Chapter 8

How Old is the Indo-European Language Family?
Illumination or More Moths to the Flame?

Quentin D. Atkinson & Russell D. Gray

1. An electric light on a summer night

The origin of Indo-European has recently been described as 'one of the most intensively studied, yet still most recalcitrant problems of historical linguistics' (Diamond & Bellwood 2003, 601). Despite over 200 years of scrutiny, scholars have been unable to locate the origin of Indo-European definitively in time or place. Theories have been put forward advocating ages ranging from 4000 to 23,000 years BP (Otte 1997), with hypothesized homelands including Central Europe (Devoto 1962), the Balkans (Diakonov 1984), and even India (Kumar 1999). Mallory (1989) acknowledges 14 distinct homeland hypotheses since 1960 alone. He rather colourfully remarks that

> the quest for the origins of the Indo-Europeans has all the fascination of an electric light in the open air on a summer night: it tends to attract every species of scholar or would-be savant who can take pen to hand (Mallory 1989, 143).

Unfortunately, archaeological, genetic and linguistic research on Indo-European origins has so far proved inconclusive. Whilst numerous theories of Indo-European origin have been proposed, they have proven difficult to test. In this chapter, we outline how techniques derived from evolutionary biology can be adapted to test between competing hypotheses about the age of the Indo-European language family. We argue that these techniques are a useful *supplement* to traditional methods in historical linguistics. This chapter is a development and extension of previous work on the application of phylogenetic methods to the study of language evolution (Gray & Atkinson 2003; Atkinson & Gray 2006; Atkinson *et al.* 2005).

2. Two theories

There are currently two main theories about the origin of Indo-European. The first theory, put forward by Marija Gimbutas (1973a,b) on the basis of linguistic and archaeological evidence, links Proto-Indo-

European (the hypothesized ancestral Indo-European tongue) with the Kurgan culture of southern Russia and the Ukraine. The Kurgans were a group of semi-nomadic, pastoralist, warrior-horsemen who expanded from their homeland in the Russian steppes during the fifth and sixth millennia BP, conquering Danubian Europe, Central Asia and India, and later the Balkans and Anatolia. This expansion is thought to roughly match the accepted ancestral range of Indo-European (Trask 1996). As well as the apparent geographical congruence between Kurgan and Indo-European territories, there is linguistic evidence for an association between the two cultures. Words for supposed Kurgan technological innovations are notably consistent across widely divergent Indo-European sub-families. These include terms for 'wheel' (*$rotho$-, *$k^w(e)k^wl$-o-), 'axle' (*aks-lo-), 'yoke' (*jug-o-), 'horse' (*$ekwo$-) and 'to go, transport in a vehicle' (*$wegh$-: Mallory 1989; Campbell 2004). It is argued that these words and associated technologies must have been present in the Proto-Indo-European culture and that they were likely to have been Kurgan in origin. Hence, the argument goes, the Indo-European language family is no older than 5000–6000 BP. Mallory (1989) argues for a similar time and place of Indo-European origin — a region around the Black Sea about 5000–6000 BP (although he is more cautious and refrains from identifying Proto-Indo-European with a specific culture such as the Kurgans).

The second theory, proposed by archaeologist Colin Renfrew (1987), holds that Indo-European languages spread, not with marauding Russian horsemen, but with the expansion of agriculture from Anatolia between 8000 and 9500 years ago. Radiocarbon analysis of the earliest Neolithic sites across Europe provides a fairly detailed chronology of agricultural dispersal. This archaeological evidence indicates that agriculture spread from Anatolia, arriving in Greece at some time during the ninth millennium BP and reaching as far as the British Isles by 5500 BP (Gkiasta *et al.* 2003). Renfrew maintains that the linguistic argument

for the Kurgan theory is based on only limited evidence for a few enigmatic Proto-Indo-European word forms. He points out that parallel semantic shifts or widespread borrowing can produce similar word forms across different languages without requiring that an ancestral term was present in the proto-language. Renfrew also challenges the idea that Kurgan social structure and technology was sufficiently advanced to allow them to conquer whole continents in a time when even small cities did not exist. Far more credible, he argues, is that Proto-Indo-European spread with the spread of agriculture — a scenario that is also thought to have occurred across the Pacific (Bellwood 1991; 1994), Southeast Asia (Glover & Higham 1996) and sub-Saharan Africa (Holden 2002). On the basis of linguistic evidence, Diakonov (1984) also argues for an early Indo-European spread with agriculture but places the homeland in the Balkans — a position that may be reconcilable with Renfrew's theory.

The debate about Indo-European origins thus centres on archaeological evidence for two population expansions, both implying very different time-scales — the Kurgan theory with a date of 5000–6000 BP, and the Anatolian theory with a date of 8000–9500 BP. One way of potentially resolving the debate is to look outside the archaeological record for independent evidence that allows us to test between these two time depths. Genetic studies offer one potential source of evidence. Unfortunately, due to problems associated with admixture, slow rates of genetic change and the relatively recent time-scales involved, genetic analyses have been unable to resolve the debate (Cavalli-Sforza *et al.* 1994; Rosser *et al.* 2000). Another potential line of evidence is contained in the languages themselves, and it is the linguistic evidence we shall turn to now.

3. The demise of glottochronology and the rise of computational biology

Languages, like genes, chronicle their evolutionary history. Languages, however, change much faster than genes and so contain more information at shallower time depths. Conventional means of linguistic inquiry, like the comparative method, are able to infer ancestral relationships between languages but cannot provide an absolute estimate of time depth. An alternative approach is Morris Swadesh's (1952; 1955) lexicostatistics and its derivative 'glottochronology'. These methods use lexical data to determine language relationships and to estimate absolute divergence times. Lexicostatistical methods infer language trees on the basis of the percentage of shared cognates between languages — the more similar the languages, the more closely they are related. Words are judged to be cognate if they

can be shown to be related via a pattern of systematic sound correspondences and have similar meanings (see Fig. 8.1 for some examples). This information can be used to construct evolutionary language trees. Glottochronology is an extension of this approach to estimate divergence times under the assumption of a 'glottoclock', or constant rate of language change. The following formulae can be used to relate language similarity to time along an exponential decay curve:

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

where *t* is time depth in millennia, *C* is the percentage of cognates shared and *r* is the 'universal' constant or rate of retention (the expected proportion of cognates remaining after 1000 years of separation: Swadesh 1955). Usually, analyses are restricted to the Swadesh word list, a collection of 100–200 basic meanings that are thought to be relatively culturally universal, stable and resistant to borrowing. These include kinship terms (e.g. mother, father), terms for body parts (e.g. hand, mouth, hair), numerals and basic verbs (e.g. to drink, to sleep, to burn). For the Swadesh 200-word list, a value of 81 per cent is often used for *r*.

Linguists have identified a number of serious problems with the glottochronological approach:

1. Much of the information in the lexical data is lost when word information is reduced to percentage similarity scores between languages (Steel *et al.* 1988).
2. The methods used to construct evolutionary trees from language distance matrices have been shown to produce inaccurate results, particularly where rates of change vary (Blust 2000).
3. Languages do not always evolve at a constant rate. Bergsland & Vogt (1962) compared present-day languages with their archaic forms and found evidence for significant rate variation between languages. For example, Icelandic and Norwegian were compared to their common ancestor, Old Norse, spoken roughly 1000 years ago. Norwegian has retained 81 per cent of the vocabulary of Old Norse, correctly suggesting an age of approximately 1000 years. However, Icelandic has retained over 95 per cent of the Old Norse vocabulary, falsely suggesting that Icelandic split from Old Norse less than 200 years ago.
4. Languages do not always evolve in a tree-like manner (Bateman *et al.* 1990; Hjelmslev 1958). Borrowing between languages can produce erroneous (or, in extreme cases, meaningless) language trees. Also, widespread borrowing can bias divergence time estimates by making languages seem more similar (and hence younger) than they really are.

These problems have led many linguists to completely abandon any attempt to derive dates from lexical data. For example, Clackson (2000, 451) claims that the data and methods 'do not allow the question "When was Proto-Indo-European spoken?" to be answered in any really meaningful or helpful way'.

Fortunately, none of these problems are unique to linguistics. It is ironic that whilst computational methods in historical linguistics have fallen out of favour over the last half-century, computational biology has thrived. In much the same way as linguists use information about current and historically attested languages to infer their history, evolutionary biologists use DNA sequence, morphological and sometimes behavioural data to construct evolutionary trees of biological species. Questions of relatedness and divergence dates are of interest to biologists just as they are to linguists. As a result biologists must also deal with the problems outlined above: nucleotide sequence information is lost when data is analyzed as distance matrices (Steel *et al.* 1988); distance-based tree-building methods may not accurately reconstruct phylogeny (Kuhner & Felsenstein 1994); different genes (and nucleotides) evolve at different rates and these rates can vary through time (Excoffier & Yang 1999); and finally, evolution is not always tree-like due to phenomena such as hybridization and horizontal gene transfer (Faguy & Doolittle 2000).

Despite these obstacles, computational methods have revolutionized evolutionary biology. Rather than giving up and declaring that time-depth estimates are intractable, biologists have developed techniques to overcome each problem. Here, we describe how these biological methods can be adapted and applied to lexical data to answer the question 'How old is the Indo-European language family?'

4. From word lists to binary matrices

In order to estimate phylogenies accurately we need to overcome the problem of information loss encountered in lexicostatistics and glottochronology. This requires a large data set with individual character-state information for each language. Lexical data are ideal because there are a large number of well-studied characters available and these can be divided into meaningful evolutionary units known as *cognate sets* (as described above, words are judged to be cognate if they can be shown to be related via a pattern of systematic sound correspondences and have similar meaning). Cognate words from different languages can be grouped into cognate sets that reflect patterns of inheritance. Owing to the possibility of unintuitive or misleading similarities between words from different languages, expert

knowledge of the sound changes involved is required in order to make cognacy judgements accurately. For example, knowledge of regular sound correspondences between the languages is required to ascertain that the English word *when* is cognate with Greek *pote* of the same meaning. Conversely, English *have* is not cognate with Latin *habere* despite similar word form and meaning.

To estimate tree topology and branch lengths accurately requires a large amount of data. Our data was taken from the Dyen *et al.* (1992) Indo-European lexical data base, which contains expert cognacy judgements for 200 Swadesh list terms in 95 languages. Dyen *et al.* (1997) identified eleven languages as less reliable and hence they were not included in the analysis presented here. Three extinct languages (Hittite, Tocharian A and Tocharian B) were added to the data base in an attempt to improve the resolution of basal relationships in the inferred phylogeny. Multiple references were used to corroborate cognacy judgements (Adams 1999; Gamkrelidze & Ivanov 1995; Guterbock & Hoffner 1986; Hoffner 1967; Tischler 1973; 1997). For each meaning in the data base, languages were grouped into cognate sets. Some examples are shown in Figure 8.1.

By restricting analyses to basic vocabulary such as the Swadesh word list the influence of borrowing can be minimized. For example, although English is a Germanic language, it has borrowed around 60 per cent of its total lexicon from French and Latin. However, only about 6 per cent of English entries in the Swadesh 200-word list are clear Romance language borrowings (Embleton 1986). Known borrowings were not coded as cognate in the Dyen *et al.* data base. For example, the English word *mountain* was not coded as cognate with French *montagne,* since it was obviously borrowed from French into English after the Norman invasion. Any remaining reticulation can be detected using biological methods such as split decomposition, which can identify conflicting signal. The issue of borrowing in lexical data is discussed in more detail by Holden & Gray (Chapter 2 this volume; see also Bryant *et al.* 2005).

We can represent the information in Figure 8.1 most simply as binary characters in a matrix, where the presence or absence of a particular cognate set in a particular language is denoted by a 1 or 0 respectively. Figure 8.2 shows a binary representation of the cognate information from Figure 8.1. Using this coding procedure we produced a matrix of 2449 cognacy judgements across 87 languages. Alternative coding methods are also possible, such as representing the data as 200 meaning categories each with multiple character states. It has been argued that semantic categories are the fundamental 'objects' of linguistic

English	here[1]	sea[5]	water[9]	when[12]
German	hier[1]	See[5], Meer[6]	Wasser[9]	wann[12]
French	ici[2]	mer[6]	eau[10]	quand[12]
Italian	qui[2], qua[2]	mare[6]	acqua[10]	quando[12]
Modern Greek	edo[3]	thalassa[7]	nero[11]	pote[12]
Hittite	ka[4]	aruna-[8]	watar[9]	kuwapi[12]

Figure 8.1. *Selection of languages and Swadesh list terms. Cognacy is indicated by the numbers in superscript.*

Meaning		here				sea				water			when
Cognate set	1	2	3	4	5	6	7	8	9	10	11	12
English	1	0	0	0	1	0	0	0	1	0	0	1
German	1	0	0	0	1	1	0	0	1	0	0	1
French	0	1	0	0	0	1	0	0	0	1	0	1
Italian	0	1	0	0	0	1	0	0	0	1	0	1
Greek	0	0	1	0	0	0	1	0	0	0	1	1
Hittite	0	0	0	1	0	0	0	1	1	0	0	1

Figure 8.2. *Cognate sets from Figure 8.1 expressed in a binary matrix showing cognate presence (1) or absence (0).*
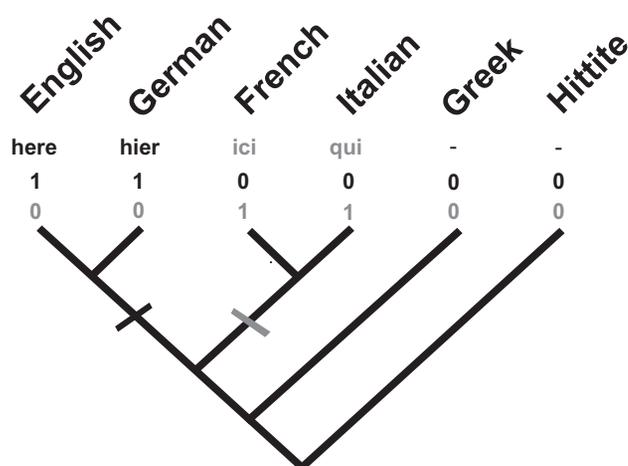

Figure 8.3. *Character states for cognate sets 5 (black) and 6 (grey) from Figure 8.1 are shown mapped onto a hypothetical tree. Implied character state changes can then be reconstructed on the tree. The black and grey bands show a likely point at which cognate sets 5 and 6 were gained. We can use this information to evaluate possible evolutionary scenarios.*

change (see Evans *et al.* Chapter 9 this volume) and that binary coding of the presence or absence of cognate sets is thus inappropriate. However, cognate sets constitute discrete, relatively unambiguous heritable evolutionary units with a birth and death (see Nicholls & Gray, Chapter 14 this volume) and there is no reason to suppose they are any more or less fundamental to language evolution than semantic categories. Further, coding the data as semantic categories makes it difficult to deal with polymorphisms (i.e. when a language has more than one word for a given meaning — e.g. for the meaning 'sea' German has both *See* and *Meer*). It also significantly increases the number of parameters required to model the process of evolution. Pagel (2000) points out that, if each word requires a different set of rate parameters, then for just 200 words in 40 languages there are 1278 parameters to estimate. We thus used a binary coding of cognate presence/absence information to represent linguistic change in our analysis.

As well as avoiding the problem of information loss, analyzing cognate presence/absence information allows us to explicitly model the evolutionary process. *Unlike lexicostatistics and glottochronology, we do not count the number of cognates shared between languages, nor do we calculate pair-wise distances between languages.* Instead, the distribution of cognates is mapped onto an evolutionary language tree (see Fig. 8.3), and likely character state changes are inferred across the whole tree.

5. Models are lies that lead us to the truth

When biologists model evolution, they lie: they lie about the independence of character state changes across sites; they lie about the homogeneity of substitution mechanisms; and they lie about the importance of selection pressure on substitution rates. But these are lies that lead us to the truth. Biological research is based on a strategy of model-building and statistical inference that has proved highly successful (Hillis

	0	1
0	$-u\pi_1$	$u\pi_1$
1	$u\pi_0$	$-u\pi_0$

Figure 8.4. *Simple likelihood rate matrix adapted for modelling lexical replacement in language evolution. This is a time-reversible model that allows for unequal equilibrium frequencies of 1s and 0s (cognate presence and absence). The model parameters are u (the mean substitution rate), and π_0 and π_1 (which represent the relative frequencies of 1s and 0s).*

1992; Hillis *et al.* 1996; Pagel 1999). The goal for biologists is not to construct a model so complex that it captures every nuance and vagary of the evolutionary process, but rather to find the simplest model available that can reliably estimate the parameters of interest.

Model choice is thus a balance between over- and under-fitting parameters (Burnham & Anderson 1998). Adding extra parameters can improve the apparent fit of a model to data, however, sampling error is also increased because there are more unknown parameters to estimate (Swofford *et al*. 1996). Depending on the question we are trying to answer, this added uncertainty can prevent us from estimating the model parameters from the data to within a useful margin of error. In many cases, adding just a few extra parameters can create a computationally intractable problem. Conversely, a model that is too simple can produce biased results if it fails to capture an important part of the process (Burnham & Anderson 1998). There is thus a compromise between biased estimates and variance inflation.

The strategy that has proved successful in biology is to start with a simple model that captures some of the fundamental processes involved and increase the complexity as necessary. For example, nucleotide substitution models range from a simple equal rates model (Jukes & Cantor 1969), to more complex models that allow for differences in transition/transversion rates, unequal character state frequencies, site specific rates, and auto-correlation between sites (Swofford *et al*. 1996). Although even the most complicated models are simplifications of the process of evolution, often the simplest substitution model captures enough of what is going on to allow biologists to extract a meaningful signal from the data. Levins (1966) gives three reasons why we should use a simple model. First, violations of the assumptions of the model are expected to cancel each other out. Second, small errors in the model should result in small errors in the conclusions. And third, by comparing multiple models with reality we can determine which aspects of the process are important.

Likelihood evolutionary modelling has become the method of choice in phylogenetics (Swofford *et al*. 1996). The likelihood approach to phylogenetic reconstruction allows us to explicitly model the process of language evolution. The method is based on the premise that we should favour the tree topology/ topologies and branch-lengths that make our observed data most likely, given the data and assumptions of our model — i.e. we should favour the tree with the highest likelihood score. We can evaluate possible tree topologies for a given model and data by modelling the sequence of cognate gains and losses across the trees.

Likelihood models have a number of advantages over other approaches. First, we can work with explicit models of evolution and test between competing models. The assumptions of the method are thus overt and easily verifiable. Second, we can increase the complexity of the model as required. For example, as explained below, we were able to test for the influence of rate variation between cognate sets and, as a result, incorporate this into the analysis using a gamma distribution. And third, model parameters can be estimated from the data itself, thus avoiding restrictive *a priori* assumptions about the evolutionary processes involved (Pagel 1997).

Likelihood models of evolution are usually expressed as a rate matrix representing the relative rates of all possible character state changes. Here, we are interested in the processes of cognate gain and loss, respectively represented by 0 to 1 changes and 1 to 0 changes on the tree (see Fig. 8.3). We can model this process effectively with a relatively simple two-state time-reversible model of lexical evolution (shown in Fig. 8.4). We extended this simple model by adding a gamma shape parameter (α) to allow rates of change to vary between cognate sets according to a gamma distribution. This was implemented after a likelihood ratio test (Goldman 1993) showed that adding the gamma-shape parameter significantly improved the ability of the model to explain the data ($\chi^2 = 108$, $df = 1$, $p < .001$). Essentially, the gamma distribution provides a number of different rate categories for the model to choose from when assigning rates to each cognate set. The gamma distribution for different values of α is shown in Figure 8.5. A suitable value for α can be estimated from the data.

Our model assumes that the appearance and disappearance of cognates is randomly distributed about some mean value. This rate can vary between cognate sets and (with the addition of rate smoothing — described below) rates of change can vary through time in a constrained way. Whilst historical, social, and cultural contingencies can undoubtedly influence
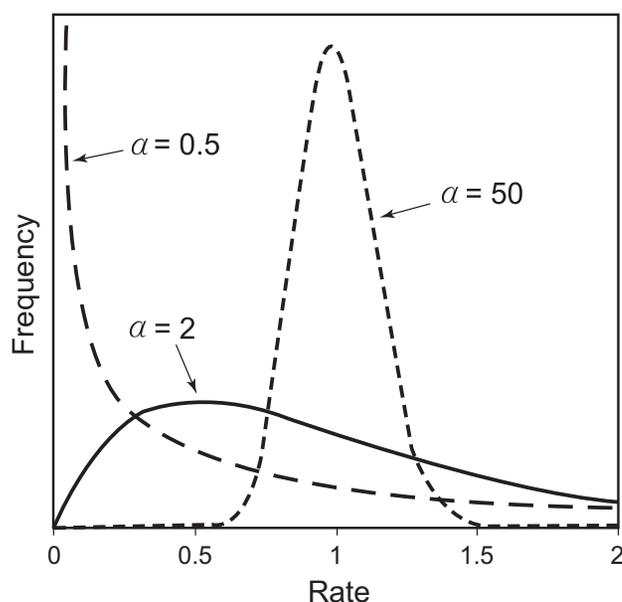
Figure 8.5. *The gamma distribution, used to model rate variation between sites. Three possible values for* α *are shown. For small values of* α *(e.g.* α *= 0.5), most cognate sets evolve slowly, but a few can evolve at higher rates. As* α *increases, the distribution becomes more peaked and symmetrical around a rate of 1 — i.e. rates become more equal (e.g.* α *= 50).*

the process of linguistic change, we explicitly reject Warnow *et al*.'s (Chapter 7 this volume) counsel of despair, that language evolution is so idiosyncratic and unconstrained that inferring divergence dates is impossible. Language evolution is subject to real-world constraints, such as human language acquisition, expressiveness, intelligibility, and generation time. We cannot help but quote Ringe *et al.* (2002, 61) on this point:

> Languages replicate themselves (and thus 'survive' from generation to generation) through a process of native-language acquisition by children. Importantly for historical linguistics, that process is tightly constrained.

These constraints create underlying commonalities in the evolutionary process that we can, and should, be trying to model.

Evans *et al*. (Chapter 10 this volume) argue that our model is 'patently inappropriate' because it assumes that all characters are independent. In biology, this is known as the I.I.D. (identically and independently distributed) assumption. Evans *et al*. correctly point out that the I.I.D. assumption is violated when individual meanings in the Swadesh word list are broken up into characters representing multiple cognate sets. Specifically, if a particular cognate set is present in a language, it will be less

likely that other cognate sets for the same meaning will also be present. However, we do not think that this lack of independence biases our results. The issue of independence will be dealt with in detail in section 10.2.

6. Bayesian inference of phylogeny

It is not usually computationally feasible to evaluate the likelihood of all possible language trees — for 87 languages there are over 1×10^{155} possible rooted trees. Further, the vast number of possibilities combined with finite data means that inferring a single tree will be misleading — there will always be uncertainty in the topology and branch-lengths. If we are to use our results to test hypotheses we need to use heuristic methods to search through 'tree-space' and quantify this phylogenetic uncertainty. Bayesian inference is an alternative approach to phylogenetic analysis that allows us to draw inferences from a large amount of data using powerful probabilistic models without searching for the 'optimal tree' (Huelsenbeck *et al*. 2001). In this approach trees are sampled according to their posterior probabilities. The posterior probability of a tree (the probability of the tree given the priors, data and the model) is related by Bayes's theorem to its likelihood score (the probability of the data given the tree) and its prior probability (a reflection of any prior knowledge about tree topology that is to be included in the analysis). Unfortunately, we cannot evaluate this function analytically. However, we can use a Markov Chain Monte Carlo (MCMC: Metropolis *et al*. 1953) algorithm to generate a sample of trees in which the frequency distribution of the sample is an approximation of the posterior probability distribution of the trees (Huelsenbeck *et al*. 2001). To do this, we used *MrBayes*, a Bayesian phylogenetic inference programme (Huelsenbeck & Ronquist 2001).

MrBayes uses MCMC algorithms to search through the realm of possible trees. From a random starting tree, changes are proposed to the tree topology, branch-lengths and model parameters according to a specified prior distribution of the parameters. The changes are either accepted or rejected based on the likelihood of the resulting evolutionary reconstruction — i.e. reconstructions that give higher likelihood scores tend to be favoured. In this way the chain quickly goes from sampling random trees to sampling those trees which best explain the data. After an initial 'burn-in' period, trees begin to be sampled in proportion to their likelihood given the data. This produces a distribution of trees. A useful way to summarize this distribution is with a consensus tree or consensus network (Holland & Moulton 2003)
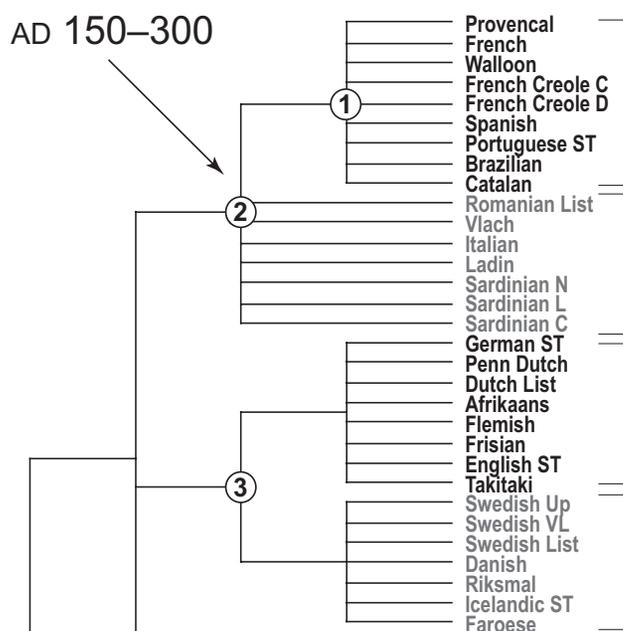
Figure 8.6. *Consensus network from the Bayesian MCMC sample of trees. Values express the posterior probability of each split (values above 90 per cent are not indicated). A threshold of 10 per cent was used to draw this splits graph — i.e. only those splits occuring in at least 10 per cent of the observed trees are shown in the graph. Branch lengths represent the median number of reconstructed substitutions per site across the sample distribution.*

depicting uncertainty in the reconstructed relationships. These graphs are, however, just useful pictorial summaries of the analysis. The fundamental output of the analysis is the distribution of trees.

The consensus network from a Bayesian sample distribution of 100 trees is shown in Figure 8.6. The values next to splits give an indication of the uncertainty associated with each split (the posterior probability, derived from the percentage of trees in the Bayesian distribution that contain the split). For example, the value 41 next to the parallel lines separating Italic and Celtic from the of the Indo-European sub-families indicates that that split was present in 41 per cent of

the trees in the sample distribution. Similarly, the split grouping Italic and Germanic languages was present in 46 per cent of the sample distribution.

7. Rate variation and estimating dates

There are at least two types of rate variation in lexical evolution. First, rate variation can occur between cognates. For example, even in the Swadesh word list, the Indo-European word for *five* is highly conserved (1 cognate set) whilst the word for *dirty* is highly variable (27 cognate sets). This is akin to site-specific rate variation in biology. Biologists can account for

AD 150–300

Figure 8.7. *The Romance languages (derived from Latin) probably began to diverge prior to the fall of the Roman Empire. We can thus constrain the age of the point on the tree which corresponds to this divergence event (2). Using this rationale 14 nodes were constrained, including an Iberian French node (1) and a Germanic node (3).*

this type of rate variation by allowing a distribution of rates. As mentioned above, we used a model of cognate evolution that allowed for gamma distributed rate variation between cognates.

Second, rates of lexical evolution can vary through time and between lineages. Clearly this will cause problems if we are trying to estimate absolute divergence times on the inferred phylogenies since inferred branch-lengths are not directly proportional to time. Again, biologists have developed a number of methods for dealing with this type of rate variation. We used the penalized-likelihood method of rate smoothing implemented in *r8s* (*version 1.7*; Sanderson 2002a), to allow for rate variation across each tree. Sanderson (2002b) has shown that, under conditions of rate variation, the penalized-likelihood rate-smoothing algorithm performs significantly better than methods that assume a constant rate of evolution.

In order to infer absolute divergence times, we first need to calibrate rates of evolution by constraining the age of known points on each tree in accordance with historically attested dates. For example, the Romance languages (derived from Latin) probably began to diverge prior to the fall of the Roman Empire. We can thus constrain the age of the node corresponding to the most recent common ancestor of the Romance

languages to within the range implied by our historical knowledge (see Fig. 8.7). We constrained the age of 14 such nodes on the tree in accordance with historical evidence (see Atkinson & Gray 2006). These known node ages were then combined with branch-length information to estimate rates of evolution across each tree. The penalized-likelihood model allows rates to vary across the tree whilst incorporating a 'roughness penalty' that costs the model more if rates vary excessively from branch to branch. This procedure allows us to derive age estimates for each node on the tree. Figure 8.8 shows the consensus tree for the initial Bayesian sample distribution of 1000 trees,[1] with branch lengths drawn proportional to time. The posterior probability values above each internal branch give an indication of the uncertainty associated with each clade on the consensus tree (the percentage of trees in the Bayesian distribution that contain the clade). For example, the value 67 above the branch leading to the Italo-Celto-Germanic clade indicates that that clade was present in 67 per cent of the trees in the sample distribution. We can derive age estimates from this tree, including an age of 8700 BP at the base of the tree — within the range predicted by the Anatolian farming theory of Indo-European origin.

A single divergence time, with no estimate of the error associated with the calculation, is of limited value. To test between historical hypotheses we need some measure of the error associated with the date estimates. Specifically, uncertainty in the phylogeny gives rise to a corresponding uncertainty in age estimates. In order to account for phylogenetic uncertainty we estimated the age at the base of the trees in the post-burn-in Bayesian MCMC sample to produce a probability distribution for the age of Indo-European. One advantage of the Bayesian framework is that prior knowledge about language relationships can be incorporated into the analysis. In order to eliminate trees that conflict with known Indo-European language groupings, the original 1000 tree sample was filtered using a constraint tree representing these known language groupings [(Anatolian, Tocharian, (Greek, Armenian, Albanian, (Iranian, Indic), (Slavic, Baltic), ((North Germanic, West Germanic), Italic, Celtic)))]. This constraint tree was consistent with the majority-rule consensus tree generated from the entire Bayesian sample distribution. The filtered distribution of divergence time estimates was then used to create a confidence interval for the age of the Indo-European language family. This distribution could then be compared with the age ranges implied by the two main theories of Indo-European origin (see Fig. 8.9). The results are clearly consistent with the Anatolian hypothesis.

Not all historically attested language splits were used in our analysis. One means of validating our
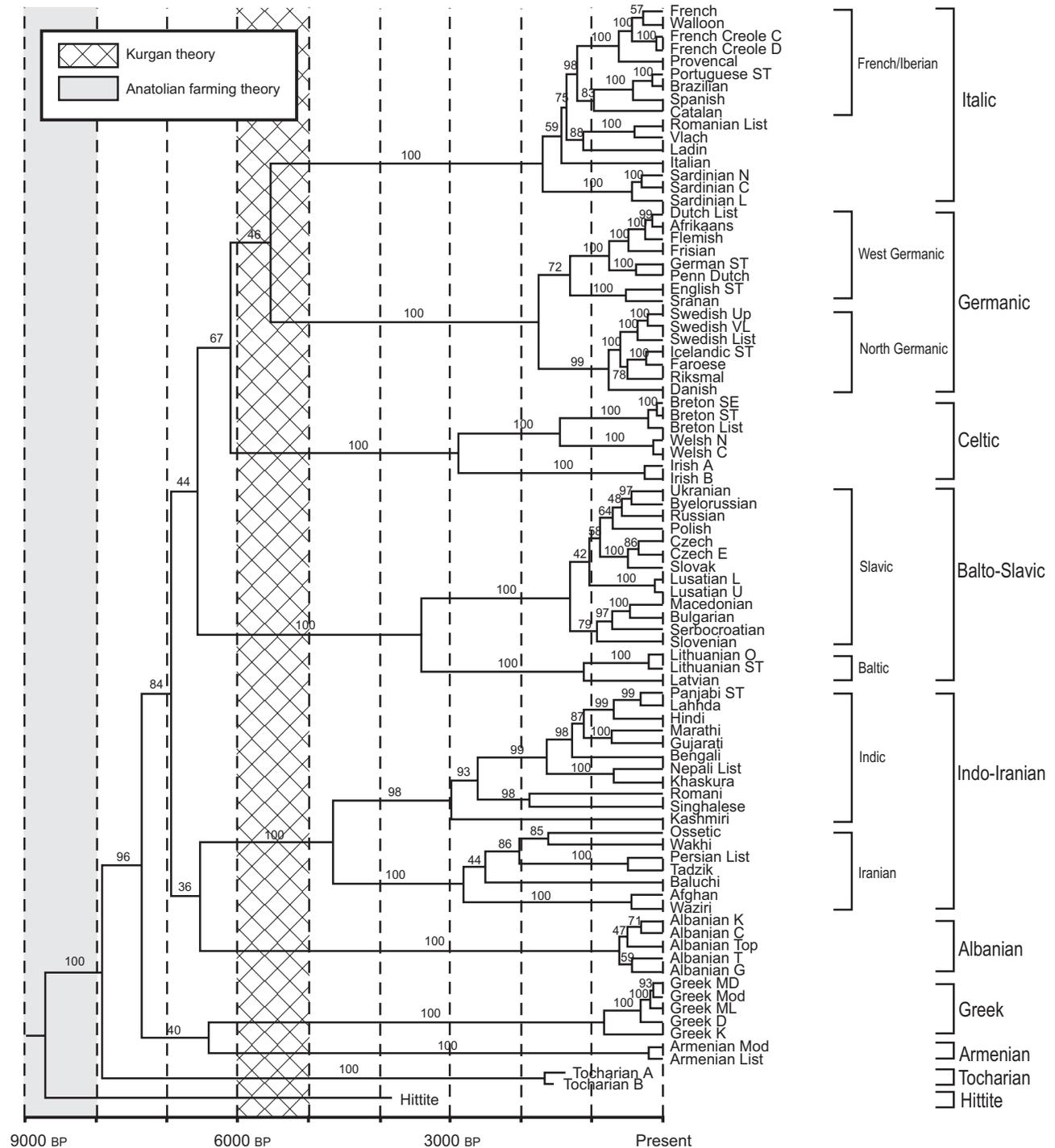
Figure 8.8. *Majority-rule consensus tree from the initial Bayesian MCMC sample of 1000 trees (Gray & Atkinson 2003). Values above each branch indicate uncertainty (posterior probability) in the tree as a percentage. Branch-lengths are proportional to time. Shaded bars represent the age range proposed by the two main theories — the Anatolian theory (grey bar) and the Kurgan theory (hatched bar). The basal age (8700 BP) supports the Anatolian theory.*

methodology is to produce divergence time distributions for nodes that were not constrained in the analysis and compare this to the historically attested time of divergence. For example, Figure 8.10 shows the inferred divergence time distributions for the North and West Germanic subgroups. The grey band in
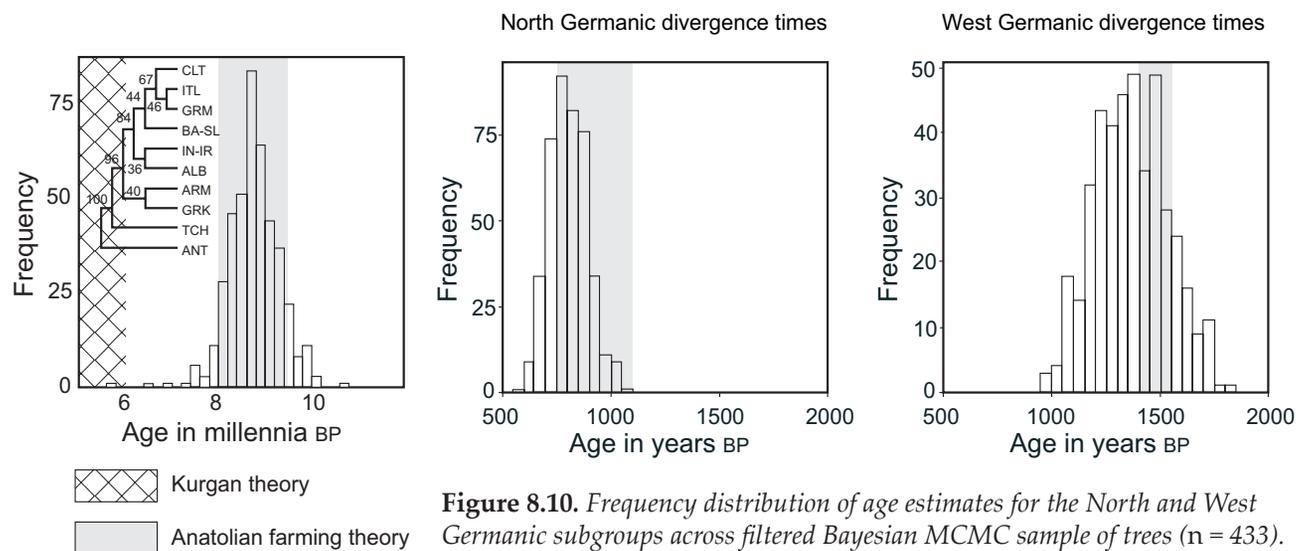
Figure 8.9. *Frequency distribution of basal age estimates from filtered Bayesian MCMC sample of trees for the initial assumption set (*n *= 435). The majority-rule consensus tree for the entire (unfiltered) sample is shown in the upper left.*

Figure 8.10. *Frequency distribution of age estimates for the North and West Germanic subgroups across filtered Bayesian MCMC sample of trees (*n *= 433). The grey bands indicate the historically attested time of divergence.*

these figures indicates the likely age of each subgroup based on the historical record. The age estimates for the North Germanic clade correspond with written evidence for the break up of these languages between AD 900 and AD 1250. Similarly, estimated ages of the West Germanic clade are consistent with historical evidence dating the Anglo-Saxon migration to the British Isles about 1500 years ago.

8. Testing robustness

A key part of any Bayesian phylogenetic analysis is an assessment of the robustness of the inferences. To do this we tested the effect of altering a number of different parameters and assumptions of the method.

8.1. Bayesian 'priors'
Initializing each Bayesian MCMC chain required the specification of a starting tree and prior parameters ('*priors*') for the analysis. The sample Bayesian distribution was the product of ten separate runs from different random starting trees. Divergence time and topology results for each of the separate runs were consistent. Other test analyses were run using a range of priors for parameters controlling the rate matrix, branch-lengths, gamma distribution and character state frequencies. The inferred tree phylogeny and branch-lengths did not noticeably change when priors were altered.
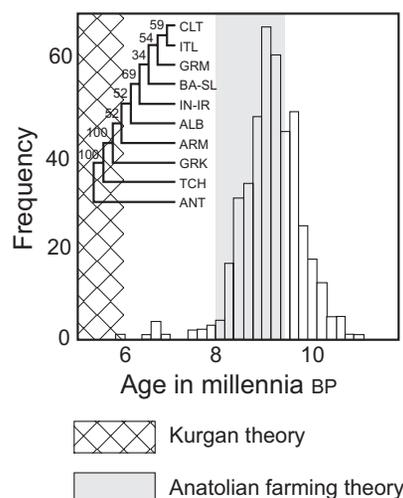

Figure 8.11. *Frequency distribution of basal age estimates from filtered Bayesian MCMC sample of trees for analysis with doubtful cognates excluded (*n *= 433). The majority-rule consensus tree for the entire sample is shown in the upper left.*

8.2. Cognacy judgements
The Dyen *et al.* (1992) data base contained information about the certainty of cognacy judgements. Words were coded as 'cognate' or 'doubtful cognates'. In the initial analysis we included all cognate information in an effort to maximize any phylogenetic signal. However, we wanted to test the robustness of our results to changes in the stringency of cognacy decisions. For this reason, the analysis was repeated with doubtful cognates excluded. This produced a similar age range to the initial analysis, indicating that our results were robust to errors in cognacy judgements (see Fig. 8.11).
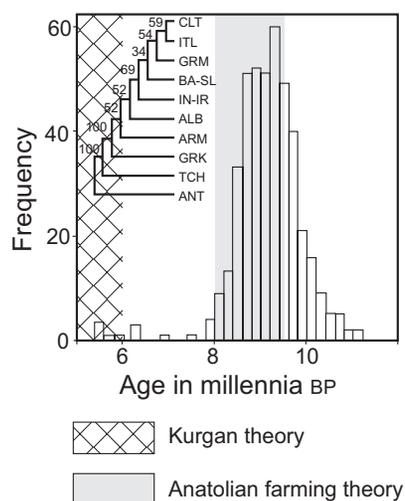
Figure 8.12. *Frequency distribution of basal age estimates from filtered Bayesian MCMC sample of trees using revised Celtic age constraint of between 1800 BP and 2200 BP. The majority-rule consensus tree for the entire sample is shown in the upper left.*
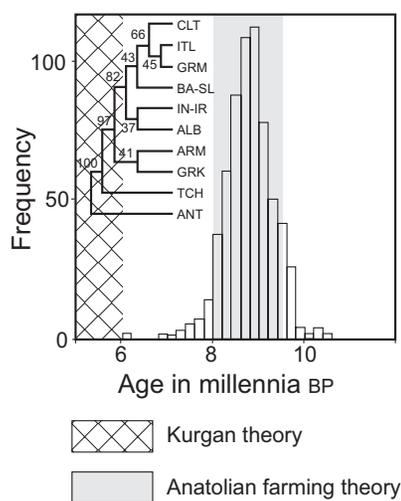

Figure 8.13. *Frequency distribution of basal age estimates from filtered Bayesian MCMC sample of trees using minimum set of topological constraints [(Anatolian, Tocharian, (Greek, Armenian, Albanian, (Iranian, Indic), (Slavic, Baltic), (North Germanic, West Germanic), Italic, Celtic))] (n = 670). The majority-rule consensus tree for the entire sample is shown in the upper left.*
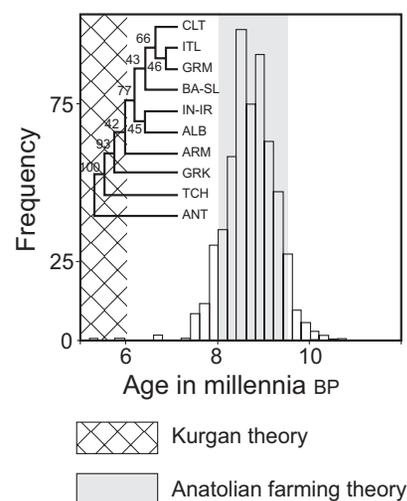
Figure 8.14. *Frequency distribution of basal age estimates from filtered Bayesian MCMC sample of trees with information about missing cognates included (n = 620). The majority-rule consensus tree for the entire sample is shown in the upper left.*

8.3. Calibrations and constraint trees

At the conference on which this volume is based, a question was raised about the age constraint used for Insular Celtic. It was suggested that whilst we used a maximum age constraint of 2750 BP, an age constraint of between 2200 BP and 1800 BP would have been more suitable. We do not wish to engage in debates about the correct age of the Insular Celtic divergence, however, a re-analysis of the data using the suggested ages serves to demonstrate the robustness of our results to variations in age constraints. Figure 8.12 shows the distribution of divergence times using the much later Celtic age constraints. Clearly, our results are robust to alterations in this age constraint. In fact, the step-by-step removal of each of the 14 age constraints on the consensus tree revealed that divergence time estimates were robust to calibration errors across the tree. For 13 nodes, the reconstructed age was within 390 years of the original constraint range. Only the reconstructed age for Hittite showed an appreciable variation from the constraint range. This may be attributable to the effect of missing data associated with extinct languages. Reconstructed ages at the base of the tree ranged from 10,400 BP with the removal of the Hittite age

constraint, to 8500 BP with the removal of the Iranian group age constraint. The results are highly robust calibration errors because of the large number of age constraints we used to calibrate rates of lexical evolution across the tree.

We also wanted to be sure that the constraint tree used to filter the Bayesian distribution of trees had not systematically biased our results. Figure 8.13 shows the divergence time distribution for the initial data set after filtering using a minimum set of topological constraints [(Anatolian, Tocharian, (Greek, Armenian, Albanian, (Iranian, Indic), (Slavic, Baltic), (North Germanic, West Germanic), Italic, Celtic))]. Again, the divergence time distribution was consistent with the Anatolian farming theory.

8.4. Missing data

Another possible bias was the effect of missing data. Some of the languages in the Dyen *et al.* (1992) data base may have contained fewer cognates because information about these languages was missing. For example, the three extinct languages (Hittite, Tocharian A & Tocharian B) are derived from a limited range of source texts and it is possible that some cognates were missed because the terms were not referred to in

the source text. This may have biased divergence time estimates by falsely increasing basal branch-lengths. Nicholls & Gray (Chapter 14 this volume) point out that we should expect fewer cognates to be present in the languages at the base of the tree anyway — the fact that Hittite has 94 cognates whilst most languages have around 200, does not necessarily imply that data is missing. Nonetheless, we tested for the effect of missing data by including information about whether or not the word for a particular term was missing from the data base. If we could not rule out the possibility that a cognate was absent from a language because it had not been found or recorded, then that cognate was coded as missing (represented by a '?' in the matrix). Encoding missing cognate information in this way means that we can account for uncertainty in the data itself using the likelihood model — the unknown states become parameters to be estimated. Analyzing this recoded data also produced an age range consistent with the Anatolian theory (see Fig. 8.14).

8.5. Root of Indo-European

Finally, we tested the effect of the rooting point for the trees. In the previous analyses, trees were rooted with Hittite. Although this is consistent with independent linguistic analyses (Gamkrelidze & Ivanov 1995; Rexová *et al.* 2003), other potential root points are possible. It could be claimed that a Hittite root biases age estimates in favour of the Anatolian hypothesis. We thus reran the rate-smoothing analysis rooting the consensus tree in Figure 8.8 with Balto-Slavic, Greek, Tocharian and Indo-Iranian groups. In all four cases the estimated divergence time *increased* to between 9500 BP and 10,700 BP.

9. Discussion

The time-depth estimates reported here are consistent with the times predicted by a spread of language with the expansion of agriculture from Anatolia. The branching pattern and dates of internal nodes are broadly consistent with archaeological evidence indicating that between the tenth and sixth millennia BP a culture based on cereal cultivation and animal husbandry spread from Anatolia into Greece and the Balkans and then out across Europe and the Near East (Gkiasta *et al.* 2003; Renfrew 1987). Hittite appears to have diverged from the main Proto-Indo-European stock around 8700 years ago, perhaps reflecting the initial migration out of Anatolia. Indeed, this date exactly matches estimates for the age of Europe's first agricultural settlements in southern Greece (Renfrew 1987). Following the initial split, the language tree shows the formation of separate Tocharian, Greek,

and then Armenian lineages, all before 6000 BP, with all of the remaining language families formed by 4000 BP. We note that the received linguistic orthodoxy (Indo-European is only 6000 years old) does approximately fit the divergence dates we obtained for most of the branches of the tree. Only the basal branches leading to Hittite, Tocharian, Greek and Armenian are well beyond this age. Interestingly, the date range hypothesized for the Kurgan expansion does correspond to a rapid period of divergence on the consensus tree. According to the divergence time estimates shown in Figure 8.8, many of the major Indo-European sub-families — Indo-Iranian, Balto-Slavic, Germanic, Italic and Celtic — diverged between six and seven thousand years ago — intriguingly close to the hypothesized time of the Kurgan expansion. Thus it seems possible that there were two distinct phases in the spread of Indo-European: an initial phase, involving the movement of Indo-European with agriculture, out of Anatolia into Greece and the Balkans some 8500 years ago; and a second phase (perhaps the Kurgan expansion) which saw the subsequent spread of Indo-European languages across the rest of Europe and east, into Persia and Central Asia.

10. Response to our critics

10.1. The potential pitfalls of linguistic palaeontology

A number of linguists have claimed that *linguistic palaeontology* offers a compelling reason why the arguments we have presented must be wrong: Proto-Indo-Europeans are claimed to have had a word for 'wheel' (*$k^w(e)k^wl$-o-*) but wheels did not exist in Europe 9000 years ago. The case is based on a widespread distribution of apparently related words for wheel in Indo-European languages. This is often presented as a knock-down argument against any age of Indo-European older than 5000 to 6000 years (when wheels first appear in the archaeological record). However, there are at least two alternative explanations for the distribution of terms associated with wheel and wheeled transport.

First, independent *semantic* innovations from a common root are a likely mechanism by which we can account for the supposed Proto-Indo-European reconstructions associated with wheeled transport (Trask 1996; Watkins 1969). Linguists can reconstruct word forms with much greater certainty than their meanings. For example, upon the development of wheeled transport, words derived from the Proto-Indo-European term *$kwel$-* (meaning 'to turn, rotate') may have been independently co-opted to describe the wheel. On the basis of the reconstructed ages shown in Figure 8.8, as few as three such semantic

innovations around the sixth millennium BP could have accounted for the attested distribution of terms related to *kʷ(e)kʷl-o- 'wheel' (one shift just before the break up of the Italic-Celtic-Germanic-Balto-Slavic-Indo-Iranian lineage, one shift in the Greek-Armenian lineage, and one shift [or borrowing] in the Tocharian lineage).

The second possible explanation for the distribution of terms pertaining to wheeled vehicles is widespread borrowing. Good ideas spread. Terms associated with a new technology are often borrowed along with the technology. The spread of wheeled transport across Europe and the Near East 5000–6000 years ago seems a likely candidate for borrowing of this sort. Linguists are able to identify many borrowings (particularly more recent ones) on the basis of the presence or absence of certain systematic sound correspondences. However, our date estimates suggest that most of the major Indo-European groups were just beginning to diverge when the wheel was introduced. We would thus expect the currently attested forms of any borrowed terms to look as if they were inherited from Proto-Indo-European — they may thus be impossible to reliably identify.

Both of these arguments are discussed in more detail elsewhere (Watkins 1969; Renfrew 1987; Atkinson & Gray 2006). It suffices to say that both the power and the pitfalls of linguistic palaeontology are well known. We are disappointed that in their rush to dismiss our paper in the media, otherwise scholarly and responsible linguists have claimed much greater certainty for their semantic reconstructions than is justifiable. This does not mean that we think there is no issue here. Ideally, we should aim to synthesize all lines of evidence relating to the age of Indo-European. Ringe (unpublished manuscript) presents a careful summary of the terms related to wheeled vehicles in Indo-European. He argues that words for 'thill' (a pole that connects a yoke or harness to a vehicle) and 'yoke' can confidently be reconstructed for Proto-Indo-European. He notes that reflexes of *kʷ(e)kʷl-o- 'wheel' have not been found in Anatolian languages but exist in Tocharian A and B and other Indo-European languages, and hence can be reconstructed for the common ancestor of all non-Anatolian Indo-European languages. Ringe claims that the specific forms of these words make parallel semantic changes or borrowing extremely implausible. It would be extremely useful to attempt to quantify just how unlikely such alternative scenarios are. Until all the assumptions of these arguments are formalized, and the probability of alternative scenarios quantified, it will remain difficult to synthesize all the different lines of evidence on the age of Indo-European.

10.2. Independence of characters

As mentioned in section 5, Evans *et al.* (Chapter 10 this volume) claim our evolutionary model of binary character evolution is 'patently inappropriate' because it assumes independence between characters when our characters are clearly not independent. However, we do not believe that any violation of independence necessarily biases our time-depth estimates. We note that the assumption of independence does not hold for nucleotide or amino acid sequence data either. For example, compensating substitutions in ribosomal RNA sequences result in correlation between paired sites in stem regions (Felsenstein 2004). However, biologists still get reasonably accurate estimates of phylogeny despite violations of this assumption. In fact, *nothing in the Evans* et al. *paper demonstrates that coding the data as binary characters, rather than the multistate characters, will produce biased results.* Pagel & Mead (Chapter 15 this volume) demonstrated that, on the contrary, binary and multi-state coded data produce trees that differ in length by a constant of proportionality. In other words, the binary and multi-state trees are just scaled versions of one another. Since we estimate rates of evolution for each tree using the branch lengths of that tree, scaling the branch lengths does not affect our results. Pagel & Meade (Chapter 15 this volume) also approximated the effect of violations of the independence assumption on the MCMC analysis by 'heating' the likelihood scores. They inferred that violations of independence would produce higher posterior probability values but would have little effect on the consensus tree topology. This means that we may have underestimated the error due to phylogenetic uncertainty but our estimates will not be biased towards any particular date.

Finally, treating cognate sets as the fundamental unit of lexical evolution does not, as Evans *et al.* (Chapter 10 this volume) argue, constitute an 'extreme violation' of the independence assumption. Almost all of the languages in the Dyen *et al.* (1992) data base contain polymorphisms, meaning that for a given language there exist multiple words of the same meaning. The polymorphisms in our data are a reflection of the nature of lexical evolution. Specifically, they demonstrate a lack of strict dependence between cognate sets within meaning categories — i.e. a word with a given meaning can arise in a language that already has a word of that meaning. Models of lexical evolution that do not allow polymorphisms (e.g. Ringe *et al.* 2002) could also be labelled as 'patently inappropriate' because they assume that for a word to arise in a language any existing words with that meaning must be concomitantly lost from the language. This is not always the case. Ringe *et al.* (2002) note that although the words 'small' and 'little' have
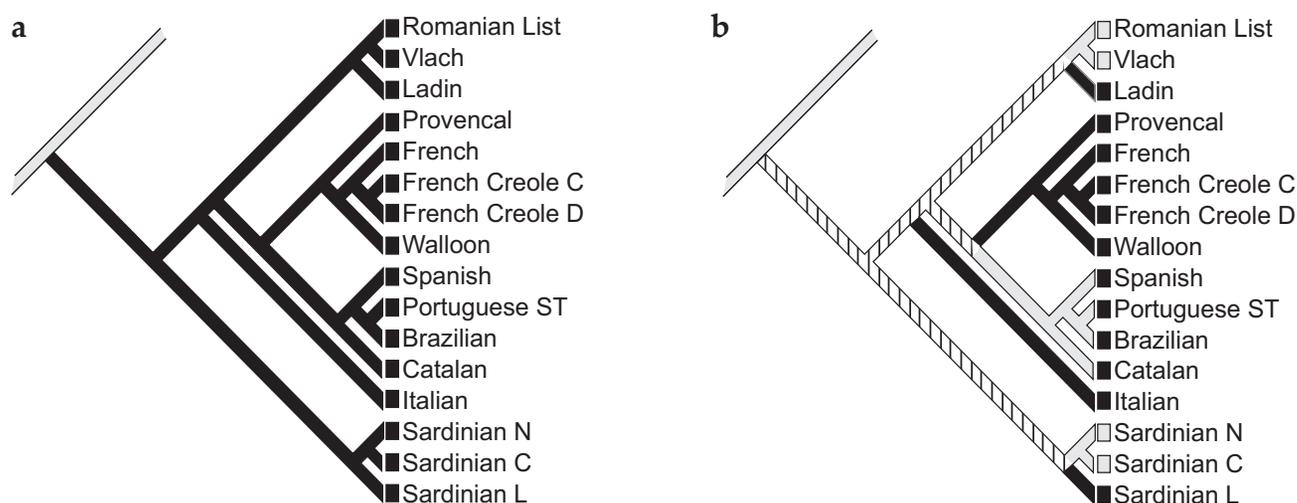
Figure 8.15. a) *Parsimony character trace for reflexes of Latin focus (originally 'hearth' but borrowed as 'fire') on the Romance consensus tree. Black indicates presence of the character, grey indicates absence and dashed indicates uncertainty. This shows borrowing across the whole Romance subgroup — evolutionary change is inferred at the base of the subgroup with no change within the subgroup, falsely inflating divergence time estimates.* b) *as with (a) but for reflexes of Latin testa (originally 'cup, jar, shell' but borrowed as 'head'). Here the borrowing is not across the whole Romance subgroup — evolutionary change is inferred within the subgroup.*

very similar meanings, they have persisted together in English for over a thousand years. Our binary coding procedure allows us to represent such polymorphisms with ease. The presence of polymorphisms means that dependencies between cognate sets are not as strong as Evans *et al.* claim. A further factor that weakens the dependencies between the cognate sets arises from the 'thinning' process that occurs in lexical evolution. The observed cognate sets do not represent the full compliment of actual cognates that arose in Indo-European (see Nicholls & Gray Chapter 14 this volume). Some cognates that existed in the past will not have persisted into present-day languages and any unique 'cognates' were not included in the analysis. This 'thinning' of the cognates also acts to reduce dependencies between characters in the analysis and thus further weakens any effect of violations of independence. Further research by Atkinson *et al.* (2005) using synthetic data has shown that violations of the independence assumption do not significantly affect date estimates.

10.3. Confidence in lexical data
From a phylogenetic viewpoint the lexicon is a tremendously attractive source of data because of the large number of possible characters it affords. However, we are aware that many historical linguists are sceptical of inferences based purely on lexical data. Garrett (Chapter 12 this volume) argues that borrowing of lexical terms, or *advergence,* within the major Indo-European

subgroups could have distorted our results. He identifies a number of cases where an ancestral term has been replaced by a different term in all or some of the daughter languages, presumably via borrowing:

> Thus Latin *ignis* 'fire' has been replaced by reflexes of Latin *focus* 'hearth' throughout Romance, and archaic Sanskrit *hanti* 'kills' has been replaced by reflexes of a younger Sanskrit form *marayati* throughout Indo-Aryan.

Garret argues correctly that, where a word has been borrowed across a subgroup after the initial divergence of the group, our method will infer that the word evolved in the branch leading up to that subgroup (see Latin *focus* example: Fig. 8.15a). This will falsely inflate the branch lengths below the subgroups and deflate branch lengths within each group. Since we estimate rates of evolution on the basis of within-group branch lengths, it is argued that we will underestimate rates of change and hence overestimate divergence lower in the tree. However, this argument requires that two special assumptions hold. First, any borrowing must occur across a whole subgroup and only across a whole subgroup. When terms are not borrowed across the whole group there is no systematic bias to infer changes in the branch leading to the group. Depending on the distribution of borrowed terms, advergence can even produce the opposite effect, falsely inflating branch lengths within subgroups and hence causing us to underestimate divergence times. It seems unlikely that all or even

most borrowed terms were borrowed across an entire subgroup. Garrett highlighted 16 instances of borrowing within Indo-European subgroups.[2] These were presumably selected because they were thought to reflect the sort of advergence pattern that would bias our results. Of these, at least 6 are unlikely to favour inferred language change at the base of a subgroup.[3] Figure 8.15b shows the example of the Romance term for 'head'.

Second, even if we accept the first assumption, we must assume that the proposed process of advergence is unique to contemporary languages. As Garrett (Chapter 12 this volume) puts it, this requires 'the unscientific assumption that linguistic change in the period for which we have no direct evidence was radically different from change we can study directly'. Rather than arguing that borrowing was rare at one stage and then suddenly became common across all of the major lineages at about the same time, it seems more plausible to suggest that borrowing has always occurred. If the same process of advergence in related languages has always occurred then the effect of shifting implied changes to more ancestral branches will be propagated down the tree such that there should be no net effect on divergence time calculation. For example, borrowing within Italic may shift inferred changes from the more modern branches to the branch leading to Italic, but borrowing between Proto-Italic and its contemporaries will also shift inferred changes from this branch to ancestral branches. This means that whilst we may incorrectly reconstruct some proto-Indo-European roots, our divergence time calculation will not be affected. We maintain that although advergence has undoubtedly occurred throughout the history of Indo-European, and that this may have affected our trees, this effect is likely to be random and there is no reason to think it will have significantly biased our results. Atkinson *et al*. (2005) analyzed synthetic data with simulated borrowing, and found that date estimates were highly robust to even high levels of borrowing.

Ringe *et al*. (2002) argue that non-lexical characters such as grammatical and phonological features are less likely to be borrowed (although they also note that parallel changes in phonological and morphological characters are possible). To avoid potential problems due to lexical borrowing they coded 15 phonological and 22 morphological characters as strict constraints in their analyses (they did not throw out the remaining 333 lexical characters). While we agree that phonological and morphological characters would be very useful, we believe there are good reasons to trust the inferences based on the lexical data in our case. The Dyen *et al*. (1992) data has had much of the known
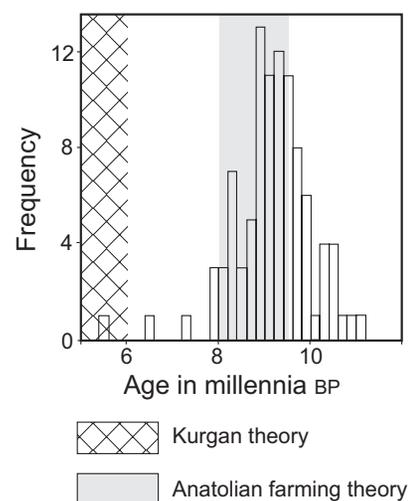

Figure 8.16. *Frequency distribution of basal age estimates from filtered Bayesian MCMC sample of trees using Swadesh 100-word list items only (*n = 97).

borrowing filtered from it. Further, the relationships we infer between Indo-European languages are remarkably similar to those inferred by linguists using the comparative method. Our results are not only consistent with accepted language relationships, but also reflect acknowledged uncertainties, such as the position of Albanian. Our time-depth estimates for internal nodes of the Indo-European tree are also congruent with known historical events (i.e. when constraints were removed step-by-step from each of the 13 internal constraint points, the reconstructed ages were within 390 years of the original constraint range: Gray & Atkinson 2003). Significantly, if we constrain our trees to fit the Ringe *et al*. (2002) typology we get very similar date estimates to our initial consensus tree topology. In short, there is nothing to indicate that either our tree typologies or date estimates have been seriously distorted by the use of just lexical data.

Determined critics might still claim that the remaining undetected lexical borrowing that undoubtedly exists in the Dyen *et al.* data (see Nicholls & Gray Chapter 14 this volume) has led us to make erroneous time-depth inferences at the root of the Indo-European tree. The Swadesh 100-word list is expected to be more resistant to change and less prone to borrowing than the 200-word list (Embleton 1991; McMahon & McMahon 2003). If undetected borrowing has biased our tree topology and divergence time estimates then the 100-word list might be expected to produce different estimates. To assess this possibility we repeated the analysis using only the Swadesh 100-word list items. Figure 8.16 shows the results of this analysis. Predictably, with a smaller data set variance in the age
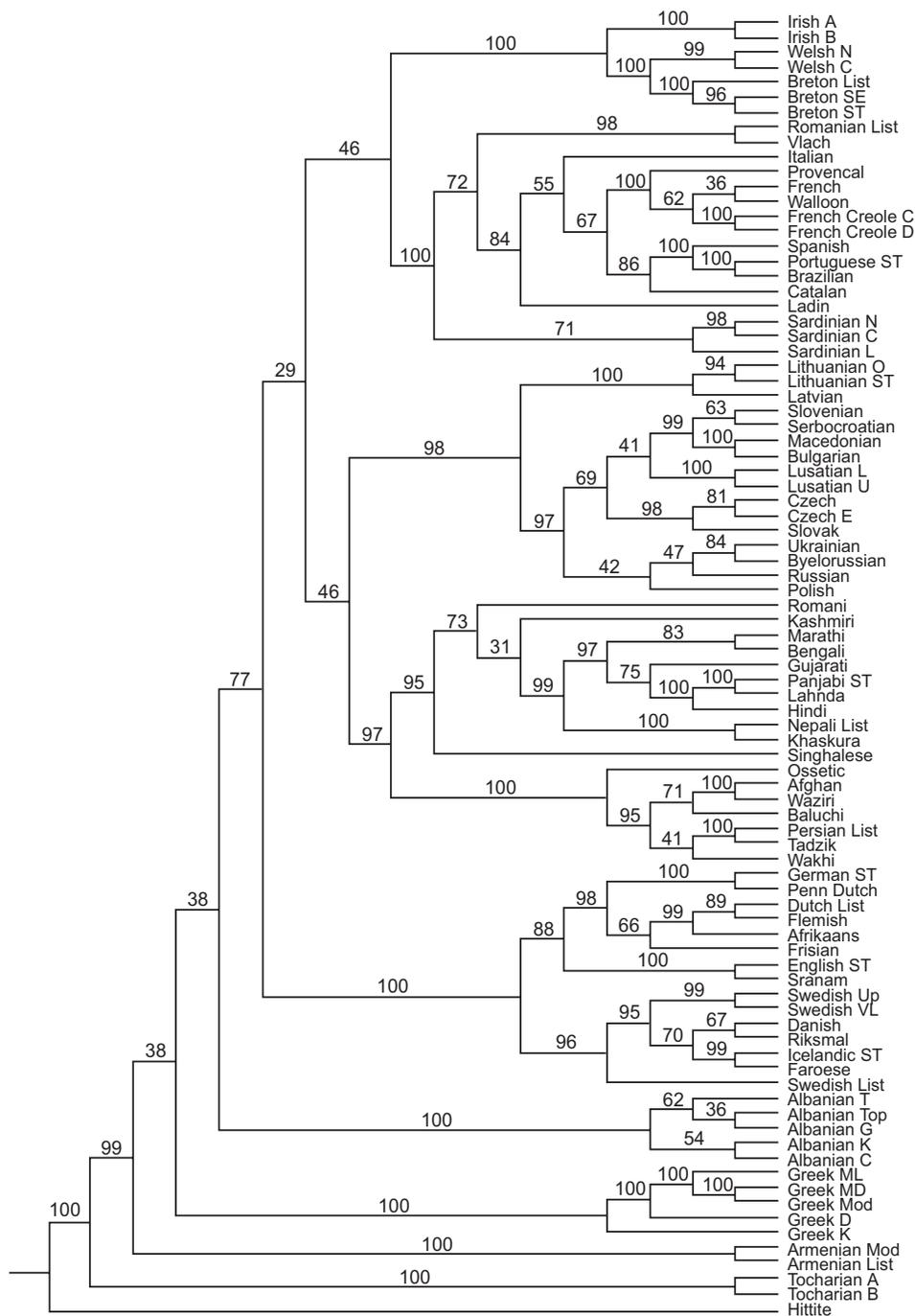
Figure 8.17. *Majority-rule consensus tree (unfiltered) for Swadesh 100-word list items only. Values above each branch express uncertainty (posterior probability) in the tree as a percentage.*

these clades. The low posterior probability values for these groups mean that we should not over-interpret the certainty of these deeper relationships, but clearly the possibility that undetected lexical borrowing is obscuring some of the deeper relationships would repay further attention. We emphasize that this possible borrowing does not appear, however, to affect our time-depth estimates for the root of the tree.

It is interesting to note that whilst our methodology produced consistent results using the Swadesh 100- and 200-word list, Tischler's (1973) glottochronological analysis was affected by the choice of word list. Tischler generated Indo-European divergence times using pair-wise distance comparisons between languages under the assumption of constant rates of lexical replacement. Using the Swadesh 200-word list, he calculated that the core Indo-European languages (Greek, Italic, Balto-Slavic, Germanic and Indo-Iranian) diverged around 5500 BP whilst Hittite diverged from the common stock around 8400 BP. This is in striking agreement with the timing depicted in Figure 8.8. However, the same calculation using the Swadesh 100-word list,

estimates increased. However, the resulting age range was still consistent with the Anatolian theory of Indo-European origin. Interestingly, the majority rule consensus tree (shown in Fig. 8.17) is *slightly* different to that obtained from the full data set. It contains a Balto-Slavic-Indo-Iranian group and an Italo-Celtic group. Ringe *et al.*'s (2002) compatibility analysis also found

produced a Hittite divergence time of almost 11,000 BP. Other inferred divergence times were also older. Tischler favoured the 200-word list results because they tended to be more consistent and were based on a larger sample size. However, the disparate 100 word list ages led Tischler to conclude that the divergence times for Hittite (and a number of other peripheral Indo-

European languages, including Albanian, Armenian and Old Irish) were in fact anomalous and he instead favoured an age for Indo-European of between 5000 and 6000 years, reflecting the break-up of the core languages. He explained the apparent earlier divergence of Hittite, Albanian, Old Irish and Armenian as an artefact of borrowing with non-Indo-European languages or increased rates of change.

11. Conclusion

The analyses we have presented here are far from the last word on the vexed issue of Indo-European origins. We expect that 'every species of scholar and would be savant who can take pen to hand' will still be drawn to the question of Indo-European origins. However, in contrast to some of the more pessimistic claims of our critics, we do not think that estimating the age of the Indo-European language family is an intractable problem. Some of these critics have argued that it is hard enough to get the tree typology correct, let alone branch lengths or divergence times. From this point of view all efforts to estimate dates should be abandoned until we can get the tree exactly right. We think that would be a big mistake. It would prematurely close off legitimate scientific inquiry. The probability of getting the one 'perfect phylogeny' from the 6.66×10^{152} possible unrooted trees for 87 languages is rather small. Fortunately we do not need to get the tree exactly correct in order to make accurate date estimates. Using the Bayesian phylogenetic approach we can calculate divergence dates over a distribution of most probable trees, integrating out uncertainty in the phylogeny. We acknowledge that estimating language divergence dates is difficult, but maintain it is possible if the following conditions are satisfied:
a) a data set of sufficient size and quality can be assembled to enable the tree and its associated branch lengths to be estimated with sufficient accuracy;
b) most of the borrowing is removed from the data;
c) an appropriate statistical model of character evolution is used (it should contain sufficient parameters to give accurate estimates but not be over-parameterized);
d) multiple nodes on the tree are calibrated with reliable age ranges;
e) uncertainty in the estimation of tree topology and branch lengths are incorporated into the analysis;
f) variation in the rate of linguistic evolution is accommodated in the analysis.

The analyses of Indo-European divergence dates we have outlined above go a long way to meeting these requirements. The Dyen *et al.* (1997) data set we used in our analyses contains over two thousand carefully coded cognate sets (condition a). Dyen *et al.* excluded known borrowings from these sets (condition b). The two state, time-reversible model of cognate gains and losses with gamma distributed rate heterogeneity produced accurate trees (i.e. congruent with the results of the comparative method and known historical relationships)[4] (condition c). When the branch lengths were combined with the large number of well-calibrated nodes (condition d), the estimated divergence dates were also in line with known historical events. The Bayesian MCMC approach allowed us to incorporate phylogenetic uncertainty into our analyses (condition e), and to investigate the consequences of variations in the priors, tree rooting, and stringency in cognate judgements. Finally, rate smoothing allowed us to estimate divergence dates without the assumption of a strict glottoclock (condition f). *We challenge our critics to find any paper on molecular divergence dates that uses as many calibration points, investigates the impact of so many different assumptions, or goes to the same lengths to validate its results.*

In the words of W.S. Holt, history is 'a damn dim candle over a damn dark abyss'. Although we see reason for careful scholarship when attempting to estimate language divergence dates, we see no justification for pessimism here. Far from dancing around the question of Indo-European origins like moths around a flame, with the light of computational phylogenetic methods we can illuminate the past.

Notes
1. Ten million post burn-in trees were generated using the MrBayes (Huelsenbeck & Ronquist 2001). To ensure that consecutive samples were independent, only every 10,000th tree was sampled from this distribution, producing a sample size of 1000.
2. The proposed borrowings were: in Romance — ear, fire, liver, count, eat, head and narrow; in Germanic — leaf, sharp and think; and in Indic — kill, night, play, suck, flower and liver. We note that this list was not intended by Garrett to be a comprehensive account of all possible borrowings.
3. Borrowings that are unlikely to favour inferred language change at the base of a subgroup or that would favour inferred language change within a subgroup are: in Romance — ear, head, narrow; in Germanic — leaf; and in Indic — flower and liver.
4. We do, however, agree that the question of model specification would repay further investigation (see Nicholls & Gray Chapter 14 this volume; Pagel Chapter 15 this volume; Atkinson *et al.* 2005).

References

Adams, D.Q., 1999. *A Dictionary of Tocharian B.* (Leiden Studies in Indo-European 10.) Amsterdam: Rodopi. Avail-

able via online data base at S. Starostin & A. Lubotsky (eds.), *Database Query to A dictionary of Tocharian B*. http://iiasnt.leidenuniv.nl/ied/index2.html.

Atkinson, Q.D. & R.D. Gray, 2006. Are accurate dates an intractable problem for historical linguistics? in *Mapping our Ancestry: Phylogenetic Methods in Anthropology and Prehistory*, eds. C. Lipo, M. O'Brien, S. Shennan & M. Collard. Chicago (IL): Aldine, 269–96.

Atkinson, Q.D., G. Nicholls, D. Welch & R.D. Gray, 2005. From words to dates: water into wine, mathemagic or phylogenetic inference? *Transactions of the Philological Society* 103(2), 193–219.

Bateman, R., I. Goddard, R. O'Grady, V. Funk, R. Mooi, W. Kress & P. Cannell, 1990. Speaking of forked tongues: The feasibility of reconciling human phylogeny and the history of language. *Current Anthropology* 31, 1–24.

Bellwood, P., 1991. The Austronesian dispersal and the origin of languages. *Scientific American* 265, 88–93.

Bellwood, P., 1994. An archaeologist's view of language macrofamily relationships. *Oceanic Linguistics* 33, 391–406.

Bergsland, K. & H. Vogt, 1962. On the validity of glottochronology. *Current Anthropology* 3, 115–53.

Blust, R., 2000. Why lexicostatistics doesn't work: the 'Universal Constant' hypothesis and the Austronesian languages, in *Time Depth in Historical Linguistics*, eds. C. Renfrew, A. McMahon & L. Trask. (Papers in the Prehistory of Languages.) Cambridge: McDonald Institute for Archaeological Research, 311–32.

Bryant, D., F. Filimon & R.D. Gray, 2005. Untangling our past: Pacific settlement, phylogenetic trees and Austronesian languages, in *The Evolution of Cultural Diversity: Phylogenetic Approaches*, eds. R. Mace, C. Holden & S. Shennan. London: UCL Press, 69–85.

Burnham, K.P. & D.R. Anderson, 1998. *Model Selection and Inference: a Practical Information-Theoretic Approach.* New York (NY): Springer.

Campbell, L., 2004. *Historical Linguistics: an Introduction*. 2nd edition. Edinburgh: Edinburgh University Press.

Cavalli-Sforza, L.L., P. Menozzi & A. Piazza, 1994. *The History and Geography of Human Genes*. Princeton (NJ): Princeton University Press.

Clackson, J., 2000. Time depth in Indo-European, in *Time Depth in Historical Linguistics*, eds. C. Renfrew, A. McMahon & L. Trask. (Papers in the Prehistory of Languages.) Cambridge: McDonald Institute for Archaeological Research, 441–54.

Devoto, G., 1962. *Origini Indeuropeo.* Florence: Instituto Italiano di Preistoria Italiana.

Diakonov, I.M., 1984. On the original home of the speakers of Indo-European. *Soviet Anthropology and Archaeology* 23, 5–87.

Diamond, J. & P. Bellwood, 2003. Farmers and their languages: the first expansions. *Science* 300, 597.

Dyen, I., J.B. Kruskal & P. Black, 1992. *An Indoeuropean Classification: a Lexicostatistical Experiment.* (Transactions 82(5).) Philadelphia (PA): American Philosophical Society.

Dyen, I., J.B. Kruskal & P. Black, 1997. FILE IE-DATA1. Available at http://www.ntu.edu.au/education/langs/ielex/IE-DATA1.

Embleton, S., 1986. *Statistics in Historical Linguistics.* Bochum: Brockmeyer.

Embleton, S.M., 1991. Mathematical methods of genetic classification, in *Sprung from Some Common Source*, eds. S.L. Lamb & E.D. Mitchell. Stanford (CA): Stanford University Press, 365–88

Excoffier, L. & Z. Yang, 1999. Substitution rate variation among sites in mitochondrial hypervariable region I of humans and chimpanzees. *Molecular Biology and Evolution* 16, 1357–68.

Faguy, D.M. & W.F. Doolittle, 2000. Horizontal transfer of catalase-peroxidase genes between archaea and pathogenic bacteria. *Trends in Genetics* 16, 196–7.

Felsenstein, J., 2004. *Inferring Phylogenies*. Sunderland (MA): Sinauer.

Gamkrelidze, T.V. & V.V. Ivanov, 1995. *Indo-European and the Indo-Europeans: a Reconstruction and Historical Analysis of a Proto-Language and Proto-Culture*. Berlin: Mouton de Gruyter.

Gimbutas, M., 1973a. Old Europe *c.* 7000–3500 BC, the earliest European cultures before the infiltration of the Indo-European peoples. *Journal of Indo-European Studies* 1, 1–20.

Gimbutas, M., 1973b. The beginning of the Bronze Age in Europe and the Indo-Europeans 3500–2500 BC. *Journal of Indo-European Studies* 1, 163–214.

Gkiasta, M., T. Russell, S. Shennan & J. Steele, 2003. Neolithic transition in Europe: the radiocarbon record revisited. *Antiquity* 77, 45–62.

Glover, I. & C. Higham, 1996. New evidence for rice cultivation in S., S.E. and E. Asia, in *The Origins and Spread of Agriculture and Pastoralism in Eurasia*, ed. D. Harris. Cambridge: Blackwell, 413–42.

Goldman, N., 1993. Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* 36, 182–98.

Gray, R.D. & Q.D. Atkinson, 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426, 435–9.

Guterbock, H.G. & H.A. Hoffner, 1986. *The Hittite Dictionary of the Oriental Institute of the University of Chicago.* Chicago (IL): The Institute.

Hillis, D.M., 1992. Experimental phylogenetics generation of a known phylogeny. *Science* 255, 589–92.

Hillis, D.M., C. Moritz & B.K. Marble, 1996. *Molecular Systematics*. 2nd edition. Sunderland (MA): Sinauer.

Hjelmslev, L., 1958. *Essai d'une Critique de la Methode dite Glottochronologique. Proceedings of the Thirty-second International Congress of Americanists, Copenhagen, 1956*. Copenhagen: Munksgaard.

Hoffner, H.A., 1967. *An English–Hittite Dictionary*. New Haven (CT): American Oriental Society.

Holden, C.J., 2002. Bantu language trees reflect the spread of farming across Sub-Saharan Africa: a maximum-parsimony analysis. *Proceedings of the Royal Society of London* Series B 269, 793–9.

Holland, B. & V. Moulton, 2003. Consensus networks: a method for visualising incompatibilities in collections of trees, in *Algorithms in Bioinformatics, WABI 2003*, eds. G. Benson & R. Page. Berlin: Springer-Verlag, 165–76.

Huelsenbeck, J.P. & F. Ronquist, 2001. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* 17, 754–5.

Huelsenbeck, J.P., F. Ronquist, R. Nielsen & J.P. Bollback, 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294, 2310–14.

Jukes, T.H. & C.R. Cantor, 1969. Evolution of protein molecules, in *Mammalian Protein Metabolism*, vol. 3, ed. M.N. Munro. New York (NY): Academic Press, 21–132.

Kuhner, M.K. & J. Felsenstein, 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution* 11, 459–68.

Kumar, V.K., 1999. *Discovery of Dravidian as the Common Source of Indo-European*. Retrieved Sept. 27th 2002 from http://www.datanumeric.com/dravidian/.

Levins, R., 1966. The strategy of model building in population biology, *American Scientist* 54, 421–31.

Mallory, J.P., 1989. *In Search of the Indo-Europeans: Languages, Archaeology and Myth.* London: Thames & Hudson.

McMahon, A. & R. McMahon, 2003. Finding families: quantitative methods in language classification. *Transactions of the Philological Society* 101, 7–55.

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller & E. Teller, 1953. Equations of state calculations by fast computing machines. *Journal of Chemical Physics* 21, 1087–91.

Otte, M., 1997. The diffusion of modern languages in prehistoric Eurasia, in *Archaeology and Language,* eds. R. Blench & M. Spriggs. London: Routledge, 74–81.

Pagel, M., 1997. Inferring evolutionary processes from phylogenies. *Zoologica Scripta* 26, 331–48.

Pagel, M., 1999. Inferring the historical patterns of biological evolution. *Nature* 401, 877–84.

Pagel, M., 2000. Maximum-likelihood models for glottochronology and for reconstructing linguistic phylogenies, in *Time Depth in Historical Linguistics,* eds. C. Renfrew, A. McMahon & L. Trask. (Papers in the Prehistory of Languages.) Cambridge: The McDonald Institute for Archaeological Research, 413–39.

Renfrew, C., 1987. *Archaeology and Language: the Puzzle of Indo-European Origins.* London: Cape.

Rexová, K., D. Frynta & J. Zrzavy, 2003. Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. *Cladistics* 19, 120–27.

Ringe, D., n.d. Proto-Indo-European Wheeled Vehicle Terminology. Unpublished manuscript.

Ringe, D., T. Warnow & A. Taylor, 2002. Indo-European and computational cladistics. *Transactions of the Philological Society* 100, 59–129.

Rosser, Z.H., T. Zerjal, M.E. Hurles *et al.,* 2000. Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *American Journal of Human Genetics* 67, 1526–43.

Sanderson, M., 2002a. *R8s, Analysis of Rates of Evolution*, version 1.50. http://ginger.ucdavis.edu/r8s/

Sanderson, M., 2002b. Estimating absolute rates of evolution and divergence times: a penalized likelihood approach. *Molecular Biology and Evolution* 19, 101–9.

Steel, M., M. Hendy & D. Penny, 1988. Loss of information in genetic distances. *Nature* 333, 494–5.

Swadesh, M., 1952. Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* 96, 453–63.

Swadesh, M., 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21, 121–37.

Swofford, D.L., G.J. Olsen, P.J. Waddell & D.M. Hillis, 1996. Phylogenetic Inference, in *Molecular Systematics*, eds. D.M. Hillis, C. Moritz & B.K. Marble. 2nd edition. Sunderland (MA): Sinauer, 407–514.

Tischler, J., 1973. *Glottochronologie und Lexicostatistik*. Innsbruck: Innsbrucker Verlag.

Tischler, J., 1997. *Hethitisch-Deutsches Worterverzeichnis*. Dresden: Probedruck.

Trask, R.L., 1996. *Historical Linguistics.* New York (NY): Arnold.

Watkins, C., 1969. *Indogermanische Grammatik III/1. Geschichte der Indogermanischen Verbalflexion.* Heidelberg: Carl Winter Verlag.