

PERFECT PHYLOGENETIC NETWORKS: A NEW METHODOLOGY FOR RECONSTRUCTING THE EVOLUTIONARY HISTORY OF NATURAL LANGUAGES

LUAY NAKHLEH
Rice University

DON RINGE
*University of
Pennsylvania*

TANDY WARNOW
University of Texas

In this article we extend the model of language evolution exemplified in Ringe et al. 2002, which recovers phylogenetic trees optimized according to a criterion of weighted maximum compatibility, to include cases in which languages remain in contact and trade linguistic material as they evolve. We describe our analysis of an Indo-European (IE) dataset (originally assembled by Ringe and Taylor) based on this new model. Our study shows that this new model fits the IE family well and suggests that the early evolution of IE involved only limited contact between distinct lineages. Furthermore, the candidate histories we obtain appear to be consistent with archaeological findings, which suggests that this method may be of practical use. The case at hand provides no opportunity to explore the problem of conflict between network optimization criteria; that problem must be left to future research.*

1. INTRODUCTION. Languages differentiate and divide into new languages by a process roughly similar to biological speciation:¹ communities separate (typically geographically), the language changes differently in each of the new communities, and in time people from separate communities can no longer understand each other.² While this is not the only way in which languages change, it is this process that is referred to when we say, for example, ‘French and Italian are both descendants of Latin’. The evolution of families of related languages can be modeled mathematically as a rooted tree in which internal nodes represent ancestral languages at the points in time at which they began to diversify and the leaves represent attested languages. Reconstructing this process for various language families is a major endeavor within historical linguistics, but it is also of interest to archaeologists, human geneticists, and physical anthropologists, for example, because an accurate reconstruction of how particular families of languages have evolved can help answer questions about human migrations, the times at which new technologies were first developed, when ancient people began to use

* This work was supported in part by the David and Lucile Packard Foundation (Warnow) and by the National Science Foundation with grants EIA 01-21680 (Warnow), BCS 03-12830 (Warnow), SBR-9512092 (Warnow and Ringe), and BCS 03-12911 (Ringe). Warnow would like to acknowledge and thank the Radcliffe Institute for Advanced Study, the Program in Evolutionary Dynamics at Harvard University, and the Institute for Cellular and Molecular Biology at the University of Texas at Austin for their support during the time this work was done. The authors would like to thank Ann Taylor for help in putting the dataset together and James Clackson, Joe Eska, and Craig Melchert for expert advice regarding data of particular languages. The software used to construct perfect phylogenetic networks was written by Luay Nakhleh but used earlier code (for perfect phylogeny reconstruction) developed by Alexander Michailov and optimized by Alex Garthwaite.

¹ We take this opportunity to point out that the similarity between biological and linguistic speciation has nothing whatsoever to do with nineteenth-century ideas about the ‘organic’ nature of language. The micro-level processes of biological descent and linguistic descent are actually quite different, but they give rise to similar large-scale patterns, and the similarities are topological—that is, mathematical (see Hoenigswald 1960:144–60, 1987, Ruvolo 1987).

² We are well aware that whether one is confronted with ‘the same language’ or ‘different languages’ is a complex matter. However, it seems difficult to dispute that two speakers who cannot understand one another at all are ‘speaking different languages’; we therefore adduce that situation as the paradigm case. What matters for cladistics is that, given enough divergence with too little effective contact, a single language will eventually become two or more different languages by any reasonable criterion.

horses, and so on (see e.g. White & O'Connell 1982, Mallory 1989, Roberts et al. 1990).³

Various researchers (e.g. Gleason 1959, Dobson 1969, 1974, Embleton 1986) have noted that if speech communities do not remain in effective contact as their languages diverge, a tree is a reasonable model for the evolutionary history of their language family, and that this tree (called a *PHYLOGENY* or *EVOLUTIONARY TREE*) can be inferred from shared unusual innovations in language structure (changes in inflection, regular sound changes, and the replacement of lexemes for basic meanings). Such techniques established the major subfamilies within Indo-European (IE) decades ago but have not been sufficient to resolve the family's evolution fully; major questions, such as whether all of the non-Anatolian branches of the family constitute a clade (the 'Indo-Hittite hypothesis') or whether Greek and Armenian are sisters, continue to be debated. More recently, techniques for using multistate characters have been devised which suggest that the vast majority of linguistic characters,⁴ provided that they are correctly chosen and coded, should be *COMPATIBLE* on the true tree (see Ringe et al. 2002:70–78 with references); in other words, each character should evolve without backmutation or parallel evolution.⁵ This condition is also expressed by saying that the tree is a *PERFECT PHYLOGENY*, that is, a phylogenetic tree that is fully compatible with all of the data. (See §2 for an extended discussion of those requirements.)

A collaboration between linguist Don Ringe and computer scientist Tandy Warnow led to a computational technique to solve the 'perfect phylogeny' problem (determining whether a perfect phylogeny exists for a given dataset); that technique was subsequently used to analyze an IE dataset compiled by Don Ringe and Ann Taylor (see the references under all three authors in the bibliography). Their initial test of the methodology largely supported the claim that a perfect phylogeny should exist, but not entirely. The Germanic subfamily especially seemed to exhibit nontreelike behavior, evidently acquiring some of its characteristics from its neighbors rather than (only) from its direct ancestors.⁶

³ Readers who have not been trained in historical linguistics also need to understand that recognition of language families is different from and independent of the reconstruction of phylogenetic trees, and that the recognition of cognates—words and affixes inherited by two or more related languages from any common ancestor—also does not depend on prior knowledge of the true tree. Cognates are recognized by the regular correspondences between their sounds that are the direct result of regular sound changes; see especially Hoenigswald 1960 for discussion of this fundamental point. Cognates cannot be reliably recognized by mere similarity. Language families are recognized by a density of putative cognates too great to be attributed to mere chance resemblance; see Ringe 1999 for some of the problems involved.

⁴ A character is a linguistic parameter in which languages can agree or differ; languages are assigned the same state of the character if they agree, but different states if they differ. Characters of interest in linguistic phylogeny are highly specific, since general characters (such as word order) typically reveal much less about shared linguistic history. For instance, the basic meaning 'hand' can be chosen as a character; IE languages that exhibit cognates of English *hand* will all be assigned a single state of that character, languages that exhibit cognates of French *main* a second state, and so on. Among phonological developments, across-the-board merger of Proto-Indo-European (PIE) *m and *mb^h can be chosen as a character; the two Tocharian languages (which share that merger) will then be assigned one state, while all the other IE languages in our database (which did not undergo the merger) will be assigned another state. On the coding of characters see further Ringe et al. 2002:71–76.

⁵ Backmutation is the reappearance at some point in the phylogenetic tree of a state that has already appeared at some earlier point in the same line of descent but was subsequently lost; in other words, a sequence of states $a \rightarrow b (\rightarrow c \dots) \rightarrow a$ in a single line of descent is backmutation. Parallel development is the appearance of the same state independently in different lines of descent.

⁶ We wish to emphasize that this appears to be an ineluctable conclusion of Ringe et al. 2002; we see no grounds for questioning it and do not revisit the problem here. Interested readers are referred to Ringe et al. 2002, especially pp. 85–92. Since the best tree found in that earlier work also figures largely in this

Consequently, though their methodology seemed promising and offered potential answers to many of the controversial problems in the evolution of IE (cf. Jasanoff 1997, Winter 1998, Ringe 2000 with references), it is clearly necessary to extend their model to address the problem of how characters evolve when diverging language communities remain in significant contact. For such cases trees are not an appropriate model of evolution; NETWORKS are needed instead to model the evolutionary history of the family.

In this article we show how to extend the perfect phylogeny approach to the case in which the language family requires a network model (that is, an underlying tree with additional ‘contact’ edges; see Fig. 3 for an example) instead of a tree model, and we test this approach on the same IE dataset analyzed by Ringe, Warnow, and Taylor. Our analysis finds several networks with a very small number of contact edges that are plausible with respect to what is known about the early linguistic geography of the IE family. The study thus leads us to conjecture that the IE family, though it did not evolve by means of clean speciation, exhibits a pattern of initial diversification that is close to treelike: the vast majority of characters evolve down the ‘genetic’ tree, and the evolution of the rest can be accounted for by positing limited borrowing between languages. It also suggests that this extended model of character evolution is plausible and that the tools we have developed may be helpful in reconstructing evolutionary histories for other datasets that are similarly close to treelike in their evolution.

The rest of this article is organized as follows. We review the model of Ringe and Warnow, and then present our extension to the case of network evolution. We next describe the data we use to represent the IE family, and then turn to our computational analysis of the data which results in the candidate networks we then consider. Comparing the candidate networks in the light of known IE history produces a set of five feasible solutions, leading to a detailed discussion of the best network that we find. We conclude with a discussion of the implications of this work for future research in IE and general historical linguistics. Notes on the formal mathematical model of language evolution on networks and the computational approach are given in Appendix A. The full set of our coded data, together with a list of characters omitted and the reasons for their omission, are made available in an online appendix at <http://www.cs.rice.edu/~nakhleh/CPHL>; a selection is given in Appendix B.

2. INFERRING EVOLUTIONARY TREES. An evolutionary tree, or phylogeny, for a language family S describes the evolution of the languages in S from their most recent common ancestor. Different types of data can be used as input to methods of tree reconstruction; QUALITATIVE CHARACTER data, which reflect specific observable discrete characteristics of the languages under study, are one such type of data. Qualitative characters for languages can encode phonological, morphological, and lexical evidence, as described immediately below. Current approaches for subgrouping used in historical linguistics explicitly select characters that appear to have evolved without backmutation or parallel development; because of this, our analysis is based on a subset of the characters (eliminating those with clear parallel development, in particular). We have also found it advisable to eliminate characters that are POLYMORPHIC (those for which at least one language exhibits more than one state) because models of linguistic evolution involving polymorphic characters that are (at least provisionally) accepted as linguistically realistic have not yet been established.

project, we also wish to emphasize that the method by which the tree was found is a heuristic using the criterion of weighted maximum compatibility to optimize the tree; thus one of the trees discussed in this paper is the result of rigorous cladistic analysis.

Experience shows that it is easy to construct a comparative dataset using only qualitative characters that evolve without backmutation—that is, characters that never change from a given state to a second state (and potentially to a third, etc.) and then back to the given state (see Ringe et al. 2002:70). The relative absence of backmutation in linguistic data is partly the result of known properties of linguistic systems and language change and partly the result of probabilistic factors. Backmutation in phonological characters is easy to avoid: since phonemic mergers are irreversible (Hoenigswald 1960: 75–82, 87–98), one can base one’s phonological characters on mergers.⁷ One might suppose that inflectional morphology is not so well behaved, since there seems to be no comparable reason why backmutation should not occur. But in fact the only cases we can find are those in which an entire inflectional category has been acquired and then later lost again; obvious examples are the innovative nominal cases of Old Lithuanian, which do not survive in the modern language, and the superlative of adjectives in Latin, which is clearly an innovation (from the point of view of Proto-Indo-European (PIE)) but does not survive in most Romance languages. It is not difficult to exclude such characters from the dataset; alternatively, one can allow for their unusual pattern of development by coding each language that lacks the inflectional category with a unique state. (In the last example given, PIE would be assigned state 1 (no superlative), Latin state 2 (superlative in *-ismo-), and the Romance languages that have lost the Latin category would be assigned not state 1 but states 3, 4, 5, and so on—a separate state for each language.) Otherwise inflectional characters do not seem to exhibit backmutation, apparently because inflectional systems are so complex and idiosyncratic that the same configuration of inflectional markers is very unlikely to arise more than once independently. Even among lexical characters we have not been able to find clear instances in which the usual word X for a given meaning was completely replaced by a different word Y which was then completely replaced by X again.⁸ In this case, however, the reason seems to be probabilistic rather than structural. When an old word is replaced, the choice of replacement is more or less open-ended; both native words and loanwords in a wide range of related meanings are reasonable candidates. The probability that a word that was the usual word for the same concept in earlier centuries would be chosen as the new replacement is probably always very small.

But if backmutation is not a problem, parallel development most certainly is. Most individual sound changes are ‘natural’ phonetic developments, in the sense that they probably originate as responses to universal phonetic pressures, and they can recur at widely separated times and places (see e.g. Ringe et al. 2002:66–68). Phonological characters must therefore be based on highly unusual sound changes, or on sets or sequences of sound changes that are not especially likely to have occurred together, and this greatly decreases the amount of phonological information that can be used for cladistic purposes. Parallel shifts in the meanings of words are also very common.

⁷ There are also some sound changes that do not usually seem to be ‘undone’ even though they do not involve merger; for instance, while the palatalization of velars by immediately following front vocalics is commonplace, we cannot find a well-authenticated instance in which palatal consonants have become velars. (For precisely that reason it is unlikely that the PIE ‘palatal’ stops, which developed into velars in numerous daughter languages, were phonetically palatal, as Michael Weiss observed to one of the authors in the late 1980s.) A much more unusual case is the Italo-Celtic sound change $*p \dots k^w > *k^w \dots k^w$, in which the latter state must be innovative because it violates a PIE constraint on the shape of roots (Ringe et al. 2002: 100).

⁸ Our lexical characters are defined semantically, each cognate set comprising a single state of the character, for the reasons outlined in Ringe et al. 2002:71, n. 8.

Fortunately the collective experience of historical linguists seems to show that most parallel developments in lexical datasets can be detected. In fact, detection of parallel semantic developments is so much a part of the everyday work of historical linguists that there is no general discussion of it in the recent literature; one can get a good idea of what is known by perusing the entries of Buck 1949. Only in inflectional morphology does parallel development appear to be uncommon (though replacement of an unusual or 'marked' inflectional affix by the default affix for that inflectional category can occur repeatedly, as Bill Poser reminds us). In the absence of a realistic stochastic model of language change, the only straightforward way to deal with this problem is to exclude from the dataset characters that exhibit parallel development anywhere in the tree, and this also reduces the amount of usable evidence substantially.⁹

Borrowing of states between significantly different languages is also a problem, but one that has been seriously overestimated in the recent literature. The assumption that 'anything can be borrowed' from one language into another has been given wide currency by Thomason & Kaufman 1988, but many who cite that work have paid too little attention to its authors' clear conviction that the borrowing of inflectional morphology is difficult and rare (cf. Thomason & Kaufman 1988:3–4, 74–76). In fact, recent work suggests that it is even rarer than Thomason and Kaufman claim. For instance, fieldwork by Maarten Mous has revealed that all speakers of the famous 'mixed' language Ma'a are actually native speakers of an ordinary Bantu language who use Ma'a as an 'inner' language to exclude outsiders, and further that the Cushitic vocabulary of Ma'a is taken from two rather distantly related Cushitic languages (Mous 1996, 1997 with references); it follows that, however Ma'a arose, it cannot be a Cushitic language that somehow borrowed Bantu morphology (pace Thomason & Kaufman 1988:223–28). Ruth King has concluded, also on the basis of fieldwork, that English influence on the French of Prince Edward Island is mediated by lexical borrowing and cannot reasonably be characterized as the borrowing of morphology and other closed-class items. King cautions that none of the claimed cases of the borrowing of morphology into a native dialect is based on evidence sufficient to prove that that is what really happened (see King 2000:44–47 with references, 2003; cf. also Appel & Muysken 1987:158–63).

Of course it is true that we find, for example, a Norse pronoun in the modern descendant of Old English, and Hebrew nouns with Hebrew plurals in a modern descendant of Middle High German. But research on language contact shows that the transfer of closed-class items from language to language typically occurs via processes quite different from the casual borrowing of foreign words into one's native language. For instance, work on the English spoken by native speakers of Yiddish shows that bilinguals with only one native language typically import closed-class items from their native language into the language they learned imperfectly as adults, not the other way around (cf. Rayfield 1970:103–7, Prince & Pintzuk 2000). King's own work has demonstrated that the apparent borrowing of English morphosyntax into the French of Prince Edward Island is actually something different, more complex, and much more interesting: English lexemes have been borrowed in the usual way, some of them bring with them

⁹ Reliably inferring phylogenies in the presence of substantial parallel development will require a realistic stochastic model of the evolution of linguistic character sets, a problem that we are addressing in other work (see Warnow et al. 2005). In this article we concentrate on borrowing as an explanation for the incompatibility of characters on a tree not because we regard borrowing as a more plausible hypothesis, but because we wish to tackle one problem at a time; since this one is easier, we have tackled it first. However, it is true (as a referee observes) that borrowing is a more economical hypothesis when a single contact event can account for a language's acquisition of multiple lexical items.

specific morphosyntactic features that are unusual in French, and the resulting perturbation of native morphosyntax gives rise to new syntactic patterns—which are NOT identical with English patterns, but are similar enough that an unsophisticated approach to syntax will not find the differences (see especially King 2000). Most importantly, it appears that the processes that give rise to the transfer of inflectional morphology are disruptive enough to leave clear diagnostic traces in the structure of a language.

Since inflectional morphology does not seem to be transferred laterally from language to language in typical contact situations (cf. Thomason & Kaufman 1988:74–76, Ross 1997:209–10), we think it reasonable to suppose that no such transfer has occurred in the absence of persuasive structural evidence to the contrary. We therefore expect to find detectable evidence of contact only among lexical characters and phonological characters in the default case. Moreover, though it is clear that sound changes do spread through diversifying dialect continua, lexical borrowing seems to be much commoner and more unconstrained; we therefore expect to find the bulk of the evidence of contact among lexical characters. So far as we can tell, this is the pattern that we do find in our IE dataset. Detectable loanwords are coded with unique states, since if they were coded with the same states as the ‘lending’ language the tree-inferring algorithm would interpret that configuration as shared inheritance of states by normal linguistic descent.¹⁰ But it is reasonable to suspect that not all loanwords are detectable as such; in particular, borrowed words that happen not to involve distinctive sound changes characteristic of either the lending language or the borrowing language can be difficult to detect.

Finally, there are good reasons for excluding polymorphic characters from the dataset. The most compelling conceptual reason is that the evolution of polymorphic linguistic characters has never been investigated in detail. In an early attempt to do so, Bonet et al. 1996 modeled polymorphism as a consequence of semantic shift, but it now seems clear that polymorphism can arise from borrowing as well. Whether those two processes are sufficient to describe the full extent of polymorphism, or whether additional sources must be taken into account, remains unclear. Thus in a real sense we do not know what constraints polymorphic characters imply about the underlying evolutionary history. Until such an understanding can be reached, it seems advisable to exclude all polymorphic characters from this analysis, and we have done so.

If backmutation and parallel development have been excluded, and if there is no undetected borrowing between lineages, an important property of character-state change follows: when the state of a qualitative character changes in the evolutionary history of the set of languages, we expect it to change to a state that exists nowhere else at that time and has not appeared earlier. We express this property by saying that all usable qualitative characters in a historical linguistic analysis should be compatible on the true evolutionary tree, provided that the characters are carefully analyzed (so that the determinations of cognate classes are correct, for example) and are properly selected (so as to properly code characters that have clearly undergone borrowing in their history). This observation, made first by Gleason (1959) and Dobson (1969) and eventually elaborated by Ringe and Warnow, is already implicit in the ‘comparative method’ as formalized by Hoenigswald (1960).

The problem of reconstructing the evolutionary history of a set of languages can thus be described as the search for a tree on which all of the characters in the dataset

¹⁰ The points discussed in this paragraph and the ones immediately above have been treated at greater length in Ringe et al. 2002.

are compatible; such a tree, if it exists, is called a perfect phylogeny. A perfect phylogeny should exist so long as the data evolve in the fashion described above AND the evolution of the language family has been treelike (i.e. with ‘clean’ speciations).

But this approach obviously cannot be relied on to reconstruct evolutionary histories for those language families in which related dialects have evolved in close contact with each other; in such cases the evolutionary divergences may not be sufficiently ‘clean’. More precisely, whereas borrowing between clearly different speechforms (i.e. ‘different languages’ or divergent dialects) is reasonably tightly constrained and clearly different from change in normal genetic descent, borrowing between closely related dialects seems to be largely unconstrained and is often indistinguishable from changes that could in principle be of very different types (see e.g. Labov 1994, Ross 1997). In such cases a tree model is inappropriate, and the evolutionary process is better represented as a network.

The initial analysis of a comparative IE dataset by Warnow and Ringe (first reported in Warnow et al. 1995; substantially revised and augmented in Ringe et al. 2002) in fact failed to find a perfect phylogeny. They demonstrated that the IE linguistic data are ‘almost perfect’: that is, an overwhelming majority of the characters were compatible with a single evolutionary tree, but by no means all characters were. The problem seemed to be the Germanic subfamily, which appeared to have remained in contact with other languages early in its evolution so that a tree was an inappropriate model of that evolution. In other words, part of the IE family, but only a part, must have evolved otherwise than through clean speciation.

3. PERFECT PHYLOGENETIC NETWORKS. In this section we show how we have extended the model of character evolution on trees to produce a model of how characters should evolve down networks. That is, we show how we can define PERFECT PHYLOGENETIC NETWORKS (PPNs) by extending the perfect phylogeny concept to the network case.

The evolution of a family of languages, when the languages evolve via clean speciation, is modeled as a rooted tree (typically bifurcating), so that internal nodes represent ancestral languages and leaves represent the languages under study. In this case, it is reasonable to orient edges from the ancestral languages toward the descendant languages so that all of the edges in the tree are directed (from the root toward the leaves); these directions are consistent with the flow of time. However, when languages evolve in such a way as to be able to borrow from each other when they have not yet diverged very much, then additional edges are needed in order to show how characters evolve in the network. Since these edges represent exchanges between languages due to contact, we call them ‘contact edges’. These edges are ‘bidirectional’, so that characters can be borrowed in both directions. Such a graphical representation is called a network rather than a tree, to reflect the inclusion of these additional edges. Let us examine the construction of a PPN and the reasons why we would wish to construct one.

Figure 1 lists states of characters for five hypothetical languages, L_1 through L_5 ; each character is binary, potentially exhibiting states 0 and 1. In Fig. 1a there are only two characters, c_1 and c_2 ; both are compatible on a single tree, so that the perfect phylogeny T shown in Figure 2 can be constructed for this dataset. The character states for each node are given; if this were a real example, the states at the terminal nodes (‘leaves’) would be coded from actual data, while those at the internal nodes would be inferred. In Fig. 1b a third character, c_3 , has been added, and because of the distributions of the states of the three characters, there exists no perfect phylogeny for this dataset. However, Figure 3 shows a network N that contains, in addition to the underlying tree T (as in Fig. 2), one contact edge between two reasonably closely related

	c_1	c_2
L_1	0	0
L_2	0	0
L_3	0	1
L_4	1	1
L_5	1	1

	c_1	c_2	c_3
L_1	0	0	0
L_2	0	0	1
L_3	0	1	1
L_4	1	1	0
L_5	1	1	0

a. Five languages $L_1 \dots L_5$ with two characters c_1 and c_2 .

b. The same five languages with a third character c_3 .

FIGURE 1.

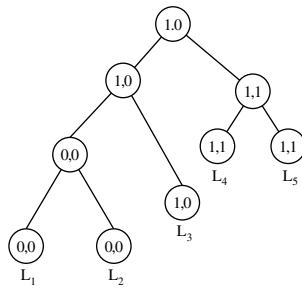


FIGURE 2. Perfect phylogeny T for the languages and character states of Fig. 1a.

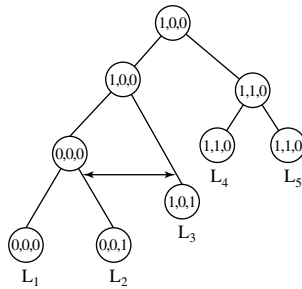


FIGURE 3. Perfect phylogenetic network N for the languages and character states of Fig. 1b.

languages, L_2 and L_3 . This network can be resolved into three alternative trees, as in Figure 4. One of these trees is the genetic tree T of Fig. 2; in each of the others, the contact edge replaces one of the edges of the genetic tree. Each character can evolve down at least one of these three trees; that is, each of the three trees potentially models the evolutionary history of one or more of the characters. The character that exhibits a borrowed state must have evolved down one of the alternative trees that includes the contact edge.

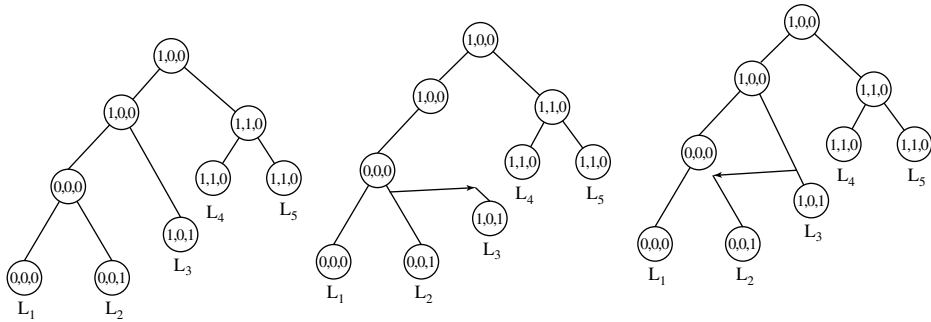


FIGURE 4. The three trees contained inside the network in Fig. 3. While the network ‘reconciles’ the evolutionary history of all three characters, each one of those characters actually evolved down exactly one of the three trees.

To motivate our model of character evolution down networks we start from the tree model. When a tree is a reasonable model of a language family’s evolution we assume that qualitative characters are compatible with the tree: when a character changes state on an edge, it changes to a new state not yet in the tree (backmutation and parallel evolution having been excluded). But a network contains several evolutionary trees, and a character can evolve down any of them. We therefore say that a qualitative character c is compatible with a network N if c is compatible with at least one of the trees contained in N .

The three characters of the network N shown in Fig. 3 are compatible with N , since each of the characters is compatible with at least one of the three trees contained in N : characters c_1 and c_2 are compatible with the first tree in Fig. 4, and character c_3 is compatible with both the second and third trees in Fig. 4.

The assumptions inherent in the methodology of linguistic cladistics, when extended to the case where languages evolve on a phylogenetic network, imply that linguistic characters should be compatible on the true phylogenetic network. Just as in the case of trees, we say that a network is a PPN for a set of languages described by a set of qualitative characters if every character is compatible with the network (as in the example given immediately above).

A PPN can deviate from the tree model in a number of ways. The greater the number of characters that must evolve along lateral (i.e. contact) edges, the greater the deviation in terms of character compatibility; the greater the number of lateral edges, the greater the deviation in topological terms. Finally, the greater the number of borrowing events (i.e. character states transmitted on contact edges), the greater the deviation in terms of what might be called LOAN PARSIMONY; this is not the same as either of the measures just mentioned, since a single item can be borrowed along more than one lateral edge.

Note that the three criteria just enumerated measure different things; hence different ways of deviating from a tree are evaluated differently depending on the criterion

chosen to evaluate the network. For example, if the wave model is a fully realistic description of how the languages have diverged, we should not expect to be able to find a clearly defined underlying genetic tree on which the vast majority of the characters are compatible;¹¹ most characters may be involved in ‘borrowing’, and so even if the number of contact edges is fairly small (because of geographic constraints, for example), the other two criteria should yield fairly poor values (i.e. the proportion of characters simultaneously compatible will be low and the number of borrowing events will be high). By contrast, if the language family diversifies in a mostly treelike fashion (most of the characters evolving down the underlying genetic tree) AND the number of contact edges is quite small, then most or all of the genetic tree should be discernible because of the clear compatibility pattern. In such a case we should find only a small number of trees, with at least one of which a large percentage of the characters is compatible; moreover, those trees could appear to be alternative genetic trees only because we might, in some cases, find it difficult to distinguish tree edges from contact edges. (For instance, such a difficulty might be encountered in clades for which the principal evidence is lexical.) As the number of contact edges increases, although it might still make historical sense to speak of a genetic tree, it can be difficult to discern the tree. Nevertheless, the high proportion of characters that are simultaneously compatible will distinguish this situation from the situation in which the wave model is the only appropriate model.

The debate about the prehistoric diversification of the IE family has to some extent focused on the two extremes: the wave model, in which there is no clear underlying genetic tree, and the Stammbaum model, in which there is no significant borrowing. Our proposal explicitly takes account of intermediate possibilities, in which there is a clear genetic tree (with which a high proportion of characters is compatible) but also some borrowing. Furthermore, our proposal allows the deviation from a tree model to be measured along several partly independent parameters. (The number of contact edges and the number of incompatible characters are independent, but the third measure, loan parsimony, involves a combination of those two.) In this article we both present our model and attempt to discover where the IE family fits within the space of possibilities defined by the model. Just how clearly can we identify an IE genetic tree? Is the evolution of IE largely treelike, or is the wave model really a better model in this case? Our analysis shows dramatic support for the claim that the diversification of IE was largely treelike: almost all (95%) of the characters evolve down our proposed genetic tree, and we need only three additional contact edges to explain all the data; thus all three criteria yield satisfyingly low scores. Finally, our proposed network is also largely consistent with known geographical and chronological constraints on IE linguistic pre-history.

Since the simplest model is the most desirable, other things being equal (‘Occam’s razor’), we want to find a PPN that optimizes all three mathematical criteria, with the smallest number of borrowing events, the smallest number of contact edges, and the highest percentage of characters compatible on the underlying genetic tree. Such a

¹¹ The wave model of language divergence, associated with the late-nineteenth-century linguist Johannes Schmidt, is conceived of as an alternative to the tree (or Stammbaum) model, couched in completely different terms. The wave model treats diverging speechforms as dialects in contact that can be mapped geographically; linguistic changes spread across the map from dialect to dialect in patterns that can overlap. It seems clear that some types of linguistic divergence are best represented by the wave model and others by the tree model; it also seems clear that there are patterns of divergence resulting from many centuries of development that can be accommodated by either model.

network would explain the evolution of all of the characters (via genetic transmission and/or borrowing) and would not need to imply either parallel evolution or backmutation. It is worth noting that a PPN always exists, because one can always construct a network in which all pairs of leaves are connected by contact edges; on such a network all characters are compatible. But since such a network fits all possible characters, it says nothing interesting about the evolutionary history of the dataset.

Finding the smallest number of borrowing events is obviously easier if one has first found (or estimated) the tree with which the largest number of characters is compatible and has added to it the minimum number of contact edges necessary to construct a PPN. Accordingly our approach involves two steps.¹²

- Given the set L of languages described by set C of qualitative characters, find or estimate the optimal genetic tree T for L . The optimal tree is compatible with all required characters—that is, all characters whose states cannot normally be acquired by contact—and compatible with a maximum number of other characters. (If T is a perfect phylogeny, or deviates very little from a perfect phylogeny, it can be found; if it deviates too much from a perfect phylogeny, existing techniques may be insufficient to prove that the best tree discovered is in fact the optimal tree. See the discussion in Ringe et al. 2002:78–80.)
- If T is a perfect phylogeny, then return T . Otherwise, add a minimum number of contact edges to T to make it a PPN.

For example, the characters c_1 and c_2 in Fig. 1b are compatible with the tree T in Fig. 2, whereas character c_3 is not. By adding one contact edge to T , we obtain the network N of Fig. 3, on which all three characters are compatible.

This is the approach that we used in this study in order to analyze the IE dataset compiled by Ringe and Taylor. Because our data were close to treelike, our analysis was able to complete in a reasonable amount of time (a few hours). We describe that analysis in the next section.

4. THE INDO-EUROPEAN DATASET. Our basic dataset consists of 294 characters for twenty-four IE languages. We first describe and explain our choice of languages and characters, then describe our coding of the characters.

The languages of our dataset are listed in Table 1, in approximate chronological order of the attestation of their subfamilies, together with the abbreviations by which they are represented in our trees.

LANGUAGE	ABBREVIATION	LANGUAGE	ABBREVIATION
Hittite	HI	Old English	OE
Luvian	LU	Old High German	OG
Lycian	LY	Classical Armenian	AR
Vedic	VE	Tocharian A	TA
Avestan	AV	Tocharian B	TB
Old Persian	PE	Old Irish	OI
Ancient Greek	GK	Welsh	WE
Latin	LA	Old Church Slavonic	OC
Oscan	OS	Old Prussian	OP
Umbrian	UM	Lithuanian	LI
Gothic	GO	Latvian	LT
Old Norse	ON	Albanian	AL

TABLE 1. The twenty-four IE languages analyzed.

¹² This is roughly similar to the approach of Alroy 1995.

They represent all ten well-attested subgroups of the IE family: Anatolian (represented by Hittite, Luvian, and Lycian); Tocharian (A and B); Celtic (Old Irish and Welsh); Italic (Latin, Oscan, and Umbrian); Germanic (Gothic, Old Norse, Old English, and Old High German); Albanian; Greek; Armenian; Balto-Slavic (Old Church Slavonic, Old Prussian, Lithuanian, and Latvian); and Indo-Iranian (Vedic, Avestan, and Old Persian). In general, we have chosen as a representative of each subgroup a language or languages that are attested relatively fully at the earliest possible date. Thus Indic is represented by Vedic, since the Rigveda and the other Vedic texts are extensive enough to provide data for most of our characters; but to represent eastern Iranian we use ‘younger’ Avestan rather than the earlier Gatha-Avestan, since the Gathas are too short a text for our purposes. Greek is represented by Classical Attic rather than the earlier Homeric not only because the attestation of Attic is far more extensive, but also because Homeric Greek is known to be an artificial literary dialect. The motivations for our other choices are similar. We use modern data for Welsh, Lithuanian, Latvian, and Albanian both because earlier data are much less accessible and because we believe that it would make little difference in those cases; in particular, even the earliest documents in Albanian and the Baltic languages are only a few centuries old. It can be proved that none of the languages in our database is the ancestor of any other, because each has undergone irreversible changes (such as phonemic mergers) not shared by any of the others.

Because our method is character-based, not distance-based, the fact that the languages of our database are not contemporaneous has no negative effect on the results; all that matters is whether the states of each character fit the branching structure of the tree. In fact it is to our advantage to use the earliest attested languages, since these are more likely to have retained character states that are informative of the underlying evolutionary history. By contrast, distance-based methods, since they are required to work from contemporaneous languages, must use comparatively less phylogenetically informative data in some cases.

In order to represent as many of the major subgroups as was feasible we decided to use some fragmentarily attested ancient languages for which only a minority of the lexical characters could be filled with actual data. The languages in question are Lycian (for which we have only about 15% of the 259-word list), Oscan (roughly 20%), Umbrian (ca. 25%), Old Persian (ca. 30%), and Luvian (ca. 40%). At the other extreme we have complete or virtually complete wordlists (including at least 99% of the lexical items) not only for the modern languages but also for Ancient Greek, Latin, Old Norse, Old English, and Old High German; we also have nearly complete wordlists (including at least 95% of the items) for Vedic, Classical Armenian, Old Irish, and Old Church Slavonic. The remaining wordlists each exhibit between about 70% and about 85% of the list items. Gaps in the data are coded with unique states, which are compatible with any tree; therefore, though they do not cause problems for our method, they do decrease the robustness of certain subgroups. That is, of course, realistic.

The inclusion of all three Baltic languages and of four Germanic languages introduces parallel development in a fairly large number of lexical characters, and that decreases the amount of evidence that this method can use. Nevertheless we have retained those languages in the database because the internal subgrouping of Balto-Slavic and of Germanic are matters of interest to the specialist community.¹³ By contrast, including only two West Germanic languages—Old English and Old High German, the northern-

¹³ We have no reason to doubt the cladistic structures of these subgroups found in Ringe et al. 2002, which are very robust and are in each case consistent with one of the standard alternative opinions.

most and southernmost respectively—potentially avoids much greater character incompatibilities, since the internal diversification of West Germanic is known to have been radically nontreelike (cf. Ringe et al. 2002:110).

Our database includes twenty-two phonological characters encoding regular sound changes (or, more often, sets of sound changes) that occurred in the prehistories of various languages in the database, thirteen morphological characters encoding details of inflection (or, in one case, of word formation), and 259 lexical characters defined by meanings on a basic wordlist. (See the online appendix at <http://www.cs.rice.edu/~nakhleh/CPHL> for a complete list of these characters and their coding.) The data were assembled by Don Ringe and Ann Taylor, who are specialists in IE historical linguistics, with the advice of other specialist colleagues. The characters were chosen in the following way.

Ringe and Taylor attempted to find sound changes and sets of sound changes unlikely to be repeated that are shared by more than one major subgroup of the family; they were able to discover only three plausible candidates, which are our first three phonological characters. The remaining phonological characters define various uncontroversial subgroups of the family. It would have been possible to find many more such phonological characters and/or to use even larger sets of sound changes for some of them, but nothing would have been gained. Because so few probably unrepeatable sound changes are shared by more than one uncontroversial subgroup, the question of whether the characters chosen might favor or disfavor construction of a phylogenetic tree did not arise. For a detailed presentation of the phonological characters see Ringe et al. 2002:113–16.

In attempting to find viable morphological characters Ringe and Taylor made a surprising discovery: the states of most morphological characters either are confined to a single major subgroup or are characteristic of the family as a whole, rendering them useless for the first-order subgrouping of the family. Several such morphological characters appear in our database, either because they are important inflectional markers or because they are useful for establishing one or more of the major subgroups. The database also includes all of the morphological characters that might aid in determining the first-order subgrouping that Ringe and Taylor were able to discover. These characters were chosen without regard to their possible compatibility with a phylogenetic tree. For details see Ringe et al. 2002:117–20.

Assembly of the lexical part of the database proceeded somewhat differently. Ringe and Taylor began with a version of the Swadesh 200-word list, since that is a standard comparative lexical set; to it they added about 120 meanings that appear to be culturally significant for older IE languages. These characters, too, were chosen without regard for their possible compatibility with a phylogenetic tree.

The database just described was the basis of the analysis reported in Ringe et al. 2002. For the present project we have modified it in the following ways. We exclude all polymorphic characters from the dataset (for the reasons outlined above); we even exclude those polymorphic characters that can be recoded as pairs of monomorphic characters (see Ringe et al. 2002:84–85). We also exclude all characters that clearly exhibit parallel development (whether or not they are compatible with any plausible tree). Those exclusions are the reason why we use fewer morphological characters, and many fewer lexical characters, than Ringe et al. 2002. (For further discussion of the reasons why individual characters were excluded see the online appendix at <http://www.cs.rice.edu/~nakhleh/CPHL>.)

We assigned character states to the languages in our dataset as follows. In the case of the phonological characters, a language either has or has not undergone a regular

sound change (or set of regular sound changes) at some point in its prehistory; it is assigned one state if it has and another if it has not, so that phonological characters normally exhibit two states each. For all other characters, states are assigned on the basis of cognation classes. Words and inflectional affixes in two or more related languages are said to be cognate if the languages have inherited them from their most recent common ancestor by direct linguistic descent. For each character, all of the members of each cognation class are assigned the same state; noncognate words and affixes are assigned unique states. We emphasize that all loanwords in a language are noncognates by definition, since they entered the language by a process other than direct, unbroken linguistic descent; thus they are assigned unique states. (But if reflexes of a word borrowed into language X appear in the daughters of X, they are coded as cognates for the clade including X and its daughters, since WITHIN THAT CLADE they have been transmitted by genetic descent.) Readers should also be aware that cognation cannot be determined by inspection; a knowledge of the principles of language change and of the individual histories of all of the languages is needed to make such a determination. More information about our coding of the data can be found in Ringe et al. 2002; here we discuss only two points of interest not noted above.

First, of the 294 characters we used in our phylogenetic analysis, 256 are *INFORMATIVE*, which means that they can help distinguish between candidate phylogenies. (An *UNINFORMATIVE* character, by contrast, is compatible with every tree; in other words, it is a ‘trivial acceptance’.) Second, a considerable number of lexical characters can reasonably be coded in more than one way, because of partial cognations between the items; an example is given in Ringe et al. 2002:82–83. (One morphological character is also double-coded for the same reason.) Double codings (or, in a couple of cases, triple or even quadruple codings) increase the number of characters without augmenting the independent available evidence. In consequence, of the 294 characters of our database, 242 are independent. This is still a very substantial number for a comparative linguistic database. Finally, we have weighted our characters in a maximally simple way. Every candidate tree is required to be compatible with all of the phonological characters but two (P2 and P3, which define the ‘satem’ subgroup and might either reflect shared descent in the strictest sense or have spread through a dialect continuum; see e.g. Hock 1986:442–44). Every candidate tree is also required to be compatible with eight of our morphological characters (M3, M5–6, M8, and M12–15).

This weighting scheme is, of course, only approximately realistic; in particular, we do not wish to imply that states of the required characters could not be transferred from lineage to lineage under any circumstances whatsoever. But since weighting those characters with large but finite values would give the same result, the simpler procedure seems preferable in this case.

5. PHYLOGENETIC ANALYSIS.

5.1. OVERVIEW. We analyzed five candidate genetic trees (identified at the beginning of §5.2). Tree A was constructed using Perfect Phylogeny software (described in Appendix A) and is optimal with respect to weighted maximum compatibility. Trees B and C can be obtained by modifying Tree A in linguistically plausible ways, but they are also among the possible trees with high compatibility weights (see Appendix A for further discussion). Trees D and E were suggested to us by our colleague Craig Melchert. We selected these trees in part because each is compatible with the vast majority of the characters—the best being compatible with more than 95% of the characters, while even the worst is compatible with 92%—and because all are also compatible with all of the morphological characters, which are expected to be the most resistant to borrow-

ing (cf. Meillet 1925:22–33, Ringe et al. 2002:65).¹⁴ For each tree we sought to add a minimum number of contact edges in order to produce a PPN; three edges sufficed for all but one of these trees (which needed more than three). We then scrutinized each of the resultant networks to consider whether the proposed episodes of contact were feasible on the basis of known constraints, both geographic and chronological, on the evolution of the IE family. This led us to reject all but five of the resultant PPNs (three on Tree A and two on Tree B). Of those five, one (on Tree A) was clearly the most plausible. Interestingly, this most plausible of numerous possible PPNs also optimized each of our mathematical optimization criteria (number of incompatible characters, number of contact edges, and number of borrowing events). Thus both mathematically and on the basis of known constraints one solution is superior to the others. That suggests strongly that the IE family evolved largely in a treelike fashion—sufficiently so that the underlying genetic tree is largely recoverable, and that specific contact episodes between neighboring linguistic communities can also be detected.

In the remainder of §5 we describe the candidate trees in detail, the differences between trees in terms of incompatibility patterns, and the PPNs based on these trees.

5.2. OUR CANDIDATE TREES. We analyzed five candidate genetic trees, three (Trees A, B, and C, shown in Figures 5 through 7) that our team has come to regard in recent years as worthy of serious investigation and two (Trees D and E, shown in Figures 8 and 9) that were suggested to us by Craig Melchert. The most important difference between these five trees is the placement of Germanic. The differing placement of Albanian is much less important, since Albanian can attach anywhere within a fairly large region of each tree with equally good fit; its variable placement is a result of the fact that it has lost nearly all of the diagnostic inflectional morphology (as well as a large proportion of inherited words on our wordlist). Thus each tree actually represents several trees which differ only with respect to exactly where Albanian attaches.¹⁵ By contrast, the variable placement of Germanic appears to reflect a major idiosyncrasy of that subgroup's evolution, leading our team to the conjecture that Germanic might not have evolved in a strictly treelike fashion (Ringe et al. 1998:407–8, Ringe et al. 2002:110–11). Our detailed analysis of these different scenarios allows us to test each of the histories conjectured for Germanic.

The differing positions of Germanic in the five trees result in different character-incompatibility patterns, as follows: for Tree A the fourteen incompatible characters are all lexical; for Tree B the nineteen incompatible characters include two phonological and seventeen lexical characters; for Tree C the seventeen incompatible characters are all lexical; for Tree D the twenty-one incompatible characters are all lexical; for Tree E the eighteen incompatible characters include two phonological and sixteen lexical characters. Interestingly, the incompatible characters for Tree A are a proper subset of the incompatible characters for Tree C. Therefore Tree C will represent a preferred hypothesis for the IE genetic tree only if we can complete Tree C to a PPN that improves upon our best PPNs for Tree A either by the remaining mathematical criteria (the number of contact edges or the loan parsimony criterion) or by significantly greater conformity to known chronological or geographic constraints on IE linguistic prehistory. The incompatible characters for Trees A and B are incomparable: twelve lexical

¹⁴ Note that our first three trees, which are based on the findings of Ringe, Warnow, and Taylor (2002), differ from trees published in their earlier work (such as Ringe et al. 1998) because their most recent publication uses a larger, and corrected, set of characters.

¹⁵ In each of the five trees in Figs. 5 through 9, Albanian can be shifted to any position within the region indicated by the thick lines (tree edges).

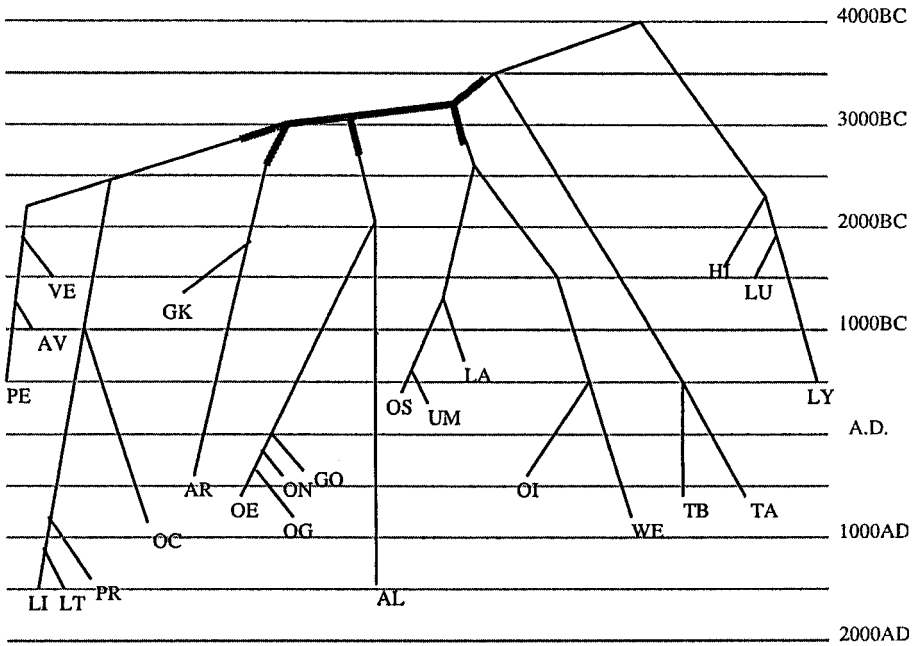


FIGURE 5. Tree A.^a

^a The thick lines represent the region within which Albanian can attach without changing the quality of the outcome (Figs. 5–9).

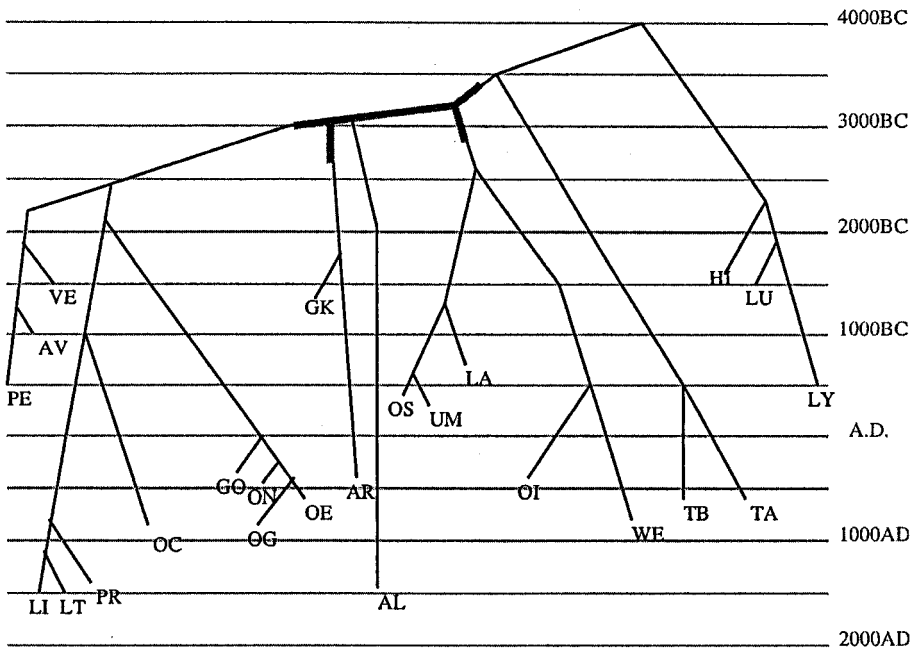


FIGURE 6. Tree B.

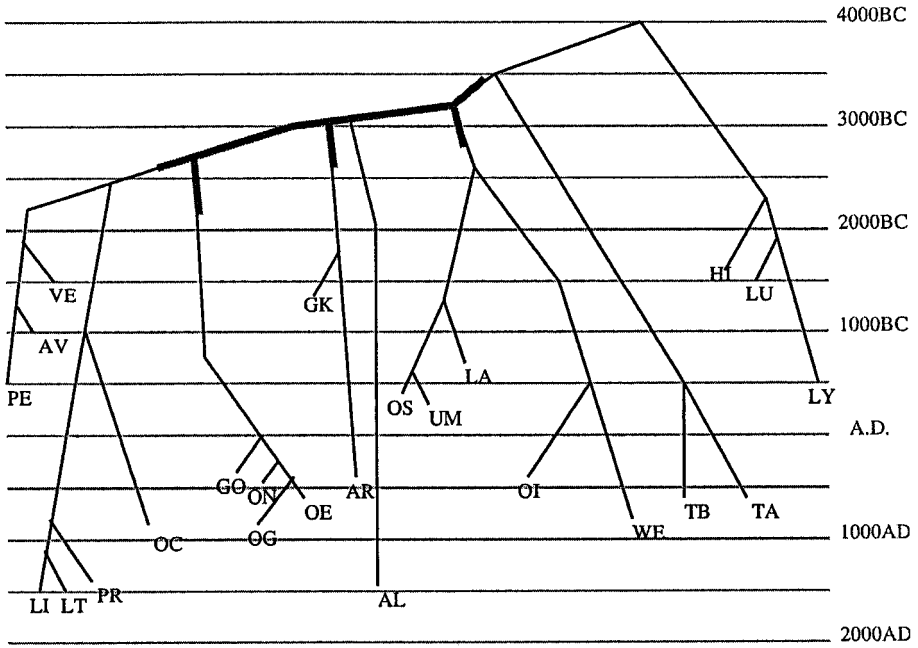


FIGURE 7. Tree C.

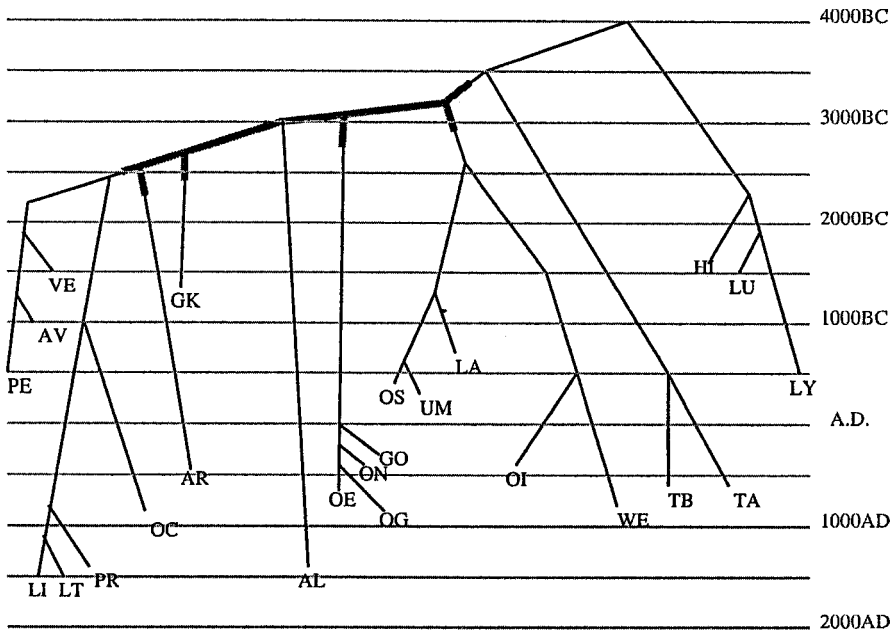


FIGURE 8. Tree D.

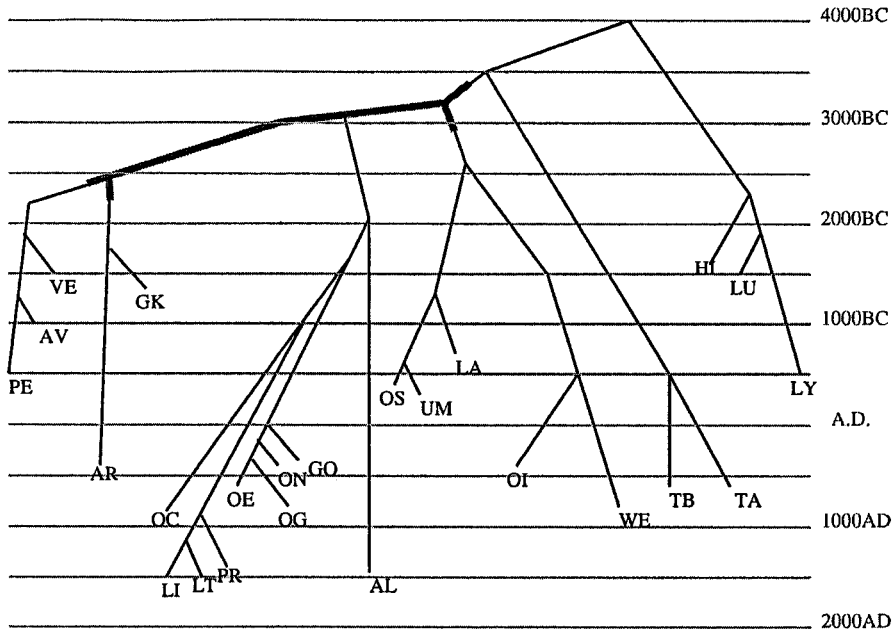


FIGURE 9. Tree E.

characters are incompatible on both trees, but two lexical characters are incompatible on Tree A and compatible on Tree B, and two phonological characters and three lexical characters are incompatible on Tree B but compatible on Tree A. Consequently, both Trees A and B remain roughly comparable candidates for the underlying genetic tree; a choice between them can reasonably be made only in the light of further analysis in which we extend them to PPNs. Tree D differs significantly from Tree A in not grouping Greek and Armenian together; Tree E differs from Tree A in its placement of Balto-Slavic. The incompatible characters for Tree A are a proper subset of the twenty-one incompatible lexical characters from Tree D; as in our comparison between Trees A and C, we will consider D a hypothesis to be preferred over Tree A only if we can complete Tree D to a PPN that improves upon our best PPN for Tree A in some way. The set of eighteen incompatible characters for Tree E, two phonological and sixteen lexical, is incomparable with the set of characters incompatible with Tree A.

5.3. CONSTRUCTING THE PPNs. The second phase of our analysis commenced after we had developed the algorithms and software to determine all of the ways we could add a minimum number of edges to a tree in order to obtain PPNs. The problem of adding a minimum number of edges to a tree to obtain a PPN is computationally difficult; consequently, we used a heuristic that allowed us to obtain solutions in a reasonable amount of time. Our technique for constructing a PPN from a tree is the following:

- Pruning the candidate tree. In this step we modified our candidate tree by replacing certain rooted subtrees (i.e. clades) by single nodes, as long as two conditions held: first, all characters are compatible below the root of the subtree, and second, there is linguistic evidence that suggests that undetected borrowing should not have occurred between a language in that clade and a language outside the clade. The subtrees of language groups that were replaced by their roots are: Germanic,

Celtic, Italic, Tocharian, Indo-Iranian, Anatolian, Greco-Armenian (for those trees in which Greek and Armenian are sisters, i.e. all of our trees other than Tree D), and Baltic. Albanian and Old Church Slavonic remain as individual languages.

- Adding the minimum number of contact edges. After the pruned tree was obtained, the algorithm searched for all of the ways we could add a minimum number of contact edges and get a PPN. This part of the analysis took eight hours on each of our candidate trees and found several networks with only three contact edges.

5.4. THE SET OF PPNs OBTAINED. Table 2 summarizes the quantitative results of our analysis of each of the five trees.

	TREE A	TREE B	TREE C	TREE D	TREE E
Number of incompatible characters	14	19	17	21	18
Minimum number of contact edges needed	3	3	3	3	>3
Number of 3-edge solutions found	16	4	1	2	0
Number of plausible 3-edge solutions	3	2	0	0	0

TABLE 2. Summary of results of phylogenetic analysis of the five IE trees in Figs. 5–9.

Our brute-force analysis produced twenty-three PPNs exhibiting only three contact edges each: sixteen PPNs based on Tree A, four based on Tree B, one on Tree C, two on Tree D, and none on Tree E. We compared the twenty-three PPNs in two ways:

- with reference to linguistic and archaeological evidence, which can render certain proposed contacts unlikely or even impossible (cf. Mallory 1989), and
- according to the mathematical criteria proposed in this paper: (i) the number of characters compatible on the genetic tree, (ii) the number of contact edges in the PPN, and (iii) loan parsimony, that is, the number of undetected borrowing events in the PPN.

Five of the twenty-three PPNs with only three contact edges are consistent with known geographical and chronological constraints on the prehistory of the family. Three of the five are based on Tree A, which is our preferred solution for the genetic tree of IE (based on the number of compatible characters); this is an encouraging convergence of results. It therefore seems reasonable to exclude PPNs with more than three edges based on any of these trees.

5.5. COMPARISON OF THE PPNs. We now compare the twenty-three PPNs we obtained. We first compare them with respect to geographic and chronological constraints, and then evaluate them with respect to mathematical criteria. Several interesting results emerge, of which the most striking is that the tree that optimizes the mathematical criteria is also the most clearly feasible with respect to the geographical and chronological constraints. We begin our discussion with our favored candidate for the genetic tree, Tree A (see Fig. 5). This is the tree found by cladistic analysis based on weighted maximum compatibility (not maximum parsimony) and published in Ringe et al. 2002.

PPNs BASED ON TREE A. The dates assigned to the terminal nodes of Tree A are obtained from historical data. Since our method is not distance-based, it is not necessary to adhere closely to the dates of the substantial corpora that underlie our lexical lists. Instead we here give the dates of the earliest documentary material from each language that shows clearly that it was already different from its closest relatives. By contrast, the dates assigned to the internal nodes are of necessity largely the result of informed guesswork, since proposed ‘glottochronological’ methods for determining the dates of prehistoric languages have proved to be unreliable (see especially Bergsland & Vogt 1962, none of whose objections have been effectively met by recent work; see Eska &

Ringe 2004 and Evans et al. 2005). Dates for a few internal nodes can be fixed with reasonable certainty.¹⁶ For instance, the complete archaeological continuity between the Yamna Horizon (up to ca. 2200 BC; Mallory 1989:211), its eastern Andronovo offshoot, and cultures known to have spoken Indo-Iranian languages allows us to place Proto-Indo-Iranian in the temporal vicinity of 2000 BC, give or take a couple of centuries (cf. the discussion of Mallory 1989:210–15, 226–29). Since Indo-Iranian is one of the most deeply embedded subgroups in the tree, it follows that all of the first-order branching must have occurred by that date. Most internal nodes, though, can be dated only within fairly wide ranges by a kind of linguistic ‘dead reckoning’ and must therefore be treated with caution.

The algorithm described above generated sixteen solutions for Tree A, three of which were consistent with known constraints on the prehistory of the IE family (cf. Mallory 1989). Those sixteen solutions are described in Table 3; the feasible solutions are in boldface in the table.¹⁷

SOLUTIONS	EDGE 1	EDGE 2	EDGE 3	min. # borrowing events in PPN
1	(PT,PS)	(PC,PBS)	(PC,PG)	19
2	(PIC,PB)	(PC,PG)	(PC,PGA)	20
3	(PT,PS)	(PI,PG)	(PI,PBS)	19
4	(PI,PG)	(PI,PGA)	(PIC,PB)	20
5	(PI,PG)	(PI,PGA)	(PG,PB)	17
6	(PT,PBSII)	(PI,PG)	(PI,PB)	19
7	(PT,PII)	(PI,PB)	(PI,PG)	18
8	(PT,PBS)	(PI,PB)	(PI,PG)	18
9	(PT,PS)	(PI,PB)	(PI,PG)	19
10	(PT,PB)	(PI,PB)	(PI,PG)	19
11	(PIC,PGA)	(PI,PB)	(PI,PG)	20
12	(PI,PB)	(PI,PGA)	(PI,PG)	20
13	(PC,PGA)	(PI,PB)	(PI,PG)	20
14	(PI,PG)	(PI,PB)	(PG,PGA)	20
15	(PI,PB)	(PAL,PGA)	(PI,PALG)	23
16	(PA,PBSII)	(PI,PG)	(PI,PB)	19

TABLE 3. The sixteen 3-edge solutions found on Tree A; the boldface rows (1, 3, and 5) correspond to the three feasible solutions. Each solution is described in terms of the three lateral edges added to Tree A to produce a PPN; the rightmost column gives the minimum number of borrowing events needed to make all characters compatible on the PPN.

The three feasible PPNs based on Tree A are solutions 1, 3, and 5 of Table 3. The first of these (solution 1; see Figure 10) is clearly more plausible than the second (solution 3; see Figure 11). Both posit a contact edge between Proto-Tocharian and Proto-Slavic. That is somewhat surprising, because it implies that an ancestor of Tocharian was still in eastern Europe in the last millennium BC, and it seems clear that by the turn of the millennium the speakers of (pre-)Proto-Tocharian were within striking distance of Xinjiang (where the Tocharian languages are actually attested from about

¹⁶ The type of reasoning employed in this and subsequent paragraphs, attempting to correlate linguistic events and historical and archaeological findings, has been commonplace in IE linguistics at least since Porzig 1954. The best summary of the relevant archaeological facts is Mallory 1989.

¹⁷ The abbreviations used for protolanguages in Tables 3 and 4 are the following: PA: Proto-Anatolian, PAL: pre-Albanian, PALG: Proto-Albano-Germanic, PB: Proto-Baltic, PBS: Proto-Balto-Slavic, PBSII: Proto-Balto-Slavo-Indo-Iranian (i.e. the ancestor of the ‘satem core’), PC: Proto-Celtic, PG: Proto-Germanic, PGA: Proto-Greco-Armenian, PI: Proto-Italic, PIC: Proto-Italo-Celtic, PII: Proto-Indo-Iranian, PS: Proto-Slavic, PT: Proto-Tocharian. Whether some of these protolanguages ever existed depends, of course, on the configuration of the true tree.

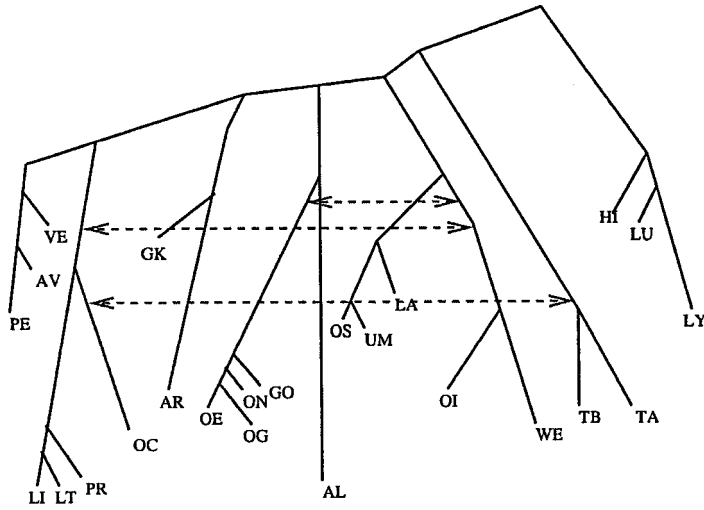


FIGURE 10. The first feasible PPN based on Tree A (corresponding to solution 1 in Table 3).^a

^a The solid black lines are the underlying tree edges that indicate genetic transmission of character states; those edges are directed. The dashed lines represent contact edges between language groups, which are bidirectional (Figs. 10–14).

the sixth c. AD). However, we know very little about the prehistoric movements of speakers of (pre-)Tocharian, and what we do know is that they were in contact with steppe-dwelling Iranian tribes; a long migration eastward in a relatively short period of time therefore does not seem out of the question.

The PPN in Fig. 10 also posits a contact edge between Proto-Celtic and Proto-Germanic, which is unobjectionable, and one between Proto-Celtic and Proto-Balto-

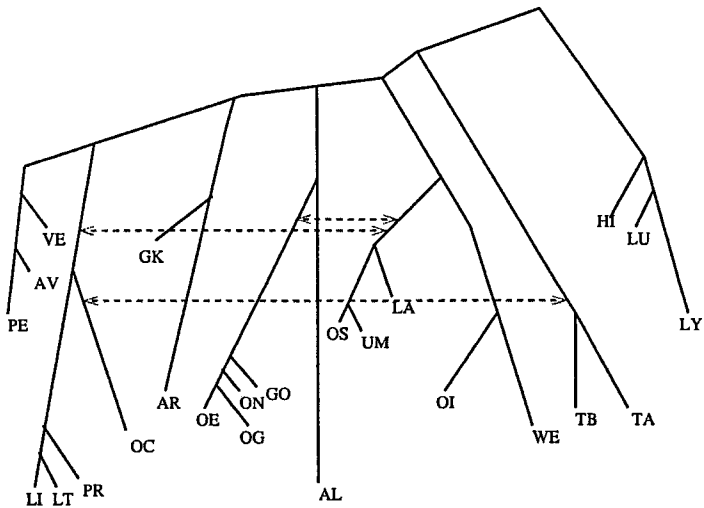


FIGURE 11. The second feasible PPN based on Tree A (corresponding to solution 3 in Table 3).

Slavic, which at first seems surprising; the PPN in Fig. 11 also posits a contact edge between Proto-Italic and Proto-Germanic, which is likewise unobjectionable, and one between Proto-Italic and Proto-Balto-Slavic, which likewise seems surprising. In other words, each of these PPNs posits that one of the westernmost subgroups of the family was, at very early periods, in contact both with Germanic and with Balto-Slavic, and it is the connections with Balto-Slavic that are unexpected. But while it is clearly out of the question for Baltic and Slavic languages to have been in contact with Celtic or Italic languages during the historical period, the linguistic geography of eastern Europe could have been very different in, say, the third millennium BC. In particular, it is possible that speakers of Proto-Italo-Celtic occupied the Hungarian plain in about 3200 BC (David Anthony, p.c.), and that Italic and Celtic began to diverge in eastern Europe; and since it also seems possible that Germanic and Balto-Slavic evolved on the other side of the Carpathians, contact between both those groups and one of the western groups might have been possible for some centuries. Proto-Celtic, the more northerly of the two western protolanguages, is clearly a more plausible candidate, and so the first PPN is therefore probably preferable to the second. (Note that the relative chronological positions of the internal nodes of all of our trees must be allowed to vary within certain limits. They cannot be fixed absolutely by archaeological data, and variation in the rate of linguistic change is still too poorly understood to render their calculation from internal evidence feasible.)

Let us now consider the third feasible PPN based on Tree A (solution 5; see Figure 12).

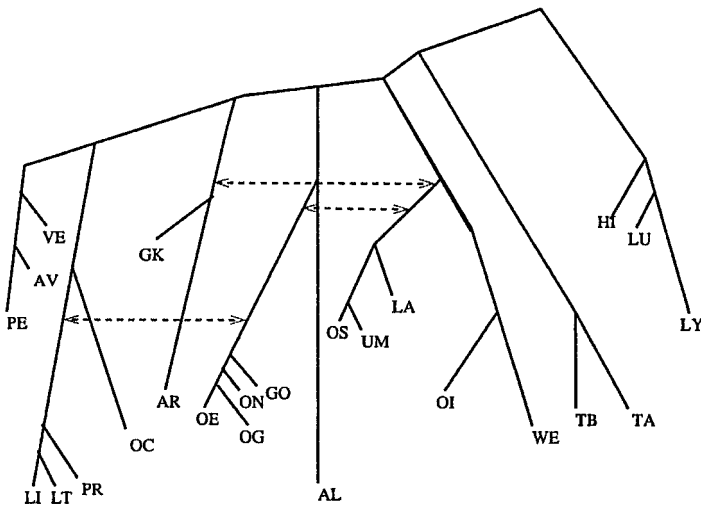


FIGURE 12. The third feasible PPN based on Tree A (corresponding to solution 5 in Table 3).

This PPN posits a contact edge between Proto-Italic and Proto-Germanic, which is (again) unobjectionable; a second contact edge (at a later date) between Proto-Germanic and Proto-Baltic, which is highly likely; and a contact edge between Proto-Italic and Proto-Greco-Armenian, which is surprising but cannot be excluded (given how little we know about the prehistoric linguistic geography of eastern Europe). Interestingly, of these three contact edges the most surprising has the smallest support: only two of the characters (‘free’ and one alternative coding of ‘young’) must use the contact edge between Proto-Italic and Proto-Greco-Armenian, compared to six on the first contact

edge and nine on the second. It is therefore possible that this third contact edge is not realistic, and that some other explanation (such as undetected parallel development) should be considered for the nontreelike evolution of ‘free’ and ‘young’—especially in view of the fact that the other two contact edges, which are very well supported, seem thoroughly realistic.

All solutions on Tree A other than the three just discussed are implausible or actually impossible for chronological and geographical reasons, as follows. The temporal constraints that we applied to find candidate contact edges (see the previous section) are the ones that are easiest to define in formal terms, but they are not the only ones that exist; while it is clearly impossible for a living language to be in real linguistic contact with a living ancestor, it is equally impossible for two living languages spoken at different periods in history to be in contact.¹⁸ Solutions 6 through 16 posit a contact edge between Proto-Italic (the ancestor of all the Italic languages), at some time after its separation from Proto-Celtic, and Proto-Baltic (the ancestor of all the Baltic languages), at some time after its separation from Proto-Slavic. Proto-Italic must have begun to diversify into the attested Italic languages well before 1000 BC, because our earliest documents from the Osco-Umbrian subgroup, from about the sixth c. BC, exhibit so many innovations not shared by Latin that it is clear that those two subgroups of Italic had been diverging for centuries. But Baltic is most unlikely to have begun diverging from Slavic by 1000 BC, because Proto-Slavic seems still to have been more or less uniform in the eighth c. AD, and Proto-Baltic and Proto-Slavic are so similar that they had probably been diverging for less than two millennia. In addition, Proto-Baltic was clearly spoken somewhere in northeastern Europe (not southwest of modern Poland); and while Proto-Italic may not have been spoken in Italy, it can hardly have been spoken anywhere to the northeast of modern Hungary. Solutions 2 and 4 posit a contact edge between Proto-Baltic and Proto-Italo-Celtic; since the latter is a direct ancestor of Proto-Italic, the chronological constraints exclude these two solutions a fortiori, though the geographical situation is considerably less clear.

PPNS BASED ON TREE B. The algorithm described above generated four three-edge solutions for Tree B, two of which were consistent with known constraints on the prehistory of the IE family. Those solutions are described in Table 4; the feasible solutions are in boldface.

The first feasible solution (solution 1 in Table 4) posits contact between (pre-)Proto-Italic and the common ancestor of Germanic and Balto-Slavic (which is reasonably plausible), between Germanic and Celtic (highly plausible), and between Tocharian and Slavic (as for two solutions on Tree A); the minimum number of borrowing events needed to make all characters compatible on this PPN is twenty-one. This solution

¹⁸ We define ‘contact’ as involving actual linguistic interchange between native speakers of significantly different dialects or languages, on the grounds that the full range of natural contact phenomena cannot occur otherwise. (Note that this follows the same principle as restricting the term ‘natural human language’ to languages spoken by native speakers.) The influence of a dead ancestor accessible only through written records (and perhaps still learned as a second spoken language by an educated elite), such as the influence of Latin on medieval or modern Romance languages, is of course a different process, for which it would be advisable to find some other term. It might be argued that a creole can be in real linguistic contact with its ancestor; for instance, though Afrikaans is in some sense descended from sixteenth- or seventeenth-century Dutch, modern Dutch is not so different that it can be called a different language, and contact between the two is clearly possible. However, we define ‘genetic descent’ so as to EXCLUDE the origin of creoles, since the latter clearly involves at least one step that is not merely normal native language acquisition (see Ringe et al. 2002:63). In any case none of the languages of our database shows evidence of creolization.

SOLUTIONS	EDGE 1	EDGE 2	EDGE 3	min. # borrowing events in PPN
1	(PI,PGBS)	(PG,PC)	(PT,PS)	21
2	(PG,PIC)	(PI,PBS)	(PT,PS)	23
3	(PI,PB)	(PC,PGA)	(PG,PIC)	25
4	(PI,PS)	(PC,PGA)	(PG,PIC)	24

TABLE 4. The four 3-edge solutions found on Tree B; the boldface rows (1 and 2) correspond to the two feasible solutions. Each solution is described in terms of the three lateral edges added to Tree B to produce a PPN; the rightmost column gives the minimum number of borrowing events needed to make all characters compatible on the PPN.

appears possible, but it is not markedly better than the first two feasible solutions on Tree A, and it is considerably less plausible than the third. The PPN that corresponds to this solution is shown in Figure 13.

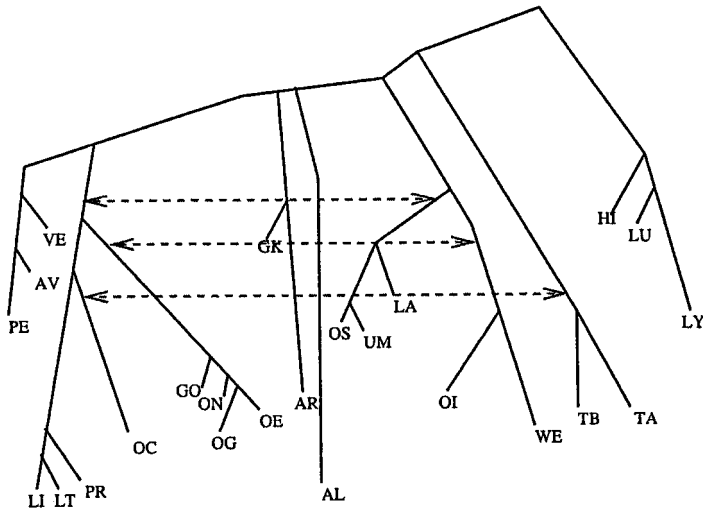


FIGURE 13. The first feasible PPN based on Tree B.

The other feasible solution (solution 2 in Table 4) posits contact between pre-Proto-Germanic and Proto-Italo-Celtic (which might be possible if the Proto-Italo-Celtic node should actually be somewhat later than we have hypothesized), between Proto-Italic and Proto-Balto-Slavic (surprising, but not necessarily out of the question), and between Tocharian and Slavic (as above); the minimum number of borrowing events needed to make all characters compatible on this PPN is twenty-three. This solution, too, cannot be summarily excluded but is not as plausible as the third solution on Tree A. The PPN that corresponds to this solution is shown in Figure 14.

It is important to note that the two feasible PPNs based on Tree B imply different chronological orderings of Proto-Italo-Celtic and Proto-Germano-Balto-Slavic. Both solutions need to be considered, since the times of internal nodes in the tree are somewhat indeterminate. Additional clarification about the dates of these internal splits would help clarify the relationships between these languages.

We now describe the two remaining solutions, noting the reasons why we can eliminate these on the basis of known constraints. The first of them (solution 3 in Table 4) posits contact between Italic and Baltic but not Slavic, which seems impossible (see above); the minimum number of borrowing events needed to make all characters compatible on this PPN is twenty-five. The second (solution 4 in Table 4) posits contact

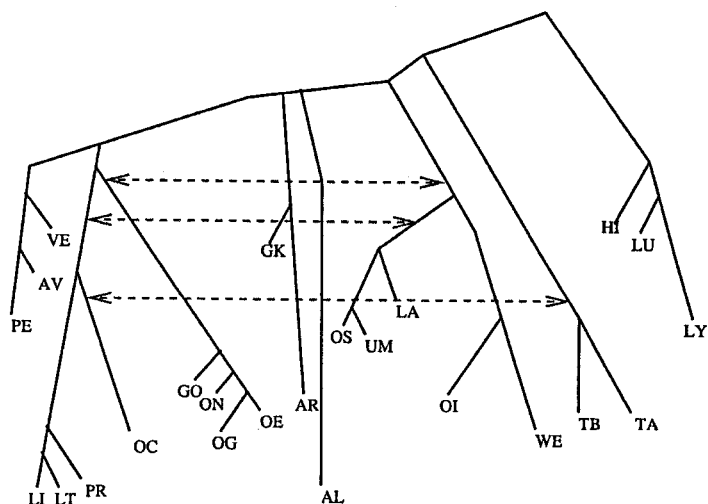


FIGURE 14. The second feasible PPN based on Tree B.

between Italic and Slavic but not Baltic, which likewise seems impossible; the minimum number of borrowing events needed to make all characters compatible on this PPN is twenty-four. Both of these solutions are unfeasible because the contacts that must be posited seem to be chronologically impossible.

PPNS BASED ON TREES C, D, AND E. The algorithm described above generated one three-edge solution for Tree C. However, that solution posits contact between Italic and Baltic but not Slavic, which seems impossible both for chronological and for geographical reasons. The minimum number of borrowing events needed to make all characters compatible on this PPN is twenty-four.

The algorithm also generated two three-edge solutions for Tree D. But like that generated for Tree C, both solutions posit a contact between Italic and Baltic but not Slavic, which seems impossible both for chronological and for geographical reasons. The minimum number of borrowing events is twenty-one.

The algorithm did not find any three-edge solutions for Tree E.

ADDITIONAL ANALYSES. Since the Greco-Armenian subgroup is very weakly supported, and since the unity of Italic has been repeatedly impugned, we reanalyzed our most promising tree, Tree A (Fig. 5), without pruning the Italic and Greco-Armenian subgroups. (Recall that one of the steps in our phylogenetic analysis involves pruning maximally compatible subtrees). Three new PPNs with three lateral edges were found (in addition to the sixteen PPNs reported above). However, all three PPNs posit a lateral edge between Italic (at some time after its separation from Celtic) and Baltic (at some time after its separation from Slavic). Hence all three solutions seem impossible for the reasons discussed above.

SUMMARY OF RESULTS. Our five different feasible PPNs for the IE family exhibit interesting similarities and differences. In the first place, it appears that solutions with three lateral edges are possible only if Germanic is the outlier within the 'core' (i.e. the residue of the family after Italo-Celtic has diverged), or if it is the nearest sister of Balto-Slavic. Four of the five solutions posit a contact episode between Slavic and Tocharian; that is an unforeseen, indeed a surprising, result. The remaining solu-

tion—the third feasible on Tree A (solution 5)—is much less surprising; in fact, the contact between Germanic and Baltic that it posits is so highly plausible that this solution is probably preferable to the others on that ground alone.

5.6. OPTIMIZATION OF MATHEMATICAL CRITERIA. Recall the three mathematical criteria by which we evaluate PPNs: (i) the number of characters compatible on the genetic tree, (ii) the number of contact edges in the PPN, and (iii) loan parsimony, that is, the number of undetected borrowing events in the PPN. Among our twenty-three PPNs, both feasible and unfeasible, the best that we can do in optimizing by criterion (i) is fourteen, since all PPNs based on Tree A have only fourteen characters incompatible on that tree, while PPNs based on the other candidate trees have more. All of the PPNs that we consider posit three contact edges, so there is no difference with respect to criterion (ii). For the third criterion, the best we can do is seventeen, which is achieved by solution 5 on Tree A; all other solutions must posit at least nineteen borrowing events, and most posit more. Thus solution 5 on Tree A optimizes all three of the mathematical criteria and is the unique optimal solution.

Note that our favored PPN—solution 5 based on Tree A—is not only optimal with respect to each of the three mathematical criteria; it is also the most credible in terms of known geographic and chronological constraints on the prehistory of IE diversification. That is, it optimizes each of the FOUR criteria. It is therefore the solution that we propose for the first-order diversification of the IE family. More research is of course needed to confirm this.

6. SUMMARY. Because our best PPN is so clearly better than the other scenarios, a closer look at it is justified. Our proposed scenario for the historical diversification of the IE language family posits Tree A (the tree found in Ringe et al. 2002 by computational cladistic methods) as the underlying genetic tree and also posits three contact edges; it is the PPN of Fig. 12. We note that our network suggests AT MOST three historically real episodes of contact between the relevant language groups. It makes sense to examine these three possible contact episodes to determine how much support our analysis suggests for each.

Two of the contact edges, both involving Germanic, do have a significant number of characters evolving down them; they are obviously necessary components of this PPN. The third edge, between Proto-Italic and Proto-Greco-Armenian, has only two characters transmitting states across it. Thus, while the contact edges that involve Germanic are well supported, the contact edge that does not involve Germanic seems debatable. It is possible that some other explanation for the evolution of the two characters involved (‘free’ and ‘young’) that does not involve borrowing can be found.

With respect to the question of whether the evolution of IE should be represented by a wave model or a tree model (for the most part), we believe we have shown that although a tree model does not fit the family’s history perfectly, there is clear evidence that the underlying history is almost entirely treelike, and that it is clearly sensible to infer a genetic tree within which some borrowing (previously undetected) has occurred.

7. CONCLUSIONS AND FUTURE WORK. It is clear that historical linguists can make good use both of well-articulated phylogenetic models and of computational methods for inferring the evolutionary histories of datasets, especially because these problems are not as simple as the structure of linguistic descent might lead us to suppose. Our analysis of the IE dataset shows that a combination of algorithmic analysis and additional linguistic considerations allows us to identify a small set of feasible scenarios that explain the character-state patterns we see in our IE data. This is highly encouraging. Perhaps equally encouraging is the fact that the evolutionary process suggested by our

best network is highly treelike: not only do we need to posit only three contact edges, but almost all of the characters evolve on the underlying genetic tree. These two properties together suggest that the construction of a genetic tree for IE is sensible and plausible. Finally, the agreement between the optimal solution with respect to the mathematical criteria we have proposed and the optimal solution with respect to both archaeological and chronological constraints suggests that the model we have developed might be sufficiently realistic to be useful to historical linguists in general.

Our new methodology has found a possible solution to an old and intractable problem in historical linguistics. On those pragmatic grounds we argue that this methodology is promising and should be pursued. Particularly important for future work is exploration of how to resolve conflicts between optimization criteria for PPNs.

APPENDIX A: ALGORITHMS.

In this appendix we describe how we obtained our candidate genetic trees and the algorithmic techniques used to complete a candidate genetic tree to a perfect phylogenetic network (PPN) containing a minimum number of contact edges.

The candidate genetic trees we have used are each based on an analysis of the character dataset under a variant of the maximum compatibility approach in which (i) we require some characters—almost all of our phonological and morphological characters, but none of the lexical characters—to be compatible on any candidate tree, and (ii) we consider incompatibility of phonological or morphological characters more problematic, in general, than incompatibility of lexical characters. This is equivalent to a weighted maximum-compatibility approach, in which each character is assigned a weight and we seek a tree (or trees, if multiple solutions exist) that maximizes the total weight of the compatible characters. For our approach three weights will suffice: infinite weight for the required phonological and morphological characters, a sufficiently large (but finite) weight for the remaining phonological and morphological characters, and a smaller weight for each of the lexical characters.

Provably obtaining optimal solutions to weighted maximum compatibility is computationally difficult, because it is an NP-hard problem. (See Garey & Johnson 1979 for a definition of NP-hardness and an explanation of its consequences.) Exact solutions are therefore not easily obtained except through computationally intensive techniques. For our own analyses of our IE datasets, we produced trees using heuristics—that is, techniques that are not guaranteed to produce optimal solutions, but that may work well in practice—and then scored each tree with respect to weighted compatibility. Those trees that had the best scores among the trees examined were used as our candidate genetic trees.

In our search for a practical solution to weighted maximum compatibility we used two different but complementary techniques, each of which produces possible solutions for the weighted maximum-compatibility problem which can then be compared with respect to their weighted compatibility scores.

Our first technique is the following. Recall that a perfect phylogeny for a dataset is a tree on which every character in the dataset is compatible. Determining whether a perfect phylogeny exists, and computing it if it does, is another NP-hard problem, for which software exists (see Kannan & Warnow 1997). Since some of our characters are required to be compatible, we use the perfect phylogeny software on the set of required characters to compute the set of all minimally resolved perfect phylogenies. This set of trees has the property that any tree that is compatible with all of the required characters is either one of the trees of the set or can be obtained by refining one of the trees of the set. We then use the remaining characters to complete the refinement of these trees, in order to find those trees whose compatible characters exhibit maximum weight. That step proceeds as follows. First we order the remaining characters consistently with the weighting scheme (characters having higher weight appearing in the sequence before characters having lower weight). Then we examine each character in turn, checking to see if it is possible to refine the tree so as to make that character compatible (in all possible ways); if it is not, the tree is not changed. In that way we produce a tree, or several trees, with which (we hope) a maximum-weight set of characters is compatible. Note that this technique always produces trees on which all of the required characters are compatible. This technique produced one tree with the best weighted compatibility score we could find (Tree A) and two trees with somewhat worse scores (Trees B and C).

Our second technique for obtaining candidate trees is based on the similarity between the weighted maximum-compatibility and maximum-parsimony optimization criteria. The maximum-parsimony criterion seeks to minimize the total number of character-state changes on the tree; it is a popular approach for phylogenetic estimation in biology, and it has also been used in historical linguistic reconstructions (see e.g. Holden 2002). Because of its popularity for molecular phylogenetics, there are very effective heuristics for maximum

parsimony available (such as PAUP*; see Swofford 1996). Like weighted maximum compatibility, maximum parsimony is an NP-hard problem, so these heuristics, too, do not provably find optimal solutions. The similarity between maximum parsimony and even unweighted maximum compatibility is obvious; but they are in fact different problems, and it is possible for them to give different solutions for the same dataset. Because of their similarity, however, trees that are optimal or near-optimal with respect to maximum parsimony can be good candidates for maximum compatibility, and possibly also for weighted maximum compatibility. We therefore use the heuristics for maximum parsimony available in the software package PAUP* to search for additional candidate trees, and we compute the set of incompatible characters on each tree that is optimal or near-optimal with respect to maximum parsimony. Unfortunately, this technique produced no trees that were as good as our Tree A in terms of the weight of the characters compatible on them.

We now turn to the algorithmic problem. We begin with a formal statement of the problem.

- **Problem:** minimum increment to a PPN.
- **Input:** a rooted tree T , leaf-labeled by a set S of languages, and a set C of qualitative characters defined on S .
- **Output:** a PPN N containing T and p additional contact edges, so that p is minimum.

In other words, we wish to find a minimum number of contact edges to be added to T so as to produce a PPN.

Recall that a network N is a PPN if, for every character c in the set C , c is compatible on at least one of the trees contained within N .

A simple brute-force method to find a minimum increment of T to a PPN is as follows.

- Let $i = 1$.
- For each way of adding i contact edges to T , determine if the resulting network is a PPN for S with respect to the character set C .
- If any such way of adding i edges produces a PPN, return all such PPNs (with i contact edges), and exit; ELSE increment i by 1 and repeat.

In other words, the algorithm begins with a tree T on the language set S , and we complete the tree to a network in all possible ways, starting with just one contact edge, and then increasing the number of contact edges by one, until we find the smallest possible number of contact edges that suffice to make a PPN. The algorithm therefore needs to be able to determine whether each character is compatible on a given network with i contact edges, a separate problem that we now formally define and analyze.

- **Problem:** determining compatibility on a network.
- **Input:** rooted phylogenetic network N and set C of characters defined on the leaves of N .
- **Output:** *Yes* if every character in C is compatible on some tree contained within N ; otherwise *No*.

Again, a straightforward algorithm will work for this problem. Given a network N with i contact edges, we compute each of the (at most) 3^i trees contained within N . Then, for each character in C , we check compatibility on each of the 3^i trees. As long as each character is compatible on at least one of these trees, we return *Yes*; otherwise we return *No*.

It is easy to determine compatibility of a character on a tree. It can be done by inspection (i.e. without a computer), since all we need to do is label every node that is on a path between two leaves with the same state, and so long as no node gets two contradictory labels, we have compatibility. More formally, however, we can use the algorithm given by Walter Fitch (1971) to determine whether a given character is compatible on a tree. The algorithm in Fitch 1971 actually computes the state assignments to the internal nodes of a fixed tree, so as to minimize the total number of changes on the tree. Thus if a character has r states at the leaves of the tree, the character is compatible on the tree if and only if Fitch's algorithm is able to assign states to the internal nodes such that only $r - 1$ changes are needed on the tree. Fitch's algorithm runs in $O(rn)$ time for an r -state character, so this costs no more than $O(n^2)$ time to test compatibility of a single character on a tree. Therefore, to check if a character is compatible on a network with i contact edges, the algorithm would use $O(n^2 3^i)$ time, since it would examine each of the 3^i trees. Since we need to do this for each character, the running time would be $O(kn^2 3^i)$, where k is the number of characters. There are

$\binom{2n-2}{i}$ possible ways of adding i edges to a tree T on n leaves, which is $O(2^{2i} n^{2i})$. Therefore it takes $O(kn^2 3^i 2^{2i} n^{2i})$ time to find if there is a PPN based on tree T with i contact edges, which is $O(k2^{4i} n^{2i+2})$.

However, we would have to do this for $i = 1, 2, \dots, p$, where p is the minimum number of contact edges we need to get a PPN. So the running time would actually be

$$k \sum_{1 \leq i \leq p} O(2^{4i} n^{2i+2}) = O(k2^{4p+1} n^{2p+3}).$$

We summarize this analysis as follows:

Theorem 1. Let T be a phylogenetic tree on a set L of n languages, and assume that each language in L is assigned a state for each character in a set C of k characters. We can solve the minimum increment to a perfect phylogenetic network (i.e. we can find a completion of T to a PPN with a minimum number of contact edges) in $O(k2^{4p+1}n^{2p+3})$ time, where p is the number of contact edges we need at a minimum.

For additional mathematical results on the computational complexity of problems related to PPNs see Nakhleh 2004.

APPENDIX B: SAMPLE OF DATA.

A full presentation of our data and its coding would be much longer than this paper. Interested readers can find complete information in the online appendix at <http://www.cs.rice.edu/~nakhleh/CPHL>. For those who wish to understand our methodology without examining all of the data in detail, we here present a sample, as follows. Phonological characters are exemplified by P1, P2, P3, P16, and P19. The first three are the only phonological characters of which derived states are shared by more than one major subgroup of the family; P16 is an example of a complex phonological character, coding a sequence of sound changes, while P19 is an example of a phonological character based on an unusual sound change not likely to have occurred more than once independently. Morphological characters are exemplified by M3, M5, M6, M8, and M12 through M15, the characters that we believe should be compatible on the true tree. The lexical characters presented here are chosen to exemplify patterns of informativeness and compatibility.

PHONOLOGICAL CHARACTERS.

P1 *p . . . k^w > *k^w . . . k^w.

Hitt.	4	Av.	1	Luv.	5	Goth.	1
Arm.	1	OCS	1	Lyc.	6	ON	1
Gk.	1	Lith.	1	TA	1	OHG	1
Alb.	1	OE	1	OPer.	1	Welsh	2
TB	1	OI	2	OPru.	1	Osc.	3
Ved.	1	Lat.	2	Latv.	1	Umb.	3
1 absent [ancestral]							3 situation obscured by merger of *p and *k ^w
2 present							4 etc. no evidence

Notes on P1.

We accept the usual view that state 1 is ancestral because it fits with the PIE morpheme structure constraint prohibiting stops at the same place of articulation from appearing both initially and finally in a root. Precisely because of that constraint, the change involves no merger, and so could not be directed from the fact that mergers are irreversible. Thus this is a case in which a more comprehensive approach to phonology contributes to our understanding of a sound change and permits us to direct a phonological character.

P2 full 'satem' development (*k^w and *k > *k; *ḳ > affricate or fricative; etc.).

Hitt.	1	Av.	2	Luv.	1	Goth.	1
Arm.	1	OCS	2	Lyc.	1	ON	1
Gk.	1	Lith.	2	TA	1	OHG	1
Alb.	1	OE	1	OPer.	2	Welsh	1
TB	1	OI	1	OPru.	2	Osc.	1
Ved.	2	Lat.	1	Latv.	2	Umb.	1
1 absent [ancestral]							2 present

P3 'ruki'-retraction of *s.

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

1 absent 3, 4 situation obscured by subsequent sound change or orthography
2 present

Notes on P2 and P3.

We assign Armenian state 1 of P2 because the merger of velars and labiovelars in that language appears to be incomplete. No such merger seems to have occurred in Albanian; see Demiraj 1997 *passim*.

We accept the arguments of Andersen 1968 for complete ruki-retraction in Baltic. The subsequent merger of *s and *š in Latvian has destroyed the evidence in that language; the consistent spelling of all fricatives as *s* in Old Prussian likewise renders it opaque on this point, whether or not it reflects a real sound change. We emphasize that the assignment of states 3 and 4 (rather than 1) to these languages presupposes the specific hypothesis about their development just described; the alternative would involve denying the unity of the Balto-Slavic subgroup from a very early date, so far as we can see, and the overwhelming agreement of the Balto-Slavic languages in a great variety of characteristics seems to us to preclude that alternative.

Strictly speaking, P3 is not directed, since it involves only allophony; it is not inconceivable that the rule existed in PIE but was lost in most daughters, surviving only when a subsequent phonemic split occurred.

P16 sequence of changes:

(a) Grimm's Law; (b) Verner's Law; (c) shift of stress to initial syllables; (d) merger of unstressed *e with *i unless *r follows immediately.

Hitt.	1	Av.	1	Luv.	1	Goth.	2
Arm.	1	OCS	1	Lyc.	1	ON	2
Gk.	1	Lith.	1	TA	1	OHG	2
Alb.	1	OE	2	OPer.	1	Welsh	1
TB	1	OI	1	OPru.	1	Osc.	1
Ved.	1	Lat.	1	Latv.	1	Umb.	1
		1 absent [ancestral]		2 present			

P19 merger of *ōw and *zw with *ww.

Hitt.	1	Av.	1	Luv.	1	Goth.	1
Arm.	1	OCS	1	Lyc.	1	ON	1
Gk.	1	Lith.	1	TA	1	OHG	2
Alb.	1	OE	2	OPer.	1	Welsh	1
TB	1	OI	1	OPru.	1	Osc.	1
Ved.	1	Lat.	1	Latv.	1	Umb.	1
		1 absent [ancestral]		2 present			

Notes on P19.

This unique sound change, characteristic only of West Germanic languages, is often overlooked; for full discussion and references see Stiles 1985:89–94. Though the only two examples are 'four' and 'you (pl., non-nominative)', they are absolutely secure, and there are no counterexamples.

MORPHOLOGICAL CHARACTERS.

M3 thematized aorist.

Hitt.	1	Av.	2	Luv.	1	Goth.	9
Arm.	2	OCS	2	Lyc.	1	ON	10
Gk.	2	Lith.	4	TA	2	OHG	11
Alb.	3	OE	5	OPer.	6	Welsh	12
TB	2	OI	2	OPru.	7	Osc.	2
Ved.	2	Lat.	2	Latv.	8	Umb.	2
		1 absent, probably not lost [ancestral]		3 etc. no evidence			
		2 present or immediately reconstructable					

Notes on M3.

Our coding of state 1 reflects an extrapolation from the findings of Cardona 1960 as follows. Cardona argued that only the thematic aorists *h₁lud^hét 'arrived' and *widét 'caught sight of' (in our reconstructions) can be reconstructed for (non-Anatolian) PIE; but the latter can be eliminated, since the contrast between the root-ablaut of Latin *vīdit* 'saw' and that of Greek $\xi(F)\iota\delta\epsilon$ 'saw' reveals an original athematic aorist. (The meaning of the Latin stem precludes *wóyde 'knows' as an etymon.) Since the Anatolian languages do not exhibit clear evidence even for simple thematic presents, which are solidly attested in all of the other branches, we suggest that their lack of simple thematic aorists is an archaism. If this argument is rejected, those languages should be assigned separate states, and the character becomes uninformative (since it can then be fitted to any tree).

The clear evidence for thematized aorists in Tocharian is essentially the relic preterit TB *lac*, TA *lāc* 'went out'; in Old Irish, the corresponding relic preterite *luid* 'went'; in the Italic languages, OLat. perf. 3sg. *-ed* and the corresponding Osco-Umbrian endings.

We reject the hypothesis that the West Germanic 2sg. indicative forms of the strong preterite, which exhibit a zero-grade root in the first three classes, reflect an inherited thematized aorist; the complete lack

of any trace of such a category in East and North Germanic (which do NOT form a clade) makes the hypothesis grossly improbable. It seems obvious that these West Germanic forms are optatives that have secondarily acquired indicative function as well (see character M15 below); a clear parallel is the later spread of the present optative 1sg. to indicative function in southern OE, though the motivation for the change must have been different (cf. Cowgill 1965).

A language can lack evidence for this category for a variety of reasons. In some the category 'aorist' is not clearly present; in Albanian, and perhaps in the Baltic languages, it has been remodeled so much that clearly thematized stems can no longer be identified; in Old Persian we happen to find no thematic stems among the few aorists attested. Similar factors underlie the absence of evidence for particular characteristics in the characters that follow.

M5 mediopassive primary marker (sg. and 3pl.).

Hitt.	1	Av.	2	Luv.	1	Goth.	2
Arm.	3	OCS	5	Lyc.	7	ON	2
Gk.	2	Lith.	6	TA	1	OHG	10
Alb.	4	OE	2	OPer.	2	Welsh	1
TB	1	OI	1	OPru.	8	Osc.	1
Ved.	2	Lat.	1	Latv.	9	Umb.	1
	1	*-r	[ancestral]		3	etc.	no evidence
	2	*-y	(= active *-i)				

Notes on M5.

On the Hittite endings see Yoshida 1990. We accept the usual view that state 1 is ancestral, because state 2 can be explained as the result of analogy with the active endings, whereas no similar explanation is available for state 1.

M6 thematic optative.

Hitt.	4	Av.	2	Luv.	7	Goth.	2
Arm.	5	OCS	2	Lyc.	8	ON	2
Gk.	2	Lith.	2	TA	1	OHG	2
Alb.	6	OE	2	OPer.	2	Welsh	3
TB	1	OI	3	OPru.	2	Osc.	3
Ved.	2	Lat.	3	Latv.	2	Umb.	3
	1	*-ih ₁ -		3	*-ā-		
	2	*-oy-		4	etc.		no evidence

Notes on M6.

We accept the hypothesis of Trubetzkoy 1926 regarding the original status of the Italo-Celtic forms in *-ā-. On the probable reflexes of this suffix among the Welsh subjunctive endings see Pedersen 1913:356–57; on the phonology of the Tocharian ending see Ringe 1996:84.

In Balto-Slavic the function of this category has been shifted; it appears variously as an imperative (e.g. in Old Church Slavonic) and in other modal functions. On the Baltic reflexes see Stang 1966:422–27, 434–35, 437–40.

M8 most archaic superlative suffix.

Hitt.	3	Av.	1	Luv.	9	Goth.	1
Arm.	4	OCS	7	Lyc.	10	ON	1
Gk.	1	Lith.	8	TA	11	OHG	1
Alb.	5	OE	1	OPer.	1	Welsh	2
TB	6	OI	2	OPru.	12	Osc.	2
Ved.	1	Lat.	2	Latv.	13	Umb.	2
	1	*-isto-		3	etc.		other, or no superlative
	2	*-ismō-					

M12 imperfect subjunctive in *-sē-.

Hitt.	1	Av.	1	Luv.	1	Goth.	1
Arm.	1	OCS	1	Lyc.	1	ON	1
Gk.	1	Lith.	1	TA	1	OHG	1
Alb.	1	OE	1	OPer.	1	Welsh	1
TB	1	OI	1	OPru.	1	Osc.	2
Ved.	1	Lat.	2	Latv.	1	Umb.	3
	1	absent	[ancestral]	2	present	3	unattested

Notes on M12.

We accept the usual view that the nonattestation of this suffix in Umbrian is accidental. State 1 is clearly ancestral because a category ‘imperfect subjunctive’ makes no sense in the organization of the PIE verb (in which tense and mood are disjunct categories).

M13 gerundive in *-ndo-.

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

M14 syncretism of 3sg. and 3pl.

Hitt.	1	Av.	1	Luv.	1	Goth.	1
Arm.	1	OCS	1	Lyc.	1	ON	1
Gk.	1	Lith.	2	TA	1	OHG	1
Alb.	1	OE	1	OPer.	1	Welsh	1
TB	1	OI	1	OPru.	2	Osc.	1
Ved.	1	Lat.	1	Latv.	2	Umb.	1
		1 absent [ancestral]		2 present			

Notes on M14.

Since syncretism is a type of conditioned merger, this morphological character can (exceptionally) be directed in the same way as most phonological characters.

M15 replacement of 2sg. indicative by optative in the strong preterite.

Hitt.	1	Av.	1	Luv.	1	Goth.	1
Arm.	1	OCS	1	Lyc.	1	ON	1
Gk.	1	Lith.	1	TA	1	OHG	2
Alb.	1	OE	2	OPer.	1	Welsh	1
TB	1	OI	1	OPru.	1	Osc.	1
Ved.	1	Lat.	1	Latv.	1	Umb.	1
		1 absent [ancestral]		2 present			

Notes on M15.

This character is directed for the same reason as the preceding: a conditioned merger (syncretism) of two inherited categories has occurred.

Our coding follows the same principle as the coding of phonological characters: those languages that do not even have such a category as ‘strong preterite’ cannot have undergone the change, and so must be assigned state 1. This would not be possible for an undirected character.

LEXICAL CHARACTERS.

1 ‘all (pl.)’.

[two characters]

Hitt.	1	Av.	5a	Luv.	8	Goth.	6a
Arm.	2	OCS	5b	Lyc.	8	ON	6a
Gk.	3	Lith.	5b	TA	3	OHG	6a
Alb.	4	OE	6a	OPer.	5a	Welsh	9
TB	3	OI	6b	OPru.	5b	Osc.	10
Ved.	5a	Lat.	7	Latv.	5b	Umb.	11
		3 *pántes		6 *ol-			
		5 *wi-		6a (*olnoy >) PGmc. *allai			
		5a (*wí-kwo- >) PIIr. *víšva-		6b PCelt. *olyoy			
		5b (*wi-so- >) PBS *visa-		8 PLuv. *pūno-			

Notes on 1 ‘all (pl.)’.

We have coded this character both by root-etymology and by derivational morphology, on the hypothesis that there is a direct historical connection between states 5a and 5b and likewise between states 6a and 6b. The broader coding is incompatible on all of our trees.

See Hübschmann 1897:416 on the Armenian word (< *sm̥-) and Stang 1966:97, 238 on the Balto-Slavic forms.

30 'dig'.

Hitt.	1	Av.	6	Luv.	11	Goth.	9
Arm.	2	OCS	7	Lyc.	12	ON	9
Gk.	3	Lith.	8	TA	5	OHG	9
Alb.	4	OE	9	OPer.	6	Welsh	10
TB	5	OI	10	OPru.	13	Osc.	15
Ved.	6	Lat.	1	Latv.	14	Umb.	16
1	*b ^h odh ₂ - ~ *b ^h edh ₂ -		9		PGmc.	*grabidi	
5	PToch. *rāpa-		10		PCelt.	*klād- ~ *klad-	
6	PIIr. *kánti						

33 'drink'.

Hitt.	1	Av.	4	Luv.	1	Goth.	6
Arm.	2	OCS	3	Lyc.	7	ON	6
Gk.	3	Lith.	5	TA	1	OHG	6
Alb.	3	OE	6	OPer.	8	Welsh	3
TB	1	OI	3	OPru.	3	Osc.	3
Ved.	3	Lat.	3	Latv.	5	Umb.	9
1	*ēh ₂ g ^{wh} ti		5		PEBalt.	*gerja	
3	*peh ₃ - ~ *pī- (pres. *pībeti)		6		PGmc.	*drinkidi	

Notes on 33 'drink'.

On the Anatolian and Tocharian forms see now Kim 2000; though considerable analogical remodeling must be posited to explain the shape of the Tocharian verb, the two do appear to be related.

37 'ear'.

[two characters]

Hitt.	1	Av.	5	Luv.	1	Goth.	2x
Arm.	2	OCS	2	Lyc.	6	ON	2x
Gk.	2	Lith.	2	TA	3x	OHG	2x
Alb.	2	OE	2x	OPer.	5	Welsh	3y
TB	3x	OI	3y	OPru.	2	Osc.	7
Ved.	4	Lat.	2	Latv.	2	Umb.	8
1	*stómṃ ~ *stṃén- ('ear' ?)		3		derivs. of	*k ^l ew- 'hear'	
2	*h ₂ éwsos		3x		PToch.	*klēwt ^o	
2x	PGmc. *ausōn- ~ *auzōn-		3y		PCelt.	*klowstā	
			5		PIr.	*gaušah	

Notes on 37 'ear'.

We employ both codings for superstate 2, but since the two derivatives of 'hear' are clearly independent we have coded them separately.

38 'earth'.

[two characters]

Hitt.	1	Av.	1	Luv.	1	Goth.	4
Arm.	2	OCS	1x	Lyc.	7	ON	4
Gk.	3	Lith.	1x	TA	1	OHG	4
Alb.	1	OE	4	OPer.	8	Welsh	9
TB	1	OI	5	OPru.	1x	Osc.	6
Ved.	1	Lat.	6	Latv.	1x	Umb.	10
1	*d ^h ég ^h ōm, *g ^h m-, loc. *g ^h d ^h sém		4		PGmc.	*erþōn-	
1x	PBS *žemjā		6		PItal.	*tersā	

Notes on 38 'earth'.

We employ both codings for superstate 1.

41 'eye'.

[two characters]

Hitt.	1a	Av.	4	Luv.	1a	Goth.	2x
Arm.	2	OCS	2	Lyc.	1a	ON	2x
Gk.	2	Lith.	2	TA	2	OHG	2x
Alb.	3	OE	2x	OPer.	4	Welsh	5
TB	2	OI	1b	OPru.	2	Osc.	6
Ved.	2	Lat.	2	Latv.	2	Umb.	7

- 1 derivs. of *sek^w- 'see' 2 *h₃ók^w
 1a PAnat. *sog^wo- 2x PGmc. *augōn-
 1b PCelt. *sok^wlis 4 PIr. *čašma

Notes on 41 'eye'.

It is most unlikely that there is any direct connection between states 1a and 1b; coding by root-etymology is therefore inadvisable. By contrast, it is very likely that state 2x replaced state 2 directly (since the Germanic word for 'eye' has clearly been deformed by lexical analogy with 'ear'); we have therefore employed both codings in that case.

58 'foot'.

- Hitt. 1 Av. 1 Luv. 1 Goth. 1
 Arm. 1 OCS 3 Lyc. 1 ON 1
 Gk. 1 Lith. 4 TA 1 OHG 1
 Alb. 2 OE 1 OPer. 1 Welsh 5
 TB 1 OI 5 OPru. 3 Osc. 1
 Ved. 1 Lat. 1 Latv. 4 Umb. 1
 1 *póds ~ *pód- ~ *ped- 4 PEBalt. *kājā
 3 PBS *nagā 5 PCelt. *traget-

63 'give'.

[two characters]

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

1 *ay-

1x PAnat. *p-ay-

2 *deh₃-

2a, 2c original pres. unclear

2b pres. *dédeh₃tí and developments of same

2bx PBS pres. *dōd- (apparently ← *dedō- < *dédeh₃-, but how?)

4 PGmc. *gibidi (*geba-)

Notes on 63 'give'.

The reduplicating syllable *de- was replaced by the productive *di- in Greek and Italic (an unremarkable parallel development); in Osco-Umbrian the stem was thematized, but in Latin the reduplication was lost by sound change in compounds and the dereduplicated form was then generalized to the simplex (see e.g. Sommer 1948:538–39).

In Celtic and Iranian this verb was confused with *d^heh₁- 'put' because the two had become very similar by sound change.

We have employed both alternative codings.

69 'hand'.

[two characters]

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

1 *g^hés_g

3 PGmc. *handuz

1x PPIr. *z^hástas < *g^héstos (remodeled, but how?) 4 PCelt. *lāmā (< *p^h₂meh₂ 'palm')

2 PBS *rankā

5 PItal. *man-

Notes on 69 'hand'.

We employ both codings, since it is likely that state 1x replaced state 1 directly.

113 'not'.

- Hitt. 1 Av. 1 Luv. 1 Goth. 1
 Arm. 2 OCS 1 Lyc. 1 ON 4
 Gk. 2 Lith. 1 TA 3 OHG 1
 Alb. 1 OE 1 OPer. 1 Welsh 1

TB 3 OI 1 OPru. 1 Osc. 1
 Ved. 1 Lat. 1 Latv. 1 Umb. 1
 1 *né and extensions 3 PToch. *ma
 2 *h₂óyu 'life'

Notes on 113 'not'.

On the Greek and Armenian forms see Cowgill 1960. This character is incompatible on Tree D.

115 'one'.

Hitt. 1a Av. 1c Luv. 5 Goth. 1d
 Arm. 2 OCS 3 Lyc. 6 ON 1d
 Gk. 2 Lith. 4 TA 2 OHG 1d
 Alb. 2 OE 1d OPer. 1c Welsh 1d
 TB 2 OI 1d OPru. 1d Osc. 7
 Ved. 1b Lat. 1d Latv. 4 Umb. 8
 1 derivatives of *oy- 'single' 2 *sem-
 1a *oyos 4 PEbalt. *vianas
 1b *óykos
 1c *óyws
 1d *óynos

Notes on 115 'one'.

There are fairly strong indications that the original meaning of state 1 was not the numeral 'one'; for instance, the Greek cognate of 1c, οἶος, means 'alone', while that of 1d, οὔνη, means 'one-spot (on dice)', and the Latin adverb 'once' is *semel*—arguably 'stranded' when the numeral from which it was derived was replaced by *oīnos > *ūnus*. We have therefore coded states 1a–d separately. On the Hittite form see Eichner 1992:34, 42–44.

We have not coded the Slavic and East Baltic forms as substates of 1 because they do not fit by regular sound correspondences; we believe that the origin of those forms remains very unclear.

Nevertheless this character is incompatible on all of our trees.

116 'other'.

Hitt. 1 Av. 4a Luv. 7 Goth. 4b
 Arm. 2 OCS 5 Lyc. 8 ON 4b
 Gk. 2 Lith. 6 TA 2 OHG 4b
 Alb. 3 OE 4b OPer. 4a Welsh 2
 TB 2 OI 2 OPru. 6 Osc. 2
 Ved. 4a Lat. 2 Latv. 6 Umb. 9
 2 *ályos 4b PGmc. *anþeraz (< *án-teros)
 4a PIlr. *anyás 6 PBalt. *kitas

Notes on 116 'other'.

We have coded states 4a, 4b separately because it is not clear that there is any direct connection between them—especially in view of the fact that there may be some root-connection with state 2 (conceivably *ál-yo-s : *án-tero-s, with an archaic consonant alternation).

119 'play'.

Hitt. 1 Av. 7 Luv. 13 Goth. 19
 Arm. 2 OCS 8 Lyc. 14 ON 20
 Gk. 3 Lith. 9 TA 15 OHG 21
 Alb. 4 OE 10 OPer. 16 Welsh 22
 TB 5 OI 11 OPru. 17 Osc. 23
 Ved. 6 Lat. 12 Latv. 18 [loan] Umb. 24

Notes on 119 'play'.

This character is uninformative because there are no cognates.

142 'short'.

Hitt. 1 Av. 3 Luv. 1 Goth. 15
 Arm. 2 OCS 7 Lyc. 11 ON 16
 Gk. 3 Lith. 8 TA 12 OHG 16
 Alb. 4 [loan] OE 9 OPer. 13 Welsh 10
 TB 5 OI 10 OPru. 14 Osc. 17
 Ved. 6 Lat. 3 Latv. 14 Umb. 18

1 PAnat. *mannenk^{wo}- 14 PBalt. *insas
 3 *mrég^{hus} ~ *mrég^héw- 16 PNWGmc. *skamm-
 10 PCelt. *birros

176 'three'.

Hitt. 1 Av. 1 Luv. 1 Goth. 1
 Arm. 1 OCS 1 Lyc. 1 ON 1
 Gk. 1 Lith. 1 TA 1 OHG 1
 Alb. 1 OE 1 OPer. 1 Welsh 1
 TB 1 OI 1 OPru. 1 Osc. 1
 Ved. 1 Lat. 1 Latv. 1 Umb. 1
 1 *tréyes, fem. *tistrés, neut. *trif₂

Notes on 176 'three'.

This character is uninformative because all of the languages share the same state.

183 'two'.

Hitt. 1 Av. 1 Luv. 2 Goth. 1
 Arm. 1 OCS 1 Lyc. 1 ON 1
 Gk. 1 Lith. 1 TA 1 OHG 1
 Alb. 1 OE 1 OPer. 1 Welsh 1
 TB 1 OI 1 OPru. 1 Osc. 3
 Ved. 1 Lat. 1 Latv. 1 Umb. 1
 1 *duóh₁

Notes on 183 'two'.

Luvian and Oscan are assigned unique states because 'two' happens not to be attested in the surviving documents of those languages. This character is uninformative because there is only one shared state.

310 'beard'.

Hitt. 1 Av. 4 Luv. 7 Goth. 11
 Arm. 1 OCS 5 Lyc. 8 ON 12
 Gk. 2 Lith. 5 TA 9 OHG 5
 Alb. 1 OE 5 OPer. 10 Welsh 13 [loan]
 TB 3 OI 6 OPru. 5 Osc. 14
 Ved. 1 Lat. 5 Latv. 5 Umb. 15
 1 *smákrú (*smékru ?) 5 *b^harsd^heh₂

Notes on 310 'beard'.

The unique distribution of state 5 is incompatible on all our trees.

341 'free'.

Hitt. 1 Av. 7 Luv. 12 Goth. 10
 Arm. 2 [loan] OCS 8 Lyc. 1 ON 10
 Gk. 3 Lith. 9 TA 13 OHG 10
 Alb. 4 OE 10 OPer. 14 Welsh 10
 TB 5 OI 11 OPru. 15 Osc. 3
 Ved. 6 Lat. 3 Latv. 16 Umb. 17
 1 PAnat. *arawos 10 *priHós 'beloved'
 3 *h₁léwd^heros

Notes on 341 'free'.

The shift of 'beloved' to 'free' might suggest loan-translation reflecting very early contact between Celtic and Germanic (cf. Feist 1939 s.v. *freis* with references). The distribution of the Mediterranean word (state 3) might also reflect a contact phenomenon of some sort.

420 'young'.

[two characters]

Hitt. 1 Av. 5 Luv. 6 Goth. 5x
 Arm. 2 OCS 5 Lyc. 7 ON 5x
 Gk. 2 Lith. 5 TA 8 OHG 5x
 Alb. 3 OE 5x OPer. 9 Welsh 5x
 TB 4 OI 5x OPru. 10 Osc. 11
 Ved. 5 Lat. 5 Latv. 5 Umb. 5
 2 *néwos 'new' and derivs. 5 *h₂yuh₁én-
 5x *h₂yuh₁ñkós

Notes on 420 'young'.

We have adopted both codings for superstate 5, since a direct replacement of the basic word by its derivative is plausible. The narrower coding is incompatible with all of our trees.

421 'tear' (noun, i.e. 'eye-water').

Hitt. 1	Av. 2b	Luv. 5	Goth. 2a
Arm. 2a	OCS 4	Lyc. 6	ON 2a
Gk. 2a	Lith. 2b	TA 2b	OHG 2a
Alb. 3	OE 2a	OPer. 7	Welsh 2a
TB 2b	OI 2a	OPru. 8	Osc. 9
Ved. 2b	Lat. 2a	Latv. 2b	Umb. 10
2a *dákru	2b *ákru		

Notes on 421 'tear'.

There is an obvious relation between the forms represented by the two large states, but its exact nature remains obscure. Though the Hittite word clearly resembles them (mainly because it ends in *-ru*), it is too different from either set to be assigned the same state.

We have coded states 2a, 2b separately, simply because there is otherwise only one shared state. This character is incompatible on all of our trees.

REFERENCES

- ALROY, JOHN. 1995. Continuous track analysis: A new phylogenetic and biogeographic method. *Biology* 44.2.152–78.
- ANDERSEN, HENNING. 1968. IE *s after *i, u, r, k* in Baltic and Slavic. *Acta linguistica Hafniensia* 11.171–90.
- APPEL, RENÉ, and PIETER MUYSKEN. 1987. *Language contact and bilingualism*. Baltimore: Edward Arnold.
- BERGSLAND, KNUD, and HANS VOGT. 1962. On the validity of glottochronology. *Current Anthropology* 3.115–53.
- BONET, MARIA; CYNTHIA A. PHILLIPS; TANDY WARNOW; and SHIBU YOOSEPH. 1996. Constructing evolutionary trees in the presence of polymorphic characters. *SIAM Journal of Computing* 29.103–31.
- BUCK, CARL D. 1949. *A dictionary of selected synonyms in the principal Indo-European languages*. Chicago: University of Chicago Press.
- CARDONA, GEORGE. 1960. *The Indo-European thematic aorists*. New Haven, CT: Yale University dissertation.
- COWGILL, WARREN. 1960. Greek *ou* and Armenian *oč'*. *Language* 36.347–50.
- COWGILL, WARREN. 1965. The Old English present indicative ending *-e*. *Symbolae linguisticae in honorem Georgii Kuryłowicz*, 44–50. Wrocław: Polska Akademia Nauk.
- DEMIRAJ, BARDHYL. 1997. *Albanische Etymologien*. Amsterdam: Rodopi.
- DOBSON, ANNETTE J. 1969. Lexicostatistical grouping. *Anthropological Linguistics* 11.216–21.
- DOBSON, ANNETTE J. 1974. Unrooted trees for numerical taxonomy. *Journal of Applied Probability* 11.32–42.
- EICHNER, HEINER. 1992. Anatolian. *Indo-European numerals*, ed. by Jadranka Gvozdanović, 29–96. Berlin: Mouton de Gruyter.
- EMBLETON, SHEILA M. 1986. *Statistics in historical linguistics*. Bochum: Brockmeyer.
- ESKA, JOSEPH F., and DON RINGE. 2004. Recent work in computational linguistic phylogeny. *Language* 80.569–82.
- EVANS, STEVEN N.; DON RINGE; and TANDY WARNOW. 2005. Inference of divergence times as a statistical inverse problem. *Phylogenetic methods and the prehistory of languages*, ed. by James Clackson et al. Cambridge: Cambridge University Press, to appear.
- FEIST, SIGMUND. 1939. *Vergleichendes Wörterbuch der gotischen Sprache*. Leiden: Brill.
- FITCH, WALTER M. 1971. Toward defining the course of evolution: Minimum change for a specified tree topology. *Systematic Zoology* 20.406–16.
- GAREY, MICHAEL R., and DAVID S. JOHNSON. 1979. *Computers and intractability: A guide to the theory of NP-completeness*. New York: Freeman and Co.
- GLEASON, HENRY A. 1959. Counting and calculating for historical reconstruction. *Anthropological Linguistics* 1.22–32.

- HOCK, HANS HENRICH. 1986. *Principles of historical linguistics*. Berlin: Mouton de Gruyter.
- HOENIGSWALD, HENRY M. 1960. *Language change and linguistic reconstruction*. Chicago: University of Chicago Press.
- HOENIGSWALD, HENRY M. 1987. Language family trees, topological and metrical. In Hoenigswald & Wiener, 257–67.
- HOENIGSWALD, HENRY M., and LINDA WIENER (eds.) 1987. *Biological metaphor and cladistic classification: An interdisciplinary perspective*. Philadelphia: University of Pennsylvania Press.
- HOLDEN, CLARE J. 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: A maximum-parsimony analysis. *Proceedings of the Royal Society of London, Series B* 269.793–99.
- HÜBSCHMANN, H. 1897. *Armenische Grammatik, I. Theil: Armenische Etymologie*. Leipzig: Breitkopf & Härtel.
- JASANOFF, JAY H. 1997. An Italo-Celtic isogloss: The 3pl. mediopassive in *-ntro. *Festschrift for Eric Hamp*, vol. 1, ed. by Douglas Q. Adams, 146–61. Washington, DC: Institute for the Study of Man.
- KANNAN, SAMPATH, and TANDY WARNOW. 1997. A fast algorithm for the computation and enumeration of perfect phylogenies when the number of character states is fixed. *SIAM Journal of Computing* 16.1749–63.
- KIM, RONALD. 2000. 'To drink' in Anatolian, Tocharian, and Proto-Indo-European. *Historische Sprachwissenschaft* 113.151–70.
- KING, RUTH. 2000. *The lexical basis of grammatical borrowing*. Amsterdam: John Benjamins.
- KING, RUTH. 2003. Language contact and linguistic structure. Paper presented at NWAVE 32, Philadelphia.
- LABOV, WILLIAM. 1994. *Principles of linguistic change, vol. 1: Internal factors*. Oxford: Blackwell.
- MAIR, VICTOR (ed.) 1998. *The Bronze Age and early Iron Age peoples of eastern Central Asia*. Washington, DC: Institute for the Study of Man.
- MALLORY, J. P. 1989. *In search of the Indo-Europeans*. London: Thames & Hudson.
- MEILLET, ANTOINE. 1925. *La Méthode comparative en linguistique historique*. Oslo: Aschehoug.
- MOUS, MAARTEN. 1996. Was there ever a Southern Cushitic language (pre-)Ma'a? *Cushitic and Omotic languages*, ed. by Catherine Griefenow-Mewis and Rainer Voigt, 201–11. Cologne: Rüdiger Köppe.
- MOUS, MAARTEN. 1997. The *ela* alternation in Mbugu: The limits of allomorphy. *Linguistics in the Netherlands* 1997.123–34.
- NAKHLEH, LUAY. 2004. *Phylogenetic networks in biology and historical linguistics*. Austin: University of Texas, Austin dissertation.
- PEDERSEN, HOLGER. 1913. *Vergleichende Grammatik der keltischen Sprachen, Zweiter Band*. Göttingen: Vandenhoeck & Ruprecht.
- PORZIG, WALTER. 1954. *Die Gliederung des indogermanischen Sprachgebiets in neuer Sicht*. Heidelberg: Winter.
- PRINCE, ELLEN F., and SUSAN PINTZUK. 2000. Bilingual code-switching and the open/closed class distinction. *University of Pennsylvania Working Papers in Linguistics* 6.3.237–57.
- RAYFIELD, J. R. 1970. *The languages of a bilingual community*. The Hague: Mouton.
- RINGE, DON. 1996. *On the chronology of sound changes in Tocharian*. Vol. 1. New Haven, CT: American Oriental Society.
- RINGE, DON. 1999. How hard is it to match CVC-roots? *Transactions of the Philological Society* 97.213–44.
- RINGE, DON. 2000. Tocharian class II presents and subjunctives and the reconstruction of the Proto-Indo-European verb. *Tocharian and Indo-European studies* 9.121–42.
- RINGE, DON; TANDY WARNOW; and ANN TAYLOR. 2002. Indo-European and computational cladistics. *Transactions of the Philological Society* 100.59–129.
- RINGE, DON; TANDY WARNOW; ANN TAYLOR; ALEXANDER MICHAILOV; and LIBBY LEVISON. 1998. Computational cladistics and the position of Tocharian. In Mair, 391–414.
- ROBERTS, R. G.; R. JONES; and M. A. SMITH. 1990. Thermoluminescence dating of a 50,000-year-old human occupation site in Northern Australia. *Science* 345.153–56.

- ROSS, MALCOLM. 1997. Social networks and kinds of speech-community event. *Archaeology and language 1: Theoretical and methodological orientations*, ed. by Roger Blench and Matthew Spriggs, 209–61. London: Routledge.
- RUVOLO, MARYELLEN. 1987. Reconstructing genetic and linguistic trees: Phenetic and cladistic approaches. In Hoenigswald & Wiener, 193–216.
- SOMMER, FERDINAND. 1948. *Handbuch der lateinischen Laut- und Formenlehre*. 2nd and 3rd edn. Heidelberg: Winter.
- STANG, CHRISTIAN. 1966. *Vergleichende Grammatik der baltischen Sprachen*. Oslo: Universitetsforlaget.
- STILES, PATRICK V. 1985. The fate of the numeral '4' in Germanic. *North-Western European Language Evolution* 6.81–104.
- SWOFFORD, DANIEL. 1996. *PAUP*: Phylogenetic analysis using parsimony (and other methods), version 4.0*. Sunderland, MA: Sinauer Associates.
- THOMASON, SARAH GREY, and TERRENCE KAUFMAN. 1988. *Language contact, creolization, and genetic linguistics*. Berkeley: University of California Press.
- TRUBETZKOY, NIKOLAI. 1926. Gedanken über den lateinischen a-Konjunktiv. *Beiträge zur griechischen und lateinischen Sprachforschung* (Festschrift for Paul Kretschmer), 267–74. Berlin: Deutscher Verlag für Jugend und Volk.
- WARNOW, TANDY; DON RINGE; and ANN TAYLOR. 1995. Reconstructing the evolutionary history of natural languages. (Institute for Research in Cognitive Science report 95-16.) Philadelphia: Institute for Research in Cognitive Science, University of Pennsylvania.
- WARNOW, TANDY; STEVEN N. EVANS; DON RINGE; and LUAY NAKHLEH. 2005. A stochastic model of language evolution that incorporates homoplasy and borrowing. *Phylogenetic methods and the prehistory of languages*, ed. by James Clackson et al. Cambridge: Cambridge University Press, to appear.
- WHITE, J. PETER, and JAMES F. O'CONNELL. 1982. *A prehistory of Australia, New Guinea, and Sahul*. New York: Academic Press.
- WINTER, WERNER. 1998. Lexical archaisms in the Tocharian languages. In Mair, 347–57.
- YOSHIDA, KAZUHIKO. 1990. *The Hittite mediopassive endings in -ri*. Berlin: Mouton de Gruyter.

Nakhleh
 Department of Computer Science
 Rice University
 6100 Main St., MS 132
 Houston, TX 77005-1892
 [nakhleh@cs.rice.edu]

[Received 4 March 2003;
 revision invited 2 September 2003;
 revision received 10 March 2004;
 accepted 15 August 2004]

Ringe
 Department of Linguistics
 619 Williams Hall
 University of Pennsylvania
 Philadelphia, PA 19104-6305
 [dringe@unagi.cis.upenn.edu]

Warnow
 Department of Computer Sciences
 The University of Texas at Austin
 Taylor Hall 2.124
 Austin, TX 78712-1188
 [tandy@cs.utexas.edu]