Chapter 36

Reconstructing the evolutionary history of natural languages

Tandy Warnow[*] Donald Ringe[†] Ann Taylor[‡]

October 23, 1995

Abstract

In this paper we present a new methodology for determining the evolutionary history of related languages. Our methodology uses linguistic information encoded as qualitative characters, and provides much greater precision than previous methods. Our analysis of Indo-European (IE) languages resolves questions that have troubled scholars for over a century.

1 Introduction

The determination of evolutionary trees for natural languages is a major endeavor within historical linguistics, but current techniques that are used to generate trees are limited either by computational problems or through the use of methods which lose information present in the primary data.

In this paper we will present a method for efficiently inferring the evolutionary history of languages known to be related (i.e. members of the same family of languages) which avoid the difficulties that have made this analysis intractable. We use primary data, and show that an appropriate optimization problem can be solved exactly for these data.

We have applied this method to the problem of inferring the evolutionary history for the IE family of languages, and have made several surprising and strikingly strongly supported findings. Our motivation for studying this particular family of languages was that little progress had been made on definitively settling the first-order subgrouping of IE, despite the fact that it is the best attested and best studied family of languages available to historical linguists. A solution to the first-order subgrouping of this family would establish the applicability of this methodology beyond question. We analyzed the IE data with particular interest in determining whether our new methodology could lay to rest the debate on two longstanding conjectures: the *Indo-Hittite hypothesis* and the *Italo-Celtic hypothesis.*

The rest of the paper is organized as follows. In Section 2 we begin with a discussion of computational aspects of inferring evolutionary trees, both from qualitative characters and distances. In Section 3 we describe the history of the methodology of reconstructing the evolutionary history of natural languages. In Section 4 we discuss our methodology and its computational complexity. The application of our methodology to IE is presented in Section 5. We conclude in Section 7 with a discussion of the contribution this method makes to historical linguistics.

2 Inferring Evolutionary Trees

An evolutionary tree, or phylogeny, for a set S of taxa (i.e., of species or languages) describes the evolution of the taxa in S from their most recent common ancestor. Data of different types can be used as input for methods of tree construction; typical are distance data (the basis of lexicostatistics, see below) and character data, which reflect specific observable characteristics of the species under study ("morphological" data in biology; there is no comparable term in linguistics, in which "morphology" is used narrowly to mean the grammatical characteristics of words). A character mathematically is a partition of the taxa (species or languages, for example) into distinct states. Thus, for example, we can define a character based upon the number of legs present, which therefore has as many states as there are different numbers of legs.

One way that evolutionary trees constructed from morphological features (and other qualitative characters) can be evaluated is called the *compatibility* criterion (see [15] for a discussion of these different criteria). We say that a character α is compatible (also called "convex") with a vector-labelled tree T if for every state of α the nodes having that state form a connected sub-

[*] Department of Computer and Information Science, University of Pennsylvania Philadelphia, PA. Email: tandy@central.cis.upenn.edu. This work was supported in part by a National Young Investigator Award from NSF, CCR-9457800, and by a grant from the Division of Linguistics at NSF, grant SBR-9512092.

[†] Department of Linguistics, University of Pennsylvania, Philadelphia, PA. Email: dringe@unagi.cis.upenn.edu. Research supported in part by a grant from the Division of Linguistics at NSF, grant SBR-9512092.

[‡] Department of Linguistics, University of Pennsylvania, Philadelphia, PA. Email: ataylor@linc.cis.upenn.edu. This work was supported by NSF grants CCR-9457800 and SBR-9512092.

graph of T. The compatibility score of a tree T is defined by: $c(T) = |\{\alpha \in C : \alpha \text{ is convex on T}\}|$. Given an input of species set S defined by character set C, finding the tree T of maximum compatibility score (or the largest subset of compatible characters) is called the *Compatibility Criteria Problem.*

Some biological characters are obviously convex, for example, the morphological character *vertebrate-invertebrate*; such characters provide indisputable information about the evolutionary history of the set of taxa. The problem with using compatibility to evaluate trees is that selecting convex characters is difficult, and many characters are not convex. For example, consider the character for the presence or absence of *wings*. Because wings arise more than once in the evolutionary history of animals, if the character *wings* is convex on a tree, then the tree is wrong. Including *wings* in a data set to be analyzed under the compatibility criterion thus introduces noise, and inherently decreases sensitivity and specificity of the method. This indicates that if the compatibility criterion is to be useful, it is crucially important that as many as possible non-convex characters be eliminated in advance. Because of these reasons and because the selection of convex characters is difficult, compatibility approaches have been largely discarded in Biology.

By contrast, we will show that properly selected information about the languages can be expressed as qualitative characters, and that these characters *will be convex on the true tree.* This means that if such a character is not convex on a tree T, then either the tree is incorrect or the scholarly judgement that the character would be convex on the true tree is false. Thus, the number of characters which are not convex on T is an accurate measure of the "badness" of the tree, since we will permit (and explicitly instruct) the linguist to remove all characters which are likely to be non-convex on the tree. As long as these judgements are made rigorously and the number of remaining characters is high, the tree that results should be indicative of what we believe is the true tree.

Therefore, the appropriate criterion for use in evaluating evolutionary trees in Linguistics is the *compatibility criterion.* Unfortunately, the following theorem shows that this is one of the harder problems to solve:

THEOREM 2.1. (FROM [23]) *The Compatibility Criteria Problem is* NP-hard *and cannot be approximated by a polynomial-time algorithm within a factor of* $|C|^{1/4-o(1)}$ *unless* $QNP = co\text{-}QR$.

The first proof showing this problem is NP-hard can be found in [8]. Details about these complexity classes can be found in [17].

Fortunately, we will also show that linguistic data is good enough that almost all the characters we work with will be compatible on the true tree, so that we will need to remove only a few characters in order to obtain what is called a *perfect phylogeny* (i.e. a tree on which *all* the characters are convex). For such data, we can find the provably optimal trees quickly.

3 Subgrouping Methodologies

Methodologies for subgrouping related languages have been debated for over a century[21]. There are two basic types of methodologies: *classical* or *traditional* methods, which are character based, and *lexicostatistical* methods, which are distance based.

3.1 Classical Methods In classical methods, languages are assigned to the same subgroup — that is, members of a set of related languages are believed to depend from the same node of the evolutionary tree — only if two conditions are met: (1) the languages in question exclusively share innovations, and (2) those innovations are unlikely to have occurred independently[16]. To use the terms we have defined above, the interpretation of character information for classical subgrouping is as follows: (1) character states which are innovations should be *convex* on the true evolutionary tree, and (2) only characters which are unlikely to be affected by parallel development are used.

3.2 Lexicostatistics An alternative method of subgrouping, lexicostatistics, was developed in the 1950's and 1960's ([10, 11, 12]). For lexicostatistical analysis one determines what proportion of the most basic vocabulary is shared by each pair of languages under investigation; it is assumed that most shared items are retained inheritances and that the proportion of basic items replaced correlates roughly with the time elapsed from the point at which the two languages in question were still a single language. The validity of these assumptions, while questioned, has not led to a complete rejection of these distance-based methods. *Lexicostatistics* refers in general to any method for constructing trees for languages based upon distances obtained in this way, with the usual method a *pair-grouping* method (i.e. it takes the closest pair and makes them siblings, computes a parent node, and recurses on the smaller set)[10].

3.3 Critique of these methodologies The major weaknesses of lexicostatistical methods are that they rely upon derived rather than primary data and most associated optimization problems are *NP-hard*[9, 14]. Indeed, it seems that the real reason that distance based methods are so popular in Linguistics is that software is available for constructing trees (optimal or not) from distance data, and the software is easy to use and fast. These software packages, however, are based upon heuristics which have unproven performance, and thus do not reliably find trees which are optimal with respect to any objective criterion. While many linguists use the software as a final tool, the best linguists[13] use it only to generate trees which can then be evaluated through the use of classical methods.

The classical methods are character based, and indicate a key insight into the correct way to do evolutionary tree construction. Determining which characters denote evolutionary information is a sophisticated and difficult matter and a fair amount of the debate in the field concerns these judgements and how to use the linguistic information to define characters. The significance of inflectional peculiarities is especially often unclear. There are other problems beyond the choice of characters, however, which involve the limited use by historical linguists of the character information. Specifically, linguists have known that all character states should (if possible) be convex on the true tree, but this understanding was never made explicit in the literature. That is, the codified knowledge only specifically indicated that states that were innovations should be convex; the realization that this implied convexity for all states of the character was never made precise.

These two types of methodologies nevertheless have some features in common. All reliable methods of subgrouping languages must start from the *comparative method*, the simple but rigorous mathematical method for reconstructing protolanguages which was codified by Henry Hoenigswald in [16]. Without the comparative method one cannot even recognize cognate vocabulary – that is, words inherited by genetically related languages from their protolanguage, as opposed to words borrowed through language contact or words that happen, through sheer chance, to be similar in sound and meaning [26].

4 Our methodology for constructing evolutionary trees from linguistic characters

Our methodology has three essential components, *encoding linguistic information using qualitative characters, an algorithm to find the optimal and near-optimal trees,* and *methods for finding the common features of the best trees.*

The encoding of the linguistic information as qualitative characters involves a great deal of linguistic scholarship, and to some degree the judgements can be open to debate as different linguists will deem different information as relevant or not relevant to the evolutionary tree, and even when agreeing to the relevance of a character they may differ in their encoding of the character states. The encoding of linguistic information as characters thus involves linguistic judgement as well as mathematical modelling. This is described in Section 4.1.

4.1 Types of Linguistic Characters

Lexical. For lexical characters, the character is the semantic slot, as for example, the meaning 'hand'. Languages which have reflexes of the same proto-lexeme for this semantic slot exhibit the same state for the character, and are said to be *cognate.*

Morphological. For morphological characters, the character is generally a grammatical feature, as for example the formation of the future stem, the way the passive is marked, the genitive singular ending of o-stem nouns and adjectives, etc. Languages in which the feature is instantiated in the same way, or by a reflex of the same proto-morpheme, exhibit the same state for the character. On occasion, we may be able to determine which states are ancestral and which are innovations.

Phonological. For phonological characters, the character is a sound change. Languages which share the same outcome (generally, those that undergo the change versus those that do not) exhibit the same state for the character. Phonological characters are not as useful as morphological and lexical ones, however, because of the high probability of independent parallel development in this area. Most sound changes are natural and the fact that two languages both undergo the same change does not, if the change is natural enough, necessarily indicate common innovation. Thus, only sound changes that are rare or fairly complex can be safely used as characters. An example of this type of sound change in the IE family is the so-called 'ruki' rule, which involves the retraction of */s/ after /r/, /u/, /k/ and /i/.

Encoding linguistic information as characters The determination of character states for morphological and phonological data is straightforward, but the determination of character states for lexical characters requires some discussion. The encoding of lexical information as characters depends upon cognation judgements, and these are accomplished through the application of the Comparative Method. The Comparative Method produces equivalence classes of cognates and not just a similarity score, and except in unusual cases (discussed later on in this paper) these judgements are entirely accurate.

Note that there is additional information present in the phonological characters (and sometimes also in the lexical and morphological ones as well) which we need to include. We can encode all these constraints as qualitative characters, using techniques that have been developed in Biology.

Selection of Characters As much as possible we have used linguistic features for our character data that are unlikely to be borrowed. Resistance to change of any kind, although desirable in that it is more likely to result in characters which narrow down the space of optimal trees, is not required.

4.1.1 Detecting and handling parallel development

Borrowing. The most obvious kind of borrowing event occurs when one language uses a word from another language. Obvious borrowings are easily detected, such as the use of *croissant* in English, and can be encoded in an appropriate manner. This allows us to keep the characters for these semantic slots in our data set, rather than discarding them, and this improve the specificity of our method. Undetected borrowings which cannot be distinguished from true cognates are a more serious problem. Fortunately, these undetected borrowings are rare, because they must occur between languages that are so similar that words in one language look like words in the other; in other words, languages that have not diverged very much from their common ancestor. If these borrowings occur in sufficient numbers, however, they create a distinct pattern in the data in which a certain language shares states for a substantial group of lexical characters with one language (or group of languages) while for the rest of the lexical characters and the majority of the morphological characters it shares states with a different language or group. This is the case of Germanic in the IE family (see Section 5 for details).

Polymorphic characters: What we are calling "polymorphism" arises in two different ways. The first such case is called *independent parallel semantic shift.* In this case two or more languages independently shift the meaning of one lexical item to another (e.g. a number of IE languages use the stem *wi:ro-*, originally meaning 'young man, warrior', in the meaning 'man'). Most of these cases are fairly obvious (e.g. the use of a root meaning 'give light' for 'moon', roots meaning 'blow', 'breathe' for 'wind', etc.). The other way polymorphism happens is when for some languages a semantic slot is associated with two (or more) lexical items, both of which can be reconstructed for the protolanguage without detectable distinction in meaning (e.g., Proto-IE 'warm': $*g^{w}her$-, $*tep$-; 'wash': $*lewh_3$-, $*neyg^{w}$-, etc.) However, this also can be explained as really a semantic shift, where the mechanism of the semantic shift is not quite as obvious. In either case, the character can have more than one state in a given language. A similar situation occurs in Biology (especially in population genetics) where a character can have more than one state in a species. In population genetics, genes often have more than one allele present in a population. These biological characters are called *polymorphic* for the same reasons.

In Linguistics, by contrast to Biology, polymorphism can be shown to arise by having two (or more) distinguishable characters merge into one. The problem of separating out the characters into (usually) two characters is a technical problem with interesting algorithmic implications. In many cases, we have been able to determine a partial separation of the states present at the leaves which must be extended in the final solution to a full separation. Our technique for handling polymorphic characters is therefore to temporarily eliminate them, find the tree that fits the remaining characters, and then verify that the polymorphic characters fit the tree (i.e. can be separated into a small number of characters convex on the tree). This verification can be accomplished in polynomial time[4].

4.2 Finding optimal trees

The best possible tree for a set of species defined by characters has every character convex on it. Such a tree is called a *perfect phylogeny.* It is not hard to see that when a perfect phylogeny exists it has a maximum compatibility score. Determining if a perfect phylogeny exists (called the *Perfect Phylogeny Problem*) is *NP-Complete*[3, 28], but by contrast with compatibility, it can be solved in polynomial time when any of the relevant parameters ($n = |S|, k = |C|$, or the maximum number r of states per character) is bounded[2, 1, 20, 18, 19]. We will show that linguistic data is "close" to perfect in the sense that the compatibility score is close to those achievable by perfect phylogenies, and that the incompatible characters can, in fact, be impugned on rigorous grounds as being inappropriate for use.

The methodology we have developed thus does more than show how to encode linguistic information as characters and then produce trees from the characters; rather, it also identifies the incompatible characters, thus enabling the linguist to reconsider the scholarly judgements and, in the course of the analysis, determine which characters we should not have included from the start. This method will be efficient if the number of initially incompatible characters is not too large. Precisely, we will define the imperfection of a data set as follows.

DEFINITION 4.1. *A set S of species defined by the character set C has* imperfection *t if the optimal tree has $|C| - t$ characters convex on it.*

We now give the key observation for an "efficient" method for finding optimal trees on data sets with very small imperfection.

THEOREM 4.1. *We can find the best tree with respect to compatibility in $O(2^{2^r}ntk^{t+2})$ time, where $n = |S|$, t is the imperfection of the input set, and $k = |C|$.*

Proof. To find a maximum cardinality subset of compatible characters, we can search among all subsets C_0 such that $C_0 \subseteq C$ in decreasing order of cardinality until we find all the largest compatible sets of characters (and the perfect phylogenies for them). This requires $O(tk^t)$ calls to [19] for a total cost of $O(2^{2^r}tnk^{t+2})$ time.

4.2.1 Computing Minimal Trees In Linguistics, as in Biology, we are interested in *minimal* trees. A tree T is said to be *minimal with respect to compatibility* if the contracting of any edge decreases the compatibility score of T. The reason we are interested in minimal phylogenies is that we wish the tree to represent the information forced by the data set, and no other. Thus, for example, a tree T of the IE family will indicate support for the *Italo-Celtic hypothesis* if and only if the leaves for Old Irish and Latin (representatives of Celtic and Italic subfamilies, respectively) are siblings and have no additional siblings, so that the parent of these leaves has no other children. If the tree is not minimal, it may falsely indicate support. Thus, the relevant information in a tree is contained in its minimal form. To compute minimal trees, we take a potential tree and contract its edges until to do so would decrease the compatibiltiy score. Determining which edges can be contracted can be done in a straightforward manner, by obtaining first a *canonical labelling* of the nodes (see [19]). Edges which can be contracted have endpoints whose labels differ only in positions for which one is a dummy state. The canonical labelling can be computed in $O(kn)$ time, where k is the number of characters and n the number of leaves. Contracting edges which can be contracted does not create new edges which can be contracted, and hence the operation of obtaining a minimal form of the tree takes no more than $O(kn)$ time.

5 The subgrouping of Indo-European

In order to test the methodology we attempted a subgrouping of IE, among the best understood of the world's language families. We selected from each of the subfamilies within IE the oldest well-attested language to represent the subfamily. Thus we have Latin (LA, 1st century B.C.E.) representing Italic, Old Irish (OI, 8th-9th cc. C.E.) representing Celtic, Hittite (HI, 16th-13th cc. B.C.E.) representing Anatolian, Vedic (VE, ca. 1000 B.C.E.) representing Indic, Avestan (8th-6th cc. B.C.E.) representing Iranian, Old English (OE, 9th-10th cc. C.E.) representing Germanic, Tocharian B (TB, 6th-8th cc. C.E.), Greek (GK, Classical Attic dialect, 5th c. B.C.E.), Armenian (AR, 5th c. C.E.), Albanian (AL, 20th c. C.E.), Lithuanian (LI, 20th c. C.E.) representing Baltic, and Old Church Slavonic (OCS, 10th c. C.E.) representing Slavic. The following is a detailed description of our findings for the IE family.

5.1 Choosing characters In order to reduce the possibility of borrowings among the lexical characters and bias on our part in choosing these characters, we used an existing basic vocabulary list of 208 semantic slots[30].[1]. Each semantic slot was treated as a single character and judgements of cognation were made on the basis of the comparative method. Once the states were encoded for each character, we detected evidence of borrowing, parallel development, and polymorphism. These included:

1. all characters for which two or more lexical roots are reconstructible for the protolanguage. (total 10)

2. other characters in which parallel semantic shift or borrowing has clearly taken place or in which the probability that it has appears to be very high (total 27)

Of the characters in (2), the directionality of the parallel semantic shift or borrowing or could be detected in all but 7 cases, so that we could include in our analysis all but 17 characters. Of the full set of characters, 49 were informative (i.e. characters that do not fit every possible tree on the leaf set).

Since nothing similar to a basic vocabulary list exists for morphological and phonological characters and since these will vary from family to family, an appropriate set of morpho/phonological characters has to be developed for each family. For the IE test we used ten Proto-Indo-European morphological items which have a reflex in most of the IE languages, and four phonological developments which we judged to be sufficiently abnormal as not to be easily repeatable. These 14 characters are: organization of the verb system, presence of the augment, presence of a thematized aorist, productive function of -ské/ó-, function of -dhí, mediopassive primary marker (sg. and 3pl.), thematic optative suffix, most archaic future stem, genitive singular of o-stem nouns and adjs., superlative suffix, satem sound change, retraction of s in "ruki"-environments, shape of oblique dual and plural case endings, and initial d-

[1] Our list has one more item than Tischler's[30] because we split the item *day* into two items, *period of 24 hours* and *period of daylight.*

in 'tears'. Of these morphological/phonological characters, ten proved to be informative.

Thus, at the end we had 49 informative lexical characters and 10 informative morphological characters.

6 Results

6.1 Finding the Best Tree The best trees found by the algorithm for the the pruned dataset each had 12 characters nonconvex. Somewhat surprisingly, the trees are identical except for the position of the subgroup including Germanic (represented by Old English) and Albanian. As we soon discovered, the position of Albanian rests on only two or three cognates, because Albanian has lost so much of the material inherited from PIE; but it is reasonable to suppose that the labile position of Germanic has something to do with the relatively large number of nonconvex characters.

We therefore removed Germanic from the data set and ran the algorithm again. We looked at all the trees with compatibility scores close to optimal (the best tree had all but four characters convex). All of these trees were strikingly similar to each other, and at this point we were quite confident that the features shared by all the trees should be true of the "true" tree.

We re-examined each of the characters that were non-convex on any one tree in the set, and discovered that many could be eliminated as showing either evidence of (previously undetected) borrowing or polymorphism. For example, *night* and *all* seemed likely to be polymorphic. The lexical characters *liver* and *tears* probably show undetected borrowing, the former between some late Anatolian language and Armenian and the latter between Iranian and Tocharian. This left only two characters that we were unable to impugn: the *medio-passive marker* and the semantic slot *ye.*

We removed the impugned characters and ran the algorithm again on the reduced set of characters. Now the best tree was in fact a unique perfect phylogeny! Not only that, but when we added back the polymorphic characters, we obtained the desired separation into two convex characters in each case.

The tree we found is given in Figure 1 (note that it is a rooted tree, because our encoding of our linguistic judgements includes the directionality constraints). The position of Albanian is not indicated in the tree, because it can be placed anywhere above the dotted lines (there is little left to tie it anywhere within the tree, as we have already noted). Because this tree is a perfect phylogeny, and the *unique* perfect phylogeny for our data, it is compatible with *all* the linguistic judgements we made, but every other tree will be incompatible with some judgement we had made. This is the best we could have hoped for.

6.2 Resolving Controversies in Indo-European We were interested in determining whether we could resolve longstanding controversies in Indo-European, specifically the Indo-Hittite and Italo-Celtic hypotheses. An examination of the tree we obtained indicates support for the Indo-Hittite hypothesis and denies the Italo-Celtic hypothesis. We then decided to consider the robustness of this information by examining the suboptimal trees as well. Since these judgements require rooted trees in order to be considered, we examined the rooted versions of the optimal and near-optimal trees.

*Indo-Hittite*The first question we examined was the robustness of the Indo-Hittite hypothesis. The only near-optimal trees that did not support the Indo-Hittite hypothesis were incompatible with the morphological character reflecting the presence or absence of the thematized aorist. Note that because this is a morphological character, it has more weight than lexical characters which are *vastly* easier to borrow. If this character is impugned, the position of the root becomes somewhat more indeterminate; that is, it can appear anywhere above the Greek/Armenian node, creating several possible ways of rooting the trees. To an Indo-Europeanist, however, none of these groupings is as convincing as that which splits Anatolian alone from the main branch, and none of them is supported by any positive evidence either. Clearly our results here point toward acceptance of the Indo-Hittite hypothesis. This is interesting in light of the raging debate which has been going on for about the last 50 years over whether the Anatolian family, here represented by Hittite, was the first to break off from the rest of the family.

*Italo-Celtic*We then considered the Italo-Celtic hypothesis, which asserts that Italic (represented by Latin) and Celtic (represented by Old Irish) should be siblings in the tree with no third sibling. Here, there were near-optimal trees which did not deny the Italo-Celtic hypothesis, but these require impugning the character *ye* in order to obtain trees in which Latin and Old Irish *could* be siblings. The methodology therefore indicates that in order to support the Italo-Celtic hypothesis, *ye* must be impugned, and at least one other character must be found that groups Italic and Celtic together against Hittite (or PIE) and one of the languages below Latin on the tree.

*The problem of Germanic*We were then in a position to return to the problem of Germanic. Even among the morphological characters there is no clear consensus about where Germanic belongs. One character groups Germanic with Balto-Slavic against everything else, including even Indo-Iranian; but three others exclude Germanic from this core area. The remaining characters are compatible with either position. In short, the

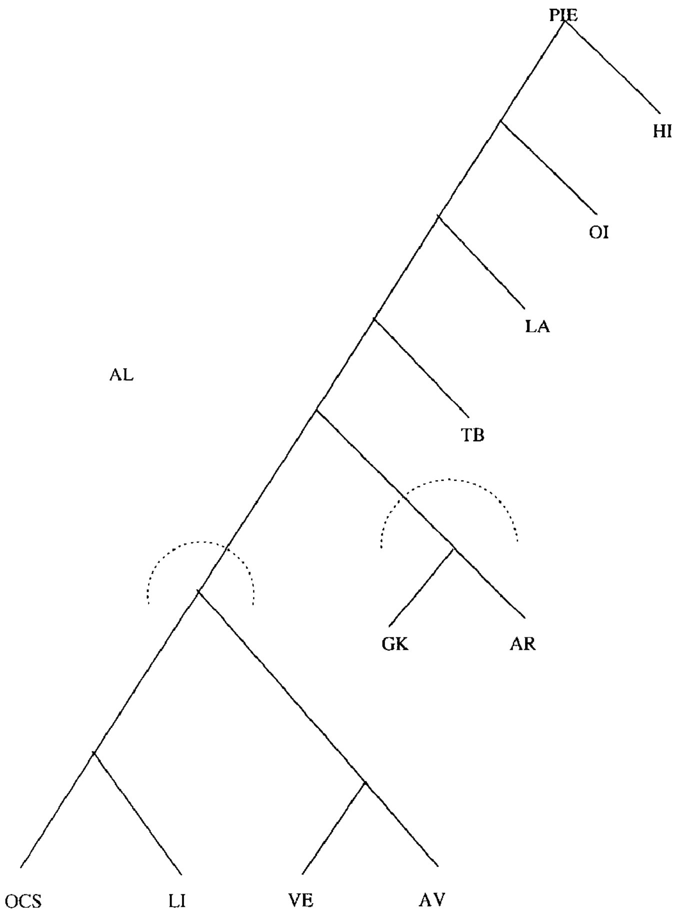

Figure 1: The tree obtained on IE without Germanic

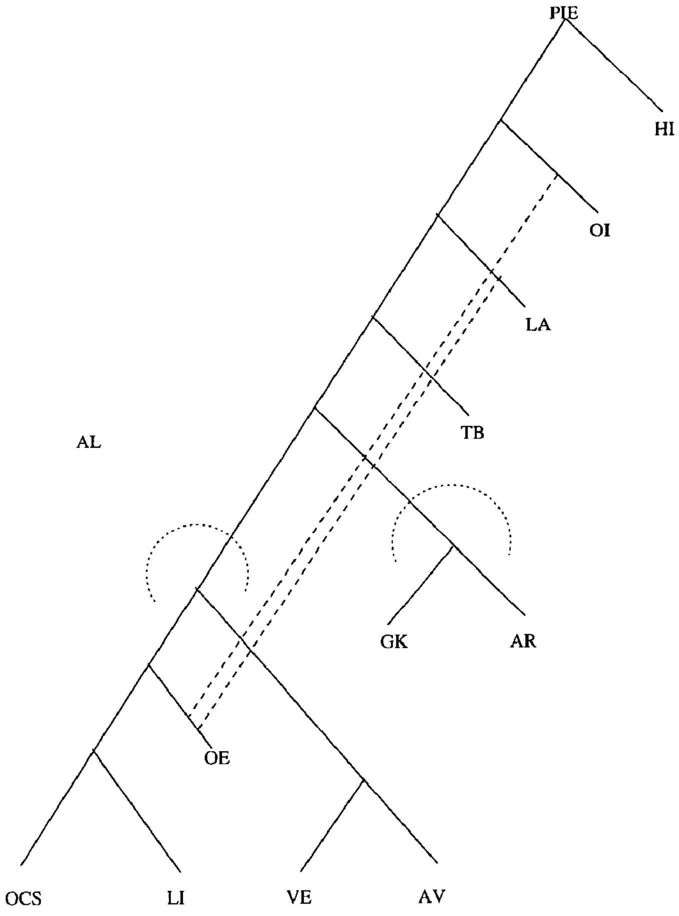

Figure 2: The Network for Indo-European

development of Germanic is not exhibiting treelike behavior. But a plausible explanation in terms of the tree can be given: it is possible that Germanic began its independent life as a sister of Balto-Slavic, then shifted its affiliations before the characteristic satem changes spread through the satem core (which by then must have been a dialect network, though still undifferentiated enough to allow innovations to spread easily within it). Moreover, it is clear enough which other branches of IE Germanic now came into contact with: the substantial proportion of its basic vocabulary that it shares with Italic and/or Celtic (but not with the satem core) must reflect lexical borrowing from those languages into Germanic before any sound changes that would betray their status as borrowings had occurred (therefore at the pre-Proto-stage for all three groups). This rather complicated state of affairs can be represented in Figure 2, in which the added edges (given in dashes) indicates development through contact rather than "genetic" transmission.

7 Conclusions

The relative merits of traditional subgrouping, lexicostatistics, and the method we have been developing can be seen by comparing the results of those methodologies as applied to a traditionally intractable problem, the first-order subgrouping of the IE language family. Traditional methods failed to produce a convincing tree, and conservative IEists settled for a description of the relations between the languages resembling a network of geographical dialects ([22]). Subsequent work along traditional lines produced no further positive results[24, 25]; arguments in favor of a more articulated tree structure supporting the Indo-Hittite hypothesis[29, 6, 7] and the Italo-Celtic hypothesis[5] were debated at length and rejected. However, many linguists continue to suspect that new arguments to support these hypotheses can be found. Lexicostatistical work made no substantial advances; even the most careful and sophisticated applications of lexicostatistics to the IE problem produced equivocal and contradictory results[30, 12], and the best-informed mathematical linguist who has attempted such work makes notably modest and reserved claims for the method[13].

By contrast, we have been able to construct a robust evolutionary tree of the IE languages, as detailed in Section 5; we have even been able to show that the Germanic subgroup of the family underwent a surprising shift in its affiliations at a very early period of its independent history–an unexpected but thoroughly plausible finding that has startling implications for the history of Germanic syntax. While a considerable number of problems remain to be solved, our promising preliminary results give us reason to hope that we have finally evolved a method which preserves the strengths of traditional subgrouping techniques — as lexicostatistics does not — while avoiding the well-known weaknesses of traditional methodology.

Although we are convinced that the evolutionary tree we have constructed is accurate, perhaps more important than the particular explanation of the evolutionary history we propose is the observation that this methodology provides an accurate and precise measure of the consistency of linguistic judgements. Furthermore, the method also helps correct the judgements which might have been mistaken. Finally, because any linguist can use this method and determine the tree that is most consistent with his or her own judgements, a distinct advantage of using this method is that it identifies the precise linguistic judgements that are incompatible with a given tree, and thus automatically focuses the debate on the scholarly details which are pertinent.

This research also raises the question of whether clear results along these lines can be obtained in biology. The problem in inferring evolutionary trees in biology tends to be that the methods available are not adequately sensitive and specific on all the data sets. By contrast, our methodology provides a way of analyzing data that is both sensitive and specific, and thus is able to produce a unique best tree (whereas many data sets in Biology have hundreds if not thousands of equally good trees).

8 Acknowledgements

The first author wishes to thank Paul Angello and ARO (Grant #DAAL0389-C-0031) whose financial support has made this research possible.

9 Bibliography

References

[1] Agarwala, R. and D. Fernandez-Baca, 1994: Fast and simple algorithms for perfect phylogeny and triangulating colored graphs, *DIMACS TR# 94-51.*

[2] Agarwala, R. and Fernandez-Baca. D. 1994: A polynomial time algorithm for the phylogeny problem when the number of states is fixed, SIAM Journal on Computing 23(6):1216-1224.

[3] Bodlaender, H., Fellows, M. and Warnow, T. 1992: Two strikes against perfect phylogeny, Proceedings of the International Congress on Automata and Language Processing.

[4] Bonet, M., Phillips, C., Warnow, T. and Yooseph, S. 1995, *Constructing evolutionary trees in the presence of polymorphic characters*, manuscript.

[5] Cowgill, Warren 1970: Italic and Celtic superlatives and the dialects of Indo-European, in Cardona, George,

Henry M. Hoenigswald, and Alfred Senn (eds.), *Indo-European and Indo-Europeans,* University of Pennsylvania Press, Philadelphia.

[6] Cowgill, Warren 1975: More evidence for Indo-Hittite: the tense-aspect systems, in Heilmann, Luigi (ed.), *Proceedings of the Eleventh International Congress of Linguists,* Mulino, Bologna.

[7] Cowgill, Warren 1979: Anatolian *hi*-conjugation and Indo-European perfect: instalment II, in Neu, Erich, and Wolfgang Meid (eds.), *Hethitisch und Indogermanisch,* Innsbrucker Beiträge zur Sprachwissenschaft, Innsbruck.

[8] W. H. E. DAY AND D. SANKOFF, *Computational complexity of inferring phylogenies by compatibility,* Syst. Zool., Vol. 35, No. 2 (1986), pp. 224–229.

[9] , W.H.E. Day, 1987: Computational complexity of inferring phylogenies from dissimilarity matrices, *Bulletin of Mathematical Biology,* 49(4), pp. 461-467.

[10] Dyen, Isidore 1962: The lexicostatistically determined relationship of a language group, *IJAL* 28:153-61.

[11] Dyen, Isidore 1975: On the validity of comparative lexicostatistics, in *Linguistic Subgrouping and Lexicostatistics*, pp. 137-149, Mouton, Paris.

[12] Dyen, Isidore, Kruskal, Joseph B. and Black, Paul 1992: An Indoeuropean Classification: A Lexicostatistical Experiment, *Transactions of American Philosophical Society* 82(5), Philadelphia, PA.

[13] Embleton, Sheila M. 1986: *Statistics in historical linguistics.* Brockmeyer, Bochum.

[14] M. Farach, S. Kannan and T. Warnow, *A Robust Model for Finding Optimal Evolutionary Trees*, Algorithmica, special issue on Computational Biology, Vol. 13, No. 1, 1995, pp. 155-179. (A preliminary version of this paper appeared at STOC.)

[15] Felsenstein, J. 1982: Numerical methods for inferring evolutionary trees, *The Quarterly Review of biology,* Vol.57, No.4.

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

[17] D. S. Johnson, "A catalog of complexity classes", in *Algorithms and Complexity,* volume A of *Handbook of Theoretical Computer Science*, Elsevier science publishing company, Amsterdam, 1990, pp. 67–161.

[18] Kannan, S. and Warnow, T. 1994: Inferring evolutionary history from DNA sequences, *SIAM Journal on Computing* 23(4):713-737.

[19] Kannan, S. and Warnow, T. 1995: A fast algorithm for the computation and enumeration of perfect phylogenies, Proceedings of ACM-SIAM Symposium on Discrete Algorithms, 1995.

[20] McMorris, F.R., Warnow, T. and Wimer, T. 1994: Triangulating vertex colored graphs, *SIAM Journal on Discrete Mathematics* 7(2):296-306.

[21] Meillet, A. 1925: *La Méthode Comparative en Linguistique Historique*, H. Aschehoug & Co., Oslo.

[22] Porzig, Walter 1954: *Die Gliederung des indogermanischen Sprachgebiets.* Carl Winter, Heidelberg.

[23] Phillips, C.A. and Warnow, T.J. *The Asymmetric Median Tree: a new model for building consensus trees,* manuscript, 1995.

[24] Ringe, Donald A., Jr. 1988: Laryngeal isoglosses in the western Indo-European languages, in Bammesberger, Alfred (ed.), *Die Laryngaltheorie,* Carl Winter, Heidelberg.

[25] Ringe, Donald A., Jr. 1991: Evidence for the position of Tocharian in the Indo-European family? *Die Sprache* 34:59-123.

[26] Ringe, Donald A., Jr. 1992: On Calculating the Factor of Chance in Language Comparison, *Transactions of American Philosophical Society* Vol.82, no.1, Philadelphia, PA.

[27] Ruvolo, Maryellen 1987: Reconstructing genetic and linguistic trees: phenetic and cladistic approaches, in Henry M. Hoenigswald and Linda F. Wiener, (eds.),*Biological Metaphor and Cladistic Classification*, pp. 193-216, University of Pennsylvania Press, Philadelphia.

[28] Steel, M. 1992: The complexity of reconstructing trees from qualitative characters and subtrees, *Journal of Classification* 9:91-116.

[29] Sturtevant, Edgar H. 1933: *A comparative grammar of the Hittite language.* Linguistic Society of America, Philadelphia.

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