CHAPTER EIGHT

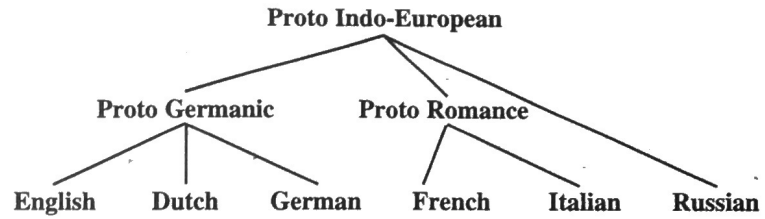
SUBGROUPING

By using the comparative method, not only can we reconstruct a proto-language, but we can use the results that it provides to determine which languages are more closely related to other languages in a family. Compare the following words in six Indo-European languages:

English	Dutch	German	French	Italian	Russian
wʌn	e:n	ains	œ̃	uno	adʲin
tu:	twe:	tsvai	dø	due	dva
θɹi:	dri:	dʁai	tʁwa	treˌ	trʲi
fɔ:	fi:r	fi:ʁ	katʁ	kwatro	ʧetirʲe
faɪv	fɛif	fynf	sɛ̃k	ʧiŋkwe	pʲatʲ

There are enough similarities even here, in the words for 'two' and 'three', for example, to suggest that we could justify putting these six languages into a single language family. However, there are other similarities that seem to suggest that English, Dutch, and German are closer to each other than they are to the other three languages. Similarly, French and Italian seem to be fairly closely related to each other, while being less closely related to the others. Finally, Russian seems to stand out on its own. What we can say here is that we have three *subgroups* of the one language family — one containing the first three languages, one containing the next two, and a final subgroup with only a single member.

We can represent subgrouping in a family tree by a series of branches coming from a single point. The family tree for the six languages described above would look something like this:



This diagram can be interpreted as meaning that English, Dutch and German are all derived from a common protolanguage (which we can call Proto Germanic) that is itself descended from the protolanguage that is ancestral to all of the other languages (which we can call Proto Indo-European). We can therefore offer a tentative definition of a subgroup by saying that it comprises a number of languages that are all descended from a common protolanguage that is intermediate between the ultimate (or highest level) protolanguage and the modern language, and which are as a result more similar to each other than to other languages in the family.

8.1 SHARED INNOVATION AND SHARED RETENTION

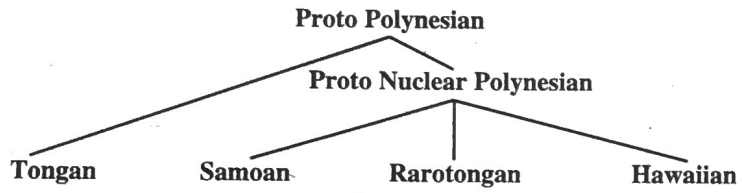
Clearly, languages that belong to the same subgroup must share some similarities that distinguish them from other languages in the family that do not belong to this subgroup. However, the simple fact that there are similarities does not *necessarily* mean that two languages belong in the same subgroup. If we say that two languages belong in the same subgroup, we imply that they have gone through a *period of common descent*, and that they did not diverge until a later stage in their development.

Similarities between languages can be explained as being due to either *shared retention* from the protolanguage, or *shared innovations* since the time of the protolanguage. If two languages are similar because they share some feature that has been retained from the protolanguage, you cannot use this similarity as evidence that they have gone through a period of common descent. The retention of a particular feature in this way is not significant, because you should expect a large number of features to be retained anyway.

However, if two languages are similar because they have both undergone the same innovation or change, then you can say that this is evidence that they have had a period of common descent and that they therefore do belong to the same subgroup. You can say that a shared innovation in two languages is evidence that those two languages belong in the same subgroup, because exactly the same change is unlikely to take place independently in two separate languages. By suggesting that the languages have undergone a period of common descent, you are saying that the particular change took place only once between the higher level protolanguage and the intermediate proto-language which is between this and the various modern languages that belong

in the subgroup. Other changes then took place later in the individual languages to differentiate one language from another within the subgroup.

If you look back to the reconstructions that you made for Proto Polynesian in Chapter 5, you will see that Samoan, Rarotongan, and Hawaiian have all undergone unconditional loss of the original phonemes /*h/ and /*ʔ/. This suggests that Samoan, Rarotongan, and Hawaiian all belong together in a subgroup of Polynesian from which Tongan is excluded. Between Proto Polynesian and the intermediate ancestor language from which these three languages are derived (but not Tongan), there was an intermediate proto-language which we can call Proto Nuclear Polynesian:



While it is shared innovations that we use as evidence for establishing subgroups, certain kinds of innovations are likely to be stronger evidence for subgrouping than other kinds. As I have just said, subgrouping rests on the assumption that shared similarities are unlikely to be due to *chance*. However, some kinds of similarities between languages are in fact due to chance, i.e. the same changes *do* sometimes take place quite independently in different languages. This kind of situation is often referred to as *parallel development* or *drift*. One good example of drift is in the Oceanic subgroup of the Austronesian family of languages (which includes all of the Polynesian languages, as well as Fijian, and the Austronesian languages of Fiji, Vanuatu, New Caledonia, Solomon Islands, and Papua New Guinea). In Proto Oceanic, word final consonants were apparently retained from Proto Austronesian. However, many present-day Oceanic languages have since apparently lost word final consonants by a general rule of the form:

C → ø / ___ #

The fact that many Oceanic languages share this innovation is not sufficient evidence to establish subgroups. Loss of final consonants is a very common sort of sound change that could easily be due to chance, and the same sound change occurs in Oceanic as well as in some languages that we would not otherwise want to call Oceanic languages. In the Enggano language, spoken on an island off the coast of southern Sumatra, final consonants were also lost, but we would not necessarily want to say that this language belongs in

the Oceanic subgroup as this language shares no other features of Oceanic languages.

In classifying languages into subgroups, you therefore need to avoid the possibility that innovations in two languages might be due to drift or parallel development. You can do this by looking for the following in linguistic changes:

(i) Changes that are particularly unusual.
(ii) Sets of several phonological changes, especially unusual changes which would not ordinarily be expected to have taken place together.
(iii) Phonological changes which correspond to unconnected grammatical or semantic changes.

For example, if Samoan, Rarotongan, and Hawaiian only shared the single change whereby /*h/ was lost, it might be possible to argue that this is purely coincidental, especially as the loss of /h/ is a fairly common sort of change anyway. However, as these three languages also share the change:

ʔ → ø

we can argue that coincidence is less likely to be the explanation and that these three languages are indeed members of a single subgroup.

If two languages share a common sporadic or irregular phonological change, this provides even better evidence for subgrouping those two languages together, as the same irregular change is unlikely to take place twice independently. One piece of evidence that can be quoted for the grouping of Oceanic languages into a single subgroup of Austronesian is the irregular loss of /*R/ that has taken place in the Proto Austronesian word ***maRi** 'come'. On the basis of evidence from the present-day Oceanic languages, we can reconstruct the form **/*mai/** 'come' in Proto Oceanic. On the basis of the reconstructed Proto Austronesian form **/*maRi/**, however, we would have expected the Proto Oceanic form to be **/*mari/** instead of **/*mai/**. Proto Oceanic appears to have lost this sound in just this single word to produce an irregular reflex of **/*maRi/**. It is highly unlikely that every single Oceanic language would have independently shifted **/*maRi/** to **/*mai/**, so we conclude instead that this irregular change happened just once, between Proto Austronesian and Proto Oceanic, and that the modern Oceanic languages reflect this irregularity as a retention from Proto Oceanic.

The Oceanic subgroup of the Austronesian family has not been established on the basis of just this single innovation, even though it is an irregular one. There are several other regular phonological changes that have also taken place at the same time. These include the following:

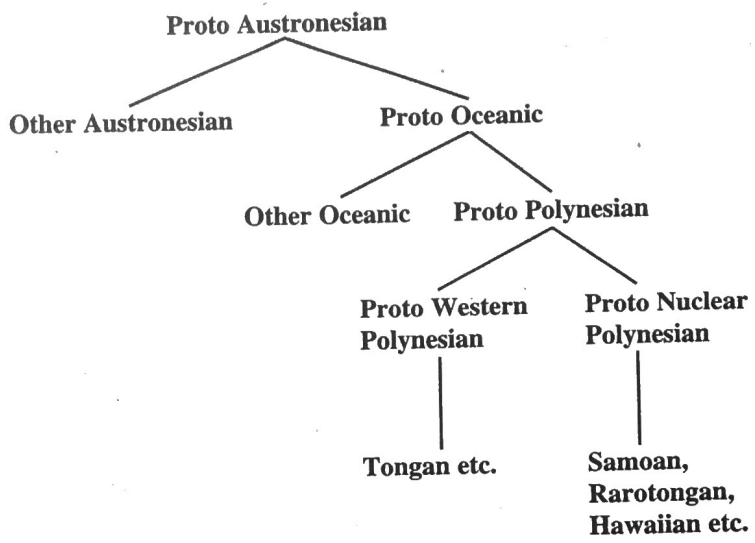
ə → o
b → p
g → k

These involve a change of schwa to /o/, as well as the devoicing of stops, so parallel development is unlikely to be the explanation. We can therefore conclude that any Austronesian language that shares all of these innovations is a member of the Oceanic subgroup.

The pair of shared innovations that I gave above in Samoan, Rarotongan, and Hawaiian are also better evidence for subgrouping than just a single change. For instance, both Tongan and Hawaiian have undergone a shift of /*s/ to /h/. It would contradict the conclusion that I just reached to say that Tongan and Hawaiian belong to a single subgroup on the basis of this shared innovation. Where there is information that is consistent with competing subgrouping interpretations, we should evaluate this and see which solution is the most reasonable one. The fact that the first conclusion was reached on the basis of a pair of shared innovations, whereas the second conclusion would have to be based on just a single innovation, makes the first conclusion a more reliable one. We must simply conclude that both Tongan and Hawaiian independently changed /*s/ to /h/ at separate times in history after the two had diverged.

Finally, if we can match phonological innovations with shared grammatical or semantic innovations, then we can argue that we have good evidence for putting the languages that share these features into the same subgroup. Although the grammatical reconstruction of Proto Austronesian is much less well developed than its phonological reconstruction, there are some linguists who argue that there are many aspects of the basic clause structure of Oceanic languages that are different from that of Proto Austronesian. If this turns out to be confirmed, then this would be further evidence for the existence of an Oceanic subgroup.

When we speak of subgroups of languages, it is possible to speak of *higher level subgroups* and *lower level subgroups*. As you have seen, languages that belong to a subgroup within a single language family have experienced a period of common descent. However, it is possible for languages within a single subgroup of a larger language family also to be subgrouped together on the basis of shared innovations. This means that we can speak of subgroups *within* subgroups. For instance, there are strong arguments for saying that the Polynesian languages represent a separate subgroup within the Oceanic subgroup, on the basis of their shared phonological, lexical, and grammatical innovations. In this kind of situation, we can speak of Oceanic being a higher-level subgroup, while the Polynesian languages constitute a lower-level subgroup. Languages that belong together in higher-level subgroups therefore diverged relatively early, while lower-level subgroups involve later developments. Of course, the Polynesian languages can be further subgrouped into even lower-level subgroups again, and I have already indicated that we can justify a subgroup consisting of Samoan, Rarotongan, and Hawaiian, as well as a Western Polynesian subgroup, of which Tongan is a member. We could represent the different levels of subgrouping as follows:



8.2 LEXICOSTATISTICS AND GLOTTOCHRONOLOGY

There is a another rather different technique for subgrouping languages that is often used with languages for which there are relatively limited amounts of data available, and that is *lexicostatistics*. Since Melanesia and Australia are areas of great linguistic diversity, and because comparatively few of these languages are well known to linguists, this is a technique that has to date been used very frequently in trying to determine the nature of interrelationships in that part of the world (though this technique is not frequently used when comparing better known languages). We therefore need to have a good understanding of how linguists have applied this technique, as well as the strengths and weaknesses of the technique as it has been applied.

Lexicostatistics is a technique that allows us to determine the degree of relationship between two languages, simply by comparing the vocabularies of the languages and determining the degree of similarity between them. This method operates under two basic assumptions. The first of these is that there are some parts of the vocabulary of a language that are much less subject to lexical change than other parts, i.e. there are certain parts of the lexicon in which words are less likely to be completely replaced by non-cognate forms. The area of the lexicon that is assumed to be more resistant to lexical change is referred to as *core vocabulary* (or as *basic vocabulary*).

There is a second aspect to this first general assumption underlying the lexicostatistical method, and that is the fact that this core of relatively change-resistant vocabulary is the same for all languages. The universal core vocabulary includes items such as pronouns, numerals, body parts,

geographical features, basic actions, and basic states. Items like these are unlikely to be replaced by words copied from other languages, because all people, whatever their cultural differences, have eyes, mouths, and legs, and know about the sky and clouds, the sun, and the moon, stones, and trees, and so on. Other concepts, however, may be *culture-specific*, or known only to people of certain cultures. The word 'canoe', for example, is culture-specific, because somebody who grew up in the desert of central Australia would be unlikely to have a word to express this meaning in their language. Similarly, the word for 'boomerang' would also be culture-specific, because not all cultures have such implements. Such words are generally found much more likely to have been copied. In fact, the English word 'boomerang' was copied from an Australian language after Captain Cook first reported seeing these over 200 years ago. Not surprisingly, he recorded the original word **/bumaraŋ/** incorrectly, and it is Captain Cook's incorrect spelling that we follow when we pronounce the word 'boomerang' in English today.

The contrast between the amount of lexical change that takes place in the core vocabulary as against the *peripheral vocabulary* (or the general vocabulary) can be seen by looking at the vocabulary of English. If you take the dictionary of English as a whole, you will find that about 50 per cent of the words have been copied from other languages. Most of these have been copied directly from French, as there has been massive lexical influence from French on English over the last 900 years. Many other words have been copied from forms that were found in ancient Latin and Greek. French has also taken many words from the same languages, which makes the lexicons of English and French appear even more similar, even with words that were not directly copied from French into English. However, if we restrict ourselves just to the core vocabularies of French and English, we find that there is much less sharing of cognate forms, and the figure for words copied from French into English in this area of the lexicon drops to as low as 6 per cent.

The second assumption that underlies the lexicostatistical method is that the actual *rate* of lexical replacement in the core vocabulary is more or less stable, and is therefore about the same for all languages over time. In peripheral vocabulary, of course, the rate of lexical replacement is not stable at all, and may be relatively fast or slow, depending on the nature of cultural contact between speakers of different languages. This second assumption has been tested in 13 languages for which there are written records going back over long periods of time. It has been found that there has been an average vocabulary retention of 80.5 per cent every 1000 years. That is to say, after 1000 years a language will have lost about a fifth of its original basic vocabulary and replaced it with new forms.

If these assumptions are correct, then it should be possible to work out the degree of relationship between two languages by calculating the degree of similarity between their core vocabularies. If the core vocabularies of two languages are relatively similar, then we can assume that they have diverged quite recently, and that they therefore belong to a lower level subgroup. If, on the other hand, their core vocabularies are relatively dissimilar, then we can assume that they must have diverged at a much earlier time, and that they therefore belong to a much higher level of subgrouping.

Different levels of subgrouping have been given specific names by lexicostatisticians, as follows:

Level of subgrouping	Shared cognate percentage in core vocabulary
dialects of a language	81–100
languages of a family	36–81
families of a stock	12–36
stocks of a microphylum	4–12
microphyla of a mesophylum	1–4
mesophyla of a macrophylum	0–1

You should note immediately that lexicostatisticians are using the term *family* in a completely different way from the way we have been using it in this textbook. I (and most other historical linguists) take the term *family* to refer to *all* languages that are descended from a common ancestor language, no matter how closely or distantly related they are to each other within that family. According to a lexicostatistical classification, however, a *family* is simply a particular level of subgrouping in which the members of that subgroup share more than 36 per cent of their core vocabularies. Languages that are in lesser degrees of relationship (but still presumably descended from a common ancestor) are not considered to be in the same family, but in the same *stock* or *phylum*.

Having outlined the assumptions behind lexicostatistics and the theory behind its application, I will now go on to show how lexicostatisticians have followed this method. The first problem is to distinguish the so-called core vocabulary from the peripheral vocabulary. I gave some indication earlier about the kinds of words that would need to go into such a list. But how long should it be? Some have argued that we should use a 1000-word list, others a 200-word list, and others a 100-word list. (Notice how the lengths of these lists all involve numbers that can easily divided by 100 to produce a percentage. One suspects that these lists are not being drawn up according to any firm linguistic criterion about what can be shown to be 'basic' as against 'peripheral' vocabulary, but merely to make the lexicostatisticians' task of calculation easier.) It would be awkward to insist on a 1000-word list for the languages of Australia and Melanesia where many languages are only very sketchily recorded and linguists do not have access to word lists of this length. Many people think that a 100-word list is too short and the risk of error is too great, so most lexicostatisticians tend to operate with 200-word lists. The most popular list of this length is known as the *Swadesh list*, which is named after the linguist Morris Swadesh who drew it up in the early 1960s. This list comprises the following items:

all	*dull*	*heart*	*neck*	*skin*	*turn*
and	*dust*	*heavy*	*new*	*sky*	*twenty*
animal		*here*	*night*	*sleep*	*two*
ashes	*ear*	*hit*	*nose*	*small*	
at	*earth*	*hold/take*	*not*	*smell*	*vomit*
	eat	*horn*		*smoke*	
back	*egg*	*how*	*old*	*smooth*	*walk*
bad	*eight*	*hundred*	*one*	*snake*	*warm*
bark	*eye*	*hunt*	*other*	*snow*	*wash*
because		*husband*		*some*	*water*
belly	*fall*		*person*	*spear*	*we*
big	*far*	*I*	*play*	*spit*	*wet*
bird	*fat*	*ice*	*pull*	*split*	*what*
bite	*father*	*if*	*push*	*squeeze*	*when*
black	*fear*	*in*		*stab/pierce*	*where*
blood	*feather*		*rain*	*stand*	*white*
blow	*few*	*kill*	*red*	*star*	*who*
bone	*fight*	*knee*	*right/correct*	*stick*	*wide*
breast	*fire*	*know*	*right side*	*stone*	*wife*
breathe	*five*		*river*	*straight*	*wind*
brother	*float*	*lake*	*road*	*suck*	*wing*
burn	*flow*	*laugh*	*root*	*sun*	*wipe*
	flower	*leaf*	*rope*	*swell*	*with*
child	*fog*	*left side*	*rotten*	*swim*	*woman*
claw	*foot*	*leg*	*rub*		*woods*
clothing	*four*	*live*		*tail*	*work*
cloud	*freeze*	*liver*	*salt*	*ten*	*worm*
cold	*fruit*	*long*	*sand*	*that*	
come	*full*	*louse*	*say*	*there*	*ye*
cook			*scratch*	*they*	*year*
count	*give*	*man/male*	*sea*	*thick*	*yellow*
cut	*good*	*many*	*see*	*thin*	
	grass	*meat/flesh*	*seed*	*think*	
dance	*green*	*moon*	*seven*	*this*	
day	*guts*	*mother*	*sew*	*thou*	
die		*mountain*	*sharp*	*three*	
dig	*hair*	*mouth*	*shoot*	*throw*	
dirty	*hand*		*short*	*tie*	
dog	*he*	*name*	*sing*	*tongue*	
drink	*head*	*narrow*	*sister*	*tooth*	
dry	*hear*	*near*	*sit*	*tree*	

Even with this list, there are problems in applying it to some of the languages of Melanesia, Australia, and the South Pacific. Firstly, it contains words like *and* and *in*, which in some of these languages are not expressed as

separate words, but as affixes of some kind. It contains the separate words *woman* and *wife*, even though in many languages both of these meanings are expressed by the same word. It contains words such as *freeze* and *ice* which are clearly not applicable in languages spoken in tropical areas. There are other words which *could* be included in a basic vocabulary for Pacific languages and which would not be suitable for other languages, for example: *canoe*, *bow and arrow*, *chicken*, *pig*, and so on. A basic vocabulary for Australian languages could, of course, include items such as *kangaroo* and *boomerang*.

Let us avoid the problem of exactly what should be considered basic vocabulary, and go on to see how we use a basic word list of this kind in a language in order to determine its relationship to another language. The first thing that you have to do is to examine each pair of words for the same meaning in the two languages, to see which ones are cognate and which ones are not. Ideally, whether a pair of words are cognate or not should be decided only after you have worked out the systematic sound correspondences between the two languages. If there are two forms which are phonetically similar but which show an exceptional sound correspondence, you should assume that there has been lexical copying, and the pair of words should be excluded from consideration. It is very important that you exclude copied (or borrowed) vocabulary when you are working out lexicostatistical figures, as these can make two languages appear to be more closely related to each other than they really are.

Let us now look at an actual problem. I will use the lexicostatistical method to try to subgroup the following three languages from Central Province in Papua New Guinea: Koita, Koiari, and Mountain Koiari. Rather than use a full 200-word list, I will make things simpler by using a shorter 25-word list and assume that it is representative of the fuller list:

	Koita	Koiari	Mountain Koiari	
1.	ɣata	ata	maraha	'man'
2.	mavi	mavi	keate	'woman'
3.	moe	moe	mo	'child'
4.	ɣamika	vami	mo ese	'boy'
5.	mobora	mobora	koria	'husband'
6.	mabara	mabara	keate	'wife'
7.	mama	mama	mama	'father'
8.	neina	neina	neina	'mother'
9.	da	da	da	'I'
10.	a	a	a	'you (singular)'
11.	au	au	ahu	'he, she, it'
12.	omoto	kina	kina	'head'
13.	hana	homo	numu	'hair'
14.	uri	uri	uri	'nose'
15.	ihiko	ihiko	gorema	'ear'

16.	meina	neme	neme	'tongue'
17.	hata	auki	aura	'chin'
18.	ava	ava	aka	'mouth'
19.	dehi	gadiva	inu	'back'
20.	vasa	vahi	geina	'leg'
21.	vani	vani	fani	'sun'
22.	vanumo	koro	didi	'star'
23.	gousa	yuva	goe	'cloud'
24.	veni	veni	feni	'rain'
25.	nono	hihi	heburu	'wind'

The first thing that you have to do is distinguish cognate forms from forms that are not cognate. One way in which you can do this is mark how many *cognate sets* there are to express each meaning. For instance, in the word for 'man' (1), there are two cognate sets, as Koita and Koiari have forms that are clearly cognate (i.e. **vata** and **ata** respectively), whereas Mountain Koiari has **maraha**. You can therefore label the first set as belong to Set A, and the second as belong to Set B:

	Koita	Koiari	Mountain Koiari	
1.	A	A	B	'man'

On the other hand, the word for 'chin' (17) is quite different in all three languages, so we would need to recognise three different cognate sets:

	Koita	Koiari	Mountain Koiari	
17.	A	B	C	'chin'

Finally, the word for 'sun' (21) is clearly cognate in all three languages, so you would need to recognise only a single cognate set:

	Koita	Koiari	Mountain Koiari	
21.	A	A	A	'sun'

I will now set out the cognate sets for each of these three languages on the basis of the information that I have just given you:

	Koita	Koiari	Mountain Koiari	
1.	A	A	B	'man'
2.	A	A	B	'woman'
3.	A	A	A	'child'
4.	A	A	B	'boy'
5.	A	A	B	'husband'
6.	A	A	B	'wife'
7.	A	A	A	'father'
8.	A	A	A	'mother'
9.	A	A	A	'I'
10.	A	A	A	'you (singular)'
11.	A	A	A	'he, she, it'
12.	A	B	B	'head'
13.	A	B	C	'hair'
14.	A	A	A	'nose'
15.	A	A	B	'ear'
16.	A	B	B	'tongue'
17.	A	B	C	'chin'
18.	A	A	B	'mouth'
19.	A	B	C	'back'
20.	A	B	C	'leg'
21.	A	A	A	'sun'
22.	A	B	C	'star'
23.	A	B	C	'cloud'
24.	A	A	A	'rain'
25.	A	B	C	'wind'

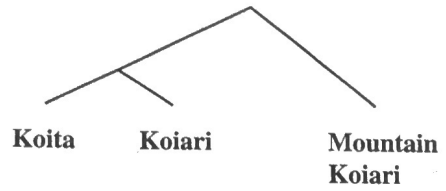
Now you need to work out the degree to which each pair of languages among the three represented above shares cognates. Firstly, examine the pair Koita and Koiari. If you count the number of pairs in these two languages which are marked as cognate (i.e which are both marked A) and those which are marked as non-cognate (i.e. in which one is marked A and the other is marked B), you will find that there are 16 forms which are shared between the two languages, and 9 which are not. From this, you can say that 16/25 of the core vocabulary of these two languages is cognate. If you do this for the remaining pairs of languages from the three languages that we are considering, you will end up with three fractions, which can be set out in the following way:

Koita		
$^{16}/_{25}$	Koiari	
$^{9}/_{25}$	$^{12}/_{25}$	Mountain Koiari

You should now convert these figures to percentages:

Koita		
64%	Koiari	
36%	48%	Mountain Koiari

Now that you have the cognate percentage figures, you need to know how to interpret them. Clearly, Koita and Koiari are more closely related to each other than either is to Mountain Koiari. On the basis of these figures, you could therefore draw a family tree of the following kind:



In terms of the degrees of relationship that I talked about earlier, these languages would all be contained within a single 'family', i.e. they share between 36 per cent and 81 per cent of their core vocabularies.

This was a rather simple example, because we considered only three languages. Although the same principles apply when we are considering cognate percentages for larger numbers of languages, the procedures for working out the degrees of relationship can become rather more complex. Let us take the following lexicostatistical figures for 10 hypothetical languages and interpret the data according to these same principles:

A								
91%	B							
88	86%	C						
68	62	64%	D					
67	65	66	63%	E				
55	51	56	53	55%	F			
57	53	54	57	56	89%	G		
23	27	36	31	32	30	29%	H	
25	28	33	29	27	34	22	88%	I
31	22	30	27	28	26	28	86	89% J

Where do you start from in a more complicated case like this? The first step is to try to find out which languages in the data are *most* closely related to each other. To do this, you should look for figures that are significantly higher than any other figures in the table, which is an indication that these particular pairs of languages are relatively closely related to each other. On this table, therefore, the sets of figures that are set in bold type are noticeable in this respect:

A								
91%	B							
88	**86%**	C						
68	62	64%	D					
67	65	66	63%	E				
55	51	56	53	55%	F			
57	53	54	57	56	**89%**	G		
23	27	36	31	32	30	29%	H	
25	28	33	29	27	34	22	**88%**	I
31	22	30	27	28	26	28	**86**	**89%** J

Communities A, B, and C are clearly very closely related to each other. Communities F and G also belong together, and so do the three communities H, I, and J.

Now you need to find out what is the next level of relationship. To make this task easier, you can now treat the subgroups that you have just arrived at as single units for the purpose of interpretation. To do this, you should relabel the units so that it is clear to you that you are operating with units at a different level of subgrouping. You can use the following labels:

ABC	I
D	II
E	III
FG	IV
HIJ	V

Now work out the shared cognate percentages between these five different lower level units, in order to fill in the information on the table below:

I				
II				
III				
IV				
V				

Where the new label corresponds to a single language on the original table, you can simply transfer the old figures across to the appropriate places on the new table:

I				
II				
63%	III			
IV				
V				

However, where the new labels correspond to a number of different communities on the original table, you will need to get the averages of the shared cognate figures in each block and enter them in the appropriate place in the new table. So, in comparing I and II, you will need to get the figures for the shared cognates of A with D, of B with D, and C with D and enter the average of those figures under the intersection of I and II. Since A and D have 68 per cent cognate sharing, B and D have 62 per cent, and C and D share 64 per cent of their cognates, the average level of cognate sharing between I and II works out at 65 per cent. So, you can now add one more figure to the table:

I
65% II
63% III
IV
V

If you do this methodically for every pair of groupings, you will end up with the following table:

I
65% II
66 63% III
54 55 55% IV
28 29 29 27% V

You should now treat this table in the same way as you treated the first table — simply look for the highest cognate figures as an indication of the next level of linguistic relationship. From these figures it seems that I, II, and III are more closely related to each other than to either IV or V, as the shared cognate percentages range above 60 per cent, whereas they are in the 20–60 per cent range for the other groups. For the next step, you should group together I, II and III in the same way and relabel them (this time as, say, X, Y, and Z) as follows:

I, II, III X
IV Y
V Z

Once again, calculate the averages of the cognate figures, which will work out to be as follows:

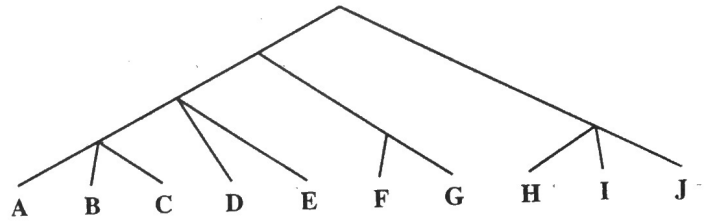
X
55% Y
29% 27% Z

It is clear from this final table that X and Y are more closely related to each other than either is to Z.

The final step in the procedure is to gather all of these facts together and represent the conclusions on a family tree that will clearly indicate the degrees of relationship between the ten original speech communities. At the lowest level of relationship, you will discover that the following units belong together, while D and E are on their own:

A, B, C
F, G
H, I, J

At the next level of relationship, you find that D and E belong to the same group as A, B, and C, while the others are all on their own. At the next level, F and G could be related to the same subgroup as the subgroup consisting of A, B, C, D, and E, with H, I, and J being a separate subgroup of their own. This situation can be represented in a family tree diagram in the following way:



Having dealt in some detail with the claim that lexicostatistics enables us to work out degrees of relationship within a language family, I will now go on to discuss a second claim that lexicostatisticians sometimes make, though most linguists are now very cautious about this. If we accept the basic assumption that languages change their core vocabulary at a relatively constant rate, we should be able to work out not only the degree of relationship between two languages, but also the actual period of time that two languages have been separated from each other. Once the percentage of cognate forms has been worked out, we can use the following mathematical formula to work out the *time depth*, or the period of separation, of two languages:

$$t = \frac{\log C}{2\log r}$$

In the formula above, t stands for the number of thousands of years that two languages have been separated, C stands for the percentage of cognates as worked out by comparing basic vocabularies, and r stands for the constant change factor mentioned earlier (the value in this formula is set at 0.805). The following table gives logarithms for numbers less than 1 for use in calculations of this type:

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.1	2.303	2.207	2.120	2.040	1.966	1.897	1.833	1.772	1.715	1.662
.2	1.609	1.561	1.514	1.470	1.427	1.386	1.347	1.309	1.273	1.238
.3	1.204	1.171	1.139	1.109	1.079	1.050	1.022	.994	.968	.942
.4	.916	.892	.868	.844	.821	.799	.777	.755	.734	.713
.5	.693	.673	.654	.635	.616	.598	.580	.562	.545	.528
.6	.511	.494	.478	.462	.446	.431	.416	.400	.386	.371
.7	.357	.342	.329	.315	.301	.288	.274	.261	.248	.236
.8	.223	.211	.198	.186	.174	.163	.151	.139	.128	.117
.9	.105	.094	.083	.073	.062	.051	.041	.030	.020	.010

So, going back to the earlier problem involving Koita, Koiari, and Mountain Koiari, if you wanted to know how long it has been since Koiari split off from Koita, you would take the cognate percentage of 64 per cent (the figure given on the table for these two languages) and convert it to a factor of one (0.64) and apply the formula:

$$t = \frac{\log.64}{2\log r}$$

$$t = \frac{\log.64}{2\log.805}$$

$$t = \frac{.446}{2 \text{ x } .217}$$

$$t = \frac{.446}{.434}$$

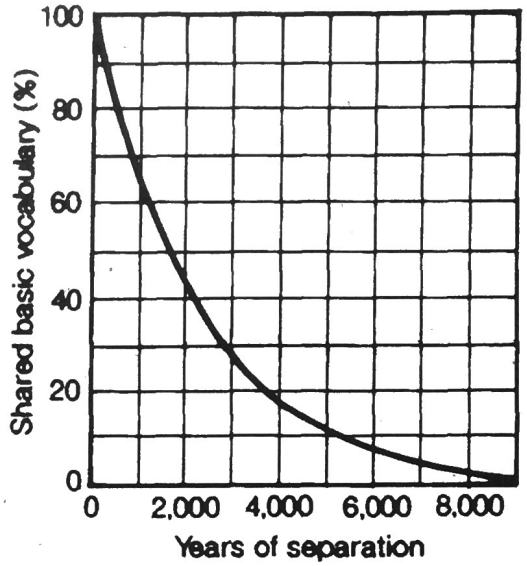
$$t = 1.028$$

This means that Koita and Koiari must have diverged 1.028 thousand years ago (i.e. 1028 years ago), which rounds off to about 1000 years. The Koiari migration from the mountains to the coast should therefore have taken place just before the French invaded the English in the famous Battle of Hastings in 1066.

This method of dating the divergence of languages is known as *glottochronology*. Following this methodology, it is possible to give approximate dates for the 'age' of the different degrees of relationships between languages. Thus:

Level of subgrouping	Years of separation
dialects of a language	less than 500 years
languages of a family	500 to 2500 years
families of a stock	2500 to 5000 years
stocks of a microphylum	5000 to 7500 years
microphyla of a mesophylum	7500 to 10000 years
mesophyla of a macrophylum	more than 10000 years

The following graph can be used to calculate roughly the time depth of any linguistic separation on the basis of the share cognate figures between two languages, if you want to avoid doing the calculations with logarithms that I have just shown you:



The techniques of lexicostatistics and glottochronology have not been without their critics. I have already hinted at a number of practical problems that are associated with these methods. Firstly there is the problem of deciding which words should be regarded as core vocabulary and which should not. Obviously, it may be possible for different sets of vocabulary to produce differing results.

Another difficulty involves the actual counting of forms that are cognate against those that are not cognate in basic vocabulary lists from two different languages. As I said earlier, ideally, copied vocabulary should be excluded from cognate counts, but to do this you need to know what the regular sound correspondences are between the two languages in order to exclude exceptional forms which are probably copied. However, since we are working with fairly short word lists, there may not be enough data to make generalisations about sound correspondences. Also, we are not likely to know much about the protolanguage if we are dealing with languages for which we have only limited amounts of data, and this will make it even more difficult to distinguish genuine cognates from copied vocabulary.

Lexicostatisticians in fact tend to rely heavily on what is often euphemistically called the *inspection method* of determining whether two forms are cognate or not in a pair of languages. What this amounts to is that you are more or less free to apply intelligent guesswork as to whether you think two forms are cognate or not. If two forms *look* cognate, then they can be given a 'yes' score, but if they are judged not to look cognate, then they are given a 'no' score.

Of course, two different linguists can take the same lists from two different languages, and since there is no objective way of determining what should be

ticked 'yes' and what should be ticked 'no', it is possible that both will come up with significantly different cognate figures at the end of the exercise. For example, I have done counts on the basis of word lists calculated by other people and have ended up with figures between 10 per cent and 20 per cent higher or lower than their count. Of course, if two different scholars compare the same pair of languages and one comes up with a figure of 35 per cent cognate sharing, and the other concludes that there is 45 per cent cognate sharing, then one is going to have to say that the two represent different families within the same stock, while the other will end up saying that they are from two languages within the same family. In glottochronological terms, this could mean a difference in time-depth of up to 600 years.

A further problem that arises in the use of lexicostatistical figures to indicate degrees of linguistic relationship is that different linguists some times use different cut-off points for different levels of subgrouping, and there is not even agreement on what sets of terminology should be used to refer to different subgroups of languages. Compare the following two systems that have been widely used for classifying Pacific languages (in which the first system was the one that I gave you earlier in this section):

Level of subgrouping	Shared cognate percentage in core vocabulary
System A	
dialects of a language	81–100
languages of a family	36–81
families of a stock	12–36
stocks of a microphylum	4–12
microphyla of a mesophylum	1–4
mesophyla of a macrophylum	0–1
System B	
dialects of a language	81–100
languages of a sub-family	55–81
subfamilies of a family	28–55
families of a stock	13–28
stocks of a phylum	5–13

You can see from the lists above that the term *family* can actually be used by lexicostatisticians with two different meanings. While I have used the term in this textbook to refer to all languages descended from a protolanguage, according to System A above, a language family refers only to languages that share more than 36 per cent of their core vocabulary, while according to System B, languages in the same family must share more than 55 per cent of their core vocabularies. Other lexicostatisticians use terms that are not included in either of these lists to refer to particularly distant degrees of lexical relationships. Scholars investigating the interrelationships in the huge

number of Austronesian languages, for example, sometimes also talk about *linkages* of languages as representing a particularly low level of subgrouping.

Another major problem in applying these figures is that they appear to be fairly arbitrary. The only figure that lexicostatisticians seem to be able to agree on reasonably consistently is that of about 81 per cent cognate similarity as representing the cut-off point between dialects and languages. It seems that as soon as speakers of two different speech traditions (and I use this term as a cover term to include both dialect and language as I do not want to make a distinction between the two at this point) have more than about a 20 per cent difference in their basic lexicons, then mutual intelligibility is lost. Beyond the language-dialect distinction, however, the choice of cognate percentage figures seems to have been based purely on whim, and has no sound scientific basis. So, we should seriously ask ourselves this question: how useful is this method of subgrouping if it has a scientific basis only at the very lowest level of subgrouping?

Apart from these practical problems, there are some more basic *theoretical* objections to these methods, which tend to destroy the validity of the underlying assumptions that I presented earlier. First, we need to question the validity of the assumption that there is a constant rate of lexical replacement in core vocabulary for all languages over time, and that this rate of replacement is 19.5 per cent every 1000 years. This figure was arrived at by testing only 13 of the world's languages, and these were languages with long histories of writing, and 11 were Indo-European languages. However, differing cultural factors can affect the speed at which lexical replacement can take place. In Chapter 7, I described how lexical replacement can be accelerated in languages in which even basic vocabulary can become proscribed by taboo. The result of lexical replacement because of taboo is that even basic vocabulary, if given sufficient time, will be subject to replacement. If languages copy words from neighbouring languages in order to avoid a forbidden word, two languages which were originally very different from each other will end up sharing a high proportion of even their core vocabularies.

There is a second theoretical problem with lexicostatistics, and that involves the interpretation of the data. Given that change is random within the core vocabulary, it is logically possible for two languages to change the same 19.5 per cent of their core vocabulary every 1000 years and to retain the remaining 80.5 per cent intact over succeeding periods. It is also possible at the other extreme for two languages that in the beginning shared the same proportions of their core vocabulary to replace 19.5 per cent of their core vocabularies every 1000 years, yet for the 19.5 per cent to be different in each successive period. The result of this will be that two pairs of languages, while separated by the same period of time, might have dramatically different vocabulary retention figures depending on which items were actually replaced. Some languages will be accidentally conservative, while others will accidentally exhibit a high degree of change. Although the time depth would

be the same, we would be forced to recognise two very different degrees of linguistic relationship.

For instance, after 2000 years, it can be expected that the range of core vocabulary retentions will be as low as 10 per cent in a few languages and as high as 80 per cent in a few others, while for most it will be around 64 per cent, as we would have predicted from the figures earlier. The languages that have retained 10 per cent and 80 per cent will *look* quite divergent, though they are in fact separated by the same time period as all of the others. We should not assume, therefore, that simply because two languages share a fairly low figure for cognate sharing, the degree of relationship is necessarily distant. This fact makes it impossible to be certain of the correctness of our interpretation of lexicostatistical data.

Finally, there are often practical difficulties in interpreting lexicostatistical data, for a wide variety of reasons, some of which have already been mentioned, and some of which may apply only in a particular situation. The data presented earlier as an illustration of a subgrouping technique was in fact highly idealised. It is not often that data from real languages produces a completely consistent picture without contradictions in interpretation. A more typical set of lexicostatistical data from a number of speech communities in the Milne Bay area of Papua New Guinea may look more like this:

Mwalakwasia								
87%	Somwadina							
82	86%	Biawa						
72	71	82%	Sigasiga					
74	73	76	78%	Lomitawa				
64	62	84	69	80%	Sipupu			
59	58	55	52	67	79%	Kelologea		
62	59	71	66	79	86	78%	Meudana	
74	72	76	65	74	76	70	71%	Kasikasi
74	74	80	67	72	62	55	66	76% Guleguleu

In this set of data, there do not appear to be many clear discontinuities or breaks between one subgroup and another. The figures seem to merge gradually into each other, producing very little hope of drawing a family tree.

READING GUIDE QUESTIONS

1. What is a subgroup?
2. What is the difference between a shared retention and a shared innovation?
3. Why can similarities between languages that are due to shared retentions not to be used as evidence for subgrouping?

4. What is drift or parallel development? How does this affect the way we go about deciding on subgroups?
5. What sorts of innovations are the best kind of evidence for subgrouping?
6. What is lexicostatistics?
7. What basic assumptions underlie the method of determining linguistic relationships by lexicostatistics?
8. What is the inspection method of determining whether two forms are cognate or not?
9. What is the difference between core and peripheral vocabulary?
10. What is glottochronology?
11. What are some problems associated with lexicostatistics and glottochronology?

EXERCISES

1. Look at the Korafe, Notu, and Binandere forms in Data Set 7. On the basis of the reconstruction of the changes from the protolanguage that you worked out in the exercises at the end of Chapter 5, would you say that Notu belongs to the same subgroup as Korafe or Binandere? Why?
2. Look back at the reconstruction of the protolanguage for Aroma, Hula, and Sinaugoro that you did in the exercises for Chapter 5. What subgrouping hypothesis can you make for these three languages on the basis of shared innovations?
3. Look at the following forms in Proto Gazelle Peninsula (New Britain, Papua New Guinea). What is the subgrouping of the four speech communities that are represented? Give the justification for the answer that you propose. (Note that the superscript vowels represent phonetically reduced sounds that are nearly voiceless, and not stressable.)

Proto Gazelle	Pila-Pila	Nodup	Vatom	Lunga-Lunga	
*rat^{u}	rat	rat^{u}	rat	rat^{u}	'basket'
*vup^{u}	vup	vuvu	vup	vuv^{u}	'fishtrap'
*ram^{u}	ram	ram^{u}	ram	ram^{u}	'club'
*vasian^{i}	vaian	vaian^{i}	vaian	vasian^{i}	'sling'
*saman^{i}	aman	aman^{i}	aman	saman^{i}	'outrigger'
*pal^{i}	pal	pal^{i}	pal	pali	'house'
*liplip^{i}	liplip	livilivu	liplip	-	'fence'
*pem^{u}	pemu	pem^{u}	pem	pem^{u}	'axe'
*pisa	pia	pia	pia	pisa	'ground'
*tirip^{u}	tirip	tirivu	tirip	tiriv^{u}	'green coconut'
*kaban^{i}	kaban	kaban^{i}	kaban	kaban^{i}	'lime'
*up^{u}	up	uvu	-	uv^{u}	'yam'
*talisa	talia	talia	talia	talisa	'nut'
*papi	pap	pavu	pap	-	'dog'

*taŋisⁱ	taŋi	taŋi	taŋi	taŋisⁱ	'cry'
*iapⁱ	iap	iavu	iap	iavⁱ	'fire'
*mulisi	muli	muli	muli	mulisⁱ	'orange'
*beso	beo	beo	beo	beso	'bird'
*lisi	li	lia	li	lisⁱ	'nits'
*sikilikⁱ	ikilik	ikilikⁱ	ikilik	sikilikⁱ	'small'
*tasⁱ	ta	tai	ta	tasⁱ	'sea'

4. The following data comes from four languages spoken in the area of Cape York in northern Queensland in Australia. Examine the reconstructed protolanguage and the descendant forms, and suggest a subgrouping hypothesis on the basis of the shared innovations. There is one set of changes which is problematic for an otherwise strong subgrouping hypothesis. What original sound is involved?

Proto Cape York	Atampaya	Angkamuthi	Yadhaykenu	Wudhadhi	
*kaʈa	ɤaʈa	aʈa	aʈa	-	'rotten'
*kantu	ɤantu	antu	antu	antu	'canoe'
*puŋku	wuŋku	wuŋku	wuŋku	-	'knee'
*ɲaŋka	ɲaŋka	aŋka	aŋka	aŋka	'mouth'
*juku	juku	juku	juku	-	'tree'
*pinta	winta	winta	winta	inta	'arm'
*puŋa	wuŋa	wuŋa	wuŋa	uŋa	'sun'
*ʈipa	lipa	jipa	jipa	-	'liver'
*wapun	wapun	apun	apu	apun	'head'
*wujpu	wujpu	ujpu	ujpu	ujpu	'bad'
*ujpuɲ	ujpuɲ	ujpuɲ	ujpuɲ	ujpuj	'fly'
*ajpaɲ	ajpaɲ	ajpaɲ	ajpaɲ	ajpaj	'stone'
*ʈalan	lalan	jalan	jala	alan	'tongue'
*paɲʈal	waɲʈaw	waɲʈa:	waɲʈa:	-	'yam'
*ɹaɲʈal	ɹaɲʈaw	jaɲʈa:	jaɲʈa:	-	'road'
*pili	wili	wili	wili	-	'run'
*ɹuŋka	ɹuŋka	juŋka	juŋka	uŋka	'cry'
*ɹa	ɹa	ja	ja	-	'throw'
*ɹupal	ɹupaw	jupa:	jupa:	-	'white'
*ɹuʈu	ɹuʈu	juʈu	juʈu	uʈu	'dead'
*pilu	wilu	wilu	wilu	ilu	'hip'
*pupu	wupu	wupu	wupu	upu	'buttocks'
*ŋampu	ŋampu	ampu	ampu	ampu	'tooth'
*maji	maji	aji	aji	aji	'food'
*ɲukal	ɲukaw	uka:	uka:	ukal	'foot'
*miɲa	miɲa	iɲa	iɲa	iɲa	'meat'
*iwuɲ	-	-	iwuɲ	iwuj	'ear'
*ɹapan	ɹapan	japan	japa	-	'strong'

5. Look at the following data from six different languages and answer the questions below:
   (i) How many language families are represented in this data?
   (ii) What are your reasons for saying this?
   (iii) What factors can you suggest to account for the similarities between languages that you say do not belong to a single family?

A	B	C	D	E	F	
mwana	mwana	umwana	baceh	anak	bata	'child'
lia	dila	lila	girjeh	triak	ijak	'cry'
ɲwa	nua	nwa	nuʃidan	minum	inum	'drink'
moto	tija	umulio	ateʃ	api	apoj	'fire'
nne	ia	ne	cæhær	əmpat	ampat	'four'
kilima	mongo	ulupili	tel	bukit	bukid	'hill'
ceka	seva	seka	xændidan	tərtawa	tawa	'laugh'
mguu	kulu	ukuulu	saq	kaki	pa	'leg'
mdo	mokoba	umulomo	læb	bibir	bibig	'lip'
mtu	muntu	umuntu	mærd	oraŋ	tau	'man'
habari	nsangu	iceevo	xæbær	kabar	balita	'news'
moja	mosi	mo	jek	satu	isa	'one'
nabii	mbikudi	umusimicisi	næbij	nabi	propetas	'prophet'
mvua	mvula	imfula	baran	huɟan	ulan	'rain'
merikebu	maswa	ubwato	mærkæb	kapal	bapor	'ship'
dhambi	masumu	icakuvifja	zamb	dosa	kasilanan	'sin'
askari	kinwani	icita	æskær	askar	suldado	'soldier'
kidonda	mputa	icilonda	zæxm	sakit	sakit	'sick'
hutoba	maloŋi	isiwi	xutbæh	xutbah	salita	'speech'
hadhithi	ŋana	icisimicisjo	hædis	cerita	istoria	'story'
hekalu	kinlongo	itempuli	hæjkil	rumah	templo	'temple'
tatu	tatu	tatu	seh	tiga	tatlo	'three'
mti	nti	umuti	dæræxt	pohon	puno	'tree'
bili	zole	vili	do	dua	dalawa	'two'

6. Examine the following data from the Māori language of New Zealand and the now extinct Moriori language of the Chatham Islands. Do a count of the shared cognates between the two and estimate, according to glottochronology when the two languages should have diverged.

Māori	Moriori	
pungarehu	*purungehu*	'ashes'
tuara	*tura*	'back'
kino	*wahike*	'bad'
kiri	*kiri*	'bark'
kopu	*takapu*	'belly'
rahi	*rahi*	'big'

manu	*manu*	'bird'
ngau	*ngahu*	'bite'
pango	*pango*	'black'
pupuhi	*puhi*	'blow'
iwi	*îmi*	'bone'
u	*u*	'breast'
manawa	*manawa*	'breath'
tuakana	*tukana*	'brother'
tamaiti	*timiti*	'child'
ao	*ao*	'cloud'
matao	*matao*	'cold'
haeremai	*haramai*	'come'
tunu	*tunu*	'cook'
kani	*motiha*	'dance'
ao	*ao*	'day'
keri	*keri*	'dig'
paru	*karupuru*	'dirt'
inu	*inu*	'drink'
maroke	*moroke*	'dry'
puhuki	*puhiku*	'dull, blunt'
nehu	*pawa*	'dust'
taringa	*tiringa*	'ear'
fenua	*fenua*	'earth'
kai	*kai*	'eat'
hua	*hu*	'egg'
waru	*tewaru*	'eight'
kanohi	*konehi*	'eye'
hinga	*hingi*	'fall'
momona	*ihara*	'fat'
papa	*papa*	'father'
piki	*piki*	'feather'
ahi	*ahi*	'fire'
ika	*ika*	'fish'
rima	*terima*	'five'
pua	*pua*	'flower'
kohu	*kohu*	'fog'
waewae	*wawae*	'foot'
fa	*tefa*	'four'
hua	*hua*	'fruit'
pai	*humaria*	'good'
tarutaru	*taru*	'grass'
marama	*marama*	'moon'
faea	*matehine*	'mother'
maunga	*maunga*	'mountain'
waha	*waha*	'mouth'

ingoa	*ingoa*	'name'
kaki	*kaki*	'neck'
ihu	*purangaihu*	'nose'
tafito	*tafito*	'old'
tahi	*tehi*	'one'
tangata	*tangata*	'person'
takaro	*hokorereto*	'play'
tika	*tikane*	'right, correct'
paiaka	*purakautimu*	'root'
mataitai	*marurua*	'salt/salty'
onepu	*one*	'sand'
fitu	*tefitu*	'seven'
poto	*poto*	'short'
waiata	*karamiha*	'sing'
kiri	*kiri*	'skin'
moe	*moe*	'sleep'
auahi	*auahi*	'smoke'
maeneene	*maene*	'smooth'
huka	*haware*	'snow'
tao	*tuparipari*	'spear'
roromi	*romi*	'squeeze'
fetu	*fetu*	'star'
noho	*noho*	'stay'
pohatu	*pohatu*	'stone'
kaha	*kaha*	'strong'
momi	*momomi*	'suck'
ra	*ra*	'sun'
huku	*huku*	'swelling'
kau	*rewa*	'swim'
hiore	*hiore*	'tail'
ngahuru	*ngauru*	'ten'
matotoru	*matotoru*	'thick'
tupuhi	*meatae*	'thin'
toru	*toru*	'three'
arero	*warero*	'tongue'
niho	*niho*	'tooth'
rakau	*rakau*	'tree'
huri	*huri*	'turn'
rua	*teru*	'two'
ruaki	*ruaki*	'vomit'
hau	*hau*	'wind'
pakau	*pakau*	'wing'
wahine	*wahine*	'woman'
mahi	*mahi*	'work'
toke	*tunga*	'worm'

7. Examine the shared cognate figures below for a number of speech communities in central Vanuatu (which are located on the map below). Can you draw a family tree that shows the degrees of relationship here?

Bongabonga										
88%	Tongariki									
32	31%	Makatea								
57	56	29%	Woraviu							
57	56	27	91%	Sesake						
53	53	28	87	86%	Nguna					
55	54	30	86	86	93%	Pwele				
56	55	29	88	87	93	94%	Siviri			
50	50	26	75	75	78	79	78%	Lelepa		
50	49	30	67	67	69	68	69	72%	Pango	
47	45	26	60	61	63	64	65	65	86%	Eratap
50	48	29	67	67	69	79	70	71	82	76% Eton

8. The historical record shows that Tok Pisin in Papua New Guinea emerged as a separate language from English in the second half of the nineteenth century. A count of shared cognates in the basic vocabularies of Tok Pisin and English reveals that out of 200 items, the two languages have common sources for 146 items. On the basis of this evidence, approximately when should the two languages have diverged? What does the result say about the assumptions of glottochronology?

9. Refer to Data Set 10 and see if you can make any judgements about the subgrouping of Sepa, Manam, Kairiru, and Sera from lexicostatistical evidence.

10. You have seen that subgrouping depends on being able to distinguish shared innovations from shared retentions from the protolanguage. Features are reconstructed in the protolanguage partly on the basis of the extent of their distribution in the daughter languages, as you learned in Chapter 5. What methodological problem do we face here?

FURTHER READING

1. Theodora Bynon *Historical Linguistics*, Chapter 7 'Glottochronology (or lexicostatistics)', pp. 266–72.
2. Robert J. Jeffers and Ilse Lehiste *Principles and Methods for Historical Linguistics*. Chapter 8 'Lexicostatistics', pp. 133–37.
3. Winfred P. Lehmann *Historical Linguistics: An Introduction*, Chapter 7 'Study of Loss in Language: Lexicostatistics', pp. 107–14.
4. Sarah Gudschinsky 'The ABC's of Lexicostatistics (Glottochronology)' in Dell Hymes (ed.) *Language in Culture and Society*, pp. 612–23.