exists: there is a similar reason, though not quite so forcible, for supposing that both the Gothick and the Celtick, though blended with a very different idiom, had the same origin with the Sanscrit; and the old Persian might be added to the same family, if this were the place for discussing any question concerning the antiquities of Persia.
> *(Sir William Jones, Third Anniversary Discourse, on the Hindus, Royal Asiatic Society, 1786)*

1. Introduction

The epigraph above, which readers will remember from Chapter 2, has had a double significance for the history of linguistics. On one hand, it provided one of the most important stimuli for research in comparative Indo-European linguistics, a field which soon became the most thoroughly investigated area of historical and comparative linguistics and which to the present has remained the most important source for our understanding of linguistic change. This is the issue that we pursued in Chapter 2.

On the other hand, Jones's statement is important because, perhaps for the first time, it offered a succinct and explicit summary of what have turned out to be the basic assumptions and motivations of comparative linguistics: accounting for similarities which cannot be attributed to chance, by the assumption that they are the result of descent from a common ancestor.

To establish this kind of account we must naturally look for languages that seem to share enough similarities to suggest that there may be a genetic rela-

tionship. In many cases, this is not all too difficult, once we accept the basic notion that languages may be genetically related to each other. To illustrate the point, consider Table 1 below.

Table 1: Vocabulary correspondences in the major European languages

	'one'	**'two'**	**'three'**	**'head'**	**'ear'**	**'mouth'**	**'nose'**
Breton	*ünan*	*dau*	*tri*	*penn*	*skuarn*	*genu*	*fri*
Welsh	*ɨn*	*dai*	*tri*	*pen*	*klɨst*	*keg*	*truɨn*
Irish	*ō̃n*	*dɔ*	*tri*	*kyan*	*kluəs*	*byal*	*srōn*
Icelandic	*eidn*	*tveir*	*þrīr*	*hō̃fuð*	*eira*	*münnür*	*nēf*
Danish	*en*	*tōʔ*	*trēʔ*	*hōðə*	*ōrə*	*monʔ*	*nāsə*
Norwegian	*ēn*	*tō*	*trē*	*hōvəd*	*ō̃rə*	*mund*	*nāsə*
Swedish	*ēn*	*tvō*	*trē*	*hō̃vud*	*ō̃ra*	*mun*	*nāsa*
Dutch	*ēn*	*tvē*	*drī*	*hōft*	*ōr*	*mont*	*nō̃s*
English	*wən*	*tUw*	*θrIy*	*hɛd*	*Iyr*	*mawθ*	*nowz*
German	*ʔains*	*tsvai*	*drai*	*kɔpf*	*ʔōr*	*munt*	*nāzə*
French	*œ̃/ün*	*dö*	*trwa*	*tēt*	*orēy*	*buš*	*ne*
Spanish	*uno*	*dos*	*tres*	*kaβeθa*	*orexa*	*boka*	*nariθ*
Portuguese	*ũ*	*doš*	*treš*	*kəbesə*	*oreʎa*	*bokə*	*nariz*
Italian	*un(o)*	*due*	*tre*	*testa*	*orɛkkyo*	*bokka*	*naso*
Rumanian	*un*	*doy*	*trey*	*kap*	*ureke*	*gurə*	*nas*
Albanian	*ñə*	*dü*	*tre*	*kokə*	*veš*	*goyə*	*hundə*
Greek	*énas*	*ðyó*	*trís*	*kefalí*	*aftí*	*stóma*	*míti*
Bulgarian	*yedan*	*dva*	*tri*	*glava*	*uxo*	*usta*	*nos*
Serbo-Croatian	*yedan*	*dva*	*tri*	*glava*	*uho*	*usta*	*nos*
Czech	*yeden*	*dva*	*tři*	*hlava*	*uxo*	*usta*	*nos*
Polish	*yeden*	*dva*	*tši*	*gwova*	*uxo*	*usta*	*nos*
Russian	*adʸin*	*dva*	*trʸi*	*galavá*	*úxo*	*rot*	*nos*
Lithuanian	*vʸíenas*	*du*	*trʸīs*	*galvá*	*ausʸís*	*burná*	*nō̃sʸis*
Latvian	*viens*	*divi*	*trīs*	*galva*	*auss*	*mute*	*deguns*
Finnish	*üksi*	*kaksi*	*kolme*	*pǟ*	*korva*	*sū*	*nenä*
Estonian	*üks*	*kaks*	*kolm*	*pea*	*kõrv*	*sū*	*nina*
Hungarian	*eǰ*	*kēt*	*hārom*	*fō̃/fey*	*fül*	*sāy*	*orr*
Turkish	*bir*	*iki*	*üč*	*baš*	*kulak*	*āız*	*burun*
Basque	*bat*	*bi*	*hirür*	*bürü*	*belari*	*aho*	*südür*

(Note: Except for French 'one', the numerals are cited without gender variation. Finnish and Estonian *ä* = [æ].))

As the table shows, even seven lexical items – if selected with care – can furnish strong evidence that the Indo-European languages of Europe (Breton – Latvian) are related to each other. The case is similar for the Uralic languages (Finnish, Estonian, and Hungarian), although the case for Hungarian may be

less obvious. Moreover, the table permits us to distinguish subgroups within, say, Indo-European: Celtic (Breton, Welsh, Irish), Germanic (Icelandic – German), Romance (French – Rumanian), Slavic (Bulgarian – Russian), and Baltic (Lithuanian and Latvian); Greek and Albanian constitute subgroups of one each.

Given the evidence of just the seven words in Table 1, there might appear to be a somewhat weaker case for a Turkish-Basque relationship. Compare the similarities in the words for 'one', 'head', and possibly also 'three'. As far as we can tell, however, the similarities are misleading. Once the basis for comparison is enlarged to, say, a hundred lexical items, it turns out that the Turkish-Basque similarities are most likely the result of chance. In fact, up to this point it has not been possible to successfully establish a genetic relationship between Basque and any other language or language group. The fact that there can be such accidental similarities raises important questions about our ability to establish genetic relationship by sheer inspection of vocabulary. (This matter is pursued further in § 2 below and in Chapter 17.)

The situation gets more complex once we introduce selected Asian languages, as in Table 2, a continuation of Table 1.

Table 2: Further data from selected Asian languages

	'one'	'two'	'three'	'head'	'ear'	'mouth'	'nose'
Hindi	*ēk*	*dō*	*tīn*	*sir/sar*	*kān*	*mũh*	*nāk*
Marathi	*ēk*	*dōn*	*tīn*	*ḍōī, muṇḍ*	*kān*	*tõī*	*nāk*
Kashmiri	*akh*	*zi?*	*trɨ?*	*kalⁱ*	*kan*	*ɨs*	*nas*
Persian	*yek*	*do*	*se*	*sær*	*guš*	*dæhɔn*	*bini*
Ossetic	*iw*	*dɨwwə*	*ərtə*	*sər*	*qus*	*dzɨx*	*fɨnd*
Armenian	*mi*	*erku*	*erekh*	*glux*	*unkn*	*beran*	*ṙəngunkh*
Tocharian B	*ṣe, sana*	*wi*	*trai, tarya*	*āśce*	*klautso*	*koyṁ*	*meli*
Tamil	*ondrɨ*	*iraṇḍɨ*	*mūndrɨ*	*talɛi*	*kāðɨ, sevi*	*vāy*	*mūkkɨ*
Kannada	*ondu*	*eraṛu*	*mūru*	*tale*	*kivi*	*bāy, mūti*	*mūgu*

(Note: Unlike the other forms, the Classical Armenian and Tocharian B forms are not given in phonetic transcription but in a transliteration of their original spelling. The Tocharian words for 'one' and 'two' distinguish between masculine and feminine forms.)

On one hand, the sets Hindi *dō* : Marathi *dōn* : Persian *do* : Osset. *dɨwwə* 'two', Hindi and Marathi *tīn* : Kashmiri *trɨ?* : Tocharian *trai/tarya* 'three', and Kashmiri *nas* 'nose' (perhaps also Hindi *nākh* 'nose') may suggest relationship to the Indo-European languages of Europe because of the phonetic similarities between the words. Note also Hindi *mũh* 'mouth' and especially

Marathi *muṇḍ* 'head' on one hand, and the Germanic words for 'mouth' on the other.

On the other hand, Hindi/Marathi *ēk*, Kashm. *akh*, Pers. *yek* 'one' look more similar to Finn. *üksi*, Est. *üks*, Hung. *eǰ*, and so do Hindi *kān*, Kashm. *kan* 'eye' to Finn. *korva*, Est. *kõrv* – as well as to Kannada *kivi*. In fact, Hindi *nākh* 'nose', perhaps also Kashm. *nas*, could just as well be considered related to Finn. *nenä*, Est. *nina* as to the words for 'nose' in the European members of the Indo-European language family. And while Hindi *mũh* and Marathi *muṇḍ* bear strong resemblances to the Germanic words for 'mouth', they are similar, too, to Kannada *mūti*.

But there are more problems. First, the evidence accumulated by more than a century of comparative linguistics suggests an especially close relationship between Indo-Aryan Hindi, Marathi, and Kashmiri on one hand and Iranian Persian and Ossetic (Iron dialect) on the other. This relationship, however, does not come out well in Table 2, except perhaps for the word for 'head'.

Further, the evidence for considering Tocharian an Indo-European language appears to be limited to one word, the numeral 'three', and there seems to be no evidence for considering Armenian an Indo-European language. Again, this conflicts with what we know as the result of extensive work in comparative Indo-European linguistics.

Finally, the evidence does correctly indicate that the Dravidian languages Tamil and Kannada (in the south of India) are not particularly closely related to any of the other language families. However, it fails to indicate that there are recurrent similarities between Dravidian and Uralic which suggest a possible relationship (see Chapter 17).

As it turns out, a number of the similarities that we just noted are accidental, just like those between Turkish *bir* and Basque *bat* 'one', and therefore do not reflect genetic relationship.

This is the case for the Hindi/Marathi and Germanic words for 'mouth'. Hindi *mũh* derives from Sanskrit *mukha-* which, if inherited, would reflect an earlier **mukho-*; Marathi *muṇḍ* goes back to Skt. *mūrdhan-*, a reflex of PIE **melədh/mḷədh-*; while the Germanic words reflect PGmc. **munþa-* which must go back to PIE **mn̥to-*. Note further that Kannada *mūti* is related to Tam. *mūñci* 'face', which appears to be older both in its phonetic shape and in its semantics. That is, as we trace these similar forms back in history we find that they become less similar. The modern similarities, thus, must be due to chance.

The similarities between the Hindi/Kashmiri/Persian and Finnish/Estonian/Hungarian words for 'one' likewise are accidental. The Indo-Aryan

words go back to Skt. *ēka-*, derived from PIE **oy-ko-* 'one, single', which in turn represents an extension by a suffix *-ko-* of the same root which, extended by the suffix *-no-*, underlies the numeral 'one' that is found in the majority of the Indo-European languages of Europe.

The similarities between Hindi and Persian/Ossetic in the word for 'head' reflect the fact that the Hindi word has been borrowed from Persian. Here, too, then, the similarities do not reflect inheritance from a common ancestor.

On the other hand, a number of genuine cognates are difficult to detect without extensive comparative research. For instance, the Tocharian and Armenian words for 'one' are ultimately related to the one found in Greek; all three of these derive – believe it or not – from PIE **sem-* 'same, similar, identical'. The Armenian and Ossetic words for 'three' are perfect cognates of the words found in the rest of the Indo-European languages. Again, given the evidence in Tables 1 and 2, this may be hard to believe; but more than a century of research has shown that the Armenian and Ossetic words are related. The Armenian form results from a change of **tr-* > *þr-*, loss of the *þ*, and prefixation of a vowel before the resulting initial *r*; the Ossetic form involves metathesis of initial **tr-* to **rt-*, an areal phenomenon in the Caucasus, plus prefixation of a vowel before initial **rt-* and other changes. The Armenian word for 'two' can likewise be related to its counterparts elsewhere in Indo-European, through a sequence of even more complex developments.

The upshot is that not all similarities – or dissimilarities – between languages in their vocabulary are indicative of genetic relationship, and that in order to establish genetic relationship we have to go significantly beyond comparing just seven vocabulary items.

2. Chance similarities, onomatopoeia, and "nursery words"

Probably any given pair of languages will offer at least some formally and semantically similar linguistic items whose similarities are simply due to chance. We have seen a few examples of this type in the preceding section, such as the Hindi, Marathi, Kannada, and Germanic words for 'mouth' (or 'head'). Even among the Indo-European languages of Table 1, some resemblances are accidental. This is certainly true for the similarity between the words for 'one' found in most of the languages and Modern Greek *énas* 'one'. The Modern Greek form reflects Ancient Greek *heîs, mía, hén* 'one (m., f.,

n.)'. As suggested earlier, this form, in turn, derives from earlier **sem-s, *sm-ia, *sem*, with a root **sem-* related to Engl. *same*. The remainder of the words for 'one' derive from a different ancestor, **oy-no-*.

Modern Greek has a word *mati* 'eye' whose phonetic and semantic resemblance to Malay *mata* is remarkable. But again, from what we know about the earlier history of these languages, the similarity between the words is due to chance. Mod. Gk *máti* goes back to earlier Gk. *ommátion*, a diminutive form of *ómma* 'eye', which in turn derives from an earlier **ok^{w}-m(e)n-*, in which **-m(e)n-* was a derivational suffix, and only the first element **ok^{w}-* originally meant 'eye'. Ironically, this element, found also in *ṓps* 'eye, face', as well as in *Kúklōps* 'Cyclops, lit. having a (single) circular eye' and related to Engl. *eye*, has disappeared from Mod. Gk. *máti*. What is left consists only of suffixal material, historically speaking.

There are similar problems for Mal. *mata*. While most Malayo-Polynesian languages have cognates of *mata* 'eye', many offer evidence for an obviously related form *kita* 'see' (e.g. Tagalog *kita*), and others testify to a form *buta* 'be blind' (e.g. Fiji *buto*). These variant forms suggest that *mata* is morphologically composite in origin, consisting of a root *-ta* meaning something like 'sight' and a prefix *ma-* which fixes the meaning of *-ta* to 'eye', while *ki-* and *bu-* alter the meaning to 'see' and 'be blind' respectively.

Similarly, Modern English and Modern Persian have phonetically and semantically virtually identical forms for 'bad': *bad* [bæd] and *bad* [bæ>d] (with a vowel slightly more retracted than that of the English word). The origin of the English word is somewhat controversial. Two derivations have been proposed, one from OE *bæddel* 'hermaphrodite, effeminate man', the other from OE *(ge)bæded* 'captured'. (Both of these involve extensive semantic shifts, generally involving pejoration.) The Persian word, on the other hand, derives from earlier Pahlavi *wad*, whose initial *w-* cannot possibly be related to the *b-* of the English form, either of its putative Old English ancestors, or any other imaginable ancestral form.

We can avoid being misled by chance similarities if we insist that our comparison be based on a very LARGE data base. For if we find striking similarities in pronunciation and meaning in, say, a thousand words, the possibility that these similarities are due to chance becomes rather remote. Note that the data base must be very large, for as (1) below shows, it is not at all difficult to find a fairly large number of chance similarities between any given pair of languages. (Sanskrit and English are of course related to each other. However, we know their linguistic histories sufficiently well to be certain that the similarities in (1) do not reflect genetic relationship. On this matter see also Chapter 17.)

(1)

Sanskrit	Mod. Engl.	
kōṇa-	*corner*	
ǰhampa-	*jump*	
taru-	*tree*	(correct cognate: Skt. *dāru-*)
tōraṇa-	*door*	(correct cognate: Skt. *dvāra-*)
krōśati	*cries*	
gati-	*gait*	
lōkati	*looks*	
marīčikā	*mirage*	
rāga-	*rag*	(musical terms)
vāmā	*woman*	

(and others, the total being at least 26 items)

In addition, certain types of vocabulary are notoriously unreliable for establishing genetic relationship. One of these is ONOMATOPOEIA. Although details may differ, onomatopoetic expressions come out remarkably similar in different languages. See for instance the rooster calls in examples (17) and (17′) of Chapter 7. In spite of their differences, Engl. *cockadoodledoo*, Germ. *kickericki*, Fr. *cocorico* share a repetition of velar stops, an inserted liquid (generally an *r*-sound), as well as, generally, accent on the last syllable.

Another area in which caution is advisable consists in vocabulary such as Engl. *dad(dy), mom(my), baby,* It. *papà, mamma, bambino,* Hindi *bāp, mã̃, baččā.* The structure, sounds, and often also the connotations of these words suggest that they are NURSERY WORDS, i.e., words and meanings which adults assign to the early babbling of infants. Like other items that occur in this kind of language, the words tend to have reduplicated consonant-vowel syllables (as in *pa-pa*). And the predominance of the vowel *a* and of the consonants *p/b, t/d, m,* and *n* in these words is a common feature of early babbling. (See Chapter 9, § 2.)

3. Similarities due to linguistic contact

Even if we eliminate chance similarities, onomatopoeia, and nursery words, we are not necessarily home free. We may be confronted with situations of the type (2). Here it appears as if English is simultaneously and equidistantly related to two quite distinct languages, French and German, with no evidence for genetic relationship between these two languages if we limit ourselves to the evidence in (2). In biology, such a dual relationship might not be entirely

unexpected, since there is such a thing as cross-breeding. But in genetic/comparative linguistics, this type of relationship is considered suspect. The suspicion always arises that such a relationship is attributable to borrowing. And in fact, in the present case we know from the history of English that the correspondences between English and French result from the secondary contact between the two languages after the Norman conquest of England.

(2)

English	French	German
calf		*Kalb*
veal	*veau*	
cow		*Kuh*
beef	*bœuf*	
swine		*Schwein*
pork	*porc*	

Even if we did not have this direct historical knowledge, we would be able to make a good case for a borrowing relation between English and French by looking at other items, such as the ones in (3).

(3)

English	French	German
to	*à*	*zu*
too	*trop*	*zu*
two	*deux*	*zwei*
twenty	*vingt*	*zwanzig*
eat	*manger*	*essen*
bite	*mordre*	*beissen*
father	*père*	*Vater* [f-]
mother	*mère*	*Mutter*
three	*trois*	*drei*
thou	*tu*	*du*

In these correspondences (which could be multiplied many times over), it is easy to see that there is a close relationship between English and German, while French generally offers very different forms. True, closer examination now reveals some recurrent similarities which also involve French (see the last four words in (3)); but it is also clear that French does not exhibit any closer affinities with English than it does with German (or vice versa). In fact, in the words for 'father' and 'mother', and in many others like them, the similarities between English and German are much more striking than those of either of the two languages to French. The special affiliation of English with French that might have been suggested by the correspondences in (2) thus turns out to be contradicted by the evidence of additional data.

What is even more significant, a comparison of the data in (2) with those in (3) shows that the English/French similarities are restricted to certain, limited, spheres of the vocabulary. On the other hand, the German/English similarities pervade the whole lexicon, including BASIC VOCABULARY. As noted in Chapter 8, borrowing tends to be limited to certain spheres of the lexicon. In addition, it is often restricted to technical vocabulary. And it has the least effect on basic vocabulary.

Cases like the English/French/German relationship are important because they provide insights that make it possible to detect borrowings in other cases, where we do not have direct historical evidence. If the similarities between two given languages are limited to certain spheres of the lexicon and if they cover little, if any, basic vocabulary, then there is a strong reason to suspect that they result from borrowing, not from genetic relationship.

Contact-induced similarities can also be found in overall structure, as the result of convergence. Recall for instance the case of the Balkans, of South Asia, or – even more strikingly – of Kupwar (Chapter 13, § 2). On the other hand, divergence in overall structure does not necessarily argue against genetic relationship. For instance, the modern Indo-European languages, though clearly related, exhibit the following basic word orders: Indo-Aryan, Iranian, and Armenian have SOV (Subject : Object : Verb); Celtic offers VSO; German, Dutch, and Frisian have a mixture of SOV and SVO characteristics; and most of the others exhibit SVO. Similarities and differences in overall structure, thus, are not a reliable guide to establishing relationship. (But see § 5 below on similarities in specific aspects of structure.)

Comparative linguists therefore usually concentrate on VOCABULARY and on correspondences that emerge from an examination of vocabulary. Moreover, since basic vocabulary is less likely to be borrowed, the evidence of such vocabulary receives the highest priority.

4. Systematic, recurrent correspondences

We can strengthen our argument for genetic relationship between given languages by showing that their similarities are not helter-skelter or sporadic, but that they are SYSTEMATIC and RECUR in large sets of words. In fact, given that sound change is overwhelmingly regular, we must expect a great degree of systematicity and recurrence in the phonetic similarities between putatively related languages.

Consider again the case of English and German. If we add the data in (4) to those in (2) and (3), we note some important phonetic differences between English words with *t* and their German counterparts. However, within these differences, we can establish a great systematicity; see the summary in (5). Moreover, even though there may be differences, the German counterparts of English *t* are phonetically similar, in that like *t*, they are dental. And if we expand our horizon to include words with English *p* and *k*, we find, allowing for some minor differences, a very similar situation; see (6). Given these facts, the conclusion becomes almost inescapable that these words, and many others like them, go back to a common ancestor and have become different through the operation of regular sound change. The ability to find such regular and systematic correspondences between languages is the cornerstone of establishing genetic relationship.

(4)

English	German
frost	*Frost*
chest	*Kiste*

(5)

English	German
t	*z* [ts] (initially and after consonant)
t	*ss* [s] (intervocalically)
t	*t* (after *s*)

(6)

English	German			
pound	*Pfund*	p-	:	pf-
penny	*Pfennig*			
ape	*Affe*	-p-	:	-f-
hope	*hoffen*			
aspen	*Espe*	-sp-	:	-sp-
wasp	*Wespe*			
cool	*kühl*	k-	:	k-
card	*Karte*			
make	*machen*	-k-	:	-x-
cook	*Koch*			

5. Shared idiosyncrasies

We can yet further improve our case if we can find shared idiosyncrasies in morphology.

Consider the English and German comparatives of Engl. *good* and its German counterpart *gut*, which are formed from what looks like a completely different lexical item – *better, best* and *besser, best-*; see (7a). Morphological relationships of this type are commonly referred to as SUPPLETION. Contrast the suppletion in (7a) with the normal pattern in Engl. *warm : warmer : warmest*, Germ. *warm : wärmer : wärmst-*. Now, (7a) demonstrates that French likewise has suppletion; but significantly, the English and German data exhibit systematic and recurrent similarities with each other, while the French forms are radically different. If we had to choose which of these patterns of suppletion must result from genetic relationship, we would surely have to opt for the patterns found in English and German. To select English and French would border on the perverse.

Similarly, the early Indo-European languages and even some modern ones exhibit striking similarities in the third person singular and plural forms of the verb 'to be', including a remarkable paradigmatic alternation between *Vs-* in the singular and *s-* in the plural. Compare (7b).

(7) a.

English	German	French
good	*gut*	*bon*
better	*besser*	*meilleur*
best	*best-*	*le meilleur*

(For the correspondences *-t- : -s-, -st- : -st-*, see (4) above. The other correspondences are similarly supported by "outside" evidence; compare the *g- : g-* in Engl. *great, give* : Germ. *gross, geben.*)

b.

Sanskrit	Latin	Mod. Germ.	Old Church Slavic
as-ti 'is'	*es-t* 'is'	*is-t* 'is'	*es-tŭ* 'is'
s-anti 'are'	*s-unt* 'are'	*s-ind* 'are'	*s-ǫtŭ* 'are'

Highly idiosyncratic paradigmatic alternations such as those in (7b) do not normally get borrowed, nor do suppletive patterns like the ones in (7a). The morphological idiosyncrasies exhibited by these patterns, therefore, combined with the fact that they involve systematic phonological correspondences, would be difficult to explain except as reflecting common heritage. In fact, evidence of patterns like (7b) no doubt contributed greatly to William Jones's proposal that "Sanscrit", Greek, Latin, and perhaps "Gothick" and "Celtick", too, are descended from a common ancestor.

6. Reconstruction

Most historical linguists believe that the ultimate proof of genetic relationship lies in reconstruction, i.e., in reversing linguistic history, as it were, by postulating linguistic forms in an ancestral or PROTO-language from which the attested forms can be derived by plausible linguistic changes. Note that "proof" here is to be understood more or less as in a court of justice, as establishing a case beyond a reasonable doubt. Moreover, to be probative, the reconstruction must be based on a large amount of lexical items and at the same time conform to a set of evaluative principles that are presented below.

For an illustration, consider the data in (8) as they bear on the reconstruction of Proto-Indo-European vowels.

(8)

	Sanskrit	Greek	Latin	Germanic		Reconstruction
a.	***i**dam*		***i**d(em)*	*Go. **i**ta*	'it, that'	***i**
	*lik**t**a-*	*é-lip-on*	*(re-)lic-tus*		'left'	
b.	*y**u**gam*	*z**u**gón*	*i**u**gum*	Go. *j**u**k*, OE *ge**o**c*	'yoke'	***u**
	*b**u**dh-*	*p**u**th-*		OE *b**u**don*, *geb**o**den*	'(a)bide, awake'	
c.	***a**sti*	***e**stì*	***e**st*	Go. ***i**st*, OE ***i**s*	'is'	***e**
	***a**tti*	*édomai*	*edō*	Go. ***i**tan*, OE *e**t**an*, *i**t**iþ*	'eat'	
d.	***a**ṣṭau*	***o**któ́*	***o**ctō*	Go. ***a**htau*	'eight'	***o**
e.	***a**ǰati*	*ágō*	*agō*	ON *aka*	'drive'	***a**
f.	*p**i**tar-*	*p**a**tḗr*	*p**a**ter*	Go. *f**a**dar*	'father'	***ə**

(Note: The data above are representative of larger sets. Given the overall available evidence, it is not possible to assume that (some of) the above sets can be "collapsed" in the process of reconstruction and that the observed differences might be attributed to special developments in the individual languages. That is, each set must be reconstructed separately for the proto-language.)

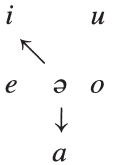
In reconstructing, we must keep in mind the following principles which determine the acceptability and plausibility of our reconstruction:

– Reconstructed items and systems and postulated linguistic changes should be NATURAL. A corollary of this principle is that postulated sound changes must be REGULAR, in conformity with the regularity principle of Chapter 4.

A further corollary is that there must be a phonetic value attached to a reconstructed sound. For instance, it would be unacceptable to reconstruct set

(8f) as *. What would be the phonetic value of such a symbol? We must assume that a reconstructed (proto-)language is essentially like any language observable today – since known languages do not have sounds like , presumably neither did Proto-Indo-European. A natural reconstruction would instead be *[ə]; for as (9) shows, [ə] can naturally change to either [a] or [i].

(9)



Moreover, it would be dubious to reconstruct *[a] for both sets (8e) and (8f) – or for sets (8d) and (8e), for that matter. To do so would require the assumption that contrary to normal expectations, sound change operates in a sporadic, irregular fashion, such that *[a] can change either to *i* or to *a* in Indo-Iranian, without any motivation for the different developments.

– The reconstruction must not violate OCCAM'S RAZOR. According to this maxim, attributed to the medieval English philosopher William of Occam, *Entia (*or *essentia) non sunt multiplicanda praeter necessitatem* 'Entities [in an argument] should not be multiplied beyond necessity.' This maxim, incidentally, is fundamental to all scientific inquiry.

In comparative reconstruction, such "entities" are (i) reconstructed items, and (ii) changes required to convert these items into the forms attested in the descendant languages.

For instance, it would be a violation of Occam's Razor if we reconstructed *[æ] for set (8a). This reconstruction would require the entirely unnecessary assumption either that all the languages innovated by changing *[æ] to [i] or that there was such a change in the ancestor language.

On the other hand, while the reconstruction *[ə] for set (8f) introduces an additional reconstructed sound, which is not attested as such in any of the daughter languages, it does so by necessity. To do otherwise would result in a violation of our first principle, since it would entail the unnecessary assumption of irregular sound change.

Finally, Occam's Razor argues that set (8d) should be reconstructed as *[o]. This reconstruction makes it possible to distinguish the behavior of this set from set (8e), in accordance with the expectation that sound change is regular. In addition, the reconstruction *[o] makes it possible to account for the attested forms with a minimum of changes. All we need to assume is that outside of Greek and Latin, the distinction between *[o] and *[a] was lost in favor of [a]. The Greek and Latin forms, then, represent unchanged out-

comes. If however we reconstructed, say, *[ɔ], then we would have to unnecessarily assume that the original sound was changed in all the languages, including Greek and Latin.

– Wherever we can, we use the OLDEST available stages of languages. This makes reconstruction simpler, since less time has passed from the time of the proto-language, and thus there has been less chance for linguistic changes to obscure the relationship between the languages.

Consider the data in (10). Most striking are the examples in (10a), where the similarities – and correspondences – between Sanskrit and Old English are quite clear, while their modern counterparts have come to differ greatly as the result of linguistic change (mainly lexical replacements). As examples such as (10b) illustrate, in some cases the relationship remains transparent, even in the modern languages. But compared to the older stages, such modern correspondences are much rarer. In fact, (10c) shows that in some cases even the oldest stages of the languages have undergone enough changes that the relationship of words that we know to be inherited from the Indo-European ancestor has become greatly obscured. Their relationship can be established only after extensive research in comparative reconstruction.

(10)		Sanskrit	Old English	Hindi	Modern English	
	a.	*asti*	*is*	*hai*	*is*	
		sa	*sē*	*vō*	*he*	
		vayam	*wē*	*ham*	*we*	
		svasar	*sweostor*	*bahan*	*sister*	
		śvaśrū	*sweogor*	*sās*	*mother-in-law*	
		vēda	*wāt*	*ǰāntā hai*	*(he) knows*	
	b.	*dvā(u)*	*twā*	*dō*	*two*	
		trayas	*þrī*	*tīn*	*three*	
		pād-	*fōt*	*pā̃v-*	*foot*	
	c.	*čakra-*	*hweogol*	*čakkā*	*wheel*	*(< *k^{w}ek^{w}lo-)*
		śṛŋga-	*horn*	*sī̃g*	*horn*	*(< *ḱer/ḱr̥-)*
		bhavati	*bēon*	*hōnā*	*be*	*(< *bhū-)*
		ṣaṭ	*seox*	*chah*	*six*	*(< *s(w)eḱs)*
		pluśi-	*flēah*	*pissū*	*flea*	*(< *pl(o)uḱ-)*

We are able to assert that the words in (10c) are in fact related to each other partly because of the evidence of cognates in the early stages of other related languages, such as Gk. *kéras* 'horn'. In other cases, such as Skt. *čakra-*, OE *hweogol*, the relationship can be demonstrated only because we have reconstructed the Proto-Indo-European ancestral language, established the sound

changes from PIE to languages such as Sanskrit and Old English, and are therefore able to show that both forms are derivable from a PIE form **kʷekʷlo-*, which is also reflected in Gk. *kúklos* 'wheel, circle'. Moreover, because we have reconstructed not just the sound system and the lexicon of Proto-Indo-European, but also its morphology, we are able to explain **kʷekʷlo-* as a morphological derivative of the independently reconstructed PIE root **kʷel-* 'move, turn' with an original meaning along the lines of 'the thing that keeps turning around'. (The morphological processes are complex, involving among other things a productive paradigmatic alternation between *el* and *l* [hence **kʷel-* beside **-kʷl-*] and a process of reduplication, which copies the initial consonant and vowel of the root [hence the initial **kʷe-* of **kʷe-kʷlo-*].)

Evidence of the type (10a), combined with that of (10c), suggests that there may be an optimal closeness of related languages that is necessary to successfully establish genetic relationship. If too many centuries of linguistic changes have increased divergence beyond that optimal stage, the evidence may become too limited, and establishing genetic relationship may become difficult or even impossible. (See also Chapter 17.)

7. What can we reconstruct and how confident are we of our reconstructions?

As seen earlier, comparative linguistics places the greatest amount of confidence in sound correspondences found in lexical comparisons, especially basic vocabulary. Nevertheless, we can reconstruct other aspects of the ancestral language, beside the lexicon. Based on the methods and assumptions illustrated in the preceding section, we can reconstruct a fair amount of the phonology of the proto-language; and using different and more sophisticated methods, we can gain a pretty good picture of the morphology of the proto-language and of aspects of its syntax.

Ironically, although we base our reconstructions on lexical evidence, lexical reconstruction in many cases is done with less confidence than the reconstruction of phonology, morphology, and syntax. Consider the case of Algonquian 'fire-water' in example (11). There is no doubt that the words for 'fire' and 'water' are inherited. Given the evidence in (11), we might feel similarly confident about reconstructing a word 'fire-water'. But appearances are deceiving. We know that the product 'fire-water', i.e., alcohol, was introduced with the arrival of Europeans, long after Proto-Algonquian was spoken. We

must therefore conclude that the words for 'fire-water' were assembled secondarily, from indigenous roots and according to inherited processes of compound formation. Moreover, we can assume that the words were not created independently, but that they were diffused through the Algonquian languages by calquing (for which see Chapter 8.)

(11)

	Fox	Cree	Menomini	Ojibwa
'fire'	*aškotēwi*	*iskotēw*	*eskōtɛ̄w*	*iškotē*
'water, liquid'	*-āpō-*		*-āpō-*	*-āpō-*
'fire-water'		*iskotēw-āp-oy*	*eskōtɛ̄w-āp-oh*	*iškotēw-āp-ō*

Examples like this show very strikingly that in some cases we are more successful in reconstructing basic morphological elements, such as the roots for 'fire' and 'water', and the morphological patterns according to which they can combine, than complete, complex words. The best we can do is to establish that Proto-Algonquian had the morphological elements and machinery to assemble a word like 'fire-water' – if the occasion had arisen. The problem is that the occasion arose only much later.

Such problems are not limited to lexical reconstruction. In syntax, too, we are much more successful in reconstructing syntactic patterns. Reconstructing specific sentences runs into even greater difficulties than reconstructing complex words. True, we may be quite certain that a speaker of Proto-Indo-European must have been able to utter a simple sentence like **pətēr (e)gʷemt* 'the father came/arrived/went'. But even for a simple sentence like this there are problems, such as the fact that Indo-Europeanists are not in full agreement as to whether we should reconstruct *gʷemt* or *egʷemt* for the form meaning 'came/arrived/went'. Moreover, there is the problem that the same idea may be expressed in more than one way. For complex sentences, the problems are obviously even greater.

The problem runs even deeper, for as the example of *gʷemt* vs. *egʷemt* illustrates, comparative linguists often disagree with each other. Their disagreement may concern matters of relatively minor detail, such as whether past-tense forms of the type *gʷemt* should be reconstructed with the "augment" **e-* for all of Proto-Indo-European or for only some dialects of the proto-language, or whether the prefix was introduced in the early stages of some of the daughter languages.

The reason for this disagreement, briefly, is this. Among the ancient Indo-European languages, the augment is limited to Indo-Iranian, Armenian, Greek, and a few other, less well attested languages. These languages were close geographical neighbors. On the other hand, Hittite, Latin, and the other early Indo-European languages show no clear traces of the augment. What is

especially embarrassing is that Hittite lacks it, since Hittite is attested earlier than either Sanskrit or Greek (or Latin). Some linguists therefore consider the augment a regional innovation, either in dialectal Proto-Indo-European (comparable to the centum : satem phenomena discussed in Chapter 11, § 4) or even later (presumably as the result of convergent developments; see Chapter 13). Other linguists argue that Hittite, Latin, and other early languages that lack the augment exhibit other innovations in verbal morphology and that the absence of the augment in these languages can therefore be considered a similar morphological innovation.

Even the reconstruction of the Indo-European sound system has been a matter of controversy and/or change of opinion. In the nineteenth century the stop system was reconstructed as in (12a), with a neat four-way contrast between voiceless, voiceless aspirated, voiced, and voiced aspirated, just as it is found in Sanskrit. (See also Chapter 2.) More recently, scholars have argued that the voiceless aspirated series of Sanskrit (indirectly attested also in Iranian) can be explained as the result of secondary developments. Occam's Razor, therefore, should prevent us from postulating it as a feature of the proto-language. As a consequence, the system in (12b) was postulated.

More recently yet it has been claimed that the system in (12b) is unnatural. The most important argument is the claim that no known languages have voiced aspirates without also having voiceless aspirates. Scholars adhering to this view therefore reconstruct the system in (12c), with voiceless stops (± aspiration), "glottalized" stops (accompanied by a glottal-stop element), and voiced stops (± aspiration) corresponding, respectively, to the voiceless, voiced, and voiced aspirated stops of (12b).

(12)

a. **Nineteenth-century reconstruction**

	Labial	Dental	Palatal	Velar	Labiovelar
voiceless	*p*	*t*	*ḱ*	*k*	*k*^{w}
voiceless aspirated	*ph*	*th*	*ḱh*	*kh*	*k*^{w}*h*
voiced	*b*	*d*	*ǵ*	*g*	*g*^{w}
voiced aspirated	*bh*	*dh*	*ǵh*	*gh*	*g*^{w}*h*

b. **Standard twentieth-century reconstruction**

voiceless	*p*	*t*	*ḱ*	*k*	*k*^{w}
voiced	*b*	*d*	*ǵ*	*g*	*g*^{w}
voiced aspirated	*bh*	*dh*	*ǵh*	*gh*	*g*^{w}*h*

c. **"Glottalic" reconstruction**

voiceless (± asp.)	*p(h)*	*t(h)*	*ḱ(h)*	*k(h)*	*k*^{w}*(h)*
glottalized	*p'*	*t'*	*ḱ'*	*k'*	*k*^{w}*'*
voiced (± asp.)	*b(h)*	*d(h)*	*ǵ(h)*	*g(h)*	*g*^{w}*(h)*

The so-called glottalic system in (12c) differs markedly from the ones in (12a) and (12b) and, if correct, would have enormous consequences for comparative Indo-European linguistics. The system is virtually identical to the one found in certain modern Armenian dialects and postulated for early Armenian by the advocates of the "glottalic theory". This has the virtue that the system is precedented and therefore can be considered natural. But another consequence is that the sound shift traditionally postulated for Armenian (see Chapter 4) can no longer be maintained. Instead, we must assume that Armenian essentially retained the stop system of Proto-Indo-European. An extension of this argument is that Grimm's Law, which, as noted in Chapter 4, is remarkably similar to the traditionally postulated Armenian sound shift, must likewise be rejected. The Germanic sound system, then, is claimed to be nearly as archaic as that of Armenian.

But there are further consequences. The striking differences between Armenian and Germanic on one hand and the rest of Indo-European on the other must now be attributed, not to innovations on the part of Armenian and Germanic, but to sound shifts in the other Indo-European languages. These shifts must be of similar proportions to the ones traditionally postulated for Armenian and Germanic. Moreover, these shifts would have to be considered independent of each other. If this assumption is correct, we would have to postulate some ten or twelve major sound shifts, instead of the two traditionally assumed for Armenian and Germanic. Such a proliferation of shifts, in turn, could be considered an argument against the reconstruction in (12c), since it would violate Occam's Razor.

Moreover, as noted in Chapter 2, the glottalic system found in some of the modern Armenian dialects may be attributed to convergence with the neighboring Caucasic languages. In this regard, note that Ossetic, an Iranian language which likewise is spoken in this region, has a similar glottalic system. But in this case, the evidence of the other Iranian languages makes it clear that the glottalic system is an innovation, no doubt the result of convergence with the other languages of the Caucasus. These facts weaken the arguments for considering the glottalic system of Armenian to be an archaism.

Finally, it has been observed that some languages do in fact have voiced aspirates without contrasting voiceless aspirates. One area in which such languages are found is part of the Indonesian archipelago. Members of the West African group of Kwa languages likewise offer such supposedly impossible sound systems.

The evidence of these languages shows that one of the most important foundations of the glottalic theory cannot be maintained, namely the claim

that languages with voiced aspirates but no contrasting voiceless aspirates are unnatural.

There are thus a number of arguments that weaken the cogency of the glottalic theory. Most Indo-Europeanists, therefore, prefer reconstruction (12b) to (12c); but proponents of the glottalic theory remain convinced that (12c) is a superior reconstruction.

Such disagreements must appear disconcerting to the non-linguist, and even to linguists working in other areas of specialization, who are unfamiliar with the often arcane arguments of comparative linguists. In principle the disagreement should come as no surprise. All reconstructions basically are HYPOTHESES about the nature of the proto-language, and by their very nature hypotheses are – well, hypothetical. True, we try to exclude questionable hypotheses by appealing to such principles as Occam's Razor and naturalness. But these are only very general guidelines. They are not simple algorithms which, if properly applied, will automatically yield correct solutions. They require judgments on the part of comparative linguists. And that is where disagreements can arise.

At the same time, we don't really have any choice; we have to develop hypotheses, even if they are "hypothetical" and sometimes controversial. If we really knew what the proto-language was like, we wouldn't have to do reconstruction.

8. Language families other than Indo-European

The present chapter, just like much of the rest of this book, so far has concentrated on Indo-European languages. This is because since their "discovery" in the late eighteenth century, these languagues have received the attention of more comparative and historical linguists than any other language family. In part this reflects the fact that until relatively recently, most linguists were speakers of Indo-European languages.

But this is not a sufficient explanation. Much of the work on the Semitic languages and the larger Afro-Asiatic family of which Semitic is a member has also been done by native speakers of Indo-European languages, and similarly, pioneering fundamental research on language families such as Bantu, Malayo-Polynesian, or the languages of the Americas has been conducted by speakers of Indo-European languages.

What is more important is that most Indo-Europeanists begin with a good foundation in the classical languages of Greek and Latin. Since these are

clearly Indo-European, it is natural for such scholars to expand their horizon (if they choose to do so) to other, related, Indo-European languages.

It may, however, also be true that the early Indo-European languages present something close to the optimal stage for comparison mentioned in § 6 above. This makes the initial task of establishing genetic relationship, as well as the job of reconstruction, relatively easy. True, even the Indo-European family includes members whose earliest attestations come from less optimal times (such as Albanian and Tocharian), or whose written attestations present other difficulties (such as Hittite). But just as in the case of examples like Skt. *čakra-*, OE *hweogol* 'wheel' in (10c) above, it is possible to at least begin to unravel the mysteries presented by such languages, because – in spite of the difficulties outlined in § 7 – we do have a fairly firm understanding of reconstructed Proto-Indo-European and therefore are able to draw on that understanding to make hypotheses as to how recalcitrant forms like *čakra-* and *hweogol* or recalcitrant languages like Albanian, Tocharian, and Hittite may be derived from Proto-Indo-European.

The following presents a brief look at the often more mixed successes of comparative linguistics as regards other language families. For selected languages the relationships are illustrated with examples of lexical correspondences. In some cases, the phonetic similarities in the correspondences are strong enough to strike even the non-specialist. In others, the similarities are more remote, but extensive comparative work makes it certain that the forms are cognate. In a few cases (especially Altaic), sets of highly dissimilar correspondences are included specifically to illustrate the problems that lead linguists to disagree on whether the languages in question are genetically related.

The highly controversial issue of whether it is possible to establish longer-range genetic relationships (as between Indo-European and Semitic or Afro-Asiatic) or even a genetic relationship between all of the world's languages is taken up in the next chapter.

As we saw in the introduction to this chapter, beside Indo-European there are members of at least two other language families in Europe. One of these families is the FINNO-UGRIC group which includes Finnish, Estonian, and Hungarian, as well as a number of other less well-known languages such as Lapp (now often called Saami, in the northern parts of Norway, Sweden, and Finland), Ostyak (also called Hanty, in western Siberia). Finno-Ugric, in turn, is part of a larger group, called URALIC, which includes Samoyed (in the northern part of the Russian Republic, east of the Ural mountains). The relationship between these languages can be illustrated by the sample correspondences below, given here in traditional transcription. Some of the forms are obviously similar, such as the words for 'winter' in all three languages, or the

words for 'fish' in Finnish and Hungarian. For others, the relationship is less obvious, such as the words for 'one' and 'two'; but a closer look yields better results. For instance, the words for 'one' all begin in vowel, and the Finnish *-k-* of this word can be related to the *-gy* = [ǰ] in Hungarian; and so forth. In fact, more than a century of comparative work on Finno-Ugric/Uralic makes it certain that all the correspondences involve genuine cognates, descended from a common ancestor.

	Finnish	Hungarian	Ostyak
'one'	*yksi*	*egy*	*it, ij*
'two'	*kaksi*	*kettő/két*	*katən, kăt*
'three'	*kolme*	*három*	*xutəm*
'fish'	*kala*	*hal*	*xut*
'heart'	*sydän*	*szív*	*sam*
'winter'	*talvi*	*tél*	*tatə, tał*

The other family represented in Table 1 above is ALTAIC, of which only one language with a literary tradition is found in Europe, namely Turkish. Turkish is part of a closely-related subgroup, called Turkic, members of which are found as far east as Central Asia and Siberia. Other members of the Altaic group include Mongol, Manchu, and Tunguz. Recent research suggests that Korean, and perhaps also Japanese, may be related to the Altaic languages. But both claims, especially the claim that Japanese is related, are controversial.

Even the Altaic group as more traditionally defined is controversial. Some scholars deny the validity of the "Altaic hypothesis" altogether and claim that Manchu and Tunguz are related to Korean and possibly to Japanese, but that there is no relationship between this group and the rest of what traditionally has been called Altaic. A comparison of the data below with those for Uralic readily illustrates how much more remote the Altaic languages are from each other. (Some of the Tunguz forms are missing, due to insufficient information.) Similarities are limited to Mong. *ǰirin* : Tung. *ǰū(r)* 'two', and Tu. *yürek* : Mong. *dzürx* 'heart'. However, semantically less exact correspondences such as Turk. *balık* 'fish' : Mong. *balgu* 'carp', and Turk. *kıš* 'winter' : Mong. *kul-de*, Tung. *kēl-di* 'cold' can easily be added.

	Turkish	Mongol	Tunguz
'one'	*bir*	*negen*	*umun*
'two'	*iki*	*qoyor/ǰirin*	*ǰū(r)*
'three'	*üč*	*gurban*	*ilan*
'fish'	*balık*	*dzagas*	
'heart'	*yürek*	*dzürx*	*mēwan*
'winter'	*kıš*	*öböl*	

Some scholars have argued for genetic relationship not just between the Altaic languages, but even between Uralic and Altaic, pointing to lexical similarities such as those below. (The glosses on the left in many cases are only approximate. For instance, the range of meanings for Alt. **al-* includes 'underside', 'frontside', 'lower part, backside, rump', and so on.) Some of the correspondences are indeed quite striking; others, such as **ñele-* : **dalag-* 'lick' are less impressive. Whatever the merits of such similarities, the Ural-Altaic hypothesis is considered even less well established than the Altaic one, and therefore even more controversial.

	Uralic/Finno-Ugric	Altaic
'under, below'	**al-*	**al-*
'tongue, language'	**kelä*	**kele*
'we'	**me-*	**min-*
'what'	**mə*	**mu*
'lick'	**ñele-*	**dalag-*
'three'	**kolme*	Mong. *gurban*

As noted in the introduction to this chapter, in addition to Indo-European, Uralic, and Altaic, Europe also is host to BASQUE, which does not seem to belong to any of the other well-established language families. It may well be that Basque had relatives in prehistoric times that died out with the coming of the Indo-Europeans to Europe in the second millennium BC, but there are no records of such languages. Many other LANGUAGE ISOLATES like Basque are found around the world, such as Sumerian in ancient Mesopotamia and Burushaski in the extreme north of South Asia. Languages like these present even greater challenges to comparative linguistics than controversial groupings like Altaic.

Language families spoken in the Asian part of Eurasia are discussed below. Illustrative correspondences are given only for selected language families. In most cases, these are the words for the numerals 'one', 'two', and 'three', but for some groups other words are cited.

The Caucasus area is home to a number of Indo-European and Altaic (Turkic) languages, including Armenian and Ossetic (an Iranian language), as well as Azeri (Azerbaijani, a Turkic language). In addition, there are three groups of CAUCASIC languages which are commonly considered "unrelated" (i.e., unrelatable) to any outside languages or language families: the Northwest Caucasic languages (e.g. Abkhaz); the Northeast Caucasic languages (e.g. Chechen-Ingush and Dagestani); and Kartvelian. The best known Kartvelian language, and the one with the longest literary attestation (since the fifth century AD) is Georgian.

The Caucasic languages are notorious for very rich consonant systems. One feature, shared by the Indo-European languages Armenian and Ossetic, is the existence of a series of glottalized stops. (See § 7 above.) Some of the languages, especially those of Northwest Caucasic, have consonant systems unexcelled in any other attested human language. Ubykh, for instance, is said to have close to 80 consonants. And some of the same languages have been claimed to have the lowest vowel inventories. Abkhaz, for example, probably only has two vowels, a low vowel *a* and a central vowel *ə*.

In the eastern area we find SINO-TIBETAN, a family that includes the Chinese language family, as well as TIBETO-BURMAN, of which Tibetan and Burmese are major members. Chinese has been attested since probably the seventeenth century BC, Tibetan since the eighth century AD, and Burmese from the twelfth century AD. Although the Sino-Tibetan family is generally considered well established, reconstructive work has not progressed very far as yet, and many aspects of the internal subgrouping of Sino-Tibetan are still uncertain. There have been proposals in the past that Thai belongs to Sino-Tibetan, but recent research suggests that it may rather be distantly related to Austronesian. The following correspondences may illustrate the relationship between Chinese, Tibetan, and Burmese. As in many other cases, some word sets exhibit much more transparent similarities than others; compare the words for 'three' and 'I' vs. the words for 'two'; but all forms can be considered cognates. (Chinese forms are from the Middle Chinese period; the Tibetan and Burmese forms come from the written forms of these languages.)

	Chinese	Tibetan	Burmese
'two'	*ñžyi-*	*gnyis*	*hnac*
'three'	*sam*	*gsum*	*sûm*
'I'	*nguo*	*nga*	*ŋa*
'name'	*myäng*	*ming*	*ə-mañ*
'tree, wood'	*syen*	*shing*	*sac*

South Asia is home to DRAVIDIAN, a family of languages spoken mainly in the south of India and parts of Sri Lanka. But one member, Brahui, is spoken much farther north, in present-day Pakistan. The major literary languages are Tamil, Malayalam, Kannada, and Telugu. The following correspondences, in traditional transcription, may illustrate the degree to which the Dravidian languages are related to each other.

	Tamil	Malayalam	Kannada	Telugu	Brahui
'one'	*oṉṟu*	*onnu*	*ondu*	*okaṭi*	*asi(ṭ)*
'two'	*iraṇḍu*	*raṇḍu*	*eraṛu*	*reṇḍu*	*ira(ṭ)*
'three'	*mūṉṟu*	*mūnnu*	*mūru*	*mūṛu*	*musi(ṭ)*

The AUSTRO-ASIATIC language family includes Mon and Khmer in present-day Kampuchea, as well as the Munda languages in Central and East-Central India.

The MALAYO-POLYNESIAN or AUSTRONESIAN languages form a far-flung, but linguistically close-knit family. They include Malay, Indonesian, Javanese, Tagalog (in the Philippines), Maori, Hawaiian, Samoan, as well as Malagasy, the language of Madagascar, just east of Africa. Compare the following correspondences.

	Indonesian	Javanese	Tagalog	Samoan	Malagasy
'one'	*satu*	*siji*	*isa*	*tasi*	*isa*
'two'	*dua*	*loro*	*dalawa*	*lua*	*rua*
'three'	*tiga*	*telung*	*tatlo*	*tolu*	*telu*

AFRO-ASIATIC, as the name suggests, extends from Africa into Asia. The group includes the SEMITIC languages (Hebrew, Arabic, as well as Assyrian and Babylonian of ancient Mesopotamia), Ancient EGYPTIAN (and its descendant, Coptic), as well as BERBER (in North Africa), CUSHITIC (including Somali), and CHADIC (including Hausa). Compare the following correspondences. (Only putatively related words are given; hence some of the blanks. The hieroglyphic script of ancient Egyptian indicates only the consonants, not the vowels.)

	Hebrew	Arabic	Egyptian	Berber	Cushitic	Chadic
'to beat'	*dɔqaq*	*daqqaqa*	*dḳw*	*dəgdəg*	*daku*	*dōka*
'bone'	*qaṣṣ*	*qs*		*ixs, iɣs*		*k'ašī*
'ear, hear'	*šmaʕ*	*samiʕa*	*sǰm*	*asim*	*māsuw*	*sim*
'heart'	*lɛ̄v*	*lubb*	*yb*	*ul*	*lēb, nibbo*	*nəfu*
'mouth'	*pɛ*	*fam*		*emi*	*(y)af*	*po*
'nose, smell'			*snsn*		*san*	*sunsunā*

In Africa, the following families are recognized, in addition of course to Afro-Asiatic.

The most widespread language family is BANTU, ranging from Swahili in Kenya and Tanzania to Setswana and Zulu in southern Africa. The Bantu languages are generally considered part of a larger family, NIGER-CONGO, which includes West African and sub-Saharan languages like Wolof, Fula, and Yoruba. An even larger putative language family is NIGER-KORDOFANIAN, which in addition to Niger-Congo embraces most of the remaining west African languages. Of these different genetic classifications, Bantu is by far the best established; Niger-Congo is a more uncertain; and Niger-Kordofanian is controversial. The correspondences below are from selected members of the Niger-Congo family.

	Bantu			Other	
	Swahili	Lingala	Setswana	Ahlõ	Efik
'one'	*m-oja*	*m-ɔkɔ*	*ŋŋwe*	*ili*	*kiet*
'two'	*m-bili/wili*	*mi-bale*	*pedi*	*iwa*	*iba*
'three'	*tatu*	*mi-sato*	*tharo*	*ita*	*ita*

In the extreme south of Africa are located the KHOISAN languages, famous for their click sounds, which are indicated by such arcane symbols as *≠k*, *≠g*, and *!k(x)*. Formerly these languages were called Bushman and Hottentot; but the names have been given up because of their negative connotations. Two languages of Tanzania, Sandawe and Hatsa, have been claimed to be distant relatives of the Khoisan languages. The following correspondences may illustrate the relationship.

	Sandawe	Khoisan	
		Naron	Khoi
'ear, hear'	*keke*	*≠kē*	*≠gai*
'four'	*haka*	*haga*	*haka*
'valley'	*Goʔa*	*!xubi*	*!kxowi*

It has been argued that the majority of the remaining African languages (including Nubian, Sudanic, and Songhai) form a single language family, called NILO-SAHARAN. But like many others, this genetic classification is controversial.

The Americas are home to a large variety of indigenous languages. According to some scholars, most of these are related to each other, and there are only three "super-families" in the Americas. But this view remains highly controversial. A more conservative approach would recognize, among others, the following groups, but would consider the genetic affiliation of many languages to be still unsettled.

ESKIMO-ALEUT is a group of languages extending from Alaska and Northern Canada to Greenland, of which Eskimo, now often referred to as Inuit, is the best-known member. As noted in Chapter 9, the term Eskimo originally is a derogatory word, apparently derived from Micmac *eskameege* 'raw fish eaters'. The term, however, is still used in technical writing and by Indigenous Americans in Alaska.

The ATHABASKAN family is named after Athabaskan, spoken in Alaska and Northwest Canada, but includes many other languages, known for their rich consonant systems, a large number of glottalized consonants, and highly complex consonant groups. Navajo, with the largest number of speakers of any Indigenous American language in the United States (some 150,000), and Apache, are also members of the Athabaskan family, though spoken much

farther south (in present-day Arizona and adjacent areas). The Athabaskan family is considered related to two other groups, the nearly extinct Eyak (Alaska), and the Tlingit group (Alaska and Northwest Canada). Some linguists argue for a larger family, "NA-DENE", which also includes Haida (Alaska and British Columbia); but that affiliation is controversial.

ALGONQUIAN is a widespread family of closely related languages, extending from the Great Lakes area to northeastern North America, and originally along the eastern seaboard as far south as Virginia. Well-known members include Blackfoot, Cheyenne, Cree, Chippewa or Ojibwa, Fox, Menomini, Ottawa, the Illinois Confederation, and Shawnee. The correspondences below, which include the (in)famous word for 'fire-water', may illustrate the relative closeness of the members of this family. Reconstruction of the linguistic ancestor, Proto-Algonquian, has made considerable progress during the past one hundred years.

	Fox	Cree	Menomini	Ojibwa
'one'	*nekoti*	*nikot-*	*nekot*	*ninkot-*
'two'	*nīšwi*	*nīso*	*nīs*	*nīš*
'three'	*neswi*	*nisto*	*nɛʔniw*	*nisswi*
'fire'	*aškotēwi*	*iskotēw*	*eskōtɛ̄w*	*iškotē*
'water'	*nepi*	*nipiy*	*nepēw*	*nimpi*
'water, liquid'	*-āpō-*		*-āpō-*	*-āpō-*
"fire-water"		*iskotēwāpoy*	*eskōtɛ̄wāpoh*	*iškotēwāpō*

Two languages spoken in California, Wiyot and Yurok, have been shown to be related to Algonquian, but at a much greater distance. The fact that Algonquian thus has relatives in California raises interesting questions about the earlier distribution of the language family, or about prehistoric migrations in North America.

IROQUOIAN is a family of languages in the eastern United States and Canada with members that bear some particularly familiar names from American history, for it comprises the members of the "Five Nations" confederacy (also known as the "Iroquois League"): Cayuga, Mohawk, Oneida, Onondaga, and Seneca, along with Tuscarora, which as a later addition, turned the confederacy into the "Six Nations". Other Iroquoian languages are Cherokee (see Chapter 3, § 5.3 for its writing system), Erie, Huron, and Wyandot. Iroquoian is sometimes classified as related to Siouan.

SIOUAN is a very far-flung family, embracing the languages of the Sioux or Dakotas, as well as Crow, Iowa, Omaha, Osage, Winnebago, and many others. The family at one time extended as far north as the Dakotas and Central Canada, as far east as Virginia and the Carolinas, and as far south as the

Gulf coast. There have been attempts to relate Siouan to HOKAN, languages spoken in the Southwest of the United States, which include Mojave, Chumash, and Yuman. But that classification is generally doubted; and there are even doubts as to whether all the Hokan languages are really related to each other or whether their similarities are mainly attributable to centuries or even millennia of mutual borrowing.

UTO-AZTECAN is a large family in the western United States, Mexico, and Central America, including Nahuatl (the language of the ancient Aztec empire), Hopi (in Arizona), and Ute (in Utah and Colorado). The following correspondences may illustrate the relationship.

	Comanche	Tübatulabal	Luiseño	Hopi	Papago	Nahuatl
'one'	*səməʔ*	*čīč*	*supúl*	*sə́ka*	*həmako*	*seem-*
'two'	*waha(h)-*	*wō*	*wə́x, wéʔ*	*lə̲́yö-m*	*gōk*	*oomi*
'three'	*pahi-*	*pāi-*	*pə́hi*	*pə́yo-m*	*vaik*	*eeyi*

MAYAN, in Mexico and Central America, is a group of fairly closely related languages, named after their most well-known member, the language of the ancient Maya civilization. As observed in Chapter 3, the Mayan civilization developed a writing system of its own, long before the arrival of the Europeans. The decipherment of the writing system has been increasingly successful in recent years.

ARAWAKAN now is found mainly in northeastern South America, but once extended into the Caribbean as well.

QUECHUA, a far-flung family with members in Peru, Ecuador, Bolivia, as well as in border areas of Argentina, Chile, and Colombia, was the language of the ancient Inca empire. The modern varieties of Quechua are very closely related to each other, as can be seen from the following correspondences. Some scholars have grouped Quechua and Aymara into a larger, "Andean" or "Quechumara" family. But like most other attempts at establishing larger genetic families in the Americas, this proposal has remained controversial.

	Ancash	Junín	Cajamarca	Amazonas	Ecuador	Ayacucho	Cuzco
'two'	*iškē*	*iškay*	*iškay*	*iškē*	*iškay*	*iskay*	*iskay*
'three'	*kimsa*	*kimsa*	*kimsa*	*kimsa*	*kimsa*	*kimsa*	*kinsa*
'six'	*hoxta*	*suʔta*	*soxta*	*suχta*	*suχta*	*soxta*	*soxta*
'language'	*qaλu*	*aλu*	*qažu*	*kadžu*	*kažu*	*kaλu*	*qaλu*

AUSTRALIA, too, is home of a large number of languages. By some estimates, at least 200 languages were spoken in Australia at the time of the European arrival. The languages have suffered an enormous degree of language death. About fifty percent of the original languages are now extinct. Many

others are dying. Some scholars claim that all the indigenous languages of Australia are related. Others class the large majority into a PAMA-NYUNGAN family, distributed over most of Australia, and assume a certain number of smaller genetic groups for the remaining languages, many of which are found in the northwest. But many details of these and other proposed genetic classifications still need to be worked out. In the meantime, Australian languages continue to die at a rapid rate; and with the languages the evidence dies out that they might contribute to a more complete understanding of Australian linguistic relationships.

In addition, we can mention various SIGN(ED) LANGUAGES and the question of their genetic affiliation. The number of such manually based languages is generally assumed to be very large – at least in the hundreds, but possibly in the thousands. In their natural state (i.e., leaving aside codes such as fingerspelled versions of spoken languages), true signed languages are unrelated to their "co-territorial" oral languages. For instance, American Sign Language (ASL) has nothing to do with American English, either historically, or structurally, or lexically. The same holds true for the relationship between British Sign Language (BSL) and British English, French Sign Language (FSL) and French, and so on. Interestingly, however, ASL and FSL are related to each other historically, and neither is related to BSL. FSL originated around 1760 through the efforts of a French teacher to the deaf, Abbé de l'Épée, and later spread to America (where it became the basis for ASL), to Russia, to Ireland (from where it spread to Australia), and to several other European countries, whose sign languages thus are related and form a language family. Through similar developments, Japanese and Korean Sign Language are related to each other. BSL and Chinese Sign Language, by contrast, constitute something like signed counterparts to oral language isolates such as Basque.

Our knowledge of relatedness among signed languages is partly based on what is known about their historical spread, but also on applying the standard methods of comparative linguistics – by comparing systematic similarities and differences in hand shapes, hand orientation, and hand movements for particular signs, in the meanings associated with these signs, and in the morphology and syntax of signed languages. Thus, just as examples in the earlier chapters have shown that sign languages are affected by the same kinds of linguistic change that are observable in oral languages, so also it is true that the principles of comparative linguistics apply equally well to signed languages as to oral languages.