Transactions of the Philological Society Volume 100:1 (2002) 59–129

INDO-EUROPEAN AND COMPUTATIONAL CLADISTICS[1]

By Don Ringe
University of Pennsylvania

Tandy Warnow
University of Texas

Ann Taylor
University of York

(Received 5 July 2001)

Abstract

This paper reports the results of an attempt to recover the first-order subgrouping of the Indo-European family using a new computational method devised by the authors and based on a 'perfect phylogeny' algorithm. The methodology is also briefly described, and points of theory and methodology are addressed in connection with the experiment whose results are here reported.

[1] The work reported here was supported by NSF grant SBR-9512092, 'Character-based methods for reconstructing the evolutionary history of natural languages', and by generous grants of hardware by Intel Corporation and of software by Microsoft to the University of Pennsylvania as part of the Eniac 2000 project. We are grateful to Sasha Mihailov for writing our perfect phylogeny software, to Alex Garthwaite for optimising the code, and to Libby Levison for constructing a user interface. Earlier versions of this paper were read at Virginia Polytechnic Institute in April 1998, to the Philological Society in May 2000, at the Universities of Cambridge and York in the spring of 2000 and at Swarthmore College in March 2001. We wish to thank the audiences of all those talks for much helpful discussion and criticism, especially James Clackson, Sean Crist, Joe Eska, Henry Hoenigswald, John Penney and Anthony Warner; discussions with Johanna Nichols have also been fruitful, and we are grateful to two anonymous reviewers for help in clearing up various obscurities and for additional references. Craig Melchert's assistance with the Anatolian data has been invaluable. Responsibility for all statements in this paper rests with the authors.

108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA.

1. Introduction

This is an interim report on work in progress. Ringe and Taylor are historical linguists, while Warnow is a computer scientist. We have each handled the most technical aspects of the project appropriate to our own disciplines; but the methodology we have developed involves much more than the application of already known algorithms to already known linguistic data, since we have encountered numerous problems not previously addressed, and the intellectual contributions of all the authors have been very varied.

2. Basic assumptions

Our methodological presuppositions and definitions are those usual in mainstream historical linguistics. However, we would like to emphasise two points that are seldom discussed by working historical linguists even though they are crucial to the enterprise.

2.1. The uniformitarian principle and its application

First, we insist on a rigorous and consistent application of the uniformitarian principle (UP). The UP holds that we can constrain our hypotheses about the structure and history of languages of the past only by reference to what we know of contemporary language structures, linguistic behaviour and changes in progress, since the recoverable information about any language or speech community of the past is always far more limited than what we can know about languages whose native speakers we can still observe; and further, that we can extrapolate into prehistory (and across gaps in the historical record) only on the basis of what we know from the study of contemporary languages and the actually documented past. Positing for any time in the past any structure or development inconsistent with what is known from modern work on living languages is unacceptable, and positing for prehistory any type of long-term development that we do not observe in documented history is likewise unacceptable, unless it can be demonstrated that there has been some relevant change in the conditions of language acquisition or use between the past time in question and later periods which can be observed or have been documented. Practically speaking, this means that 'the perspective . . . of the historical linguist', aiming 'to describe and analyse linguistic results of language contact situations' (Thomason and Kaufman 1988: 36) or of any other kind of linguistic development, uninformed by recent findings in sociolinguistics, studies of language acquisition or bilingualism, experimental phonetics, and so on, can never be adequate. The following applications of the UP are especially important to our work.

Languages replicate themselves (and thus 'survive' from generation to generation) through the process of native-language acquisition by children. Importantly for historical linguistics, that process is tightly constrained; for instance, the system of morphosyntactic categories is normally mastered by age four, and native acquisition of a language is virtually impossible after the onset of puberty (see e.g. Slobin ed. 1985, 1992). Moreover, it appears that every successful acquisition of a native language gives rise to a robust grammatical 'signature' which persists throughout life. The most important details and their consequences can be summarised as follows.

Recent research on native-language acquisition by children shows that the contrastive system of sounds, the inflectional morphology and the basic syntax of a native language are acquired in the first six or seven years of life, and that 'mixed' grammars are not acquired even in multilingual environments (see e.g. Fantini 1985, Meisel 1989 with references). Recent research in sociolinguistics shows that, while most linguistic structures can be borrowed between closely related dialects, natively acquired sound systems and inflections are resistant to change later in life; attempts to acquire a non-native phonemic contrast, phonological rule or inflectional category are at best only partly successful (cf. e.g. Labov 1994: 518–526). Borrowing between speechforms that are not very similar appears to be even more severely constrained, as one would expect. Studies of the bilingual situations in which borrowing occurs show that the phonology and morphosyntax of one's native language are typically carried over into a language learned later in life, but not usually the other way round (cf. e.g. Rayfield 1970: 103–107, Prince and Pintzuk 2000).

Much of the language-contact literature appears to challenge the results summarised in the preceding paragraph; one often encounters claims that practically anything can be borrowed into one's native language in a suitable bilingual situation. But such claims are almost always made by inference from the *results* of language contact; published data bearing on the *process* of borrowing (if that is what it is) are either wholly lacking or do not effectively exclude other analyses, such as code-switching or the results of imperfect learning of a second language (Appel and Muysken 1987: 158–163, King 2000: 44–47 with references). We therefore remain unconvinced that morphosyntax, for example, can be borrowed *into a native language*, except (probably) in one situation: morphosyntactic structures even of very different languages can apparently be borrowed into a community's native language in the context of community-wide bilingualism persisting for many generations. An obvious example is the successful borrowing of Hebrew noun plurals into Yiddish (a Germanic language) – which might show, in addition, that the lending language need not be native, so long as a large proportion of the community uses it fluently.[2] It seems likely that some first-language learners in such situations misinterpret frequent code-switching as monolingual behaviour and thus learn foreign morphosyntax as part of their native language, and that, given enough time, their analysis can become dominant in the community. However, we must also note that the *only* study in depth of such a process in progress concludes that even morphosyntactic borrowing of this kind is mediated by lexical borrowing: in effect, 'core' lexemes are borrowed and bring their morphosyntax with them (King 2000; cf. also Kroch 1994: 191–193). Other apparent examples of the systematic borrowing of phonology or

[2] Of course it is possible that the male members of traditional East European Jewish communities technically had a native passive command of Hebrew, since it was usual for boys to begin Hebrew school at three – well within the developmental window in which a native language must be acquired – but we are not aware of any cogent study of the question. Note that the corresponding failure of English to borrow such Latin and Greek plurals as *alumni, data* and *phenomena* (currently being reinterpreted as singulars in colloquial English) seems to be a direct result of the fact that familiarity with the Classical languages, which was never more than minimal for many educated speakers of English, is no longer enforced by the educational system in any English-speaking country.

morphosyntax can actually have resulted from the importation of native structures into an imperfectly learned second language (on which see further below).

We therefore think it reasonable to adopt the hypothesis that phonology and morphosyntax are normally excellent indicators of a person's native language. That hypothesis will be crucial to our discussion in the following section.

A final point of relevance to our application of the UP is the following. It has long been known that the loss of contact between diversifying dialects of a language can be either abrupt or gradual. A sequence of abrupt separations can easily be modelled as an evolutionary tree, but if dialects lose contact only gradually, they can borrow linguistic material from their nearest neighbours in overlapping patterns that render modelling of their diversification as a tree unrealistic.

We will return to all these points below.

2.2. Linguistic descent

Our second methodological point is related to the first. We adopt a precise definition of LINGUISTIC DESCENT:

> A language (or dialect) Y at a given time is said to be descended from language (or dialect) X of an earlier time if and only if X developed into Y by an unbroken sequence of instances of native-language acquisition by children.

Since mixed grammars are not known to result from native-language acquisition, it follows that any language or dialect has only one ancestor at any given point in the past, unless there has been a significant discontinuity in its transmission from generation to generation; this is the usual view among historical linguists (cf. Thomason and Kaufman 1988: 11). To judge from the aggregate of languages whose histories are actually documented for at least a few centuries, such discontinuities appear to be infrequent; that is also the standard view among historical linguists (cf. Thomason and Kaufman 1988: 3, Ross 1997: 209–210).

Moreover, most examples of discontinuity in transmission seem to fall into two categories that are reasonably well understood. One easily recognisable class of languages with discontinuous transmission histories are creoles according to the strict definition – that is, native languages which are descended from pidgins.[3] But a moment's consideration of the degree of social dislocation necessary to give rise to a whole community of pidgin-speaking households will show that creoles ought to be relatively rare. (Even widespread colonial slavery in recent centuries has given rise to only a few dozen creoles among the 6,000 or so languages still spoken.) The other obvious class of languages with anomalous histories are those descended from an imperfectly learned second language which became the community norm. Far less attention has been paid to this phenomenon, but it is not necessarily rarer than creolisation. Kroch, Taylor and Ringe (2000) argue that at least one Middle English dialect of the northeast midlands is descended from the English learned imperfectly by Scandinavian settlers who were obliged to learn the language during the reconquest but had too little contact with native speakers to enable their children to learn a 'normal' dialect natively. A surprising number of examples of the supposed borrowing of foreign morphosyntax by native speakers can be reinterpreted rather easily as the influence of a native language on a second language. For instance, the apparent borrowing of inflectional morphology from Turkish into the Asia Minor dialects of modern Greek (Thomason and Kaufman 1988: 215–222, recapitulating Dawkins 1916) is almost certainly the result of imperfect learning of Greek as a second language by clergy who had become native speakers of Turkish, the 'contaminated' dialect of the clergy subsequently becoming the norm in isolated villages. Dawkins actually cites a remarkable document of 1437 informing us that the Greek Orthodox village clergy were then Turkish-speaking (Dawkins 1916: 1 fn. 1), and in early work he had been led by phonological patterns betraying non-native learning to the

[3] It seems to us counterproductive to use the term 'creolisation' more loosely. We emphasise that Thomason and Kaufman (1988) have *not* demonstrated that there is any kind of continuum between creolisation as here defined and other types of contact phenomena; they have merely shown that the results of some other episodes of contact are not always clearly distinguishable from the results of creolisation several centuries after the fact if one investigates them from a largely system-internal perspective. That strikes us as an especially powerful argument in favor of the UP (as we interpret it).

conclusion that various Greek dialects in Italy and Turkey had some such episodes in their histories (Dawkins 1910: 270, 289). His rejection of that hypothesis – apparently on the grounds that there was no explicit documentary evidence for it – is a lamentable example of abandoning the UP. The case of Ma'a (Thomason and Kaufman 1988: 223–228) can be linguistically similar, the language having been 'saved' from overwhelming Bantu influence by timely migration (ibid. pp. 225–226) when too many of the younger generation were already native Bantu speakers. Though more research on such phenomena is urgently needed, it already seems clear that community-wide imperfect second-language acquisition is a phenomenon of some importance for linguistic history (though the number of languages which it can be shown to have affected is still small).

There remains a tiny handful of languages that exhibit unarguably mixed grammars but do not seem to be typical creoles – perhaps only Michif and Mednyj Aleut (see Thomason and Kaufman 1988: 228–238). Strictly speaking, the origins of such languages will not be definitively explained until such a development in progress can be observed and studied, though the suggestion that they were deliberately created by fluent bilinguals as a kind of 'code' to exclude outsiders is inherently plausible (cf. Mithun 1999: 596–602 with references).

From the above considerations, taken all together, it follows that the tree model of linguistic speciation is normally appropriate, *if* the loss of contact between diverging dialects has been relatively abrupt and no discontinuities of transmission can be demonstrated for any of the languages in question. And since we know that inflectional morphology and the phonemic system of a native dialect are learned very early and are resistant to subsequent change, it follows that morphology and phonology provide better information about linguistic descent (in our precise sense) than lexical evidence (see already Meillet 1925: 22–33).

3. Traditional subgrouping and its shortcomings

The research that we are reporting has developed a new computational method for subgrouping the languages of a provable family. It is *not* intended to replace already existing methods, but to supplement them. As is well known, traditional subgrouping is logically coherent and methodologically unobjectionable: in order to subgroup a particular subset of the family's languages together, one demands that they exclusively share clear and linguistically significant innovations which are unusual enough that they could not reasonably have arisen more than once independently. To put it in a biologist's terms,[4] one recognises a clade by the presence of unique synapomorphies, rigorously excluding any traits that might conceivably be analogous rather than homologous. This is so clearly correct that we have no intention of even questioning it; on the contrary, we incorporate it into our own methodology.

However, the attempt to apply the traditional criteria for subgrouping rigorously encounters severe practical problems which are too often overlooked or downplayed. Examples from phonology, inflectional morphology and the lexicon can be adduced to illustrate these problems.

Traditional subgrouping tends to rely on phonology because phonemic mergers are clearly innovations. But though the set of sound changes in each line of descent is unique, the individual changes are usually so 'natural' that they can easily be repeated in different lines of descent; that is, they are the products of phonetic pressures that operate in all languages and frequently give the same results in cases widely separated in space and time. That is true whether one states the changes in phonetic or phonemic terms. It is easy enough to find a phonetic change that has occurred in no fewer than four languages widely separated in space and time:

Proto-Greek **ti* > South Greek *si* (e.g. **dídōti* '(s)he's giving' > Attic *dídǭsi*, cf. Doric *dídǭti* with no change; **triākátioi* 'three hundred' > Attic *triākósioi*, cf. Doric *triākátioi* with no change; isogloss dating to the second millennium BCE, cf. Risch 1955: 66, 75).

[4] Serious interpenetration of biological and linguistic concepts in an evolutionary context began at least a decade and a half ago (see Hoenigswald and Wiener ed. 1987); more recent examples can be found in earlier reports of our work (see the References), in Lass (1997: 113–123) and in McMahon (2000: 137–176). A 'synapomorphy' is a shared innovation; a 'homology' is a shared trait inherited from a common ancestor, and shared traits which are not homologous are said to be 'analogous'.

Proto-Indo-European (PIE) _*ti, *d^{h}i_ > _*ti, *t^{h}i_ > pre-Proto-Tocharian _*si_ (e.g. _*h$_{1}$id^{h}í_ 'go!' > _*it^{h}í_ > _*isí_ > _*yəṣə́_ → PT _*pəyəṣə́_ > TA _piṣ_, TB _paṣ_, cf. Jasanoff 1987: 108–112, Ringe 1996: 47–48, 80, 88; the change occurred after the devoicing of aspirates, which in turn occurred after aspirate dissimilation, so that it must have been completely independent of the Greek change).

Proto-Finno-Ugric _*ti_ > Finnish _si_ (e.g. _*käti_ 'hand' > _käsi_, cf. Mari _kit_, Mansi _kāt; *weti_ 'water' > _vesi_, cf. Mari _ßüt_, Mansi _wit_; see e.g. Fromm and Sadeniemi 1956: 26–27, 39–40, Laanest 1982: 22–23, 102–103).

Proto-Polynesian _*ti_ > Tongan _si_ (e.g. _*^{ʔ}oti_ 'be finished' > _^{ʔ}osi_, cf. Maori _oti; *tiro_ 'look at' > _sio_, cf. Maori _tiro_; Biggs 1978: 703).

It is equally easy to find a merger that occurred no fewer than three times independently, namely the merger of short _e_ and _i_ in vowel systems with four or five qualitatively different vowels and a systematic distinction of length:

PIE _*e, *i_ > Proto-Tocharian _*ə_ with palatalisation of the preceding consonant (Ringe 1996: 124–126):
PIE _*léymon-_ ~ _*limn-´_ 'lake' → _*límn̥_ > PToch. _*l^{y}ə́mə_ > TA _lyäm_, TB _lyam_
PIE subj. _*léǵ^{h}eti_ '(s)he will lie down' > PToch. _*l^{y}ə́śə_ > TB _lyaśäm_

Proto-Germanic _*e, *i_ > Gothic _aí_ (= [ɛ]) before _r_ and _h_, but _i_ elsewhere (Braune and Ebbinghaus 1973: 18, 23):
PGmc. _*fiskaz_ 'fish' > Goth. _fisks_
PGmc. _*fedwōr_ 'four' > Goth. _fidwor_
PGmc. _*firnijaz_ 'ancient' > Goth. _faírneis_
PGmc. _*erþō_ 'earth' > Goth. _aírþa_

Proto-Algonkian _*e, *i_ > Cree _i_ (Bloomfield 1946: 86):
PA _*noočpinatamwa_ 'he pursues it' > Cree _noospinatam_
PA _*pe^{ʔ}tenamwa_ 'he takes it by error' > Cree _pistinam_
PA _*elenyiwa_ 'human being' > Cree _iyiniw_

The probability of parallel development is thus relatively high for most apparently shared sound changes, and the probability of historically shared development is correspondingly low. Of course not all sound changes are equally likely to recur repeatedly in historically unconnected cases; some, at least, seem rare enough that it might be worth trying to use them as potential indicators of shared history. Changes that give rise to unusual segment types come immediately to mind, but experience seems to show that changes with unusual constraints on their conditioning environments are much more common and potentially very useful (since odd conditioning environments are not very likely to recur by chance).[5] But if one chooses to use such individual sound changes for subgrouping, one must do so with a clear appreciation of the risks involved, not only because the possibility of parallel development can never be absolutely excluded, but also because our estimates of the probabilities involved must remain very approximate until we have a fairly complete catalogue of sound changes for at least a few language families and linguistic areas. Of course one can avoid the problem altogether by employing sets of phonemic changes rather than single changes as evidence for clades, since the probabilities of independent repetition plummet when multiplied (as Sarah Thomason pointed out to us years ago; see the Appendix for examples). But an unavoidable consequence of that strategy is that sound changes provide much less information for subgrouping than might be supposed.

In the domain of inflectional morphology such parallel development seems to be much less prevalent, apparently because inflectional systems are such tightly integrated idiosyncratic constructs that conditions which would give rise to similar changes are unlikely to recur in different languages. Unfortunately, if we have to work from the ultimate outcomes of the changes – the terminal nodes, or LEAVES, of the tree – we often cannot discover which inflectional markers are ancestral and which represent innovations. A notorious example is the optative suffix for thematic verb stems (i.e., those ending in **-e-* ~ **-o-*) in the more archaic Indo-European (IE) languages. The reconstructable shapes of this suffix for different subfamilies are the following:

[5] This probabilistic approach strikes us as more realistic than attempting to use only sound changes which could not possibly have been repeated, if only because such a categorical negative can never be proved.

Anatolian	Tocharian	Italic, Celtic	other subgroups[6]
[lacking]	*-ih$_1$-	*-ā-	*-oy-

The majority suffix is *-_oy_-, which is analysable into the thematic vowel (*-_o_-) and a recognisable optative marker (*-_y_- < *-_ih_$_1$-, with the 'laryngeal' consonant lost by regular sound change); precisely because it is analysable, this suffix is likely to be an innovation. But which of the alternatives is the original suffix? In Tocharian the thematic vowel is dropped and the usual optative suffix added; in Italic and Celtic a completely opaque suffix appears in place of the thematic vowel. Anatolian lacks a category 'optative', and the verb system of Anatolian is so different from that of the other subfamilies that it is reasonable to wonder whether such a category ever existed in Anatolian. That is as far as reliable philological reasoning will take us; we have no idea whether PIE had a thematic optative like that of Tocharian or that of Italic and Celtic, or used a completely different formation, or had no optative at all.

So in the case of our phonological evidence for subgrouping innovations are generally obvious, but we often cannot discover which are truly homologous; and in the case of our morphological evidence homologies are generally obvious, but it often isn't clear which ones are innovative! Lexical evidence is still more problematic for a number of well-known reasons. As with morphological evidence, it is often unclear which words are innovations; parallel semantic development is rampant, so that semantic innovations (even when detectable) are often not homologous; words borrowed from other languages can work their way into a language's basic vocabulary over time, and if they were borrowed from a closely related language the borrowing may not be easy to detect.

What we need is a method which is just as rigorous as the traditional method but can make use of more of the available evidence.

[6] The subgroups in question are Indo-Iranian, Greek, Germanic and Balto-Slavic (this optative suffix is the ancestor of the Lithuanian 'permissive' and the Slavic imperative suffixes); no trace of it survives in Armenian or Albanian, both of which preserve an organisation of the verb that strongly resembles the Greek system. On the sound change by which the laryngeal was lost in *-_oy_- see Beekes (1969: 239–242, especially p. 241), Mayrhofer (1986: 131 with references); on the Italic and Celtic suffix see Trubetzkoy (1926), which still strikes us as the most convincing analysis.

4. The absence of backmutation

The crucial step in the construction of such a methodology is the observation that, in any area of linguistic structure in which categorical distinctions are made, BACKMUTATION[7] is either impossible or vanishingly rare. In other words, we simply do not find cases in which the contrast between two elements A and B in a structured system is eliminated from the language, then at a later stage of the language's descent (in the strict sense defined above) reintroduced in precisely the same distribution that it originally exhibited. (Of course it is possible for the contrast to be reborrowed from a closely related dialect, but in such cases the distribution is usually altered; see e.g. Labov 1994: 518–526.) An exact reversal of a phonemic change (i.e., a sound change involving phonemic merger; see Hoenigswald 1960: 75–79, 86–98) within a single line of descent is literally impossible, as every historical linguist knows. Exact reversals of changes in inflectional morphology, or reversals of total shifts in the meanings of words (such that a word which originally meant only *x* came to mean only *y*, then reverted to meaning only *x* again) might in theory be possible, but they are so improbable that we have been unable to find any examples. The only clear exception to this principle is that a completely new item can be acquired and then lost again; an obvious example are the innovative cases of Old Lithuanian, which have not survived in the modern language.

Therefore, if we choose our data with care, backmutation is effectively excluded from the true evolutionary tree of a language family. If we can also manage to exclude all loanwords and to identify and sequester all parallel innovations, the true tree defined by the remainder of the data becomes a mathematically interesting object with properties that we can exploit in order to recover it from linguistic information present in its leaves. In order to explain how that is done we need to introduce some technical terms of computational cladistics.

[7] We have used this term of evolutionary biology because it is completely unambiguous – unlike, for example, 'reversal' (Lass 1997: 119).

5. Character-based cladistics

Following observations by Henry Gleason and Annette Dobson, we organise our data as characters, such that every character is a linguistic property which languages can instantiate in a variety of ways, and languages which instantiate the character in the same way are assigned the same state of that character. In principle, each character represents an identifiable point of grammar or lexical meaning which evolves formally over the course of the language family's development, and each state of the character ought to represent an identifiable unique historical stage of development – a true homology. (See Gleason 1959, Dobson 1969 for further discussion.) We employ lexical, morphological and phonological characters, as follows.

Each meaning on a basic wordlist is a character, since each language can be expected to have some word that expresses the meaning; languages are assigned the same state if and only if they exhibit true cognates in that meaning.[8] A simple example of a lexical character is the basic meaning 'hand' for the set of related languages {English, German, French, Spanish, Italian, Russian}. The data are the following:

Eng.	hand	Fr.	main	Ital.	mano
Ger.	Hand	Span.	mano	Russ.	ruká

[8] The obvious alternative – treating each cognate set as a character and assigning the same state to each language that exhibits a cognate – fails for the following reason. States must be assigned on the principle that each state should have arisen only once in the evolutionary history of the family; that is, it should be either the original state or a unique derived state, since otherwise the useful mathematical properties of a perfect phylogeny will fail to hold (see further below). It follows that *each* language which exhibits no evidence for a particular character must be assigned a *unique* state, unless we can be reasonably sure that the absence of any instantiation of that character arose by a single historical event in two or more of the languages. Since cognates can easily be lost, the latter condition never obtains; thus if the characters are cognate sets, each language exhibiting no cognate must be assigned a unique state. The result is a character with only one shared state (all the rest being unique). But *a character with that pattern of states can be fitted to any tree*, since we can always posit that each unique state is confined to a single leaf-node; technically we say that such a character is UNINFORMATIVE. It follows that if we code our lexical characters by cognate set we will have no usable lexical evidence. The problem of 'partial cognates', raised by an anonymous reviewer, will be dealt with below in the context of multiple coding of lexical characters.

Since the English and German words are cognate (i.e., descended from the same protoform, namely Proto-Germanic **handuz*, by direct linguistic inheritance), those languages must be assigned the same state for this character; the three Romance languages must likewise be assigned a second state (since their words are all descendants of Latin *manus*) and Russian must be assigned a third:

Eng.	1	Fr.	2	Ital.	2
Ger.	1	Span.	2	Russ.	3

(Note that the identities of the states matter – that is, it matters that English and German share the same state but Russian does not, and so on – but the numbers are purely arbitrary; any clear system of notation will do.) Since inflectional affixes and other inflectional markers exhibit cognations in the same way, morphological characters are similar. For instance, the character 'future tense' would have to be assigned the following states for the same six languages:

Eng.	1	Fr.	3	Ital.	3
Ger.	2	Span.	3	Russ.	4

In this case the three Romance languages share a cognate formation (reflecting the late Latin construction infinitive + *habeō*, still preserved as a phrase in Sardinian and Sicilian), but none of the other languages exhibit cognations. (The English and German constructions are parallel but not identical, since the auxiliaries employed are not the same, and in fact they reflect parallel historical development rather than a genuine homology.)[9]

Phonological characters, which reflect regular sound changes (or sets of sound changes), are coded differently: there are normally

[9] Of course this is not the only possible analysis; as an anonymous reviewer reminds us, some may prefer to maintain that English, German and Russian all lack a 'future tense' (though the coding will not change, since languages lacking a category must be assigned unique states – see the preceding footnote). But the coding must be based on *some* analysis, and it is reasonable that it be based on the analysis of the researchers. Colleagues who wish to propose an alternative coding based on an analysis of their own are of course welcome to do so; in fact, we think it would be instructive to run our software on several different codings of the same data, in order to determine how different their consequences for rigorous subgrouping really are. (See further fn. 12 below.)

only two states, since a language either has undergone a sound change or has not. Moreover, because mergers are irreversible[10] we are usually able to say with confidence which state is ancestral. Thus phonological characters provide most of our evidence for where in the tree the ROOT NODE, representing the ancestor language, lies.

The central insight on which our methodology depends can now be stated: given that backmutation is easily excluded, *if* all loanwords are coded with unique states and all characters exhibiting parallel development are shelved (temporarily), *every state of each remaining character will be* CONVEX *on the true evolutionary tree.* In other words, each node of the tree (both internal nodes and terminal nodes) will be assigned exactly one state, and for every two nodes sharing a given state for a given character, all the nodes on the unique path in the tree between those two nodes will also be labelled with the same shared state. Alternatively, we can say that each state of each character will occupy a connected subgraph of the tree; in plain English, the areas of the tree defined by single states will never be discontinuous nor overlapping. Crucially, *that will be true no matter where in the tree the root node lies*; it is that fact which allows us to dissociate the topology of the tree from its rooting, and to make use of morphological and lexical characters for which we do not know the ancestral state (i.e., the state of the root node).

If all the states of a character are convex on a particular tree, the character too can be said to be convex on, or COMPATIBLE with, the tree. Figure 1 illustrates a lexical character, 'hand', fitted to one of the 'best' trees for IE currently returned by our software (on which see further below). Note that every node, including the internal nodes, can be assigned a single state, and that for every two nodes sharing a given state of this character, all the nodes on the unique path in the tree between those two nodes also share that state. (The states of the starred nodes are indeterminate; that is, each of those nodes could be assigned any one of two or more states without loss

[10] The usual formulation – that mergers are irreversible 'by linguistic means' – is probably too weak, considering Labov's finding that phonemic contrasts are not usually borrowed successfully from other dialects of one's native language (see above). Complete replacement of one dialect by another within the window of native acquisition (in school, for example) is of course analogous to 'language shift' and cannot reasonably be analysed as the reversal of a merger within a single line of linguistic descent.

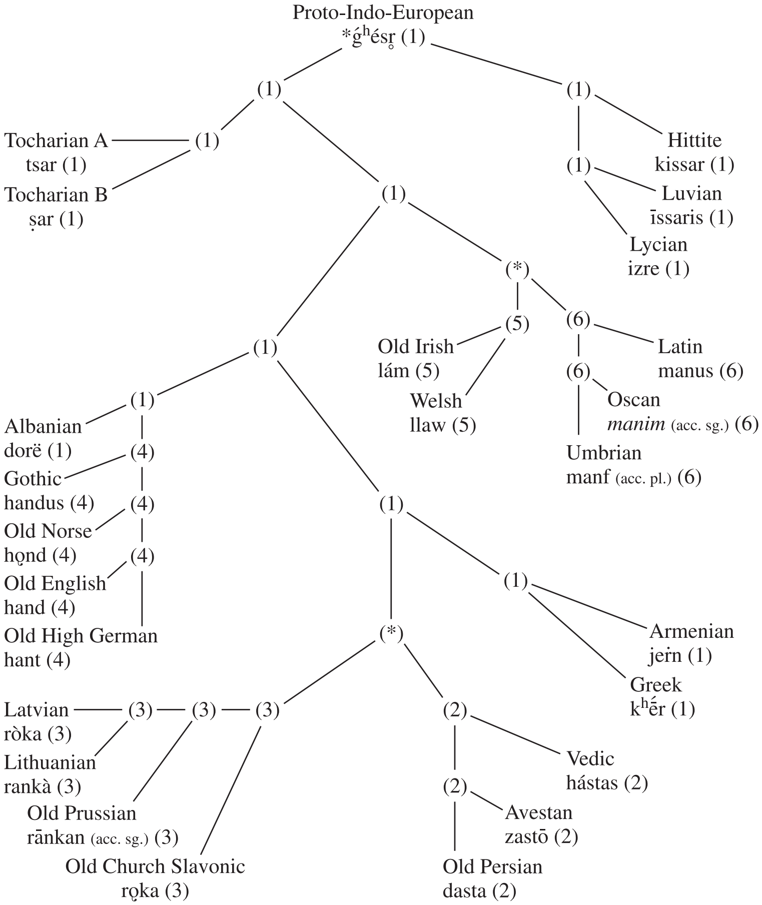

Figure 1. The lexical character 'hand' on one of our current 'best' trees.

of convexity. For instance, the node which immediately dominates both the Celtic and Italic subgroups could be assigned state 1, state 5 or state 6.) Thus this character is compatible with this tree.

Figure 2, on the other hand, illustrates a lexical character, 'one', which is incompatible with the same tree, since *either* state 2 (shared by Tocharian and Graeco-Armenian, and almost certainly reflecting

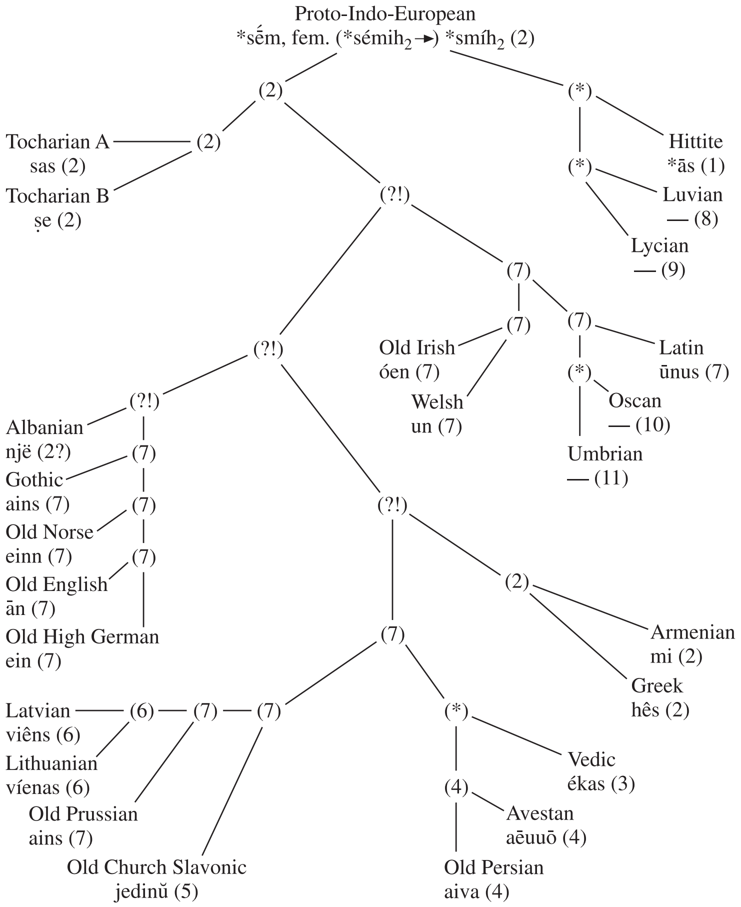

Figure 2. The lexical character 'one' on one of our current 'best' trees.

the ancestral word) *or* state 7 (shared by Italo-Celtic, Germanic and Old Prussian) must occupy a discontinuous subgraph – or else all the nodes marked '(?!)' must be assigned both state 2 and state 7, so that those states overlap.

In order to understand the significance of character compatibility, the reader is invited to think of these trees as mobiles made of string which can be picked up at any node or along any edge (i.e., any link joining two nodes), the place at which the tree is picked up being a hypothesised root node from which the whole tree depends. Consider the tree in Figure 1. For the sake of argument let us suppose (contrary to all the evidence!) that the root-node for IE were Vedic Sanskrit, or fell somewhere within the Indo-Iranian branch. If we imagine picking the tree of Figure 1 up at or near the Vedic node, the ancestral state will be hypothesised to be 2 rather than 1; yet all the states will still be convex on the tree, so that the character will still be compatible with the tree. The last statement will be true no matter where we pick the tree up. On the other hand, we can easily convince ourselves by experiment that no possible rerooting of the tree in Figure 2 will make the character compatible with it. Character compatibility is completely independent of rooting; in mathematical terms it has nothing to do with the temporal and developmental directionality which rooting imposes on an evolutionary tree. It is precisely this property that allows us to make use of morphological and lexical characters for which the rooting is unknown; earlier methodologies, by contrast, could not accommodate characters for which the rooting is unknown in the inference stage.

A tree with which all characters are compatible is called a PERFECT PHYLOGENY (PP); the true evolutionary tree of a language family should be a PP if all loanwords and all parallel development can be excluded – that is, if it were possible to do the relevant philological work perfectly![11] Finding PPs from character data of the leaves of a tree is a known problem in computational cladistics, called (naturally) the perfect phylogeny problem.

Unfortunately the PP problem is NP-hard (Bodlaender et al. 2000). In other words, it is believed that there can be no algorithm

[11] A consequence of this line of reasoning is that character compatibility, rather than parsimony, ought to be the appropriate optimisation criterion for linguistic evolutionary trees found by the analysis of character data. In other words, it should be the case that a 'bad' character – one incompatible with a particular tree – is a serious problem, requiring a non-trivial explanation, even if its non-convexity is very small and limited to a couple of nodes; and if that is true, the 'badness' of a particular tree – the margin by which it fails to be a perfect phylogeny – is best estimated by simply counting the characters that are incompatible with it, not by trying to work out how the number of evolutionary changes which the tree demands can be minimised.

that will return the correct answer for all data inputs in polynomial time calculated from the size of the input. (See Garey and Johnson 1979 for the classic discussion of this phenomenon.) Even without examining why that is so in mathematical terms, it is not very surprising intuitively if one considers the foolproof way to find an answer to the problem without regard to how much time it will take, namely exhaustive search of all possible trees. The number of distinct rooted binary-branching trees is given by the expression

$$(2\mathrm{n} - 3)(2\mathrm{n} - 5) \cdots 5 \cdot 3 \cdot 1,$$

where *n* is the number of terminal nodes representing the actually observed taxa (biological species, languages, manuscripts, etc.), while the number of distinct unrooted binary-branching trees is given by the expression

$$(2\mathrm{n} - 5) \cdots 5 \cdot 3 \cdot 1;$$

both numbers thus increase exponentially as additional taxa are added (Embleton 1986: 28–29 with references). For instance, while there is only one unrooted binary-branching tree for three leaves, and only three distinct unrooted binary-branching trees for four leaves, there are 15 for five leaves, 105 for six leaves, 945 for seven leaves, 10,395 for eight leaves, and so on. If we want to find the true tree for only twelve taxa (e.g., for a family of only twelve languages) we must examine well over 650 million such trees.

However, a part of the PP problem has been solved. Agarwala and Fernández-Baca (1994) developed an algorithm for the PP problem in the special case in which the maximum number of states per character is bounded. Their algorithm, which runs in time $O(2^{3\mathrm{r}}(\mathrm{nk}^3 + \mathrm{k}^4))$, was subsequently improved by Kannan and Warnow (1997) to an $O(2^{2\mathrm{r}}\mathrm{nk}^2)$ algorithm, where *n* is the number of taxa, *k* is the number of characters, *r* is the maximum number of states per character, and *O* is a constant.

Our software implements the algorithm of Kannan and Warnow. The algorithm itself does not produce a tree if no PP can be found, but the software does return a tree under those circumstances by doing a greedy accumulation of compatible characters (i.e., by accessing the characters serially and building a tree as it proceeds). It thus returns a tree which is a PP for some subset of the character set if it cannot find one for the whole set. We will return to that point in the following section.

6. A STRATEGY FOR HANDLING RECALCITRANT DATA

The foregoing is the mathematical core of our methodology, but it is not the whole methodology – somewhat to our surprise. Once we were able to state the problem precisely we began to encounter a variety of unforeseen practical problems with the linguistic data. The most obvious has been alluded to above: though we can certainly exclude backmutation from our data, it is far from clear that we can identify all parallel development, so that characters exhibiting parallel development can be left out of the data when we first run the PP software and dealt with later by other, more appropriate means. To be sure, traditional philological work has been able to identify an impressively large number of parallel developments among the phonological and lexical characters, and we have used those results with gratitude; but it is logically impossible to be certain that our predecessors found *all* the relevant instances (or that we have not overlooked one or more by mistake). Moreover, though most words borrowed from foreign languages can be identified as such in a language's basic vocabulary, there is always the possibility that a few will fail to exhibit the usual diagnostics of loanwords by sheer chance, especially if they were borrowed from closely related languages. That has been known for a long time, though there has never been an effective way of dealing with the problem.

But if we cannot sequester all parallel developments and remove all loanwords, which can be incompatible with a PP, we face a serious difficulty. As we noted at the end of the last section, if our software cannot find a PP for the whole dataset it returns one for a subset of the characters. That tree may or may not be the best tree constructible for the dataset – that is, the one with which the largest subset of characters is compatible – and there is no easy way to determine by how much it misses being the best tree. In principle, the alternatives are complete success and a failure the gravity of which cannot readily be estimated. One must construct a strategy to circumvent this problem.

A successful strategy must begin with a maximally rigorous coding of the character data that employs all information available from reliable traditional work in the field; that will maximise one's chances of identifying and sequestering clear cases of parallel development. All words known to be borrowed from other languages must be coded with unique states (thus making them compatible with any tree).[12] When parallel developments have been removed and loanwords have been isolated to the greatest extent possible, one inputs the data to software which implements the PP algorithm. The possible outcomes are the following:

1) At least one PP is returned (i.e., complete success).
2) Many characters are incompatible with the (ostensibly best) tree returned, so that it is completely unclear what the actual best tree is (i.e., complete failure).
3) Only a few characters are incompatible with the tree returned.

If the third outcome occurs, there is an obvious way to find the best tree. One identifies the incompatible characters; the remainder will be compatible with a PP. One then runs the software on each possible subset of the characters which is at least as large as that residual set until the largest subset of characters compatible with a PP is found. The PP for that subset is the best discoverable tree.

Of course the strategy just described is computationally acceptable only if the number of incompatible characters is small and the total number of characters reasonably small, because the number of

[12] It seems appropriate at this point to emphasise what our strategy does *not* involve. Clearly we are not attempting to construct a method which will allow us to input raw data to an algorithm and derive a completely mechanical solution, in the belief that that is somehow more 'objective' than using the results of traditional philological work; such a strategy would be justified only if we believed that the traditional approach were on the wrong track, and we do not believe that (see section 3). Of course one result of the decision to make use of traditional philology is that the results of our method can be no better than the philological judgments on which they are based. In fact we think our methodology is actually best used to test complex sets of hypotheses for consistency and work out their consequences as rigorously as possible; it would be naïve to suppose that the output of any run of the software, even if a PP were returned, should settle an argument about subgrouping without any further discussion.

subsets on which the software must be run increases exponentially according to the expression

$$\binom{\mathrm{k}}{\mathrm{t}} + \binom{\mathrm{k}}{\mathrm{t}-1} + \cdots + \binom{\mathrm{k}}{0}$$

(where *k* is the total number of characters and *t* is the number of characters incompatible with the tree returned).[13] But this strategy will give the correct result for *all* datasets that meet the (rather restrictive) conditions stated.

The reader must bear in mind that the above discussion is couched in terms of *generally applicable* strategies. Particular datasets can exhibit configurations of character states that will make other strategies feasible in dealing with *them*; but those strategies cannot be generalised to all cases, nor even to all cases meeting certain mathematical requirements (like that discussed above). In effect, the structure of a particular dataset can give the investigator a ‘lucky break’. Interestingly, that point will be relevant in the following sections.

7. The Indo-European experiment

The remainder of this paper will report the results of a large experiment on the IE family of languages in the context of which the methodology described above evolved. A number of methodological problems in addition to those already noted were encountered as the experiment proceeded; each will be discussed in the relevant context below.

7.1. Selection and organisation of data

We chose the IE family not only because it is the best-researched family of human languages and the area of specialisation in which two of us were trained, but also because it poses an interesting problem in subgrouping. The well-attested languages of the family

[13] Readers unfamiliar with this notation should be aware that the parentheses containing upper and lower items *x* and *y* respectively are read ‘*x* choose *y*’, and that that expression means ‘the number of different ways there are of choosing *y* items from a total of *x*’.

subgroup	*language*	*dialect*	*earliest date reasonably well attested*
Anatolian	Hittite	–	ca. 1400 B.C.E.
Anatolian	Luvian	cuneiform	ca. 1400 B.C.E.
Anatolian	Lycian	–	ca. 400 B.C.E.
Indo-Iranian	Old Indic	Early Vedic	ca. 1000 B.C.E.
Indo-Iranian	Avestan	'younger'	ca. 500 B.C.E.?
Indo-Iranian	Old Persian	–	ca. 500 B.C.E.
Greek	(Greek)	Classical Attic	ca. 400 B.C.E.
Italic	Umbrian	–	ca. 200 B.C.E.
Italic	Oscan	–	ca. 100 B.C.E.
Italic	Latin	Classical	ca. 100 B.C.E.
Germanic	Gothic	–	ca. 350 C.E.
Germanic	Old High German	East Franconian	ca. 900 C.E.
Germanic	Old English	Late West Saxon	ca. 1000 C.E.
Germanic	Old Norse	Old Icelandic	ca. 1200 C.E.
Armenian	(Armenian)	Classical	ca. 500 C.E.
Celtic	Old Irish	–	ca. 800 C.E.
Celtic	Welsh	modern standard	
Tocharian	Tocharian A	–	ca. 800 C.E.
Tocharian	Tocharian B	–	ca. 800 C.E.
Balto-Slavic	Old Church Slavonic	–	ca. 1000 C.E.
Balto-Slavic	Old Prussian	–	ca. 1400 C.E.
Balto-Slavic	Lithuanian	modern standard	
Balto-Slavic	Latvian	modern standard	
Albanian	(Albanian)	modern standard	

Figure 3. Languages in our database.

fall into ten subfamilies which are clear and uncontroversial, but there has never been any consensus on how those ten robust subfamilies are related to one another cladistically. That problem – the first-order subgrouping of IE – is what we hoped to solve.

As work proceeded we steadily enlarged and corrected our database; the latest revision was completed after all previous publications and presentations.[14] We now have data from twenty-four languages representing all ten robust subgroups at the earliest respective dates from which we have a respectably large amount of linguistic documentation. They are listed in Figure 3.

Currently our database includes 370 characters, namely 22 phonological characters, 15 morphological characters, and 333 lexical characters. (See the Appendix for a summary presentation

[14] The principal author ran the software on the latest revision of the database; thus all errors in the results reported here are his responsibility.

of the phonological and morphological characters and a sample of the more interesting lexical characters.) We would have been more than glad to include a much larger number of morphological characters, since they are among the best indicators of linguistic descent (see above), but the structure of the data has defeated us in that respect. In the IE family most distinctive morphological innovations are characteristic of only one of the traditionally recognised subgroups, and most shared characteristics are inherited. We believe that we have employed virtually all the useful morphological characters; in fact, colleagues have persuaded us to omit several that we originally attempted to use, and one of those that remains is suspect (see further below).

On the other hand, many of the lexical characters (and at least one of the morphological characters) can legitimately be coded in more than one way, typically because of cognations between parts of lexemes. The reasons for this procedure are perhaps best explained by illustration. Consider again the character 'hand'. The actual comparative data are given in Figure 4.

The most straightforward coding according to cognation classes is given in Figure 5 (cf. the tree in Figure 1); the ancestral form for each cognation class follows.

But there is uncontroversially some relation between the PIE protoform for state 1 and the Indo-Iranian protoform for state 2; it appears that Indo-Iranian did inherit the PIE word but replaced its second syllable by a process the details of which can no longer be recovered. It follows that states 1 and 2 should together occupy a

Hittite	kissar	Luvian	īssaris
Armenian	jeṙn	Lycian	izre
Greek	k^{h}ḗr	Tocharian A	tsar
Albanian	dorë	Old Persian	dasta
Tocharian B	ṣar	Old Prussian	rānkan [acc. sg.]
Vedic	hástas	Latvian	ròka
Avestan	zastō	Gothic	handus
OCS	rǫka	Old Norse	hǫnd
Lithuanian	rankà	OHG	hant
Old English	hand	Welsh	llaw
Old Irish	lám	Oscan	*manim* [acc. sg.]
Latin	manus	Umbrian	manf [acc. pl.]

Figure 4. Data for the lexical character 'hand'.

Hitt.	1	Av.	2	Luv.	1	Goth.	4
Arm.	1	OCS	3	Lyc.	1	ON	4
Gk.	1	Lith.	3	TA	1	OHG	4
Alb.	1	OE	4	OPer.	2	Welsh	5
TB	1	OI	5	OPru.	3	Osc.	6
Ved.	2	Lat.	6	Latv.	3	Umb.	6

Cognation classes:

1 PIE *ǵ^{h}ésr̥
2 Proto-Indo-Iranian *ź^{h}ástas
3 Proto-Balto-Slavic *rankā
4 Proto-Germanic *handuz
5 Proto-Celtic *lāmā
6 Proto-Italic *man-

Figure 5. Coding of the character ‘hand’.

connected subgraph of the tree. Therefore it is also reasonable to adopt an alternative coding which combines those two states, as in Figure 6.

So as not to lose any of the subgrouping information present in this character, we employ both codings, thus duplicating the character. This causes no problems until one attempts to calculate how many characters are compatible with a particular tree, or how many force a particular subgroup; then one must be careful not to treat the duplicates as though they were independent.

A further peculiarity of many of the lexical characters (and one morphological character) does cause serious difficulties: they are POLYMORPHIC, in that at least one language must be assigned more than one state. The existence of polymorphic characters is hardly surprising. Even in a language’s basic vocabulary there are real synonyms which can be used interchangeably. To cite an obvious modern English example, there is effectively no difference between the adjectives *little* and *small* in their most basic meaning; though

Hitt.	1	Av.	1	Luv.	1	Goth.	4
Arm.	1	OCS	3	Lyc.	1	ON	4
Gk.	1	Lith.	3	TA	1	OHG	4
Alb.	1	OE	4	OPer.	1	Welsh	5
TB	1	OI	5	OPru.	3	Osc.	6
Ved.	1	Lat.	6	Latv.	3	Umb.	6

Cognation classes:

1 PIE *ǵ^{h}ésr̥
3 Proto-Balto-Slavic *rankā
4 Proto-Germanic *handuz
5 Proto-Celtic *lāmā
6 Proto-Italic *man-

Figure 6. Alternative coding of the character ‘hand’.

there are certainly circumstances in which one would use *little* rather than *small* or vice versa, for the basic meaning 'not large' either one will do and neither one is clearly preferred. So far as we can tell, that has been true for more than a millennium – Old English (OE) exhibits the same polymorphism – so there is nothing unnatural or unstable about such a situation. But consider the consequences of that fact for our coding of the character 'small'. If only one of the Old English words exhibited cognates in the same meaning in any other language, then the one that had no cognates would have to be assigned a unique state. But unique states are compatible with any tree; therefore that unique state could simply be omitted, and the character would not be *effectively* polymorphic. A significant proportion of the polymorphic characters in our data can be resolved into tractable monomorphic characters in precisely that way. Many others, though, including this one, cannot be so resolved. Unfortunately for us, both OE words have cognates in the same meaning in Old High German (*lȳtel* is cognate with *luzzil*, and *smæl* is cognate with *smal*); so both those states of this character must be assigned to both languages, and the character is irresolvably polymorphic.

Most unfortunately, polymorphic characters exhibit different mathematical properties, so that a completely different tree-construction algorithm is needed to handle them. Such an algorithm does exist (see Bonet, Phillips, Warnow and Yooseph 1996), but it has never been implemented, so there is no available software. Under the circumstances we have a range of options for handling polymorphic characters.

1) In some cases we can recode polymorphic characters as pairs of monomorphic characters in such a way as to preserve all the subgrouping information present. This must be done with extreme caution, so as not to introduce into the coding any additional assumptions;[15] in effect, it can be used only if the polymorphism is confined to a subgroup the evolutionary structure of which is completely uncontroversial. By this means

[15] Namely, assumptions about the configuration of the tree, which is the object of the investigation! Our methodological assumptions, which have been discussed above, are of course of a different order.

we have been able to eliminate the polymorphism of only 16 characters (two of which still exhibit obvious parallel development).

2) If the polymorphism is confined to a known subfamily (the internal structure of which is not necessarily uncontroversial), we could suppress one of the states contributing to the polymorphism, since we are interested chiefly in recovering the first-order subgrouping of the family. We have not adopted this strategy, since it fails to use some information that is clearly present in the data.
3) We can set the irresolvably polymorphic characters aside for the purposes of running our software, find the best tree on the monomorphic characters, and then fit the polymorphic characters to that tree, observing reasonable linguistic constraints on the polymorphism that must be posited for internal nodes (so that a character which forces us to posit too much polymorphism on those nodes will be judged incompatible with the tree). We have adopted this tactic, even though we are not satisfied with it; until software to handle polymorphic characters can be developed, this is the best that we can do.

Because we have employed alternative codings (see above) and have resolved some polymorphic characters into pairs of monomorphic characters, our actual total of working characters is 467.

7.2. Computational analysis and output

Before running our software we set aside all irresolvably polymorphic characters and attempted to exclude all the characters known to exhibit parallel development. (Though we removed 55 characters on the latter grounds, it appears that we may have missed a couple of cases of independent innovation; see section 7.5 below for further discussion.) The residue, on which the software can be run with a reasonable hope of finding a PP, amounts to 322 characters – still a respectably large number.

Our software does not return a PP for this dataset. In fact the result is not even close to a PP; some 18 characters are incompatible with the 'best' tree returned.[16] Since the number of sets of 304 or more characters that can be chosen from a total of 322 is 141,638,934,175,332,712,152,972,803,380 (or about 1.4 × 10^{29}),[17] exhaustive search of all such sets is clearly infeasible. Thus we have no reliable means of discovering what the genuinely best tree for this dataset is; in computational terms our result is a total failure.

However, we must bear in mind that this particular dataset may exhibit properties that provide an opportunity for further progress in this particular case (even though we may not be able to generalise our findings to other cases of interest); thus it is worthwhile to examine the ostensibly best tree and the characters that are incompatible with it. The tree is given in Figure 7.[18]

In this and the following sections we will refer to the large subgroup that includes Graeco-Armenian, Balto-Slavic and Indo-Iranian as the 'core' subgroup, so as to simplify our statements of the distribution of states across the tree.

The incompatible characters, all of which are lexical, are the following; they are grouped according to the pattern of incompatibility which their states exhibit.

1) Germanic and Italic share a state in:
   one alternative coding of 'neck' (against a state shared by Iranian and Celtic);
   'straight' (against a state shared by Indo-Iranian and Celtic);
   one coding of 'suck' (against a state shared by Celtic and much of the core);
   one coding of 'break' (against a state shared by Celtic, Armenian and Vedic);
   'pour' (against a state shared by Tocharian and Greek).

[16] Apparently because so many characters are incompatible, running the software on this dataset on a dedicated state-of-the-art machine takes about eight days.

[17] We are grateful to Beth Randall for writing a computer program to calculate these sums.

[18] Our software actually returns unrooted trees; we have rooted this one on the edge connecting Anatolian with the rest of the family for reasons that will be discussed below. It should be emphasised that in this and all other trees in this paper *only the branching structure is significant*; distances of any kind are a completely meaningless byproduct of the need to fit the tree onto the page, and whether or not a branch is straight likewise has no meaning at all. (Thus the trees in Figures 1 and 7, for instance, are identical.)

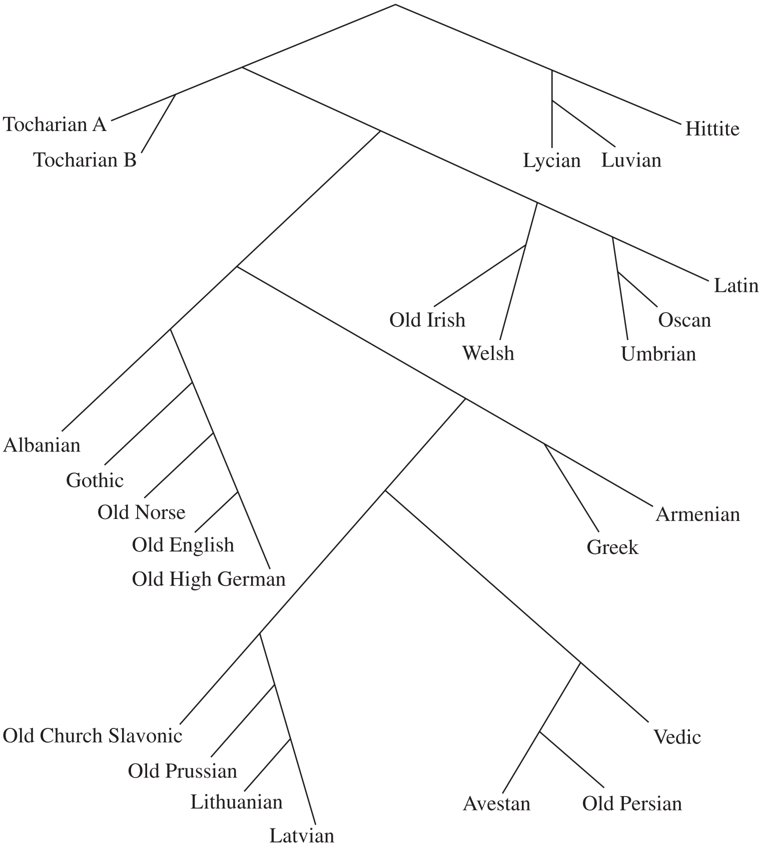

Figure 7. The apparent best tree for the entire Indo-European dataset.

2) Germanic and Celtic share a state in:
one coding of ‘all’ (against a state shared by Tocharian and Greek);
one coding of ‘breast’ (against a state shared by Tocharian and Indo-Iranian);
‘free’ (against a state shared by Italic and Greek);
one coding of ‘young’ (against a state shared by Italic and most of the core).

3) Germanic and Baltic share a state in:
   one coding of 'float' (against a state shared by Slavic and many other languages);
   'arm' (against a state shared by Tocharian and Indo-Iranian).
4) Germanic and Slavic share a state in:
   one coding of 'pig' (against a state shared by Tocharian, Greek and Indo-Iranian).
5) Germanic and Balto-Slavic share a state in:
   one coding of 'thousand' (against a state shared by Greek and Indo-Iranian).
6) Germanic, Latin and Balto-Slavic share a state in:
   'beard' (against a state shared by Anatolian, Albanian, Armenian and Vedic).
7) Germanic, Italo-Celtic and Old Prussian share a state in:
   'one' (against a state shared by Tocharian, Graeco-Armenian and perhaps Albanian).
8) Germanic, Italo-Celtic and Graeco-Armenian share a state in:
   'tears' (against a state shared by Tocharian, Indo-Iranian and Baltic).
9) Old Norse agrees with Latin against the rest of Germanic in 'head'.
10) Slavic and East Baltic, but not Old Prussian, share an irregular phonological innovation in 'nine'.

The last two incompatibilities are internal to uncontroversial subgroups and have no bearing on the first-order subgrouping of the family. Astonishingly, Germanic is implicated in all sixteen of the others; moreover, in only one case – that listed under (8) – does Germanic share a state with a subgroup not found in northern or western Europe in the last millennium B.C.E. The obvious inference is that there is not necessarily anything 'wrong' with these characters, but there might be something very peculiar about Germanic.

The anomalous behaviour of Germanic first became evident at an early stage of our investigation, when we were running a much smaller set of characters and only twelve languages. Because Germanic shared states with various subgroups not adjacent in any of the 'best' trees returned, its position in the tree shifted from run to run of the software; it was variously grouped together with Balto-Slavic and Indo-Iranian, or with Greek and Armenian, or with Italic and Celtic, at widely different positions within the tree. Since then we have greatly expanded and corrected our dataset; we have reconsidered the coding of numerous characters, recognising previously undetected parallel developments, adopting new alternative codings, and devising the system of 'split coding' to recode some polymorphic characters as pairs of monomorphic characters (see above). While the position of Germanic in the tree seems in consequence to have stabilised, the pattern of states shared with disparate subgroups is still the same.

If the pattern in which Germanic shares states with other subgroups is in some sense anomalous, and if the anomaly is responsible for the vast majority of incompatibilities in our dataset, an experiment suggests itself: what happens if we remove the Germanic languages from the dataset and run our software on the remaining languages?

If Germanic is removed, we are able to find a tree with which only four characters are incompatible.[19] Such a tree is given in Figure 8.

The four characters incompatible with this tree, all of which are lexical, are the following:

the distribution of states for 'beard' groups Latin and Balto-Slavic against Anatolian, Albanian, Armenian and Vedic;

'one' groups Italo-Celtic and Old Prussian against Tocharian, Graeco-Armenian and perhaps Albanian;

'tears' groups Tocharian, Indo-Iranian and Baltic against Italo-Celtic and Graeco-Armenian;

'nine' groups Slavic and East Baltic against Old Prussian and virtually all the non-Balto-Slavic languages.

Note that the first two are characters in which Germanic contributes to the incompatibility of the entire set; thus whatever the correct explanation for the peculiar behaviour of Germanic turns out to be, it will probably account for the anomalies in 'beard' and 'one'. Of the remaining two incompatible characters, the last represents a problem internal to Balto-Slavic (as we remarked above).

[19] The running time for this dataset is less than 24 hours.

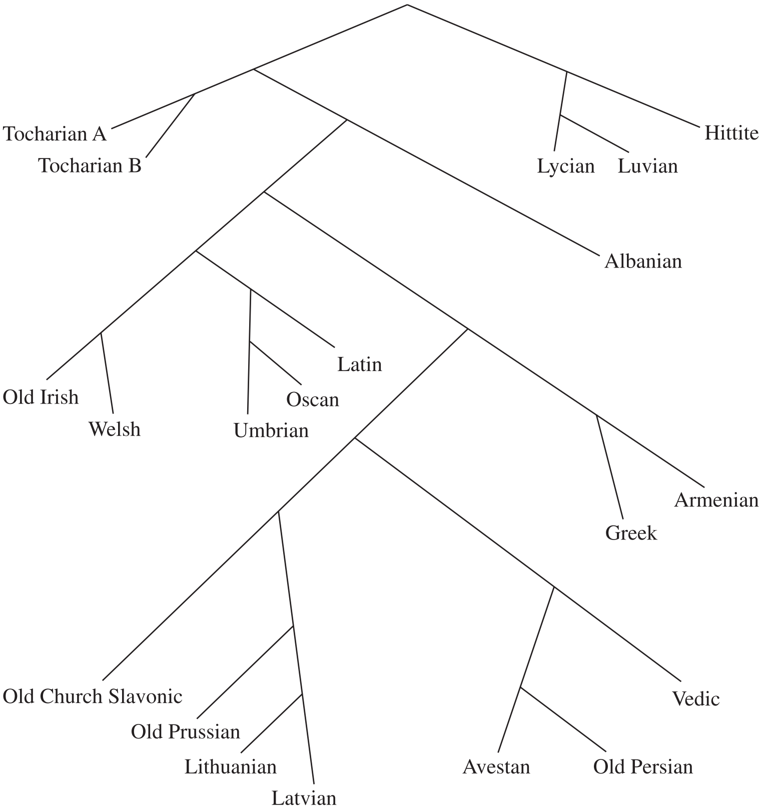

Figure 8. One of the best trees with Germanic omitted.

Moreover, those two characters are very similar linguistically: the initial consonant of 'nine' has been replaced by that of 'ten' in Slavic and East Baltic, while the initial consonant of 'tears' has been dropped in Tocharian, Indo-Iranian and Balto-Slavic. Though we have cautiously treated those changes as shared innovations (which therefore create incompatibilities with the tree), it is possible that they are in part contact phenomena, a lexical irregularity having been borrowed from one dialect or language into another that was at the time still reasonably closely related. Thus it seems possible, at least, that we have plausible explanations for all these incompatibilities.

But is this the best tree without Germanic? The number of sets of 318 characters that can be chosen from 322 is 445,197,684 – still too large to render feasible the generally applicable strategy for finding the best tree (since an exhaustive search of more than 400 million sets simply takes too long). However, the distribution of states in the incompatible characters shows clearly that no better tree can be constructed by accepting any of those characters and rejecting some others. In the first place, note that each of the four incompatible characters shows a different distribution of states, only the first two being partly similar. Thus it cannot be the case that adjusting the tree to fit all four of these characters, or any three, will give a tree which is a plausible candidate for the best tree. Moreover, it turns out that adjusting the tree to fit any one of these characters, or the first two, amounts to constructing a tree which is incompatible with a larger number of characters and/or at least one phonological or morphological character (which are better indicators of linguistic descent). For instance, in a tree that is compatible with the character 'nine' Slavic and the East Baltic languages must constitute a subgroup within Balto-Slavic, the Baltic language Old Prussian being the 'outlier' which is less closely related to the other members of the subgroup. But such a tree is incompatible with our morphological character M14 (reflecting a startling syncretism of 3SG, 3DU and 3PL verb endings shared by all the Baltic languages, but not by Old Church Slavonic); it is also incompatible with the lexical characters 'hear', 'sea', 'daughter-in-law' and 'wood', in all of which OCS preserves an inherited word which the Baltic languages have replaced with a new lexeme; and it is incompatible with one alternative coding of 'long', reflecting the fact that OCS preserves an inherited initial **d*- (found also in Indo-Iranian) which the Baltic languages have lost. Further, the tree accommodating 'nine' is incompatible with 'bee' and with one alternative coding each of 'green', 'sun' and 'yellow' (as we have coded them), reflecting the fact that the Baltic languages share a distinctive derivative of an inherited root while OCS does not; though each of these last characters could be impugned on the grounds that OCS exhibits a unique derivative (for which we did not code), which could conceivably have replaced the one still shared by the Baltic languages, it is unlikely that that is the correct explanation for all four. In sum, it is clear that a tree which accommodates 'nine' will exhibit greater non-convexity than a tree which with which 'nine' is incompatible.[20] Similar arguments demonstrate that adjusting the tree to fit any of the other incompatible characters gives a tree substantially worse than the one returned.

It is also true that all the polymorphic characters and characters exhibiting known parallel development can be fitted to this tree observing linguistically reasonable constraints (though that is not particularly impressive, given that long-term polymorphism and parallel development are common and natural among lexical characters).

We are therefore confident that we have the best tree obtainable for our dataset – once Germanic has been removed. More exactly, we have one of the best trees, for the following reason.

[20] One polymorphic morphological character appears at first glance to support the tree suggested by 'nine': for M7, reflecting the genitive singular ending of nominal o-stems, Old Prussian exhibits a reflex of PIE *-*osyo* (the PIE pronominal ending), as do numerous non-Balto-Slavic languages, while OCS and the East Baltic languages exhibit reflexes of *-*ā* < PIE *-*e-ad* (the ablative singular ending). However, there are at least two historical scenarios consistent with our best tree that can account for that distribution of states. Though genitive and ablative have undergone syntactic merger throughout Balto-Slavic, it is not certain that the merger had already occurred by the Proto-Balto-Slavic stage; conceivably it was a parallel development within the independent histories of the languages, and in that case different daughters can have generalised the same ending independently. (After all, precisely the same syntactic merger occurred independently in Greek as well.) Alternatively, it is possible that the merger had occurred in Proto-Balto-Slavic, that the genitive ending was generalised in the pronouns and the ablative ending in nouns, and that different daughters levelled in favor of one ending or the other independently. This particular character is interesting because of what it implies about polymorphic characters in general. Though the overt polymorphism is confined to Latin, parallel development and a number of other problems are strongly suspected in various other languages of the family; notoriously, no ending for this category is securely reconstructable for PIE. We have found this situation to be fairly typical of polymorphic characters: the overt polymorphism is, so to speak, just the tip of the iceberg. We suspect that polymorphic characters will never be very useful in reconstructing linguistic evolutionary trees, simply because they reflect an irreducibly 'messy' aspect of linguistic evolution.

7.3. Indeterminacies in the tree

Like other computational algorithms that construct phylogenetic trees, the PP algorithm which our software implements constructs a tree in which all branchings are binary and assigns a definite position as a leaf-node to every input language. In some cases those decisions are arbitrary, since the data do not fully determine a particular branching or position. The indeteminacies must be discovered by inspection of the distribution of states on the tree returned at every pair of internal nodes that are joined by an edge. The best tree returned by our software for this dataset includes two indeterminacies, only one of which can be resolved by further analysis.

In the trees in Figures 1, 2, 7 and 8 Luvian and Lycian form a subgroup within Anatolian, while Hittite is the outlier within that subgroup. However, our software typically returns a tree in which Hittite and Luvian are grouped together, with Lycian as the outlier. This surprised us at first, since traditional work overwhelmingly finds that Luvian is very close to the ancestor of Lycian. We therefore examined the distribution of states among the Anatolian languages for every character and discovered that the grouping of Hittite and Luvian is an artefact of the algorithm's obligatory binary branching: in fact there are no monomorphic characters that group any two of the Anatolian languages against the third. The real subgrouping of Anatolian returned by the algorithm on monomorphic characters is the unresolved ternary branching given in Figure 9.

In this case two characters which cannot be used when running the software – one because it is polymorphic, the other because it exhibits obvious parallel development – force the resolution that we have presented in the other trees in this paper. For 'year', which is polymorphic in Gothic, both Luvian and Lycian exhibit reflexes of a unique derivative **utsis*, which has replaced the inherited term **wet-*; reflexes of the latter appear not only in Hittite but also in Vedic, Greek and Albanian, thus forcing Hittite into peripheral position within the Anatolian subgroup. For 'father' both Luvian and Lycian exhibit reflexes of a 'nursery' term **dáda*, whereas Hittite, Albanian, Gothic and Old Church Slavonic exhibit reflexes of an older nursery term

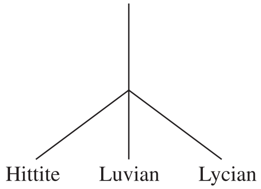

Figure 9. Subgrouping of Anatolian by the monomorphic characters alone.

**átta* (still attested in the meaning 'dad' in Greek and Latin). The latter character could not in itself be probative; after all, the older nursery term must have replaced the inherited word **ph*$_2$*tḗr* independently several times in the history of the family, so it is at least conceivable that Luvian and Lycian carried out a similar replacement independently. But the distribution of states does support that of 'year', which does not seem to be impugnable. In sum, the indeterminate subgrouping of Anatolian can be resolved, though the evidence for its resolution is very slender.

The other indeterminacy in our tree is much more serious. Readers will have noted that the position of Albanian is not the same in Figures 7 and 8: in Figure 7 it would constitute a subgroup with the core languages against Italo-Celtic even if Germanic were removed, whereas in Figure 8 its position in the tree is higher than that of Italo-Celtic. We have found that to be typical: different runs of the software assign different positions to Albanian.[21]

[21] The reasons for this will become clear from the following discussion, but note that there is also a further factor which tends to give the opposite result. In analysing a dataset as large as ours, our software is constrained to operate in a 'greedy' fashion, accessing the characters one by one and building the tree as it proceeds. Though it is possible to 'fix' individual characters, forcing the software to make any tree returned compatible with them (and we have in fact fixed all the phonological and many of the morphological characters), it is obvious that the order in which the characters are accessed might occasionally influence the output. That problem can largely be obviated by randomising the order of characters repeatedly and running the software with a variety of different orders, and we have done so. But the number of characters that offer information about the position of Albanian is so small that they can reappear in the same relative order in multiple randomisations. We have therefore been obliged to reckon with the possibility that to the extent that different runs assign the *same* position to Albanian, that is purely an artefact of how the software accesses the data. No other language in our database poses a problem of this kind.

Examination of the distributions of states in which Albanian is implicated reveals the following picture. The phonological characters and a few of the morphological characters place Albanian outside all the generally recognised subgroups, the 'satem' group (which includes Indo-Iranian and Balto-Slavic) and Italo-Celtic. Since Albanian exhibits unique states for all the morphological characters which define large subgroups in our tree, all the remaining evidence is lexical. A considerable number of lexical characters confirm its exclusion from all the generally recognised subgroups (since most of the inherited words which Albanian still preserves are widely attested in the family rather than being characteristic of some particular subgroup). In addition, the lexical characters 'not' and one alternative coding of 'day [= 24 hours]' exclude Albanian from our tentative Graeco-Armenian subgroup (on which see section 7.5). 'Drink' groups Albanian with a large majority of the languages against Anatolian and Tocharian; thus Albanian must belong to the residual group that excludes those two divergent subgroups, and the highest position it could occupy in the tree is the position it does occupy in Figure 8. One alternative coding of 'leave' groups Albanian with Germanic against Latin, Graeco-Armenian and the satem group; that character is responsible for the position of Albanian in Figure 7. 'Worm' groups Albanian with the satem languages and Celtic against Latin and Germanic (!); see section 7.6 on possible explanations for this distribution. The remaining half-dozen characters that offer any putative information at all about the subgrouping of Albanian exhibit rampant polymorphism and/or parallel development, so that shared states cannot be taken as evidence of shared ancestry. The most we can say is that Albanian cannot occupy a position higher in the tree than in Figure 8 and cannot be a member of the Italo-Celtic, the Graeco-Armenian or the satem subgroup.

7.4. Rooting the tree

Our PP algorithm returns unrooted trees: the branching structure is fully defined, but there is no indication of where the root node, representing the protolanguage, is located within the tree. This is because the algorithm exploits only information about the distribution of states; information regarding the direction of replacement of one state by another within a character (if there is any) is not accessed. It is therefore necessary to root the tree by some other means.

A mechanical means of rooting the tree which will give reliable results under any circumstances is the following.

1) Create a new taxon representing the protolanguage.
2) For each character for which the ancestral state is known, assign that state to the protolanguage.
3) For all other characters, assign a unique state to the protolanguage.

The algorithm will then return a tree with the protolanguage as one of the leaf nodes. The actual root node will of course be an internal node, namely the node at which the leaf node representing the protolanguage is joined to the rest of the tree.

If the pattern of ancestral states were complex enough, the approach just described would be indispensable; but in fact the pattern of ancestral states in our database is simple and straightforward, so that it is easy to work out by hand where the root node must fall. We do not venture to suggest an ancestral state for any of the lexical characters *as input for determining the rooting of the tree* (though of course once the tree has been rooted on the basis of the phonological and morphological characters it will automatically give the ancestral states for various lexical characters, as in Figure 1). The derived states of the phonological characters and of morphological characters M11 through M15 exclude the root node from all the generally recognised subgroups, as well as from Italo-Celtic and the satem group. (See the Appendix for our coding of the phonological and morphological characters). The only other morphological characters for which we believe we know the ancestral state are M5 and M3. The ancestral state of M5 defines that end of the unrooted tree in which Anatolian, Tocharian and Italo-Celtic are located; thus the root node must fall somewhere within that portion of the tree, or between that part of the tree and the remainder. Examination of the tree in Figure 8 shows that one of the following must then be true:

1) Anatolian is one first-order subgroup of the family, and all the other languages together constitute the other first-order subgroup; or
2) Anatolian and Tocharian together constitute one first-order subgroup of the family, and all the other languages together constitute the other first-order subgroup; or
3) Anatolian, Tocharian and Italo-Celtic together constitute one first-order subgroup of the family (with Anatolian and Tocharian more closely related within it), and all the other languages together constitute the other first-order subgroup; or
4) Tocharian is one first-order subgroup of the family, and all the other languages together constitute the other first-order subgroup; within this residual group, Anatolian is one subgroup and all the other languages together constitute the other; or
5) Italo-Celtic is one first-order subgroup of the family, and all the other languages together constitute the other first-order subgroup; within this residual group, Anatolian and Tocharian together constitute one subgroup, and all the other languages together constitute the other.

Character M3 decides in favor of the first alternative. As Cardona (1960) demonstrates, the thematic aorist is an innovative inflectional category, almost certainly not present in PIE; yet at least one such aorist, $*h_1lud^h\acute{e}t$ ‘(s)he arrived’, is attested in Tocharian and Celtic. Anatolian exhibits no thematic aorists, and does not certainly exhibit simple thematic verb stems of any kind; we therefore suggest that Anatolian preserves the ancestral state of this character. It follows that the root node must fall between Anatolian and the rest of the tree. Admittedly this coding of M3 is open to dispute, since it rests on an inference regarding the prehistory of Anatolian with which we expect some qualified colleagues to disagree. On the other hand, the first of the five alternatives outlined above is by far the most probable in any case, since it is difficult at best to find any innovations shared by Anatolian and any other generally recognised subgroup. In any case, our rooting of the tree is supported in detail by exactly one character.

In sum, our rooting of the evolutionary tree of IE is highly probable but weakly supported by the character data (though data of other kinds, such as the pattern of innovations in the verb system argued for in Ringe 2000, may offer further support for it). We do not expect our colleagues to regard this as 'the last word' on an issue which remains a matter of disagreement among specialists.

7.5. The robustness of the tree

Though the best trees which our software returns (both with and without Germanic) are for the most part stable, the mere existence of the indeterminacies noted in section 7.3 above shows that we do need to ask how robust each subgrouping in the tree is – that is, how many characters actually force each branching node. This is not only an important question, but also one to which the 'raw' data are likely to give a misleading answer. Recall that some 322 characters remain to be input to the software after most polymorphic characters and all characters believed to exhibit parallel development have been removed. In only two of these characters ('three' and 'who') all the languages share the same state, and there is only one ('play') for which each language exhibits a unique state – fortunately for our investigation, since characters with those distributions of states obviously offer no information about the subgrouping of the family. But another 36 well-behaved monomorphic characters have only one state that is shared by more than one language, all the other states being unique. Those characters too are compatible with any tree, since each unique state can be confined to a leaf node and the sole shared state assigned to all the internal nodes; in the technical terminology of computational cladistics, they are UNINFORMATIVE. In other words, though more than 300 characters are compatible with our best tree, nearly forty of them would be compatible with any tree.

But the worst news is yet to come: the vast majority of our well-behaved monomorphic characters simply define one or more of the ten uncontroversial subgroups of the family, contributing nothing to their higher-order subgrouping. It is clear that the higher-order subgrouping of the IE family has remained an unsolved problem for so many generations partly because the evidence is genuinely meagre. In the following paragraphs we will describe in detail the evidence for specific internal nodes in our best tree; it will be seen that the evidence in particular cases ranges from modest to severely limited. What that means will be considered further at the end of this section.

We have already seen that the status of Anatolian as one of the first-order subgroups of the family depends on the single character with which we have rooted the tree. If that rooting is correct, the position of Tocharian will be fixed by characters in which it shares a state with Anatolian against Italo-Celtic and other languages. There are four such characters, all lexical:

'die', in which the Luvian group and Tocharian A share a state (*wel-) against Latin, Welsh, Armenian and the satem group (*mer-, meaning 'disappear' in Hittite);[22]

'drink', in which Anatolian and Tocharian share a state (*$\acute{\bar{e}}h_2g^{wh}ti$; see Kim 2000) against all the other major subgroups except Armenian and Germanic (*peh_3- ~ *$p\bar{\imath}$-, orig. pres. *$p\acute{\imath}beti$, meaning 'swallow' in Anatolian);

one alternative coding of 'give', in which Anatolian and Tocharian share a state (*ay-) against all the other major subgroups except Albanian and Germanic (*deh_3-, orig. pres. *$d\acute{e}deh_3ti$, meaning 'take' in Anatolian; orig. meaning *'trade'?);

'many', in which Hittite and Tocharian share a state (*$me\acute{g}h_2$-, meaning 'big' in most subgroups) against Old Irish, Germanic, Greek and Iranian (*$p\acute{e}lh_1u$- ~ *$p\underset{\circ}{l}h_1\acute{e}w$-).[23]

It is at first disappointing that the position of Tocharian is fixed only by lexical characters, which in general provide the least secure evidence for subgrouping. Moreover, the last of these characters is doubtful evidence for the position of Tocharian: the word shared by Tocharian and Anatolian probably meant 'big' in the protolanguage

[22] The other Anatolian, Tocharian and Italo-Celtic languages exhibit unique states. This character is polymorphic, but the polymorphism is confined to Germanic. We have employed alternative codings of 'die', so that technically there are two characters that support the position of Tocharian by this configuration of states, but of course that is precisely the sort of duplication that must be eliminated in determining the robustness of the tree (see above). The reconstructable meanings of the roots strongly suggest that the word meaning 'die' in the 'nuclear' branches of the family was originally a euphemism.

[23] This character is polymorphic, but the state that gives rise to the polymorphism is shared only by Old Church Slavonic and several Germanic languages; it does not impinge on the question at issue here.

– note that it is also the basis of a derivative meaning ‘older’ in Tocharian B – and a shift to ‘much’ (plural ‘many’) could have occurred independently in the two subgroups. On the other hand, the remaining characters are all basic verbs, which we expect to be better than usual lexical indicators of genetic descent. On balance, we can say that the position of Tocharian in our tree is modestly supported by the evidence.[24]

Our Italo-Celtic subgroup is forced by four characters in which Italic and Celtic share a state against either Tocharian or Anatolian and at least one other subgroup (so that the last common ancestor of Italic and Celtic must be off the line of descent linking those subgroups with the core subgroups of the family). The characters are of all types:

phonological character P1, encoding the change of the sequence **p . . . k^w* to **k^w . . . k^w* – a regular sound change shared by Italic and Celtic, but not by Tocharian, nor by any of the core subgroups;[25] the PIE root-constraint prohibiting two oral stops at the same place of articulation within a root guarantees that the Italo-Celtic state is innovative;[26]

morphological character M11, encoding the partial replacement of the abstract noun suffix **-ti-* (attested in all the subgroups except Albanian) by the Italo-Celtic suffix complex **-ti-Hen-* (but see further below);

[24] For one other basic verb, ‘make’, Tocharian A shares a state with Anatolian (**$h_1$yeh$_1$-*, meaning ‘throw’ in Greek and Latin); but no two of the other subgroups share a state, so that character does not force the position of Tocharian. Much more often we find that Tocharian shares states of lexical characters with Italo-Celtic and the other subgroups against Anatolian; characters exhibiting that distribution of states include at least ‘breast’, ‘eye’, ‘four’, ‘moon’, ‘mother’, ‘tongue’, ‘brother’, ‘carry’, ‘pour’ and ‘sister’. In those cases we cannot determine whether it is Anatolian or the non-Anatolian subgroup that has innovated, with the result that the ancestral state is unrecoverable.

[25] None of the relevant words is attested in Anatolian (Craig Melchert, p. c.). The merger of labials and labiovelars in the Osco-Umbrian subgroup of Italic also eliminates evidence from those languages. Welsh does provide evidence, however, since inherited **p* had been lost in Celtic long before the British Celtic shift of **k^w* to **p*.

[26] The few exceptions to this generalisation involve apical stops (as would be expected on typological grounds); the most secure example is **tewd-* ‘push, knock’ (which might be onomatopoeic).

'lake', in which Latin and Old Irish share a state (**lóku* ~ **l̥kéw*-) against Tocharian and Greek (**léymon*- ~ **limn*-´, loc. **limén*);

'sing', in which Italic and Celtic share a state (**kan*-) against Tocharian and Old Church Slavonic (**peyH*-).

The quality of this evidence is rather uneven, the morphological character being especially vulnerable. The problem is not merely that it reflects a point of derivational morphology, which is not nearly so reliable an indicator of descent as inflectional morphology. A further difficulty is that Armenian exhibits a derivational suffix -*owt^{h}iwn* of very similar function which likewise includes the inherited suffix *-*ti*- and a further n-stem formative. It is not likely that Armenian and Italo-Celtic actually share an innovation in this instance; note that the Armenian suffix includes yet a third piece of morphological material before *-*ti*-, suggesting strongly that its history was more complex.[27] But the fact that the Armenian suffix is even partly comparable to the Italo-Celtic one raises the possibility of independent parallel development in Italic and Celtic, substantially weakening the evidence of the character. On the other hand, it does not appear that the phonological and lexical characters can be impugned. Note in particular that though the root **kan*- appears in nominal derivatives in a wide range of IE languages, it is only in Italic and Celtic that it appears as a verb (Rix et al. 1999 s.v.); and the fact that the nominals in question usually refer to animals with distinctive voices suggests that **kan*- originally denoted something other than human singing – in which case its shift to that meaning can be a shared Italo-Celtic innovation. In sum, the evidence for Italo-Celtic is quite slender but fairly solid.[28]

The position of Italic and Celtic in the tree, regardless of whether they constitute a subgroup, is also reasonably firm. On the one hand, the same lexical characters which force Tocharian up the tree (see above) force Italic and Celtic down. On the other hand, morphological character M5 excludes Italo-Celtic from the core,

[27] We are not fully convinced by Godel's argument that *ardiwnk^h* 'agricultural products, deed, demonstration' exhibits the same suffix complex *-*ti*-*Hen*- without the preceding material (see Olsen 1999: 490 with references). We are grateful to James Clackson for calling this datum to our attention.

[28] Other researchers have come to the same conclusion from very different assessments of the evidence; see e.g. Cowgill (1970) and Jasanoff (1997).

forcing it up toward Tocharian; the evidence of this character is qualitatively excellent, since it is a 'clean' and well-understood inflectional character that should reflect genetic descent straightforwardly.

The evidence for a Graeco-Armenian subgroup is significantly poorer than for Italo-Celtic. Half a dozen lexical characters ostensibly support such a subgroup:

'day [= 24 hours]', in which Greek and Armenian share a state (_*ā́mr̥_) crucially against Latin and Vedic (_*dyḗws_);[29]
'husband', in which Greek and Armenian share a state (_*h$_2$nḗr_) crucially against Tocharian and Indo-Iranian (_*pótis_);[30]
'not', in which Greek and Armenian share a state (_*h$_2$óyu_) against nearly all the other languages (which exhibit _*né_ except for Tocharian, which exhibits _*ma_);
'wind', in which Greek and Armenian share a state (_*h$_2$ónh$_1$mos_) against virtually all the other languages (which exhibit derivatives of _*h$_2$weh$_1$-_ 'blow', of which the most ancient is the participle _*h$_2$wéh$_1$n̥ts_);
'grind', in which Greek and Armenian share a state (_*h$_2$elh$_1$-_) against Hittite, Italo-Celtic, Germanic and Balto-Slavic (_*molh$_2$-_ ∼ _*melh$_2$-_);
'young', in which Greek and Armenian share a state (_*néwos_ and one of its derivatives) against Italic and the satem group (_*h$_2$yu-Hén-_; if a derivative of the latter is included, also against Celtic and Germanic).

But two of these characters can be dismissed out of hand. In the case of 'husband' the supposed Graeco-Armenian state is simply the most archaic inherited word for 'man' (i.e., 'adult male human being'), and its extension to 'husband' can easily have been a parallel innovation; in the case of 'young', the inherited word for 'new' has

[29] In this case as in several others below, alternative codings are ignored so as not to multiply artificially the characters supporting particular subgroups. In the 'broader' coding of this character (in which derivatives are coded together with the basic root-noun), the Graeco-Armenian subgroup is defined against a more conservative group including not only Latin and Vedic but also Hittite, Albanian, Osco-Umbrian and Balto-Slavic.

[30] Other groups of languages share other states that do not contribute to forcing a Graeco-Armenian subgroup.

similarly been extended, perhaps independently. It appears that we have overlooked, and thus failed to exclude, at least two characters exhibiting parallel development.[31] Possibly 'wind' is a similar case, since the etymological meaning of the Graeco-Armenian word is transparently 'breath' (cf. **h₂énh₁ti* '(s)he breathes'), though in that case the semantic shift is unusual enough to suggest a shared development. The other characters are not so easily disposed of, however. The Graeco-Armenian word for 'day' is attested nowhere else and in no other meaning. The peculiar words for 'not' appear to reflect an archaic noun meaning 'life' (!!; Cowgill 1960); the semantic shift that must be posited in that case depends on syntax and pragmatics to a degree that renders shared development the only really likely explanation, and for that reason 'not' is the best support for Graeco-Armenian – unless, of course, one rejects Cowgill's etymology. 'Grind' is weaker, since probable cognates of the Graeco-Armenian root can be found in Middle Iranian languages (Clackson 1994: 90 with fn. 20, p. 219) and are clearly the usual words for 'grind' in at least some (see especially Bailey 1979 s.v. *ārr-*); though lexical borrowing cannot completely be excluded (nor can chance resemblance in the case of a root whose reflex almost everywhere is the short sequence *al(V)-*), it is equally likely that this particular root was not a purely Graeco-Armenian innovation. The upshot is that Graeco-Armenian is supported by at least two lexical characters, one of which ('not') has some morphosyntactic content, and probably by three such characters. But though the quantity of the evidence is comparable to that for Italo-Celtic, the absence of any phonological or inflectional character makes it qualitatively poorer. In sum, though we think that Clackson (1994) has overstated his case in denying any evidence for Graeco-Armenian, we readily admit that the evidence is disappointingly meagre; in effect, he and we seem both to be quite close to the line that divides our positions, even though we are on opposite sides of it.

Finally, we need to consider the status of the satem group, including Indo-Iranian and Balto-Slavic – that is, the sister

[31] Though it is of course annoying to discover such an oversight, it is encouraging to find that it *can* be discovered after the fact by a careful analysis of one's results. The discussion of the robustness of other subgroups shows that none is crucially supported by a similar oversight.

subgroup of Graeco-Armenian within the core of the family. It is forced by three characters:

phonological character P2, encoding the complete merger of velar and labiovelar stops and the fronting of palatals, in which only Indo-Iranian and Balto-Slavic share the innovative state;[32]

phonological character P3, encoding the retraction of **s* after high vocalics, rhotics and dorsals, with an identical distribution of states;

'all', in which Indo-Iranian and Balto-Slavic share a state (**wi*-, variously extended) against several other groupings (crucially Tocharian and Greek; also Germanic and Celtic, and the Luvian group) in the 'broader' alternative coding.

Again the evidence is slender but includes phonological characters, rendering it fairly solid. In this case, however, the distribution of evidence for the states of phonological character P2 raises the question of what it actually means to posit an internal node in an evolutionary tree of this kind. We will address that question in the following section.

We must also ask why the evidence for virtually all the larger, non-traditional subgroups that our algorithm posits is so slender. As we suggested above, the evidence is fairly sparse no matter what method of subgrouping one employs, but there is also a further, and very important, factor at work. That is best demonstrated by considering the distribution of states of the character 'fear' among the core subgroups of our tree. In the broader of the two alternative codings of this character, Balto-Slavic and Indo-Iranian exclusively share a state (**b^{h}eyH*-); in both codings Greek and Armenian exclusively share another (**dwey*-). Thus the character neatly divides these languages into Graeco-Armenian and the satem grouping; yet we have cited it as evidence for neither subgroup in the discussion above. That is because it does not *force* either subgroup; it is possible to accommodate this distribution of states in a tree in which one or the other of those larger subgroups is not posited, as can be seen from Figure 10.

[32] Though both Armenian and Albanian front the PIE palatal stops, neither exhibits a complete merger of the other dorsals; see the discussions of Demiraj (1997: 63–65) and Olsen (1999: 805–808 with references).

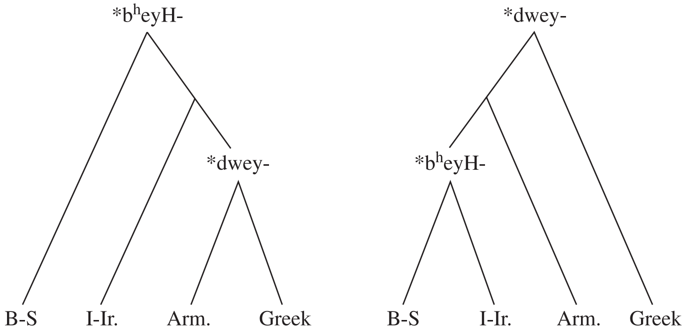

Figure 10. Two alternative subgroupings of the core permitted by the character ‘fear’.

Cases like this one are fairly common in our data, and some are spectacular. For instance, perusal of the morphological characters will show that Italic and Celtic exclusively share states of characters M6 (the thematic optative, in which the Italo-Celtic state is *-*ā*-) and M8 (the superlative, in which the Italo-Celtic state is *-*ism̥o*-). In both cases the core subgroups agree in exhibiting a different state (*-*oy*- and *-*isto*- respectively). But because Tocharian exhibits other states for both (*-*ih*$_1$- and null respectively; and note that Anatolian has a null state for both characters), these characters do not force the Italo-Celtic subgroup: it is possible that the common ancestor of Italic, Celtic and the core – or even the common ancestor of the entire family – exhibited the states actually attested only in Italic and Celtic, and that those states were replaced by the (obviously innovative) states of the core languages. There is no positive reason to believe that, and most specialists will find it utterly implausible; indeed, the states shared by Italic and Celtic are sometimes taken as indications of an Italo-Celtic subgroup. But the point is that the implausible alternative *cannot be excluded* on rigorous mathematical grounds, and for that reason the algorithm does not use these characters in constructing Italo-Celtic.

It can thus be seen that a major reason for the paucity of evidence for various subgroups is that our methodology accepts as ‘evidence’ only those pieces of evidence that are mathematically ineluctable. Taking that into account, one might argue that the fact that we have found any evidence at all for, say, a Graeco-Armenian subgroup is highly significant. On the other hand, the quantity and quality of the evidence that supports (but does not force) each controversial subgroup are also highly variable. Some half-dozen additional characters offer ancillary support for Italo-Celtic, and (as we have seen) two are exceptionally clear monomorphic inflectional characters. About a dozen and a half additional characters support the satem grouping, but all are lexical, and many are also polymorphic or exhibit obvious parallel development; this pattern tends to reinforce the impression that the unity of the satem group is rather loose, probably reflecting extensive borrowing between already differentiated dialects. Graeco-Armenian is supported by only a handful of further lexical characters, and so remains by far the weakest of the larger subgroups our methodology has found.

7.6. The meaning of internal nodes and of the Stammbaum *generally*[33]

Exactly what it means to posit an internal node in a linguistic evolutionary tree requires some further discussion. There is a widespread and long-standing conviction among historical linguists that an evolutionary tree (or *Stammbaum*) is hardly ever an appropriate model for the diversification of a language family; for instance, the introductory sections of Porzig (1954) give the impression that the tree model has been completely superseded by a network model representing diversifying dialects that share innovations in overlapping patterns (not representable by a tree). This conviction could be correct only if (1) the speciation of languages never proceeded by an abrupt and final separation of parts of a speech community and (2) each internal node of the tree were constrained to represent a virtually undifferentiated dialect – that is, had to be interpreted as a linguistic unity in the strictest possible sense. But it should be uncontroversial that condition (1) is not met; at least part of the diversification of the Oceanic family, for instance, must have consisted of abrupt losses of contact between groups of speakers,

[33] The reader will find in this section numerous points of contact with Ross (1997), which we did not become aware of until after this paper had been submitted for publication. We are grateful to an anonymous reviewer for that reference.

and other instances of 'clean' speciation are known from the historical record. It is the second condition that needs to be considered in depth.

It is easy enough to suggest a rationale for the requirement that two languages not be said to descend from the same ancestor unless they are descended from precisely the same dialect: how else can we constrain the notion of common descent? For historical linguists who believe that practically any part of the lexicon or grammar can be borrowed between different speechforms – even different languages – the partial 'mixing' of speechforms in contact constitutes a problem for the tree model that can be avoided only by requiring that no mixing at all be present in an ancestral node (even if that means abandoning the tree model altogether). But as we noted at some length in section 2, modern research on bilingualism does not suggest that absolutely anything can be borrowed into one's native language. On the contrary, bilingual speakers normally borrow lexemes into their native language, but import their native phonemic system and morphosyntax into an imperfectly learned (i.e., non-native) second language; only from dialects closely related to one's native dialect are new phonology and morphosyntax normally incorporated into one's native dialect to a degree approaching complete success. To be sure, the effects of imperfect second-language learning strongly resemble 'borrowing of morphosyntax' after the fact, if the imperfectly learned second language eventually becomes a community norm. But in our view such a pattern in the data reveals a discontinuity of transmission which should exclude the language in question from any strict 'family tree'.[34]

The importance of this consideration is that it places a natural and rather narrow constraint on what an internal node in a linguistic evolutionary tree can represent. So long as the phonology and morphosyntax that have to be posited for any node are internally consistent (or nearly so), that node can be taken to represent a group of closely related and mutually intelligible dialects – a genuine linguistic unity, in fact a speech community in the broad sense.

[34] Somewhat paradoxically, this means that we wish to strengthen the general conclusion of Thomason and Kaufman (1988) – namely, that the importance of contact phenomena in historical linguistics has been underestimated – by rejecting some of their specific arguments.

Not only do we not need to require that no internal variation have been present, we should assume that such variation *was* present in the absence of evidence to the contrary (since variation is ubiquitous in languages still spoken natively); but that by no means vitiates the claim that the node in question represents a single coherent language of the past.

Recent proposals that various subgroups of IE arose by 'convergence' need to be evaluated in the light of these considerations. For instance, Garrett (1999), whose discussion is unusually clear and well articulated, suggests that various subgroups of IE arose by borrowing of innovations among closely related dialects which were not very different from other, neighbouring dialects. That is certainly plausible (whether or not one accepts his arguments in detail), but it does not necessarily lead to the rejection of a *Stammbaum* for the IE family. Significantly more extreme convergence hypotheses are, in our view, not plausible because they appear to violate the UP.

Of course it is not hard to think of scenarios that will still pose considerable problems for the tree model. Very large geographic dialect continua, in which adjacent dialects are mutually intelligible but more distant ones are not, cannot be represented insightfully by the tree model; a network model (about which we will say more below) must be employed instead. Imperfectly learned second languages which by accident become standard in a community and are then learned natively cannot be accommodated in a tree representing only genetic descent (nor, a fortiori, can strict creoles); the genuine infiltration of one grammar by another that can occur when bilingualism persists for centuries poses further problems that must be addressed on an ad hoc basis. But when all these types of cases have been set aside, it is clear that there will still be many cases of linguistic diversification that can be modelled successfully as evolutionary trees.

Unfortunately it does not follow that the interpretation of a *Stammbaum* is always straightforward. In the simplest case two or more sister nodes and their mother can represent a relatively abrupt separation of a speech community (as defined above) into descendant communities. But it is also possible (and usually more likely) that the diversification was gradual, at first yielding a continuum of dialects trading innovations with close neighbours, and that only dialects from distant parts of the original continuum survived to leave descendants; if the overlapping pattern of innovations that occurred in the dialect continuum is not apparent from an examination of the few survivors, we will be led to posit a 'clean' speciation, and thus a tree, which reflects not the original diversification of the languages but the end product of a complex series of events of differentiation and survival. This will be relevant in our discussion of the position of Germanic in the following section.

There is at least one other type of complication that must be borne in mind in considering what a branching structure in a *Stammbaum* can represent: the character states that appear to define the subgroup can actually reflect innovations that spread through the subgroup after its member dialects had diversified significantly (but were still, of course, mutually intelligible). There is some likelihood that our satem grouping is such a case. While we agree with Andersen (1968) that the retraction of **s* (character P3) was originally uniform in Balto-Slavic (as it unarguably is in Indo-Iranian), there is some evidence that the characteristic pattern of development of dorsal stops (character P2) spread from Indo-Iranian to Balto-Slavic after they had begun to diverge (as suggested by Hock 1986: 442–444 with references p. 667).[35]

Thus even the clear results of our methodology require intelligent interpretation. The position of Germanic can be understood, if at all, only by carefully considering the unique pattern of data that characterises it.

[35] The difficulty is that some palatal stops emerge as velars in Balto-Slavic – that is, they develop according to the pattern typical of more westerly European languages; and even after we have eliminated those Slavic forms that could easily be loans from pre-Proto-Germanic, such as OCS *svekry* 'mother-in-law' and *gǫsĭ* 'goose' (plausible because other words must have been borrowed later from Proto-Germanic – e.g. OCS *mlěko* 'milk', which has actually undergone Grimm's Law!), we are left with a residue of examples like Lithuanian *akmuõ* 'stone' which are difficult to explain as borrowings. It is likely that the fronting of PIE palatal stops and the merger of velars and labiovelars reflect different linguistic events, and it seems at least possible that the fronting spread through a diversifying dialect continuum in such a pattern that dialects very far from its point of origin (such as the dialect that would eventually become Proto-Baltic) underwent the change inconsistently. Palatals that were unaffected in those dialects could then have merged with the velars.

7.7. The problem of Germanic

There are at least two scenarios that might have given rise to the peculiar pattern of data involving Germanic. One is that the diversification of the IE family must be modelled at least in part as a network rather than a tree (as discussed in the previous section). We know what happens when we apply our methodology to such a case because we deliberately set out to do so early in our research, so as to see what pattern would emerge and to learn to recognise the results of this phenomenon. We attempted to find the internal subgrouping of the West Germanic subfamily. Our IE database includes only two West Germanic languages, Old English and Old High German, and they are at the extremes of the West Germanic subgroup. What we did in the earlier experiment was to construct a database including several West Germanic languages and two North Germanic languages to serve as an outgroup, and we used modern languages, both because the early data for many West Germanic languages are too sparse and because we wanted to see what two millennia of development in contact would lead to. The results were a mess. The three best trees we could find were all very bad, all about equally bad, all very different, and each impugned by a quite different set of non-convex characters. The failure was total, and we were not able to find a better tree by omitting any one language. It is possible that the position of Germanic in the IE family is a problem of this sort, but only if it occupied so central a position in the family during its early diversification that its removal from the data would resolve the remainder into a relatively clean tree. Whether that is plausible either mathematically or archaeologically is unclear to us.

But there is also a more interesting possibility, because the pattern of character states exhibits an interesting structure. All the inflectional characters that give any precise information about the position of Germanic – namely M5, M6 and M8 – place it in the large subgroup that also includes Balto-Slavic, Indo-Iranian and Greek; and since those are the characters that are the most reliable indicators of genetic descent, it appears that Germanic should be placed in what we are calling the core of the family – the residue after the departure of Anatolian, Tocharian and Italo-Celtic. Of the

16 lexical characters in whose incompatibility the position of Germanic seems to be implicated,[36] three are non-convex on any plausible tree (namely 'beard', 'one' and 'tears'; see section 7.2). Of the remaining 13, Germanic shares a state with Baltic, Slavic or the whole Balto-Slavic subgroup in four, conformably to its probable position in the core; in the other nine, Germanic shares a state either with Italic or with Celtic. This split distribution of character states leads naturally to the hypothesis that Germanic was originally a near sister of Balto-Slavic and Indo-Iranian (possibly before the satem sound changes spread through that dialect continuum, if that is what happened); that at a very early date it lost contact with its more easterly sisters and came into close contact with the languages to the west; and that that contact episode led to extensive vocabulary borrowing at a period before the occurrence in any of the languages of any distinctive sound changes that would have rendered the borrowings detectable. The states of 'beard' and 'one' shared by Germanic can also owe part of their anomalous distributions to the same episode of contact. In short, we are led to posit an episode of intensive language contact between Germanic and the western languages well before the known periods of intensive contact with Celtic that have been established by earlier researchers. Unfortunately it is still unclear to us how this hypothesis can be tested.

In sum, it is clear that the development of Germanic exhibits some characteristics which cannot realistically be modelled with a 'clean' evolutionary tree, but it is not clear what historical developments have given rise to those anomalies.

8. Conclusions

The most interesting result of our experiment is that we have been able to construct a stable evolutionary tree for most of the IE family, contrary to one widespread view of how the family diversified. Since a tree is more tightly constrained mathematically than a network, our hypothesis is in principle more easily falsifiable than, say, that of

[36] We do not count 'head', in which the incompatibility appears to be a problem internal to Germanic.

Porzig (1954), and so should be preferred if it can be shown to be accountable to all the relevant data. Respectable support for Italo-Celtic and the persistent recalcitrance of Germanic are also interesting results.

On a methodological level, we have learned that the problem of subgrouping by character compatibility is much more difficult than we had initially supposed. Though the principles are clear, numerous practical problems intervene, including widespread character polymorphism, rampant parallel development of states, and possibly undetectable lexical borrowing. On the other hand, it is also reasonably clear that the idiosyncratic pattern of data that led us to discover the problematic behaviour of Germanic is the signature, in a character database, of some specific type of linguistic event. Though for a mathematical methodology it amounts to a 'lucky break' (since one would not necessarily expect a difficult dataset to exhibit any such pattern), it is fortuitous only in the sense that we did not foresee it. One desideratum of future research is a better understanding of the linguistic events that give rise to such patterns so as to identify them correctly in other cases.

The most important direction for future research is also clear: we need to devise appropriate methods for inferring non-treelike networks of linguistic diversification from character data, and a means of deciding whether a tree or a network is appropriate in difficult cases. Work on those problems is proceeding with all deliberate speed.

Don Ringe
Department of Linguistics
University of Pennsylvania
Philadelphia, PA 19104–6305
USA
Email: dringe@unagi.cis.upenn.edu

Tandy Warnow
Department of Computer Sciences
University of Texas at Austin
USA
Email: tandy@cs.utexas.edu

Ann Taylor
Department of Language and Linguistic Sciences
University of York
Heslington, York YO10 5DD
UK
Email: at9@york.ac.uk

Appendix: partial report of character data

The phonological characters and their states:

P1 $*p \ldots k^{w} > *k^{w} \ldots k^{w}$
1, absent [ancestral]; 2, present; 3, obscured by merger;[37] 4 &c., no evidence

P2 full ‘satem’ development of dorsals
1, absent [ancestral]; 2, present

P3 ‘ruki’-retraction of *s
1, absent [ancestral]; 2, present; 3, 4, obscured by merger or orthography

P4 lenition of stops after long vowels and unstressed vowels (only; Melchert 1994: 60–63)
1, absent [ancestral]; 2, present

P5 medial $*k^{w} > *g^{w}$ unless *s follows immediately (Melchert 1994: 61–62)
1, absent [ancestral]; 2, present

P6 ‘limited’ Cop’s Law (*éC- > *éCC-; Melchert 1994: 62)
1, absent [ancestral]; 2, present

P7 word-initial *ye- > *e-
1, absent [ancestral]; 2, present

P8 merger of *i, *e, *u and merger of *a, *ō:
1, absent [ancestral]; 2, present

P9 *mbh > *m (but not $*nd^{h} > *n$, etc.; Ringe 1996: 42–43)
1, absent [ancestral]; 2, present

[37] As an anonymous reviewer emphasises, this coding does (intentionally) imply that the merger of labials and labiovelars was a historically shared change in the Osco-Umbrian group. Note that our coding for character P3 carries a reverse implication for the loss or non-writing of distinctions between sibilants in the Baltic languages.

P10 **d* > ∅ before conss., affrication of other **d*, merger of palatalised **d* with palatalised dorsals (Ringe 1996: 64–65, 146–150)

1, absent [ancestral]; 2, present

P11 **tsk* > **tk*, but **kst* > **kəst* (Ringe 1996: 71–72)

1, absent [ancestral]; 2, present

P12 merger of all non-high vowels and syllabic nasals

1, absent [ancestral]; 2, present

P13 Bartholomae's Law (rightward assimilation of aspiration)

1, absent [ancestral]; 2, present; 3, no evidence

P14 merger of voiceless aspirated stops and preconsonantal voiceless stops as fricatives

1, absent [ancestral]; 2, present

P15 development of acute vs. circumflex contrast in non-final heavy syllables

1, absent [ancestral]; 2, present

P16 sequence of changes: (a) Grimm's Law; (b) Verner's Law; (c) initial-syllable stress; (d) merger of unstressed **e* with **i* except before **r*

1, absent [ancestral]; 2, present

P17 sequence of changes: (a) loss of intervocalic **j* unless **i* precedes and does *not* follow immediately; (b1) **əi* > **ai*, and (b2) **ōV* > **ō̄* (Cowgill 1959, Þórhallsdóttir 1993)

1, absent [ancestral]; 2, present

P18 merger of final non-nasalised **ō* with **u*; **ē* > **ā* in stressed syllables, but merges with **ai* in unstressed syllables

1, absent [ancestral]; 2, present

P19 merger of **ðw* and **zw* with **ww* (Stiles 1985: 89–94)

1, absent [ancestral]; 2, present

P20 merger of **ē* with **ī*; merger of **ō* with **ū* in final syllables, but with **ā* elsewhere

1, absent [ancestral]; 2, present

P21 **p* > **k* before obstruents, **b* before liquids, **w* before nasals and after **s*, ∅ elsewhere

1, absent [ancestral]; 2, present

P22 syncope of short vowels in final syllables next to **s* and after semivowels

1, absent [ancestral]; 2, present

Matrix of character states:

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11
Hitt.	4	1	1	2	2	2	2	1	1	1	1
Arm.	1	1	1	1	1	1	1	1	1	1	1
Grk.	1	1	1	1	1	1	1	1	1	1	1
Alb.	1	1	1	1	1	1	1	1	1	1	1
TB	1	1	1	1	1	1	1	2	2	2	2
Vedic	1	2	2	1	1	1	1	1	1	1	1
Av.	1	2	2	1	1	1	1	1	1	1	1
OCS	1	2	2	1	1	1	1	1	1	1	1
Lith.	1	2	2	1	1	1	1	1	1	1	1
OE	1	1	1	1	1	1	1	1	1	1	1
OIr.	2	1	1	1	1	1	1	1	1	1	1
Latin	2	1	1	1	1	1	1	1	1	1	1
Luv.	5	1	1	2	2	2	2	1	1	1	1
Lyc.	6	1	1	2	2	2	2	1	1	1	1
TA	1	1	1	1	1	1	1	2	2	2	2
OPer.	1	2	2	1	1	1	1	1	1	1	1
OPru.	1	2	3	1	1	1	1	1	1	1	1
Latv.	1	2	4	1	1	1	1	1	1	1	1
Goth.	1	1	1	1	1	1	1	1	1	1	1
ON	1	1	1	1	1	1	1	1	1	1	1
OHG	1	1	1	1	1	1	1	1	1	1	1
Welsh	2	1	1	1	1	1	1	1	1	1	1
Osc.	3	1	1	1	1	1	1	1	1	1	1
Umb.	3	1	1	1	1	1	1	1	1	1	1

Matrix of character states (continued)

	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	P22
Hitt.	1	1	1	1	1	1	1	1	1	1	1
Arm.	1	1	1	1	1	1	1	1	1	1	1
Grk.	1	1	1	1	1	1	1	1	1	1	1
Alb.	1	1	1	1	1	1	1	1	1	1	1
TB	1	1	1	1	1	1	1	1	1	1	1
Vedic	2	2	1	1	1	1	1	1	1	1	1
Av.	2	2	2	1	1	1	1	1	1	1	1
OCS	1	1	1	2	1	1	1	1	1	1	1
Lith.	1	1	1	2	1	1	1	1	1	1	1
OE	1	1	1	1	2	2	2	2	1	1	1
OIr.	1	1	1	1	1	1	1	1	2	2	1
Latin	1	1	1	1	1	1	1	1	1	1	1
Luv.	1	1	1	1	1	1	1	1	1	1	1
Lyc.	1	1	1	1	1	1	1	1	1	1	1
TA	1	1	1	1	1	1	1	1	1	1	1
OPer.	2	3	2	1	1	1	1	1	1	1	1
OPru.	1	1	1	2	1	1	1	1	1	1	1
Latv.	1	1	1	2	1	1	1	1	1	1	1
Goth.	1	1	1	1	2	2	1	1	1	1	1
ON	1	1	1	1	2	2	2	1	1	1	1
OHG	1	1	1	1	2	2	2	2	1	1	1
Welsh	1	1	1	1	1	1	1	1	2	2	1
Osc.	1	1	1	1	1	1	1	1	1	1	2
Umb.	1	1	1	1	1	1	1	1	1	1	2

The morphological characters and their states:

M1 organisation of the verb system

1, one stem per lexeme (1a, two conjugations; 1b, single conjugation); 2, present/aorist/perfect contrast present or reconstructable; 3, present/subjunctive/preterite contrast, the former two largely parallel; 4, present/preterite/infinitive contrast; 5, present/preterite contrast, the latter in two conjugations (‘strong’ vs. ‘weak’); 6, present/subjunctive/future/preterite contrast; 7, present/subjunctive/preterite contrast, the latter two usually sigmatic

M2 augment

1, present (including relics); 2 &c., absent

M3 thematised aorist

1, absent, probably primitively [ancestral]; 2, present or immediately reconstructable; 3 &c., no evidence

M4 productive function of *-_sḱé/ó_-

1, iterative; 2, inchoative; 3, causative; 4 &c., other or none

M5 mediopassive primary marker (sg and 3pl; see especially Yoshida 1990)

1, *-_r_ [ancestral]; 2, *-_y_ (= active *-_i_); 3 &c., no evidence

M6 thematic optative

1, *-_i_h_1-; 2, *-_oy_-; 3, *-_ā_-; 4 &c., no evidence

M7 genitive singular of o-stem nouns and adjectives [polymorphic character] (see especially de Simone 1980: 81–83, Koch 1991: 114)

1, *-_os_; 2, *-_osyo_; 3, *-_ī_; 4, replaced by ablative *-_e_-_(_h_2_)ad_; 5, replaced by *-_onso(C)_; 6, replaced by i-stem *-_eys_; 7 &c., no evidence

M8 most archaic superlative suffix

1, *-_isto_-; 2, *-_ism̥o_-; 3 &c., other or none

M9 athematic dative pl ending & M10 athematic instrumental pl ending [polymorphic set][38]

[38] A polymorphic set is a set of two or more characters that have ‘traded’ states because of parallel semantic shift. Typically all the characters of such a set (occasionally all but one) exhibit parallel development and must thus be sequestered for the purposes of running the software. A polymorphic set can be modelled as a single character of which every state is polymorphic and the component states of each polymorphic state are ordered.

Shared states only: 1, PAnat. *-*os*; 2, PAnat. abl. *-*ti*; 4, *-b^h*i*, with extensions and remodellings: 4x, *-b^h*is*; 4y, *-b^h*yos*; 4z, *-b^h*os*; 5, LOC PL *-*su*; 10, endings with *-*m*-: 10a, *-*mos*; 10b, *-*mīs*; 10c, *-*mus*; 10d, *-*mis*

M11 abstract noun suffix *-ti- and extensions

1, *-*ti*- only [ancestral]; 2, *-*ti*- and *-*ti-Hen*-; 3 &c., insufficient evidence

M12 imperfect subjunctive in *-*sē*-

1, absent [ancestral]; 2, present; 3, no evidence

M13 gerundive in *-*ndo*-

1, absent [ancestral]; 2, present

M14 syncretism of 3SG, 3DU and 3PL

1, absent [ancestral]; 2, present

M15 replacement of 2SG indicative by optative in the strong preterite

1, absent [ancestral]; 2, present

Matrix of character states:

	M1	M2	M3	M4	M5	M6	M7	M8
Hitt.	1a	2	1	1	1	4	1	3
Arm.	2	1	2	4	3	5	2	4
Grk.	2	1	2	1	2	2	2	1
Alb.	2	3	3	5	4	6	7	5
TB	3	4	2	3	1	1	5	6
Vedic	2	1	2	6	2	2	2	1
Av.	2	1	2	7	2	2	2	1
OCS	2	5	2	8	5	2	4	7
Lith.	4	6	4	9	6	2	4	8
OE	5	7	5	10	2	2	2	1
OIr.	6	8	2	11	1	3	3	2
Latin	2	9	2	2	1	3	2/3	2
Luv.	1b	10	1	12	1	7	8	9
Lyc.	1b	11	1	13	7	8	1	10
TA	3	12	2	3	1	1	5	11
OPer.	2	1	6	14	2	2	2	1
OPru.	4	13	7	15	8	2	2	12
Latv.	4	14	8	16	9	2	4	13
Goth.	5	15	9	17	2	2	9	1
ON	5	16	10	18	2	2	2	1
OHG	5	17	11	19	10	2	2 (?)	1
Welsh	7	18	12	20	1	3	3	2
Osc.	2	19	2	21	1	3	6	2
Umb.	2	20	2	22	1	3	6	2

Matrix of character states (continued)

	M9	M10	M11	M12	M13	M14	M15
Hitt.	1	2	1	1	1	1	1
Arm.	3	4(x?)	1	1	1	1	1
Grk.	5	4	1	1	1	1	1
Alb.	6	7	3	1	1	1	1
TB	8	9	1	1	1	1	1
Vedic	4y	4x	1	1	1	1	1
Av.	4y	4x	1	1	1	1	1
OCS	10a (10c?)	10b	1	1	1	1	1
Lith.	10c	10d	1	1	1	2	1
OE	10d	10d	1	1	1	1	2
OIr.	4	11	2	1	1	1	1
Latin	4z	12	2	2	2	1	1
Luv.	1	2	1	1	1	1	1
Lyc.	1	2	4	1	1	1	1
TA	13	14	1	1	1	1	1
OPer.	15	4x	1	1	1	1	1
OPru.	10a	16	1	1	1	2	1
Latv.	10c	10c (?)	1	1	1	2	1
Goth.	10d	17	1	1	1	1	1
ON	10d	18	1	1	1	1	1
OHG	10d	10d	1	1	1	1	2
Welsh	19	20	5	1	1	1	1
Osc.	4z	21	2	2	2	1	1
Umb.	4z	22	2	3	2	1	1

Some interesting lexical characters and their non-unique states:

1 'all (pl.)'

3, *pā́ntes; 5, *wi- with extensions: 5a, *wí-ḱwo- > PIIr. *víśva-; 5b, *wi-so- > PBS *visa-; 6, *ol- with extensions: 6a, *ol-noy > PGmc. *allai; 6b, PCelt. *ol-yo-; 8, PLuv. *pūno-; both codings for superstates 5 & 6 employed

24 'cold' [polymorphic]

5, PToch. *k^wərośce; 7, derivs. of *ow-; 7x, PCelt. *ougros; 8, *ḱolHtos; 10, PGmc. *kaldaz (< *gol-); states 7 and 7x coded separately, as the latter has a root-extension

28 'day [= 24 hrs.]'

1, *dyḗws and derivs; 1w, PAnat. *díwots; 1x, *dit-; 1y, *deyn- ~ *din-; 1z, POU *dyēklo-; 2, *ā́mr̥; 3, PToch. *kawnə; 5, PGmc. *dagaz; both codings for superstate 1 employed

29 'die' [polymorphic, in Gmc. only][39]

2, *mer-, orig. pres. *mr̥yétor; 2x, PEBalt. pres. *mirsta; 6, PWGmc. *sterba-, *stirbidi; 8, *wel-; 10, PGmc. *dawja-, *dawidi; 10x, denom. *dauþna-; both codings for superstates 2 & 10 employed

33 'drink' (pres. stem)[40]

1, *ḗh_2g^{wh}ti; 3, *peh$_3$- ~ *pī-, orig. pres. *píbeti; 5, PEBalt. *gerja; 6, PGmc. *drinkidi

34 'dry'

5, PToch. *asarë; 6, *sawsos; 6x, PIIr. *suškas; 7, PWGmc. *drūg- ~ *drug-; 8, derivs. of *ters- 'be dry': 8a, pre-Celt. *tērs-; 8b, PGmc. *þursu- ~ *þurzu-; both codings for superstate 6, but 8a, 8b coded separately (prob. independent derivs)

46 'fear' [polymorphic, in Gmc. only]

2, *dwey-; 4, *prek- with extensions: 4a, PToch. *praska- ~ *pərska- ← <[41] pres. *pr̥(k)-skéló-; 4b, derivs of PGmc.

[39] We give the 3SG present indicative of verbs whenever it is reconstructable; the shape of the stem is also given if it is not immediately clear from the shape of the 3SG.

[40] In the case of suppletive verbs we code only for the present stem, having found by experience that the attempt to code for all stems multiplies polymorphic characters without yielding any clear information about subgrouping.

[41] Shaftless arrows indicate development by regular sound change; arrows with shafts indicate developments of other kinds.

*_furhtaz_ 'fearful' < adj. *_pr̥któs_; 5, *_b^heyH-_ with extensions: 5a, perf. *_b^heb^hóy(H)e_; 5c, PBalt. pres. *_bijā_; 5b, 5d, unique derivs; 6, perf. *_h_2eh_2óg^he_ 'be upset'; both codings for superstate 5 employed, but 4a, 4b coded separately (because unmediated replacement is unlikely)

59 'four'
1, PAnat. *_mǣu-_; 2, *_k^wetwóres_; 2x, PGmc. *_fedwōr_ with unexpected *_f-_; both codings for superstate 2 employed

63 'give' (pres. stem)
1, *_ay-_; 1x, PAnat. cpd. *_p-ay-_; 2, *_deh_3-_, various press.: 2a, 2c, orig. pres. unclear; 2b, *_déde$h_3$ti_ and developments of same; 2bx, PBS *_dōd-_; 4, PGmc. *_geba-_, *_gibidi_; both codings for superstates 1 & 2 employed

68 'hair' [polymorphic, in Gmc. only]
7, derivs of *_wel-_: 7a, satem *_wolḱos_; 7b, PCelt. *_woltos_; 9, PNWGmc. *_hārą_; 17 PGmc. *_skuftą_; 7a, 7b coded separately (only remotely related)

89 'lake' [polymorphic, in Gmc. only]
3, *_léymon-_ ~ *_limn-´_; 7, PBS *_ežeran_; 8, *_móri_ 'sea'; 9, PGmc. *_saiwiz_; 10, PItCelt. *_lóku_ ~ *_l̥kéw-_

90 'laugh'
2, *_ǵelh_2-_; 4, PToch. *_kër-_; 7, derivs of *_smey-_ 'smile'; 9 PGmc. *_hlahja-_, *_hlahidi_

92 'left(-hand)'
5, PToch, *_ś(uw)āl(i)y-_ (?; see Pinault 1999); 6, *_sewyós_; 8, PNWGmc. *_winistraz_; 18, POU *_nertro-_ (orig. 'lower')

97 'long'
5, PToch. *_pərkrë_ (< *_b^hr̥ǵ^{h}rós_ 'tall'); 6, *_dl̥$h_1$$g^h$ós_; 6x, PBalt. *_ilgas_ (with unexpected loss of *_d-_); 7, *_long^{h}os_; both codings for superstate 6 employed[42]

113 'not'
1, *_né_ and extensions; 2, *_$h_2$óyu_ 'life'; 3, PToch. *_ma_

124 'right(-hand)'
3 derivs. of *_deḱs-_; suffixes: 3c, *_-ino-_; 3e, *_-(i)tero-_; 3f, *_-wo-_ with n-stem extension; others unique; only substates coded (otherwise uninformative)

[42] We reject any connection of the isolated Persian _dræng_ with state 7.

143 ‘sing’
5, *peyH-; 9, PGmc. *singwidi; 10, PItCelt. *kaneti

161 ‘stone’
6, *h_2éḱmō; 7, PGmc. *stainaz; 9, PItal. *lapid-

198 ‘wind’
1, derivs. of *$h_2$weh_1- ‘blow’: 1a, *$h_2$wéh_1-n̥t-s (Melchert 1994: 54 with refs); 1b, ‘post-laryngeal’ *h_2wēntós; 1c, PIIr. *váatas < *$h_2$wéh_1-n̥t-o-s; 1d, PBS *vētras; 1e, PEBalt. *vējas; 2 *$h_2$ónh$_1$mos ‘breath’ (Olsen 1999: 27; ablaut adjusted in Greek); both codings for superstate 1 employed; states 1b, 1c differ only in phonology, and the apparent replacement of 1b by 1c is illusory (resulting from parallel phonological developments outside of IIr.)

201 ‘with’ [polymorphic, in IIr. only]
5, PToch. *śəlë; 6, PIIr. *smát; 7 PIIr. *sad^{h}á; 8, PBS *sVn (vowel problematic, Stang 1966: 32); 9, *med^{h}i or *metí; 10, *kom; on the etymologically ambiguous Albanian word see Demiraj (1997: 274–275 but also 55)

348 ‘grind’
1, *molh_2- ~ *melh_2-; 2, *$h_2$elh_1- (or *alh_1-)

368 ‘make’ (pres. stem) [polymorphic, in Gmc. only]
1, *$h_1$yé$h_1$ti; 2, derivs. of *werǵ-: 2a, *worǵeyeti; 2b, *wr̥ǵyéti; 6, PIIr. *kr̥náuti; 8, PEBalt. *darā; 9, PWGmc. *makōþi; 10, PCelt. *gnī- (Pedersen 1913: 544–546); 11, PItal. *fakyo-, *fakit; 2a, 2b coded separately (indep. derivs)

Matrix of character states:

	1	24	28	29	33	34	46	59	63	68	89
Hitt.	1	1	1w	1	1	1	1	1	1x	1	1
Arm.	2	2	2	2	2	2	2	2	2a	2	2
Grk.	3	3	2	3	3	3	2	2	2b	3	3
Alb.	4	4	1x	4	3	4	3	2	3	4	4
TB	3	5	3	5	1	5	4a	2	1	5	3
Vedic	5a	6	1	2	3	6x	5a	2	2b	6	5
Av.	5a	7/8	4	2	4	6x	5a	2	2b	7a	6
OCS	5b	9	1y	2	3	6	5b	2	2bx	7a	7
Lith.	5b	8	1y	2x	5	6	5c	2	2bx	8	7
OE	6a	10	5	6	6	7	4b	2x	4	9	8/9
OIr.	6b	7x	6	7	3	8a	6	2	5	7b	10
Latin	7	11	1	2	3	9	7	2	2b	10	10
Luv.	8	12*	7	8	1	10*	8	1	1x	11	11*
Lyc.	8	13*	8*	8	7*	11*	9*	3*	1x	12*	12*
TA	3	5	3	8	1	5	4a	2	1	13	3
OPer.	5a	14*	9	2	8*	6x	10	4*	2b	14*	13*
OPru.	5b	8	1y	9	3	6	5c	2	2bx	15	7
Latv.	5b	7/8	1y	2x	5	6	5d	2	2bx	16	7
Goth.	6a	10	5	10x	6	8b	4b/6	2x	4	17	8+9‡
ON	6a	10	5	10	6	8b	11	2x	4	9/17	14
OHG	6a	10	5	6/10	6	7/8b	4b	2x	4	9	9
Welsh	9	7x	10†	2	3	12†	12	2	2c	7b	15
Osc.	10	15*	1z	11*	3	13*	13*	2	2b	18*	16*
Umb.	11*	16*	1z	12*	9*	14*	14*	2	2b	19*	17*

* unique state assigned because evidence is lacking
† unique state assigned because the word is a loan
‡ compound

	90	92	97	113	124	143	161	198	201	348	368
Hitt.	1	1*	1	1	1	1	1	1a	1	1	1
Arm.	2	2	2	2	2	2	2	2	2	2	2a
Grk.	2	3	3	2	3a	3	3	2	3	2	3
Alb.	3	4	4	1	3b	4†	4	3	4	3	4
TB	4	5	5	3	4	5	5	1b	5	4	5
Vedic	5	6	6	1	3c	6	6	1c	6/7	5	6
Av.	6*	6	6	1	3c	7*	6	1c	6/7	6*	6
OCS	7	6	6	1	3c	5	6	1d	8	1	7
Lith.	8	7	6x	1	3c	8	6	1e	8	1	8
OE	9	8	7	1	5	9	7	1b	9	7	2b/9
OIr.	10	9	8	1	3d	10	8	4	10	1	10
Latin	11	10	7	1	3e	10	9	1b	10	1	11
Luv.	12*	11	9	1	6	11*	10*	5*	11*	8*	1
Lyc.	13*	12*	10*	1	7*	12*	11*	6*	12	9*	1
TA	4	5	5	3	8	5	12	1b	5	10*	1
OPer.	14*	13*	6	1	9*	13*	6	7*	7	11*	6
OPru.	15*	14*	6x	1	10	14	13	1d	8	12*	12
Latv.	7	15	11	1	11	15	6	1e	13	1	8
Goth.	9	16	7	1	3f	9	7	1b	9	1	2b
ON	9	8	7	4	12	9	7	1b	9	1	13
OHG	9	8	7	1	3f	9	7	1b	9	1	2b/9
Welsh	16	17	12	1	3g	10	14	1b	14	1	10
Osc.	17*	18	13*	1	3e	16*	15*	8*	10	13*	11
Umb.	18*	18	14*	1	3e	10	9	9*	10	1	11

* unique state assigned because evidence is lacking
† unique state assigned because the word is a loan

References

Agarwala, Richa and Fernández-Baca, D., 1994. 'A polynomial-time algorithm for the perfect phylogeny problem when the number of character states is fixed', *SIAM Journal on Computing* 23, 1216–1224.

Andersen, Henning, 1968. 'IE **s* after *i, u, r, k* in Baltic and Slavic', *Acta Linguistica Hafniensia* 11, 171–190.

Appel, René and Muysken, Pieter, 1987. *Language Contact and Bilingualism*, Baltimore: Edward Arnold.

Bailey, H. W., 1979. *Dictionary of Khotan Saka*, Cambridge University Press.

Beekes, Robert S. P., 1969. *The Development of the Proto-Indo-European Laryngeals in Greek*, The Hague: Mouton.

Beekes, Robert S. P., 1985. *The Origins of the Indo-European Nominal Inflection*, Innsbruck: Innsbrucker Beiträge zur Sprachwissenschaft.

Biggs, Bruce, 1978. 'The history of Polynesian phonology', in S. A. Wurm and Lois Carrington (ed.), *Second International Conference on Austronesian Linguistics: Proceedings, Fascicle 2*, Canberra: Australian National University, 691–716.

Bloomfield, Leonard, 1946. 'Algonquian', in Cornelius Osgood (ed.), *Linguistic Structures of Native America*, New York: Viking, 85–129.

Bodlaender, H. L., Fellows, M. R., Hallett, Michael T., Wareham, H. Todd and Warnow, Tandy, 2000. 'The hardness of perfect phylogeny, feasible register assignment, and other problems on thin colored graphs', *Theoretical Computer Science* 244, 167–188.

Bonet, Maria, Phillips, Cynthia A., Warnow, Tandy and Yooseph, Shibu, 1996. 'Constructing evolutionary trees in the presence of polymorphic characters', *SIAM Journal of Computing* 29, 103–131.

Braune, Wilhelm and Ebbinghaus, Ernst, 1973. *Gotische Grammatik*, 18th ed., Tübingen: Niemeyer.

Cardona, George, 1960. *The Indo-European Thematic Aorists*, Ph.D. Dissertation, Yale University.

Clackson, James, 1994. *The Linguistic Relationship Between Armenian and Greek*, Oxford: Blackwell.

Cowgill, Warren, 1959. 'The inflection of the Germanic *ō*-presents', *Language* 35, 1–15.

Cowgill, Warren, 1960. 'Greek *ou* and Armenian *oč'*', *Language* 36, 347–350.

Cowgill, Warren, 1965. 'The Old English present indicative ending -*e*', in *Symbolae Linguisticae in Honorem Georgii Kuryłowicz* (no ed.), Wrocław: Polska Akademia Nauk, 44–50.

Cowgill, Warren, 1970. 'Italic and Celtic superlatives and the dialects of Indo-European', in George Cardona, Henry M. Hoenigswald and Alfred Senn (ed.), *Indo-European and Indo-Europeans*, Philadelphia: University of Pennsylvania Press, 113–153.

Dawkins, R. M., 1910. 'Modern Greek in Asia Minor', *Journal of Hellenic Studies* 30, 109–132, 267–291.

Dawkins, R. M., 1916. *Modern Greek in Asia Minor*, Cambridge University Press.

De Simone, Carlo, 1980. 'L'aspetto linguistico', in Conrad M. Stibbe et al., *Lapis Satricanus*, The Hague: Staatsuitgeverij, 71–94.

Demiraj, Bardhyl, 1997. *Albanische Etymologien*, Amsterdam: Rodopi.

Dobson, Annette J., 1969. 'Lexicostatistical grouping', *Anthropological Linguistics* 11, 216–221.

Dobson, Annette J., N.d. 'Unrooted trees for numerical taxonomy', unpublished paper. [Internal evidence shows that this paper was written in the 1970s.]

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

Fantini, Alvino E., 1985. *Language Acquisition of a Bilingual Child: a Sociolinguistic Perspective (to Age Ten)*, San Diego: College-Hill Press.

Fromm, Hans and Sadeniemi, Matti, 1956. *Finnisches Elementarbuch, I: Grammatik*, Heidelberg: Winter.

Garey, M. and Johnson, D. S., 1979. *Computers and Intractability: a Guide to the Theory of NP-Completeness*, New York: Freeman and Co.

Garrett, Andrew, 1999. 'A new model of Indo-European subgrouping and dispersal', in Steve S. Chang, Lily Liaw and Josef Ruppenhofer (ed.), *Proceedings of the Twenty-Fifth Annual Meeting of the Berkeley Linguistics Society*, Berkeley: BLS, 146–156.

Gleason, Henry A., 1959. 'Counting and calculating for historical reconstruction', *Anthropological Linguistics* 1, 22–32.

Hajnal, Ivo, 1995. *Studien zum mykenischen Kasussystem*, Berlin: de Gruyter.

Hock, Hans Henrich, 1986. *Principles of Historical Linguistics*, Berlin: Mouton de Gruyter.

Hoenigswald, Henry M., 1960. *Language Change and Linguistic Reconstruction*, University of Chicago Press.

Hoenigswald, Henry M. and Wiener, Linda F. (ed.), 1987. *Biological Metaphor and Cladistic Classification: an Interdisciplinary Perspective*, Philadelphia: University of Pennsylvania Press.

Hübschmann, H., 1897. *Armenische Grammatik, I. Theil: Armenische Etymologie*, Leipzig: Breitkopf and Härtel.

Jasanoff, Jay H., 1987. 'Some irregular imperatives in Tocharian', in Calvert Watkins (ed.), *Studies in Memory of Warren Cowgill*, Berlin: de Gruyter, 92–112.

Jasanoff, Jay H., 1997. 'An Italo-Celtic isogloss: the 3pl. mediopassive in **-ntro*', in Douglas Q. Adams (ed.), *Festschrift for Eric Hamp*, Washington, D.C.: Institute for the Study of Man, Vol. I, 146–161.

Kannan, Sampath and Warnow, Tandy, 1997. 'A fast algorithm for the computation and enumeration of perfect phylogenies when the number of character states is fixed', *SIAM Journal of Computing* 26, 1749–1763.

Katz, Joshua T., 1998. *Topics in Indo-European Personal Pronouns*, Ph.D. Dissertation, Harvard University.

Kim, Ronald, 2000. '"To drink" in Anatolian, Tocharian, and Proto-Indo-European', *Historische Sprachwissenschaft* 113, 151–170.

King, Ruth, 2000. *The Lexical Basis of Grammatical Borrowing*, Amsterdam: Benjamins.

Koch, John T., 1991. 'Gleanings from the Gododdin and other Early Welsh texts', *Bulletin of the Board of Celtic Studies* 38, 111–118.

Kroch, Anthony, 1994. 'Morphosyntactic variation', in Katharine Beals et al. (ed.), *Papers from the 30th Regional Meeting of the Chicago Linguistic Society: Parasession on Variation and Linguistic Theory*, Chicago: CLS, 180–201.

Kroch, Anthony, Taylor, Ann and Ringe, Don, 2000. 'The Middle English verb-second constraint: a case study in language contact and language change', in Susan C. Herring, Pieter van Reenen and Lele Schøsler (ed.), *Textual Parameters in Older Languages*, Amsterdam: Benjamins, 353–391.

LAANEST, ARVO, 1982. *Einführung in die ostseefinnischen Sprachen*, translated by Hans-Hermann Bartens, Hamburg: Buske.

LABOV, WILLIAM, 1994. *Principles of Linguistic Change, Volume 1: Internal Factors*, Oxford: Blackwell.

LASS, ROGER, 1997. *Historical Linguistics and Language Change*, Cambridge University Press.

MAYRHOFER, MANFRED, 1986. 'Lautlehre (segmentale Phonologie des Indogermanischen)', in Warren Cowgill and Manfred Mayrhofer, *Indogermanische Grammatik, Band I*, Heidelberg: Winter, 73–181.

MCCONE, KIM, 1991. *The Indo-European Origins of the Old Irish Nasal Presents, Subjunctives, and Futures*, Innsbruck: Innsbrucker Beiträge zur Sprachwissenschaft.

MCMAHON, APRIL, 2000. *Change, Chance and Optimality*, Oxford University Press.

MEILLET, ANTOINE, 1925. *La Méthode comparative en linguistique historique*, Oslo: Aschehoug.

MEISEL, JÜRGEN M., 1989. 'Early differentiation of languages in bilingual children', in Kenneth Hyltenstam and Loraine K. Obler (ed.), *Bilingualism Across the Lifespan*. Cambridge University Press, 13–40.

MELCHERT, H. CRAIG, 1994. *Anatolian Historical Phonology*, Amsterdam: Rodopi.

MITHUN, MARIANNE, 1999. *The Languages of Native America*, Cambridge University Press.

OLSEN, BIRGIT ANETTE, 1999. *The Noun in Biblical Armenian*, Berlin: Mouton de Gruyter.

PEDERSEN, HOLGER, 1913. *Vergleichende Grammatik der keltischen Sprachen, Zweiter Band*, Göttingen: Vandenhoeck and Ruprecht.

PETERS, MARTIN, 1991. 'Ein tocharisches Auslautproblem', *Die Sprache* 34, 242–244.

PINAULT, GEORGES-JEAN, 1999. 'Tocharian and Indo-Iranian: relations between two linguistic areas', unpublished paper read at Cambridge, December 1999.

PORZIG, WALTER, 1954. *Die Gliederung des indogermanischen Sprachgebiets*, Heidelberg: Winter.

PRINCE, ELLEN F. and PINTZUK, SUSAN, 2000. 'Bilingual code-switching and the open/closed class distinction', *University of Pennsylvania Working Papers in Linguistics* 6.3, 237–257.

PUHVEL, JAAN, 1991. *Homer and Hittite*, Innsbruck: Innsbrucker Beiträge zur Sprachwissenschaft.

RASMUSSEN, JENS, 1985. 'Der Prospektiv – eine verkannte indogermanische Verbalkategorie?', in Bernfried Schlerath and Veronica Rittner (ed.), *Grammatische Kategorien: Funktion und Geschichte*, Wiesbaden: Reichert, 384–399.

RAYFIELD, J. R., 1970. *The Languages of a Bilingual Community*, The Hague: Mouton.

RINGE, DON, 1991. 'Evidence for the position of Tocharian in the Indo-European family?', *Die Sprache* 34, 59–123.

RINGE, DON, 1996. *On the Chronology of Sound Changes in Tocharian, Vol. I*, New Haven: American Oriental Society.

RINGE, DON, 2000. 'Tocharian class II presents and subjunctives and the reconstruction of the Proto-Indo-European verb', *Tocharian and Indo-European Studies* 9, 121–142.

RINGE, DON, WARNOW, TANDY, TAYLOR, ANN, MICHAILOV, ALEXANDER and LEVISON, LIBBY, 1998. 'Computational cladistics and the position of Tocharian', in Victor Mair (ed.), *The Bronze Age and Early Iron Age Peoples of Eastern Central Asia*, Washington, D.C.: Institute for the Study of Man, 391–414.

Risch, Ernst, 1955. 'Die Gliederung der griechischen Dialekte in neuer Sicht', *Museum Helveticum* 12, 61–76.

Rix, Helmut, et al., 1998. *Lexikon der indogermanischen Verben*, Wiesbaden: Reichert.

Ross, Malcolm, 1997. 'Social networks and kinds of speech-community event', in Roger Blench and Matthew Spriggs (ed.), *Archaeology and Language I: Theoretical and Methodological Orientations*, London: Routledge, 209–261.

Slobin, Dan Isaac (ed.), 1985. *The Crosslinguistic Study of Language Acquisition, Volume 1: the Data*, Hillsdale (NJ): Lawrence Erlbaum.

Slobin, Dan Isaac (ed.), 1992. *The Crosslinguistic Study of Language Acquisition, Volume 3*, Hillsdale (NJ): Lawrence Erlbaum.

Stang, Christian, 1966. *Vergleichende Grammatik der baltischen Sprachen*, Oslo: Universitetsforlaget.

Stiles, Patrick V., 1985. 'The fate of the numeral "4" in Germanic', *Northwestern European Language Evolution* 6, 81–104.

Taylor, Ann, Warnow, Tandy and Ringe, Don, 2000. 'Character-based reconstruction of a linguistic cladogram', in John Charles Smith and Delia Bentley (ed.), *Historical Linguistics 1995, Volume 1: General Issues and Non-Germanic Languages*, Amsterdam: Benjamins, 393–408. [Reports results as of 1995.]

Thomason, Sarah Grey and Kaufman, Terrence, 1988. *Language Contact, Creolization, and Genetic Linguistics*, Berkeley: University of California Press.

Tischler, Johann, 1973. *Glottochronologie und Lexikostatistik*, Innsbruck: Innsbrucker Beiträge zur Sprachwissenschaft.

Trubetzkoy, Nikolai, 1926. 'Gedanken über den lateinischen a-Konjunktiv', in *Beiträge zur griechischen und lateinischen Sprachforschung: Festschrift Kretschmer* (no ed.), Berlin: Deutscher Verlag für Jugend und Volk, 267–274.

Þórhallsdóttir, Guðrún, 1993. *The Development of Intervocalic* *j *in Proto-Germanic*, Ph.D. Dissertation, Cornell University.

Warnow, Tandy, Ringe, Don and Taylor, Ann, 1996. 'Reconstructing the evolutionary history of natural languages', *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 1996.

Watkins, Calvert, 1966. 'Italo-Celtic revisited', in Henrik Birnbaum and Jaan Puhvel (ed.), *Ancient Indo-European Dialects*, Berkeley: University of California Press, 29–50.

Winter, Werner, 1998. 'Lexical archaisms in the Tocharian languages', in Victor Mair (ed.), *The Bronze Age and Early Iron Age Peoples of Eastern Central Asia*, Washington, D.C.: Institute for the Study of Man, 347–357.

Yoshida, Kazuhiko, 1990. *The Hittite Mediopassive Endings in* -ri, Berlin: de Gruyter.