Perfect Phylogenetic Networks: A New Methodology for Reconstructing the Evolutionary History of Natural Languages

Luay Nakhleh
Dept. of Computer Science
Rice University
Houston, TX 77005
nakhleh@cs.rice.edu

Don Ringe
Dept. of Linguistics
University of Pennsylvania
Philadelphia, PA 19104
dringe@unagi.cis.upenn.edu

Tandy Warnow
Dept. of Computer Sciences
University of Texas
Austin, TX 78712
tandy@cs.utexas.edu

Abstract

In this paper we extend the Ringe-Warnow model of language evolution to include the case where languages remain in contact, trading linguistic material, as they evolve. We describe our analysis of an Indo-European 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.

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 which is referred to when we say, for example, "French and Italian are both descendants of Latin." The evolution of related languages can be mathematically modeled 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 evolved can help answer questions about human migrations, the times at which new technologies were first developed, when ancient people began to use horses, etc. (see e.g. Mallory 1989, Roberts et al. 1990, White and O'Connell 1982).

Various researchers (e.g. Gleason 1959, Dobson 1969, (n.d.), Embleton 1986) have noted that if 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).

[1] We take this opportunity to point out that the similarity of biological and linguistic speciation has nothing whatever to do with 19th-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 1960:144-160, Hoenigswald 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.

Such techniques established the major subfamilies within Indo-European but have not been sufficient to resolve the family's evolution fully; major questions, such as the Indo-Hittite hypothesis, or whether Greek and Armenian are sisters, continue to be debated. More recently, techniques for using multi-state characters have been devised which suggest that essentially all linguistic characters, provided that they are correctly chosen and coded (see below!), should be "compatible" on the true tree (see Ringe, Warnow and Taylor 2002:70-8 with references); in other words, each character should evolve without backmutation or parallel evolution. This condition is also expressed by saying that the tree is a "perfect phylogeny", i.e. a phylogenetic tree which is fully compatible with all the data. (See further below 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 Indo-European (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 non-treelike behavior, evidently acquiring some of its characteristics from its neighbors rather than (only) from its direct ancestors.[3] Consequently, though their methodology seemed promising and offered potential answers to many of the controversial problems in evolution of IE (cf. Jasanoff 1997, Winter 1998, Ringe 2000 with references), it is necessary to extend the model to address the problem of how characters evolve when diverging language communities remain in significant contact. For these cases, trees are not an appropriate model of evolution; "networks" are needed instead to model the evolutionary history of the family.

In this paper we show how to extend the perfect phylogeny approach to the case where the language family requires a network model (that is, an underlying tree with additional "contact" edges; see Figure 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 has found several networks with a very small number of "contact" edges which are plausible with respect to what is known about the IE family's geography. The study thus leads us to conjecture that the IE family, though it did not evolve through clean speciation, exhibits a pattern of initial diversification that is close to tree-like: 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 which are similarly close to tree-like in their evolution.

The rest of this paper is organized as follows. We review the model of Ringe and Warnow in Section 2, and we present our extension to the case of network evolution in Section 3. In Section 4 we describe the data we use to represent the IE family. In Section 5 we describe our computational analysis of the data which results in the candidate networks we then consider. In Section 5.5, we compare the candidate networks in the light of known IE history, thus producing our set of five feasible solutions. In Section 6 we discuss the best network we found in detail. We conclude in Section 7 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

[3]We wish to emphasize that this appears to be an ineluctable conclusion of Ringe, Warnow and Taylor 2002; we see no grounds for questioning it and will not revisit the problem here. Interested readers are referred to Ringe, Warnow and Taylor 2002, especially pp. 85-92.

and the computational approach are given in an appendix. The full set of our coded data, together with a list of characters omitted and the reasons for their omission, will be made available in a further online appendix.

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 upon a subset of the characters (eliminating those with clear parallel development, in particular). We also think 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 which are (at least provisionally) accepted as linguistically realistic have not yet been established.

Experience shows that it is easy to construct a comparative dataset using only qualitative characters that evolve without backmutation–that is, characters which never change from one state to a second state (and potentially to a third, etc.) and then finally back to the first state (see Ringe, Warnow and Taylor 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.[4] 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 *-ism.o-), and the Romance languages that have lost the Latin category would be assigned not state 1 but states 3, 4, 5, etc.–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

[4]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 discussion of a more idiosyncratic example see Ringe, Warnow and Taylor 2002:100.

a different word Y which was then completely replaced by X again.[5] 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 likelihood that a word which 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 which can recur at widely separated times and places (see e.g. Ringe, Warnow and Taylor 2002:66-68). Phonological characters must therefore be based on highly unusual sound changes or on clusters of sound changes which are not especially likely to have occurred together, and that greatly decreases the amount of phonological information which can be used for cladistic purposes. Parallel shifts in the meanings of words are also very common. Fortunately the collective experience of historical linguists seems to show that most parallel development 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. But 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 that also reduces the amount of usable evidence substantially.[6]

Borrowing of states between significantly different languages is also a problem, but one that has been greatly overestimated in the recent literature. The assumption that "anything can be borrowed" from one language into another has been given wide currency by Thomason and Kaufman 1988. But in a devastating critique of their work Ruth King has argued persuasively that that assumption has never been proved (see King 2000:44-47 with references, 2003; cf. also Appel and Muysken 1987:158-163). 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 instead of searching in the literature on language contact for processes that might give rise to such unexpected outcomes, Thomason and Kaufman assume without discussion that inflectional morphology and "function words" can be borrowed into one's native language in much the same way as major lexemes. That is certainly not a responsible application of the uniformitarian principle. Still worse, there is now some research on language contact which shows that the transfer of "closed-class" items from language to language typically occurs via processes quite different from the borrowing of foreign words into one's native language. 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 round (cf. Rayfield 1970:103-107, Prince and 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 specific morphosyntactic features which are unusual in French, and the resulting perturbation of native morphosyntax gives rise to new syntactic patterns–which are *not* identical

[5]Our lexical characters are defined semantically, each cognate set comprising a single state of the character, for the reasons outlined in Ringe, Warnow and Taylor 2002:71, fn. 8.

[6]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 which we are addressing in other work.

with English patterns, but are similar enough that an unsophisticated approach to syntax will not find the differences (see especially King 2000).

In light of the above, we accept the hypothesis that borrowing into native dialects from languages or dialects that are not closely related is tightly constrained, lexemes being virtually the only type of linguistic unit that is borrowed outright. Since it is quite clear that other types of contact phenomena between significantly different speechforms are untypical and relatively uncommon (cf. Thomason and Kaufman 1988:3, Ross 1997:209-210), we expect to find detectable evidence of contact only among lexical characters in the default case. So far as we can tell, that is what 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.[7] But it is reasonable to suspect that not all loanwords are detectable as such, a matter that will occupy us at greater length below.

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, so that 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 the study. We have therefore excluded polymorphic characters from this analysis.

If the constraints just described are met, 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 which exists nowhere else at that time and has not appeared earlier. We express that 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 which 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 the characters in the dataset 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 tree-like (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 evolution may not be sufficiently "clean". More precisely, whereas borrowing between clearly different speech forms 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 which 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 an Indo-European dataset by Warnow and Ringe (first reported in

[7]The points discussed in this paragraph and the ones immediately above have been treated at greater length in Ringe, Warnow and Taylor 2002, to which the reader is referred.

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

(a)

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

(b)

Figure 1: (a) Five languages $L_1, \ldots, L_5$, with two characters c_1, and c_2. (b) The same five languages with a third character c_3.

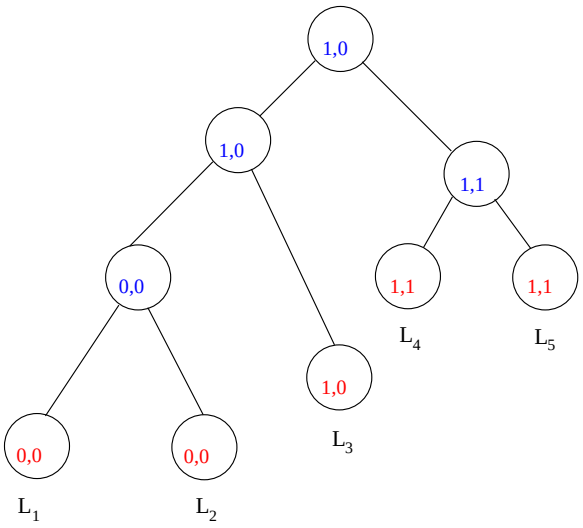

Figure 2: A perfect phylogeny T for the languages and character states of Figure 1(a).

Warnow, Ringe and Taylor 1995; substantially revised and augmented in Ringe, Warnow and Taylor 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, but by no means all. 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" 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 towards the descendent languages, so that all the edges in the tree are directed (from the root towards 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". Furthermore, they are "bi-directional", 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. Figure 3 gives an example of one such network.

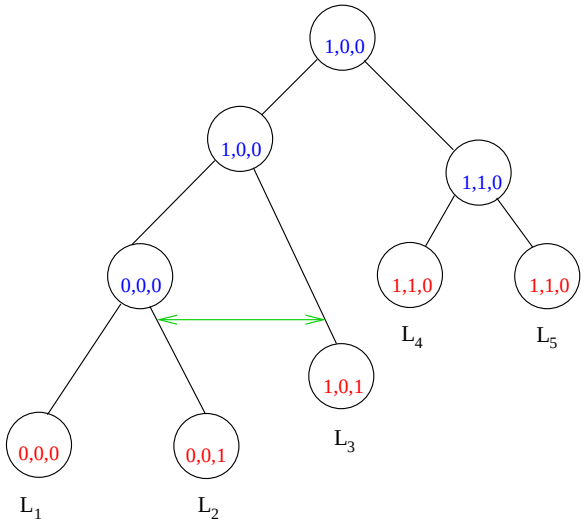

Figure 3: A perfect phylogenetic network N for the languages and character states of Figure 1(b).

We begin our discussion of how to extend the notion of character compatibility from trees to networks by observing that each character will evolve down one of the "trees contained in the network." Figure 3 shows a network N that contains, in addition to the underlying tree, one contact edge between two reasonably closely related languages. The tree is characterized by a 3-tuple of characters c_1, c_2, c_3; each character has two states, 0 and 1. These are the languages and characters of Figure 1(b), for which there does not exist a perfect phylogenetic tree. 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. Figure 4 shows the three possible trees within the network down each of which characters can evolve–that is, each of which potentially models the evolutionary history of one or more of the characters.

To motivate our model of character evolution down networks we begin with the tree model. When trees are reasonable models 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 (there being no 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 the network.

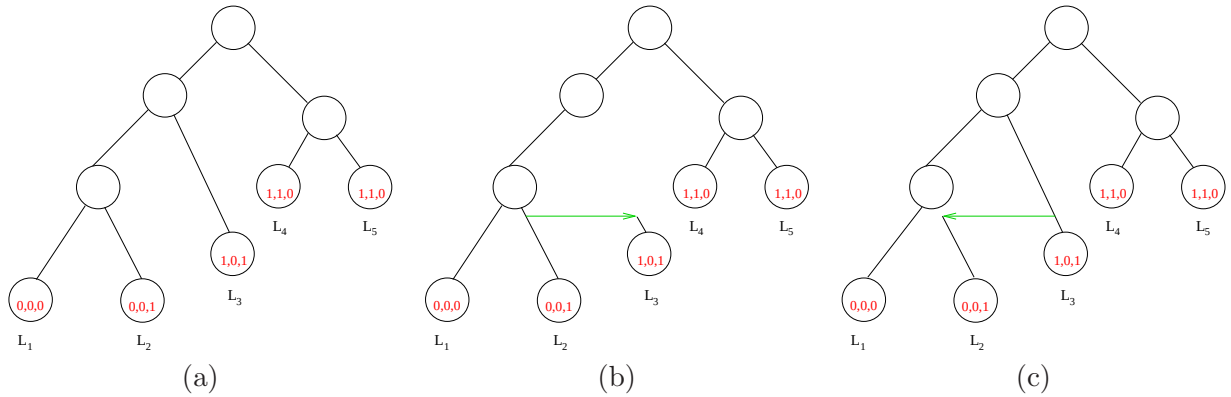

Figure 4: The three trees contained inside the network in Figure 3. While the network "reconciles" the evolutionary history of all three characters, each one of those characters actually evolved down exactly one of the three trees.

The three characters of the network N shown in Figure 3 are compatible with N, since each of the characters is compatible with at least one of the three trees contained inside N: characters c_1 and c_2 are compatible with the tree in Figure 4(a), and character c_3 is compatible with both trees in Figures 4(b) and 4(c).

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 will say that a network is a "perfect phylogenetic network" for a set of languages described by a set of qualitative characters if every character is compatible with the network (e.g., the network N in Figure 3 is a perfect phylogenetic network for the languages and characters in Figure 1(b)).

A perfect phylogenetic network 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.

We note that these different criteria measure different things, and hence different ways of deviating from a tree will be evaluated differently depending on the criterion chosen. For example, if the wave model is appropriate, we should not be able to find a clearly defined underlying genetic tree on which the vast majority of the characters are compatible; most characters may be involved in "borrowing", and so even if the number of contact edges is somewhat small (because constrained by geographical considerations, for example), the other two criteria should yield fairly poor values (i.e., the proportions of characters simultaneously compatible will be low, and the number of borrowing events will be high). On the other hand, if the language family evolves 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 the genetic tree should be mostly discernible because of the compatibility pattern: we should find only a small number of trees with each of which a large percentage of the characters is compatible, and those trees should differ only because of the

difficulty in distinguishing tree edges from contact edges. As the number of contact edges increases, although it might still make 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 that in which a wave model is appropriate.

The debate about Indo-European's history has to some extent focused on these 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 some borrowing. Furthermore, our proposal allows for 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, amounts to a combination of those two.) In this paper we both present this model and attempt to discover where Indo-European (IE) fits within 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 a wave model really a better model? As we will see, our analysis shows dramatic support for the claim that Indo-European evolution is largely treelike: almost all (95%) of the characters evolve down our proposed genetic tree, and we only need 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 history.

Since the simplest model is the most desirable, other things being equal ("Occam's Razor"), we will want to find a perfect phylogenetic network 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 network would explain the evolution of all 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 perfect phylogenetic network 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 perfect phylogenetic network. Accordingly our approach involves two steps:

- Given the set L of languages described by set C of qualitative characters, find or estimate the optimal "genetic tree" T for L. (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, Warnow and Taylor 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 perfect phylogenetic network.

(This is roughly similar to the approach of Alroy 1995.) For example, the characters c_1 and c_2 in Figure 1(b) are compatible with the tree T in Figure 2, whereas character c_3 is not. By adding one contact edge to T, we obtain the network N of Figure 3, on which all three characters are compatible.

This is the approach that we used in this study in order to analyze the Indo-European dataset compiled by Ringe and Taylor. Because our data were close to tree-like, 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 IE Dataset

Our basic dataset consists of 294 characters for 24 IE languages. We will first describe and explain our choice of languages and characters, then describe our coding of the characters.

The languages are listed in Table 1.

Table 1: The 24 IE languages analyzed.

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	PR
Umbrian	UM	Lithuanian	LI
Gothic	GO	Latvian	LT
Old Norse	ON	Albanian	AL

As can be seen, they represent all ten well-attested subgroups of the IE family (namely Anatolian, Tocharian, Celtic, Italic, Germanic, Albanian, Greek, Armenian, Balto-Slavic, and Indo-Iranian). To represent each subgroup we have chosen a language or languages that are attested relatively fully at as early a date as possible. For instance, Indic is represented by early Vedic, since the Rigveda and other very early texts are extensive enough to provide us with data for most of our characters; but we have used "younger" Avestan rather than the earlier Gatha-Avestan to represent eastern Iranian, since the Gathas are too restricted for our purposes. Greek is represented by Classical Attic rather than Homeric, both because our attestation of Attic is far more extensive and because the Homeric language is known to be an artificial literary dialect. Similar decisions have been made in the other cases. We have used modern data for Welsh, Lithuanian, Latvian, and Albanian because earlier data are much less accessible and because we judged that it would make little difference in those cases. 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 practicable we were obliged 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 wordlist), Oscan (ca. 20%), Umbrian (ca. 25%), Old Persian (ca. 30%), and Luvian (ca. 40%). At the other extreme we have complete or virtually complete ($\geq 99\%$) wordlists 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 ($\geq 95\%$) wordlists for Vedic, Classical Armenian, Old Irish, and Old Church Slavonic. Coverage of the remaining wordlists ranges from about 70% to about 85%. 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–which is, of course, realistic.

The inclusion of three Baltic languages and of four Germanic languages introduces parallel development in a considerable number of lexical characters, thus decreasing the amount of usable evidence. We have retained the full set of languages in the database because the internal subgrouping of Balto-Slavic and of Germanic are matters of ongoing debate in the specialist community.[8] On the other hand, the inclusion of only two West Germanic languages–Old English and Old High German, the northernmost and southernmost respectively–potentially avoids much greater character incompatibilities, since the internal diversification of West Germanic is known to have been radically non-treelike (cf. Ringe, Warnow and Taylor 2002:110).

Our database includes 22 phonological characters encoding regular sound changes (or, more often, sets of sound changes) that have occurred in the prehistory of various languages, 13 morphological characters encoding details of inflection (or, in one case, word formation), and 259 lexical characters defined by meanings on a basic wordlist. (See the online appendix 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 as follows.

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. For a detailed presentation of the phonological characters see Ringe, Warnow and Taylor 2002:113-116.

In attempting to find viable morphological characters Ringe and Taylor made a startling discovery: the states of most morphological characters are either 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 the morphological characters which might aid in determining the first-order subgrouping that Ringe and Taylor were able to discover. For details see Ringe, Warnow and Taylor 2002:117-120.

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;

[8] We have no reason to doubt the cladistic structures of these subgroups found in Ringe, Warnow and Taylor 2002, which were very robust and are consistent with one of the standard alternative opinions, and we will not revisit the question here.

to it they added about 120 meanings that appear to be culturally significant for older IE languages.

The database just described was the basis of the analysis reported in Ringe, Warnow and Taylor 2002. For the present project we have modified it in the following ways. We have excluded all polymorphic characters from the dataset (for the reasons outlined above); we even exclude those polymorphic characters which can be recoded as pairs of monomorphic characters (see Ringe, Warnow and Taylor 2002:84-5). We have also excluded 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, Warnow and Taylor 2002. (For further discussion of the reasons why individual characters were excluded see the online appendix.)

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 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. 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 the languages is needed to make such a determination. More information about our coding of the data can be found in Ringe, Warnow and Taylor 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.) Secondly, 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, Warnow and Taylor 2002:82-3. (One morphological character is also double-coded.) 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 the phonological characters, or with all 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 1986:442-4 with references p. 667). Every candidate tree is also required to be compatible with eight of our morphological characters (M3, M5-6, M8, and M12-15).

5 Phylogenetic Analysis

5.1 Overview

We analyzed five candidate genetic trees, three (Trees A, B, and C) that our team has come to consider in recent years (shown in Figures 5–7), and two (Trees D and E) that were suggested to us by our colleague Craig Melchert (shown in Figures 8 and 9). We selected these trees because they are each compatible with the vast majority of the characters (the best of these trees is compatible with more than 95% of the characters, and the worst is compatible with 92% of the characters), but are also compatible with all the morphological characters, which are expected to be the most resistant to borrowing (cf. Meillet 1925:22-33, Ringe, Warnow and Taylor 2002:65). 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.

For each tree, we sought to add a minimum number of contact edges in order to produce a perfect phylogenetic network (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 contact was feasible on the basis of known constraints (geographic and chronological) on the evolution of the IE family. This led us to reject all but 5 of the resultant perfect phylogenetic networks (three on Tree A, and two on Tree B). Of these, one (on Tree A) was the most believable. Interestingly, this most believable of the numerous PPNs we obtained 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 all others, and suggests strongly that the IE family evolved largely in a tree-like fashion–sufficiently so that the underlying genetic tree is largely recoverable, and so that specific contact episodes between neighboring linguistic communities can also be detected.

In the remainder of this section 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) that our team has come to consider in recent years (shown in Figures 5–7), and two (Trees D and E) that were suggested to us by our colleague Craig Melchert (shown in Figures 8 and 9).

The main difference between these five trees is the placement of Germanic. The differing placement of Albanian in these trees is 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 the diagnostic inflectional morphology and a large proportion of its inherited vocabulary. Thus each tree actually represents several trees which differ only with respect to exactly where Albanian attaches.[9] In contrast, the variable placement of Germanic appears to reflect a major idiosyncrasy of that subgroup's evolution, one that led to the conjecture that Germanic may not have evolved in a strictly tree-like fashion (Ringe et al. 1998:407-8; Ringe,

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

Warnow and Taylor 2002:110-1). Our detailed analysis of these different scenarios allows us to test each of the conjectured histories for Germanic.

The differing positions of Germanic in the five trees result in different character incompatibility patterns, as follows: for Tree A the 14 incompatible characters are all lexical; for Tree B the 19 incompatible characters include two phonological and 17 lexical characters; for Tree C the 17 incompatible characters are all lexical; for tree D the 21 incompatible characters are all lexical, and for Tree E the 18 incompatible characters include two phonological and 16 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 which 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 history. The incompatible characters for trees A and B are incomparable: 12 lexical 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 light of a further analysis in which we extend them to PPNs.

Tree D (in Figure 8) differs significantly from Tree A in not grouping Greek and Armenian together; Tree E (in Figure 9) differs from Tree A in its placement of Balto-Slavic. There were 21 incompatible characters for Tree D, all of which are lexical. The incompatible characters for Tree A are a proper subset of the incompatible characters from Tree D; as with our comparison between Trees A and C, we will only consider D a preferred hypothesis over Tree A if we can complete Tree D to a PPN which improves upon our best PPN for Tree A in some way. There were 18 incompatible characters for Tree E, 2 phonological and 16 lexical; the set of incompatible characters is incomparable to the set of characters incompatible with Tree A.

5.3 Constructing the Perfect Phylogenetic Networks

The second phase of this analysis commenced after we had developed the algorithms and software to determine all the ways we could add a minimum number of edges to a tree in order to obtain perfect phylogenetic networks. The problem of adding a minimum number of edges to a tree to obtain a perfect phylogenetic network is computationally difficult; consequently we used a heuristic which allowed us to obtain solutions in a reasonable amount of time. Our technique for constructing a perfect phylogenetic network from a tree is as follows:

- **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*, *Greek-and-Armenian* (for those trees in which Greek and Armenian are sisters - i.e. for all our trees other than Tree D), and *Baltic*. Albanian and Old Church Slavonic remain as individual languages.

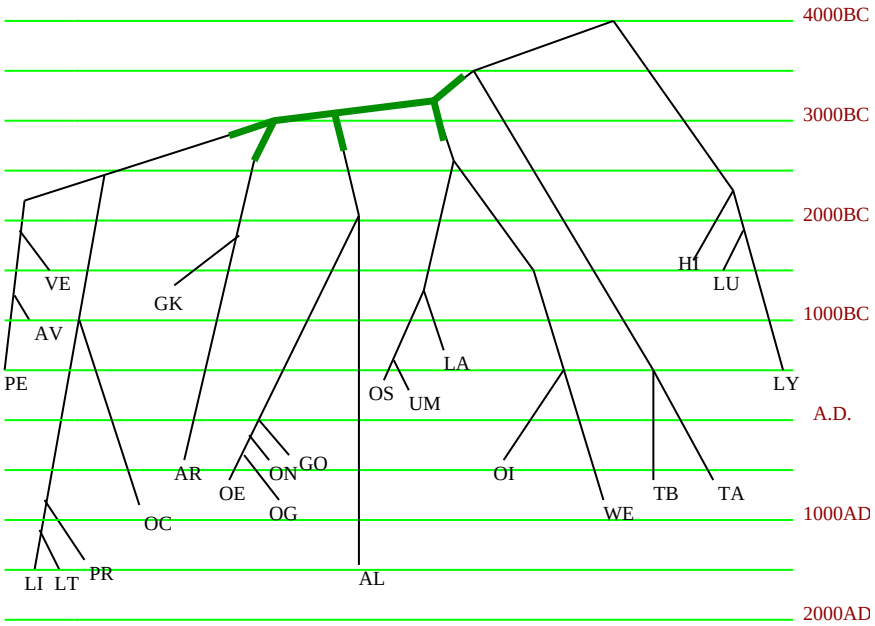

Figure 5: Tree A. The thick lines represent the region within which Albanian can attach without changing the quality of the outcome.

- **Adding the minimum number of contact edges.** After the pruned tree was obtained, the algorithm searched for all the ways we could add a minimum number of contact edges and get a perfect phylogenetic network. This part of the analysis took 8 hours on each of our candidate trees and found several networks with only three contact edges.

5.4 The set of PPNs we obtained

Table 2 summarizes the quantitative results of the analysis on each of the five trees. In the next section we analyze the solutions found on each of the five trees.

Table 2: Summary of the results of the phylogenetic analysis of the five IE trees described in Figures5–9.

	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 solutions	3	2	0	0	0

The result of our brute-force analysis produced 23 PPNs, each with only three contact edges; 16 of these were based on Tree A, 4 were based on Tree B, one was based on Tree C, two on Tree D, and none on Tree E. We compared these 23 PPNs in two ways:

- with reference to linguistic and archaeological evidence, which can render certain proposed

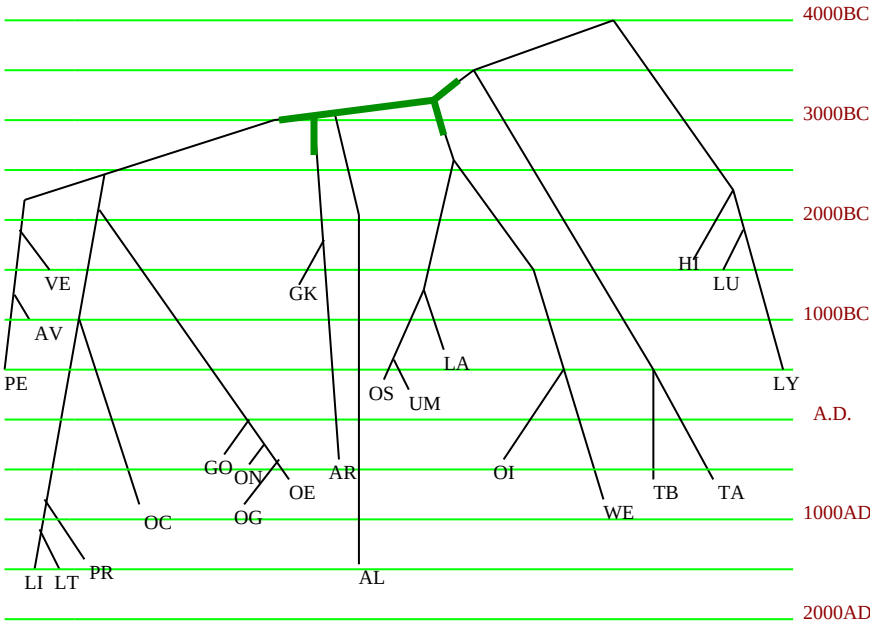

Figure 6: Tree B. The thick lines represent the region within which Albanian can attach without changing the quality of the outcome.

contacts unlikely or even impossible (cf. Mallory 1989), and

- according to the mathematical criteria we proposed earlier in the paper: (1) the number of characters compatible on the genetic tree, (2) the number of contact edges in the PPN, and (3) loan parsimony, i.e. the number of undetected borrowing events in the PPN.

Of the 23 PPNs with three contact edges that we found, five were consistent with known geographical and chronological constraints. Since three of these were based on Tree A (our preferred solution for the genetic IE tree, based upon the number of compatible characters), it seems reasonable to exclude PPNs based on any of our trees with more than three contact edges.

5.5 Comparison of the PPNs

In this section we compare the 23 PPNs we obtained. We present the comparison with respect to geographic and chronological constraints first, and then the evaluation with respect to mathematical criteria. We observed several interesting things, the most striking of which is that the tree that optimized the mathematical criteria was also the most feasible with respect to the geographical and chronological constraints.

Finding the feasible solutions We begin our discussion with our favored candidate for the genetic tree, Tree A, shown in Figure 5. This was previously published in Ringe, Warnow and Taylor 2002.

PPNs based on Tree A The dates assigned to the terminal nodes of Tree A (shown in Figure 5) are obtainable from historical data. Since our method is not distance-based, it is not necessary

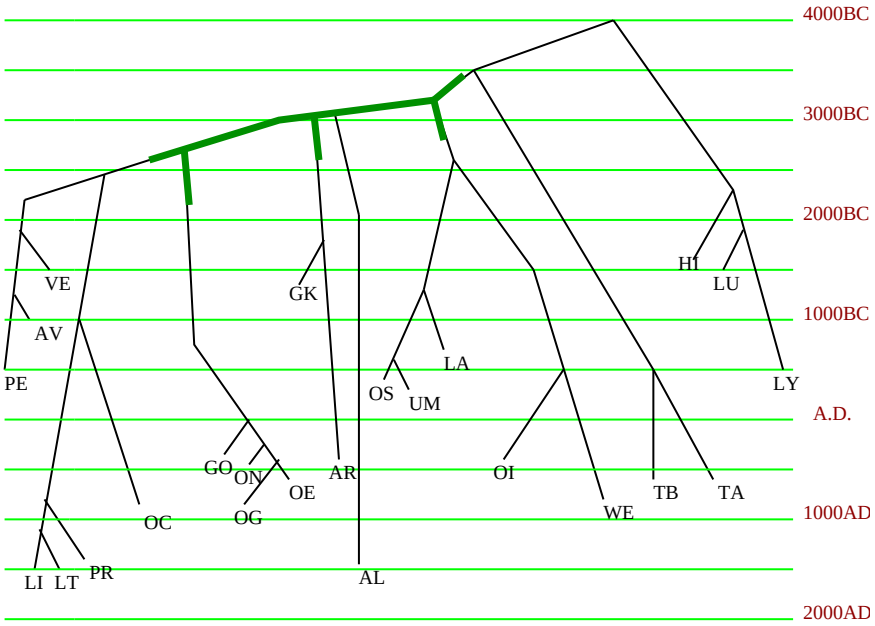

Figure 7: Tree C. The thick lines represent the region within which Albanian can attach without changing the quality of the outcome.

to use the dates of the substantial corpora which underlie our lexical lists. Instead we have here employed the dates of the earliest documentary material of each language which shows clearly that it was already different from its closest relatives. Dates assigned to the internal nodes are necessarily the result of informed guesswork, since proposed "glottochronological" methods for determining the dates of prehistoric languages have proved to be unreliable (see especially Bergsland and Vogt 1962, none of whose objections have been effectively met by recent work, and Eska and Ringe, forthcoming). Dates for a few internal nodes can be fixed with reasonable certainty.[10] For instance, the complete archaeological continuity between the Yamna Horizon (up to ca. 2200 B.C., 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 B.C., give or take a couple of centuries (cf. the discussion of Mallory 1989:210-5, 226-9). Since Indo-Iranian is one of the most deeply embedded subgroups in the tree, it follows that all 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 16 solutions for Tree A, three of which were consistent with known constraints on the history of the IE family (cf. Mallory 1989). Those 16 solutions are described in Table 3 (the feasible solutions are highlighted in the table).

Three feasible PPNs based on Tree A were found; these are solutions 1, 3, and 5 in Table 3. Of these, the first feasible PPN (solution 1, see also Figure 10) is clearly more plausible than the second feasible (solution 3, see also Figure 11).

Both posit a contact edge between Proto-Tocharian and Proto-Slavic. That is somewhat sur-

[10]The reasoning 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.

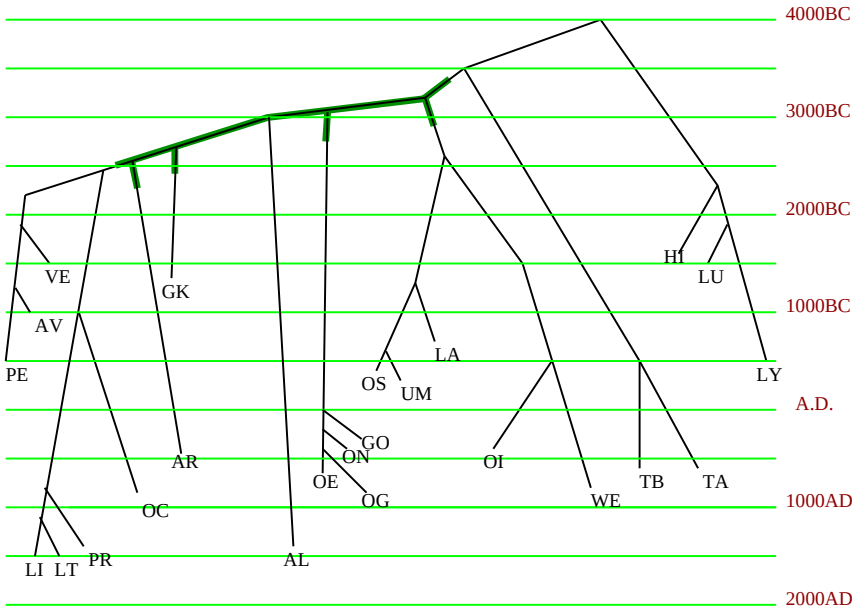

Figure 8: Tree D. The thick lines represent the region within which Albanian can attach without changing the quality of the outcome.

prising, because it implies that an ancestor of Tocharian was still in eastern Europe in the last millennium B.C.E., and it seems clear that by the turn of the millennium they were within striking distance of Xinjiang (where the Tocharian languages are actually attested from about the 6th c. C.E.). However, we know very little about the prehistoric movements of speakers of 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 Figure 10 also posits a contact edge between Proto-Celtic and Proto-Germanic, which is unobjectionable, and one between Proto-Celtic and Proto-Balto-Slavic, which is surprising; the PPN in Figure 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 is likewise surprising. In other words, each of these PPNs posits that one of the "western" 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 can have been very different in, say, the 3rd millennium B.C.E. In particular, it is possible that speakers of Proto-Italo-Celtic occupied the Hungarian plain in about 3200 B.C.E. (David Anthony, personal communication), 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

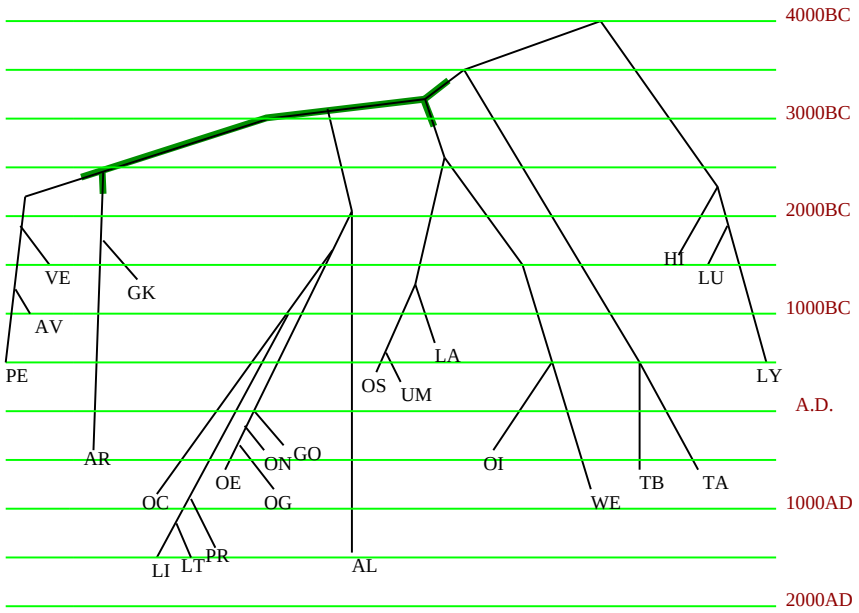

Figure 9: Tree E. The thick lines represent the region within which Albanian can attach without changing the quality of the outcome.

positions of the internal nodes of all 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.)

We now consider the third feasible PPN based on Tree A (solution 5, see Figure 12). 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, however, of these three contact edges, the questionable contact edge has the smallest support: only two of the characters ('free' and one alternative coding of 'young') must use that contact edge, compared to 6 on the first contact edge and 9 on the second. It is therefore possible that this questionable contact edge between Proto-Italic and Proto-Greco-Armenian is not realistic, and that some other explanation (such as undetected homoplasy?) should be found for the non-treelike evolution of 'free' and 'young'. The other two contact edges, which are very well supported, do seem realistic.

With the exception of the three aforementioned solutions, all other solutions are implausible or actually impossible for chronological and geographical reasons, as follows. The temporal constraints which we applied to find candidate contact edges (see the previous section) are the ones which 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 contact with its living ancestor,[11] it is equally impossible for two living languages spoken at different periods in history to be in contact. Solutions 6 through

[11]Literary contact with a dead ancestor, such as contact between medieval or modern Romance languages and Latin, is of course a completely different process; it does not come into question in preliterate societies of the sort under consideration here.

Table 3: The 16 3-edge solutions found on Tree A; the highlighted rows correspond to the three feasible solutions (1, 3, and 5). Each solution is described in terms of the three lateral edges added to Tree A to produce a PPN. The abbreviations are explained in Table 4. The rightmost column summarizes the minimum number of borrowing events needed on each of the 16 PPNs in order to make all characters compatible.

	Solutions			min # borrowing events on PPN
	Edge 1	Edge 2	Edge 3	
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

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 diverge into the attested Italic languages well before 1000 B.C.E., because our earliest documents from the Osco-Umbrian subgroup, in about the 6th c. B.C.E., exhibit so many innovations not shared by Latin that it is clear that those two halves of Italic had been diverging for centuries. But Baltic is most unlikely to have begun diverging from Slavic by 1000 B.C.E., because Proto-Slavic seems still to have been more or less uniform in the 8th c. C.E., 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 3-edge solutions for Tree B, two of which were consistent with known constraints on the history of the IE family. Those solutions are described in Table 5.

The first solution (solution 1 in Table 5) posits contact between (pre-)Italic and the ancestor of Germanic and Balto-Slavic (reasonably plausible), between Germanic and Celtic (highly plausible), and Tocharian and Slavic (as for two solutions on Tree A); the minimum number of borrowing events

Table 4: Abbreviations for proto languages.

Abbreviations	
PT	Proto-Tocharian
PG	Proto-Germanic
PI	Proto-Italic
PC	Proto-Celtic
PIC	Proto-Italo-Celtic
PB	Proto-Baltic
PS	Proto-Slavic
PGA	Proto-Greco-Armenian
PBS	Proto-Balto-Slavic
PBSII	Proto-Balto-Slavic and Indo-Iranian
PII	Proto-Indo-Iranian
PAL	pre-Albanian
PALG	Proto-Albanian and Germanic
PA	Proto-Anatolian

Table 5: The 4 3-edge solutions found on Tree B; the two feasible solutions highlighted. Each solution is described in terms of the three lateral edges added to Tree B to produce a PPN. The abbreviations are explained in Table 4. The rightmost column summarizes the minimum number of borrowing events needed on each of the 4 PPNs in order to make all characters compatible.

	Solutions			min # borrowing events on PPN
	Edge 1	Edge 2	Edge 3	
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

needed to make all characters compatible on this PPN is 21. This solution appears possible, but it is not markedly better than the first two solutions on Tree A, and it is considerably less plausible than the third. The perfect phylogenetic network that corresponds to this solution is shown in Figure 13.

The next solution (solution 2 in Table 5) posits contact between pre-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 Italic and Balto-Slavic (surprising, but not necessarily out of the question), and Tocharian and Slavic (as above); the minimum number of borrowing events needed to make all characters compatible on this PPN is 23. 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 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.

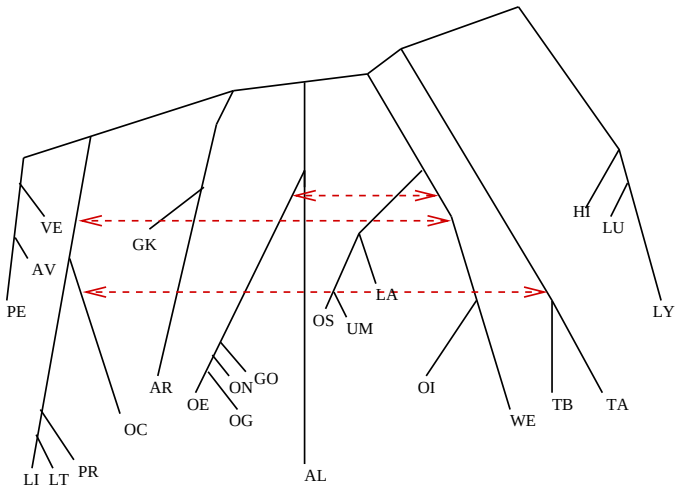

Figure 10: The first feasible perfect phylogenetic network based on Tree A (corresponding to Solution 1 in Table 3). The solid black edges correspond to the underlying tree edges and indicate vertical transmission; those edges are directed. The dashed lines represent contact between language groups, and they are bi-directional.

We now describe the two remaining solutions, and why we can eliminate these on the basis of known constraints. The first solution (solution 3 in Table 5) posits contact between Italic and Baltic but not Slavic, which seems impossible; the minimum number of borrowing events needed to make all characters compatible on this PPN is 25. The second solution (solution 4 in Table 5) posits contact 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 24. Both these solutions are thus infeasible.

PPNs based on Tree C The algorithm described above generated one 3-edge solution for Tree C. However, the 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 24.

PPNs based on Tree D The algorithm described above generated two 3-edge solutions for Tree D. However, both solutions posit a contact between Italic and Baltic but not Slavic, which seems impossible both for chronological and for geographical reasons.

PPNs based on Tree E The algorithm described above did not find any 3-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 re-analyzed Tree A (Figure 5) without pruning the Italic and Greco-Armenian groups (recall that one of the steps in our phylogenetic analysis involves pruning maximally compatible subtrees). In this case, three new PPNs with three lateral edges were found (in

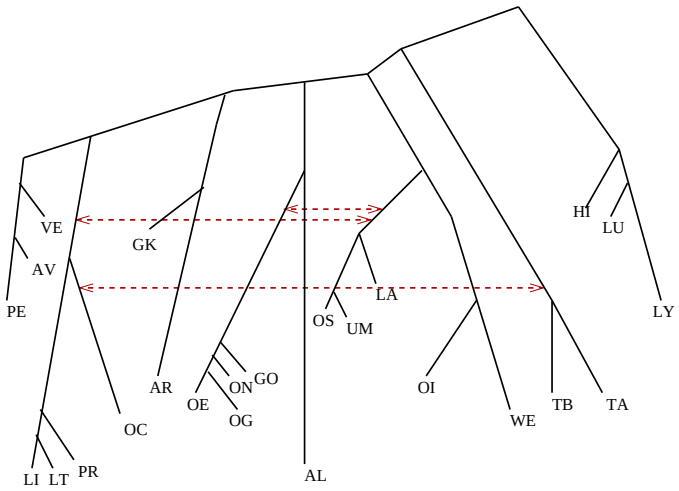

Figure 11: The second feasible perfect phylogenetic network based on Tree A (corresponding to Solution 3 in Table 3). The solid black edges correspond to the underlying tree edges and indicate vertical transmission; those edges are directed. The dashed lines represent contact between language groups, and they are bi-directional.

addition to the 16 PPNs reported earlier). 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, the three solutions are implausible, as discussed earlier.

Summary of the Results Our five different plausible 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 solution–the third on Tree A (solution 5) –is much less surprising; in fact, the contact between Germanic and Baltic which it posits is so highly plausible that this solution is probably preferable to the others on that ground alone.

5.6 Comparison of all 23 PPNs with respect to mathematical criteria

Recall the three mathematical criteria by which we evaluate PPNs: (1) the number of characters compatible on the genetic tree, (2) the number of contact edges in the PPN, and (3) loan parsimony, i.e. the number of undetected borrowing events in the PPN.

Of the 23 PPNs – feasible and infeasible together – the best that we can do with respect to criterion (1) is 14 (since all PPNs based on Tree A have only 14 characters incompatible on them, while PPNs based on the other candidate trees have more). All the PPNs we consider posit three contact edges, so there is no difference with respect to criterion (2). For the third criterion, the best we can do is 17, which is achieved by solution 5 on Tree A; all other solutions must posit at least 19 borrowing events (and usually more). Thus solution 5 on Tree A optimizes all three of the mathematical criteria, and is the unique optimal solution.

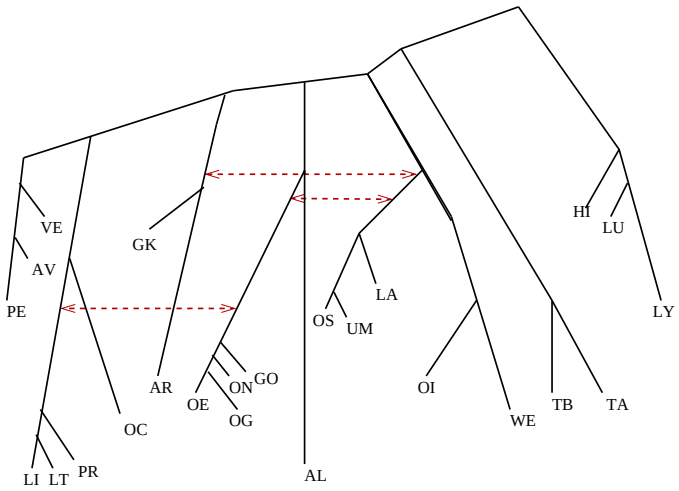

Figure 12: The third feasible perfect phylogenetic network based on Tree A (corresponding to Solution 5 in Table 3). The solid black edges correspond to the underlying tree edges and indicate vertical transmission; those edges are directed. The dashed lines represent contact between language groups, and they are bi-directional.

5.7 Discussion

Thus our favored PPN–solution 5 based on Tree A–is optimal with respect to each of the three mathematical criteria and is also most credible in terms of known geographic and chronological constraints. That is, it optimizes each of the four criteria. Our third PPN on Tree A is therefore the solution that we propose for the first-order evolution of the IE family. More research will of course be needed to confirm this. Because this PPN is so clearly better than the other scenarios, a closer look at it is justified. That is the focus of our next section.

6 Our proposed solution to IE evolution

Our best solution for the historical diversification of the Indo-European language family posits Tree A (the tree found in Ringe, Warnow and Taylor 2002) as the underlying genetic tree and three contact edges; our proposed solution is thus the perfect phylogenetic network of Fig. 12. We note that our network suggests *at most* three 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 of the evolution of the two involved characters ('free' and 'young') that will 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

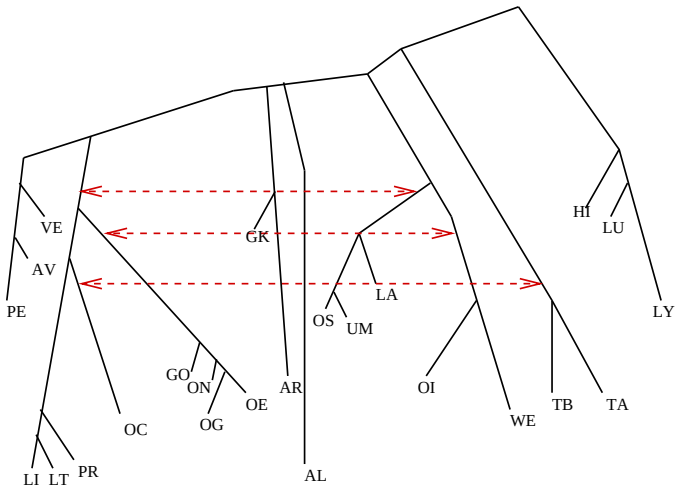

Figure 13: The first perfect phylogenetic network on Tree B. The solid black edges correspond to the underlying tree edges and indicate vertical transmission; those edges are directed. The dashed lines represent contact between language groups, and they are bi- directional.

entirely treelike, and that it may be sensible to infer a genetic tree on which some borrowing (previously undetected) has occurred.

7 Conclusions and Future Work

It is clear that historical linguists can make good use of both models and methods for inferring evolutionary histories of datasets, and that 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 with additional linguistic considerations allows us to determine a small set of feasible scenarios which 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 tree-like: not only do we need to posit only three contact edges, but almost all the characters evolve on the underlying genetic tree. These two properties together suggest that the construction of a genetic tree for Indo-European is sensible, and even plausible. Finally, the agreement between the optimal solution to the mathematical criteria we posed, and the optimal solution with respect to both archaeological and chronological constraints, suggests that the model we have posed may be sufficiently realistic to be useful to historical linguistic inquiry in general. Our new methodology has found a possible solution to an old and intractable problem in historical linguistics, and we think it is reasonable, on those grounds, to argue that this methodology is promising and should be pursued.

8 Acknowledgments

This work was supported in part by the David and Lucile Packard Foundation (T. Warnow) and by the National Science Foundation with grants EIA 01-21680 (T. Warnow), BCS 03-12830 (Warnow), and BCS 03-12911 (Ringe). T. Warnow would like to acknowledge and thank The Radcliffe Insti-

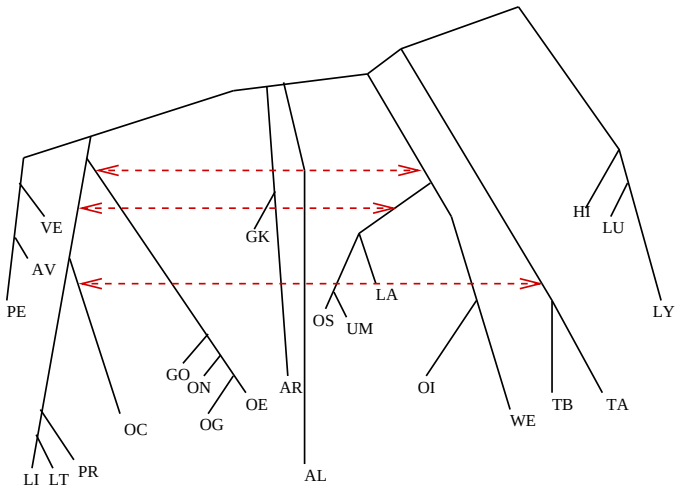

Figure 14: The second perfect phylogenetic network on Tree B. The solid black edges correspond to the underlying tree edges and indicate vertical transmission; those edges are directed. The dashed lines represent contact between language groups, and they are bi- directional.

tute for Advanced Study, the Program in Evolutionary Dynamics at Harvard University, and The Institute for Cellular and Molecular Biology at UT-Austin for their support during the time this work was done. The authors would like to thank Ann Taylor for her help in putting the dataset together, and James Clackson, Joe Eska, and Craig Melchert for their help with the data. 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.

References

Adams, Douglas Q. (ed.). 1997. Festschrift for Eric Hamp. Vol. I. Washington: Institute for the Study of Man.

Alroy, John. 1995. Continuous track analysis: a new phylogenetic and biogeographic method. Systematic Biology 44(2):152-78.

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.

Blench, Roger, and Matthew Spriggs (edd.). 1997. Archaeology and language I: theoretical and methodological orientations. London: Routledge.

Buck, Carl D. 1949. A dictionary of selected synonyms in the principal Indo-European languages. Chicago: U. of Chicago Press.

Dobson, Annette J. 1969. Lexicostatistical grouping. Anthropological Linguistics 11.216-21.

—. N.d. Unrooted trees for numerical taxonomy. Unpublished.

Embleton, Sheila M. 1986. Statistics in historical linguistics. Bochum: Brockmeyer.

Eska, Joseph E., and Don Ringe. Forthcoming. A failed attempt at linguistic phylogeny.

Fitch, Walter M. 1971. Toward defining the course of evolution: minimum change for a specified tree topology. Systematic Zoology 20.406-16.

Gleason, Henry A. 1959. Counting and calculating for historical reconstruction. Anthropological Linguistics 1.22-32.

Hock, Hans H. 1986. Principles of historical linguistics. Berlin: Mouton de Gruyter.

Hoenigswald, Henry M. 1960. Language change and linguistic reconstruction. Chicago: U. of Chicago Press.

—. 1987. Language family trees, topological and metrical. Hoenigswald and Wiener (edd.) 1987:257-67.

Hoenigswald, Henry M., and Linda F. Wiener (edd.). 1987. Biological metaphor and cladistic classification: an interdisciplinary perspective. Philadelphia: U. of Pennsylvania Press.

Jasanoff, Jay H. 1997. An Italo-Celtic isogloss: the 3pl. mediopassive in **-ntro.* Adams (ed.) 1997:146-61.

King, Ruth. 2000. The lexical basis of grammatical borrowing. Amsterdam: Benjamins.

—. 2003. Language contact and linguistic structure. Paper presented at NWAVE 32, Philadelphia, October 2003.

Labov, William. 1994. Principles of linguistic change. Vol. I: internal factors. Oxford: Blackwell.

Mair, Victor (ed.). 1998. The Bronze Age and early Iron Age peoples of eastern Central Asia. Washington: 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.

Nakhleh, L. 2004. Phylogenetic Networks in Biology and Historical Linguistics. PhD Dissertation, The University of Texas at Austin.

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. U. 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. 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, ed. by Victor Mair, 1998:391-414.

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 345.153-6.

Ross, Malcolm. 1997. Social networks and kinds of speech- community event. Blench and Spriggs (edd.) 1997:209-61.

Ruvolo, Maryellen. 1987. Reconstructing genetic and linguistic trees: phenetic and cladistic approaches, ed. by Henry Hoenigswald and Wiener 1987:193-216.

Thomason, Sarah Grey, and Terrence Kaufman. 1988. Language contact, creolization, and genetic linguistics. Berkeley: U. of California Press.

Warnow, Tandy, Don Ringe, and Ann Taylor. 1995. Reconstructing the evolutionary history of natural languages. IRCS Report 95-16. Philadelphia: Institute for Research in Cognitive Science, U. of Pennsylvania.

White, J. P., and J. 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. Mair (ed.) 1998:347-57.

9 Appendix

In this appendix we describe the algorithmic techniques used to complete a candidate genetic tree to a perfect phylogenetic network, containing a minimum number of contact edges.

We begin with a formal statement of the algorithmic problem:

- **Problem**: Minimum Increment to a PPN
- *Input:* a rooted tree T, leaf-labelled by a set S of languages, and a set C of qualitative characters defined on the set S.
- *Output*: a perfect phylogenetic network 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 add to T so as to produce a perfect phylogenetic network.

Recall that a network N is a perfect phylogenetic network 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 perfect phylogenetic network is as follows:

- Let $i = 1$
- For each way of adding i contact edges to T, determine if the resultant network is a perfect phylogenetic network for S with respect to the character set C.
- *If* any such way of adding i edges produces a perfect phylogenetic network, return all such perfect phylogenetic networks (with i contact edges), and exit; *else* increment i 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 edge, and then increasing, until we find the smallest possible number of contact edges that suffice to make a perfect phylogenetic network. The algorithm thus has to be able to determine whether each character is compatible on a given network with i contact edges, a separate problem which 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, and otherwise *No*

Again, a straightforward algorithm will work for this problem. Given N, we compute each of the trees contained within N; a simple calculation shows that there are at most 3^i such trees, since N has i contact edges. 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*, and otherwise we return *No*.

It is easy to determine compatibility of a character on a tree – this can be done by inspection (i.e., without a computer), since all that needs be done is that we label every node that is on a path between two leaves with the same state, and as long as no node gets two contradictory labels,

we have compatibility. More formally, however, we can use the algorithm given by Walter Fitch (Fitch 1971) for maximum parsimony, and which also solves it. Thus, compatibility of a character on a tree can be solved in $O(n^2)$ time, where n is the number of leaves (the $O(n^2)$ notation means that the running time is bounded from above by some constant times n^2). Thus, 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 would do this for each character, the running time would be $O(kn^2 3^i)$, where k is the number of characters. There are $\binom{\binom{2n-2}{2}}{i}$ 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 upon tree T) with i contact edges, which is $O(k2^{4i} n^{2i+2})$.

However, we would have to do this for $i = 1, 2, \ldots, p$, where p is the minimum number of contact edges we need to add to get a perfect phylogenetic network. 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 perfect phylogenetic network 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 perfect phylogenetic networks, see Nakhleh 2004.