Institute for Research in Cognitive Science

Reconstructing the evolutionary history of natural languages

Tandy Warnow
Donald Ringe
Ann Taylor

University of Pennsylvania
3401 Walnut Street, Suite 400C
Philadelphia, PA 19104-6228
June 1995

Site of the NSF Science and Technology Center for Research in Cognitive Science

IRCS Report 95-16

Reconstructing the evolutionary history of natural languages

Tandy Warnow
Department of Computer and Information Science
University of Pennsylvania

Donald Ringe
Department of Linguistics
University of Pennsylvania

Ann Taylor
Department of Linguistics
University of Pennsylvania

June 8, 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, so that prospective trees can be evaluated according to various optimization criteria, much as is done in the practice of inferring evolutionary history for biological species. By contrast with biology, however, we find that the linguistic data support evolutionary trees with extremely good compatibility scores, and that for such data it is possible to find optimal trees quickly. We have applied this method to the classification of Indo-European (IE) languages; we have been able to resolve one longstanding open problem (the *Indo-Hittite hypothesis*), and have indicated exactly what needs to be established in order to resolve another longstanding open problem (the *Italo-Celtic hypothesis*). We have also discovered rather surprising facts about the history of Germanic within this family. Thus, this method provides an ability to resolve difficult questions in Historical Linguistics that have proved resistent to traditional character-based methodologies and to the more recent distance based approaches of *lexicostatistics*. The results of our methodology also indicate weaknesses in methods currently accepted and practiced in historical linguistics. One of our more important results is the ability to detect and handle loan words that are not distinguishable from true cognates by traditional methods. Finally, this methodology permits the linguist to develop and test assumptions about the evolutionary relevance of different linguistic characters.

1 Introduction

A set of languages is said to be *genetically related* if it meets the following criteria: (1) all the languages of the set were once a single language – called a *protolanguage* – and (2) that protolanguage diversified into the languages of the set through the regular process of first-language transmission, in which children learn their native ("first") language from the adults of their immediate speech community. The patterns of similarities between languages that genetic relationship gives rise to are largely different from those caused by *contact* between adult speech communities; in particular, they appear chiefly in the most fundamental areas of languages' structure, namely their grammars and their most basic vocabulary. If a large number of speech communities, all originally speaking a single language, gradually diversify in continual contact for centuries, the pattern of relationships between them is perhaps best modelled as what linguists call a network (just a general directed graph), but if the languages lose contact as they diversify, or if the network of diversifying dialects is sampled at points sufficiently distant from one another, the pattern of relationships observed is best modelled as a tree. The use of trees to model the evolution of languages in traditional historical linguistics is thus largely justified by the observed facts.

The determination of evolutionary trees for natural languages is a major endeavor within historical linguistics. Much is known about the evolution of some language families, such as Indo-European (IE), while for others essentially nothing has been determined. Even for the IE family, however, which is the most extensively studied, the early evolutionary history is not resolved beyond a very rudimentary separation into subfamilies (Germanic, Italic, etc.); to a large extent this is due to difficulties in interpreting data correctly, determining which information really is relevant to evolutionary history, and to the computational difficulties of exploring the tree space. Thus, while it has been possible for linguists to check their scholarly interpretations of the data against a hypothetical tree, the difficulties in exploring the exponentially sized tree space and the arguments about interpretation of data have led to giant impasses.

In this paper we will present a method for efficiently inferring evolutionary history of languages known to be related (i.e. members of the same family of languages). The method has three parts:

1. We show how to encode information about languages (interpreted by the scholar) as qualitative characters so that hypothetical trees can then be evaluated according to various objective criteria, such as those used in evolutionary tree construction in Biology when based upon biomolecular sequences.

2. We show how to efficiently find the optimal and near-optimal trees with respect to the compatibility criterion (a standard criterion for evaluating evolutionary trees in use for certain kinds of biological data, and appropriate to the linguistic context) when the data supports a tree which has extremely good scores.

3. We have developed methods appropriate to linguistic trees for finding consensus and agreement trees, which can then be applied to determine the common features of the best trees for the data set. In this way we can establish those aspects of the evolutionary history which have strong support as opposed to those which have weak support.

We have applied this method to the problem of inferring 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 methodol could lay to rest the debate on two longstanding conjectures: the *Indo-Hittite hypothesis* and the *Italo-Celtic hypothesis.* The former affirms that the Anatolian subfamily is one of two first-order subgroups of IE. In terms of the tree, then, this asserts that the root should have two children, one of which is Anatolian (represented in our database by Hittite, its best attested member). The latter claims that the Italic and Celtic subfamilies are siblings in the tree.

The implications of this methodology go beyond settling open problems within a single family of languages, however; we have shown that standard methods in use in Historical Linguistics need to be modified. For example, the incidence of undetectable borrowing (that is, borrowing that has occurred so early as to be indistinguishable from true cognation) is higher than was previously thought, but can be detected and handled through an appropriate use of our methodology.

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 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 6 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. The taxa in S label the leaves of the tree; the internal nodes of the tree represent ancestors, and the tree is rooted at the most recent common ancestor of all the taxa in S. The topology of the tree represents the chronology of speciation events (points at which a species splits into two or more species). 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 qualitative character (or simply character) is a function $c : S \rightarrow Z$ where Z is the set of integers. Thus a character defines a partition of the set of species S into equivalence classes, each class characterized by a single state of the character.

Given a set S of n taxa (whether biological or linguistic) described by a set C of k characters, we can represent the information by a $n \times k$ matrix M of character state information, in which M_{ij} is the state of species s_i for the j^{th} character. In this way each species is represented by a vector of character states. An evolutionary tree T for S has its leaves labelled by the vectors representing S, and its internal nodes also labelled by vectors of character states. Thus, for each character $\alpha \in C$, the evolutionary tree T for (S, C) defines an extension of $\alpha : V(T) \rightarrow Z$. When the tree T is given, we will denote by $\alpha_i = \{v \in V(T) : \alpha(v) = i\}$; otherwise (in the absence of such a tree) we will let α_i denote the set $\{s \in S : \alpha(s) = i\}$.

Evolutionary trees based upon characters are typically evaluated using either *parsimony* or *compatibility* (see [19] 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 subgraph 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 (S, C), finding the tree T of maximum compatibility score is called the *Compatibility Criteria Problem*. The other criterion in popular use is the parsimony criterion. The parsimony score of a tree T is the sum $\sum_{e \in E(T)} H(e)$, where $H(e)$ is the hamming distance on the edge e (that is, the number of positions in which the endpoints of e differ). Finding the tree with minimum parsimony score for a given data set is called the *maximum parsimony problem*, and is *NP-hard*[9, 11, 21].

In evaluating evolutionary trees for languages we have found the compatibility criterion more relevant than the parsimony criterion as a first comparison. If 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. Since this methodology explicitly eliminates all characters for which we believe there is a non-negligible probability of parallel development, the fact that a character is not convex on the tree under consideration is much more significant than the precise number of extra evolutionary steps required by that character on that tree. Thus, the number of characters which are not convex on T is a more accurate measure of the

"badness" of T than the exact number of extra transitions which occur on the tree. For this reason, the compatibility score of a tree is more significant than the parsimony score when evaluating evolutionary trees for natural languages. However, the parsimony score of a tree is useful in letting us compare two trees with equivalent compatibility scores.

We now consider the computational complexity of the compatibility criteria problem.

Independent Set Problem

Input: Graph $G = (V, E)$ and an integer k.

Question: Does there exist a subset $V_0 \subseteq V$ of size k such that for all $\{v, w\} \subseteq V_0, (v, w) \notin E$?

Theorem 1 (from [32]) *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-QR.*

Proof: We sketch here a reduction from Independent Set similar to that in [8]. Let $(G = (V, E), k)$ be an input to the Independent Set problem. Let $S = V \cup E \cup \{r\}$, where r denotes an added element not in $V \cup E$. For each vertex $v \in V$ let $c_v : S \rightarrow \{0, 1\}$ be defined by $c_v^{-1}(1) = \{v\} \cup \{(v, w) : (v, w) \in E\}$. It is known [22] that when the species set includes a root (that is, a species r with $c(r) = 0$ for all $c \in C$), then a pair of binary characters α and β are compatible if and only if $\alpha^{-1}(1)$ and $\beta^{-1}(1)$ are either disjoint or one contains the other. As a result, c_v and c_w are compatible if and only if $(v, w) \notin E$. Thus the graph G has an independent set of size k if and only if the character set $\{c_v : v \in V\}$ has a set of k pairwise compatible characters. Since pairwise compatibility of binary characters ensures setwise compatibility[22], G has an independent set of size k if and only if the character set has a compatible subset of k characters. Thus Compatibility Criteria is *NP-hard*.

Since this is a linear reduction, the Compatibility Criteria problem is as hard to approximate as Maximum Independent Set. Bellare and Sudan [3] have proved that the Maximum Clique (and therefore Maximum Independent Set, since it is just Clique on the complemented graph) on a graph with n nodes cannot be approximated to within a factor of $n^{1/4-o(1)}$ unless QNP = co-QR. See Johnson [24] for more details. ■

Later in this paper we will show that linguistic data supports a tree with almost perfect data (that is, a tree T which has compatibility score at least $k - t$ for very small t), and that on such data we can find the provably optimal trees in time $O(2^{2r} n k^{t+2})$.

3 Subgrouping Methodologies

Methodologies for subgrouping related languages have been debated for over a century[30]. There are two basic types of methodologies: *classical* or *traditional* methods, which are character based, and *lexicostatistical* methods, which are distance based. 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 codified in [23]. 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 [35].

It is also important to use the earliest well-attested stages of languages in attempting to construct evolutionary trees for them, because natural language change steadily erodes and obliterates original features of the protolanguage as the daughter languages develop over time. Using more recent versions of languages will not rule out the correct tree, but can lead to characters which fit not only the correct tree but many trees, and thus leads to under-differentiated trees. This is for example the problem with the placement of Albanian in the IE tree; because our Albanian data is from the 20th century and Albanian is not a very

conservative language, there are very few characters which group Albanian with any other language in the family. As a result, Albanian can be fit almost anywhere in a given evolutionary tree with equally good scores.[1] Thus, using more ancient well-attested forms of languages allows us to retain information and thus increases the probability of selecting the correct tree.

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[23]. 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 ([13, 14, 15]). 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, but the usual method makes languages which have the smallest distance between them into siblings, a new parent language is created, and the method applied recursively[13].

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*[12, 17]. 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[16] 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.

[1] The situation is different for Lithuanian because, although the Lithuanian data are likewise from the 20th century, the language is unusually conservative.

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.

The input to an algorithm for evolutionary tree construction from linguistic data is, as we will show, a combination of character state and directionality information, which we have encoded as qualitative characters in which certain aspects of the output trees are required while others are desirable but not forced. We will define an optimization criterion, called *directed compatibility*, for trees constructed from such data. Having defined our optimization criterion, we will then show that we can efficiently find all the optimal and near-optimal trees for this criterion. Our algorithm is given in Section 4.2.

The motivation we have for looking at all the optimal and near-optimal trees is that while we believe the true tree will have a good score, we cannot be sure that it is absolutely the best tree and not, for example, the second best tree. Instead, by examining all the close to optimal trees we can with high confidence be sure that the correct tree is among these trees, and we can also be confident therefore that the features which are true about most (or perhaps all) of the near-optimal trees will be true about the true evolutionary tree. Thus our algorithm for finding the optimal tree is extended to find all trees within some specified bound of optimum. This permits us to apply consensus and agreement methods to the profile of near-optimal trees, so that we can infer the common features. This is described in Section 4.3.

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. Determining which words are cognate is accomplished through the application of the Comparative Method, which was codified by Henry Hoenigswald in [23]. 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. Thus, for semantic slots, we can define equivalence classes and hence represent lexical information as characters.

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.

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/.

4.1.1 Encoding linguistic information as characters

The selection of characters requires determining which linguistic information is "genetic" rather than indicative either of chance relationship or historical contact. This is accomplished through adhering strictly to the comparative method and using only basic vocabulary as the basis of the lexical characters, because we must use properties of language which are passed genetically and are resistant to borrowing. 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.

The encoding of linguistic data is in many cases quite straightforward. Reliable cognation judgements can be made through a rigorous application of the comparative method. In the absence of independent parallel development the characters derived in this manner will be compatible with the true evolutionary tree, so that a perfect phylogeny will exist for the data set. However, because independent parallel development of character states does occur, we have developed techniques for detecting and handling it.

4.1.2 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 (the use of *croissant* in English, for example), and can be treated as lexical innovations. In this way, when the directionality of the borrowing is clear, we can still use the lexical character in the analysis of the data set by assigning a unique state to this character for the language doing the borrowing, and using cognation judgements for the remaining languages. 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).

Independent parallel semantic shift. A second type of innovation which can create false cognates is 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.). When the directionality of the shift can be established (an obvious case is words for 'animal' being based upon words for 'live' or 'breathe'), then we can again include the character in our analysis by encoding all languages exhibiting innovations arising from parallel semantic shift with their own unique states. This allows us to include these characters in our analysis. As there is no way to predict what kind of semantic shift a language will undergo, undetected cases will no doubt remain.

Same choice among alternative roots. In 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*$^{\mathrm{w}}$*her-*, **tep-*; 'wash': **lewh*$_3$*-*, **neyg*$^{\mathrm{w}}$*-*, etc.) Although the choice of one or other of these alternatives may represent an innovation on the part of a subgroup of languages, since there are a small number of choices the probability that the languages independently chose the same alternative is high. For this reason all characters with two states reconstructible to the protolanguage are eliminated.

4.1.3 Other encoding problems

Unlike linguists who employ lexicostatistics, we need not eliminate a lexical character because some of the languages lack a word for that semantic slot. Lack of words for semantic slots is a fairly prevalent problem, however, for a variety of reasons. For extinct languages this is generally because the word is simply not attested in our sometimes limited corpus. It could also happen, however, that a language simply does not encode a particular semantic slot by means of a single lexical item, as for example a tropical language might lack a word meaning 'ice' or 'snow'. A character which is missing one or more states can still be used, however, as long as each missing lexical item is coded as a separate state.

Similarly, for morphological characters the problem is how to encode loss. Take as an example the IE augment, which marks the past tenses in some archaic IE languages. All the languages which have the augment clearly exhibit the same state for this character, but because loss is an easily repeatable independent innovation, those languages which do not have the augment cannot be assumed to exhibit a single state. Rather, without any evidence to the contrary, it must be assumed that each could have lost it independently. We encode this by having each of the languages which do not have the augment exhibit a unique state for this character. This encoding does not force us to group the languages lacking the augment together in one subtree.

Some of the linguistic data gives directionality between character states. For example, some words can be clearly shown to be later forms of others because of regular sound changes. This implies a directionality in the evolutionary history of the states of the character representing the semantic slot for those words. Sometimes the information is more limited and only indicates that a particular character state is ancestral, without indicating anything about the relationship of the other states to each other. All such information allows us to eliminate certain unrooted trees from consideration, and for those trees which are not eliminated, it identifies a region within the tree in which the root (protolanguage) must be placed.

To summarize, the directionality constraints we observe have one of the following two forms:

1. The state of the root may be known for some character α.

2. For some character α, we may know that state α_i occupies a subtree above state α_j (i.e. the path from the root r in T to the subtree of nodes labelled by α_j passes through at least one node v such that $\alpha(v) = i$). Both α_i and α_j are required to be convex on T.

The additional constraints we are considering above are considered absolute constraints as opposed to desirable constraints. This means that any tree we wish to consider as an evolutionary tree for our data set *must* have the properties stated above. We now show how to encode these directionality constraints as undirected characters.

Lemma 1 *We can encode the additional directionality information by adding at most one character per information item, and one additional species.*

Proof: Let x indicate the added species. For each of the input characters $c \in C$, we will set $c(x)$ to be a state unused by any other species unless otherwise specified below. To encode a constraint of type (1) which says that state i of character α is ancestral, we set $\alpha(x) = i$. We do not need to add any additional characters to the data set for this kind of constraint. For a constraint of type (2) requiring that α_i be above α_j, we add a character β and set $\beta(x) = \beta(s) = 0$ for all s such that $\alpha(s) = i$, and $\beta(s) = 1$ for all s such that $\alpha(s) = j$. All remaining species are set to unique states (one for each species). We note that the convexity requirement of α_i and α_j translates directly into a convexity requirement of β, so that we can transfer the convexity requirement to the newly added character. It can then be verified that these added characters are compatible with a tree T if and only if T rooted at x satisfies these directionality constraints. ■

4.2 Finding optimal trees

Our optimization criterion is defined as follows. In each case we will assume that we are given a set C of characters with additional character set C' encoding the directionality constraints as defined above.

Definition: The *directed compatibility score* of T with respect to $C \cup C'$ is $-\infty$ if one of the directionality constraints is violated, and otherwise it is the number of characters in $C \cup C'$ which are convex on T.

Even though we will be examining trees primarily with respect to directed compatibility, we also need to define the directed parsimony score of a tree.

Definition: The *directed parsimony* score of T with respect to $C \cup C'$ is ∞ if one of the directionality constraints is violated, and otherwise it is the parsimony score of T.

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 minimum parsimony score and a maximum compatibility score. Determining if a perfect phylogeny exists (called the *Perfect Phylogeny Problem*) is *NP-Complete*[4, 37], but by contrast with the parsimony and compatibility problems, 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, 29, 25, 26]. We will show that linguistic data is "close" to perfect in the sense that the compatibility and parsimony scores are close to those achievable by perfect phylogenies. Precisely, we will define the imperfection of a data set as follows.

Definition: 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 2 *We can find the best tree with respect to directed compatibility in* $O(2^{2r}nc^2 + 2^{2r}nk^{t+2})$ *time, where* $n = |S|$, t *is the imperfection of the input set,* $c = |C'|$ *and* $k = |C|$.

Proof: We begin by ensuring that there is at least one perfect phylogeny consistent with C' by running the perfect phylogeny algorithm of [26] on C'. This costs us $O(2^{2r}nc^2)$ time. If these characters are compatible, then we can search among all subsets C_0 such that $C' \subseteq C_0 \subseteq C \cup C'$ in decreasing order of cardinality until we find a set which supports a perfect phylogeny. This requires $O(k^t)$ calls to [26] for a total cost of $O(2^{2r}nk^{t+2})$ time for this second phase. The total of the two phases is thus as stated above. ■

The problem of inferring the optimal trees with respect to parsimony and/or directed parsimony is more complicated. If the best tree T with respect to parsimony (or directed parsimony) has parsimony score $p(T)$, then letting $t = p(T) - \sum_{c \in C}(r_c - 1)$ we note that T has imperfection bounded from above by t. (Here r_c is the number of states attained on S for character c, so that $\sum_{c \in C}(r_c - 1)$ is the parsimony score that a perfect phylogeny would attain, were it to exist.) Thus in particular T has compatibility score bounded from below by $|C| - t$, and so we can use the greedy algorithm described above to find T, provided that we can explicitly examine *all* trees with compatibility scores above a given threshold.

Unfortunately the number of all such trees is not necessarily polynomial even for bounded r and t. This has not been a problem on our linguistic data sets in which there are very few trees with optimal or near-optimal compatibility scores, so that each such tree can be examined. At this point the question is then how to set the labels of the internal nodes of each fixed tree T so as to obtain a minimum directed parsimony

score for T. To do this we use Fitch's algorithm[20]. This algorithm takes as input a leaf-labelled tree (where the labels are vectors in Z^k) and assigns labels from Z^k to the internal nodes so as to obtain a minimum parsimony score, and does so in $O(nk)$ time. Optimizing for the parsimony criterion on a fixed topology ensures an optimum score for both the directed parsimony as well as the directed compatibility criteria, and given a labelled tree it is straightforward to compute the directed parsimony and/or directed compatibility scores.

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 (directed) compatibility* if the contracting of any edge decreases the (directed) compatibility score of T. Similarly we will say that a tree T is *minimal with respect to (directed) parsimony* if the contraction of any edge increases the (directed) parsimony 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 in *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.

To compute minimal trees (whether with respect to directed parsimony or directed compatibility) is not difficult. Since the definition of minimality does not imply that the score is minimal, only that edge contractions change the score for the worse, we can take any tree T as a starting point and simply contract all unnecessary edges. That is, we identify (and contract) all edges whose contractions do not change the directed parsimony score (for example), and contract all such edges. Identifying these edges requires one application per edge of the algorithm described in the section above for computing the directed parsimony score of a leaf-labelled tree. Since each fixed topology costs us $O(nk)$ time, and there are $O(n)$ edges, this is $O(n^2k)$ time. At the end of this process the tree which results is minimal with respect to directed parsimony. A similar process can construct trees minimal with respect to directed compatibility.

To find trees which are minimal with respect to directed compatibility and which also have optimal (or near-optimal) compatibility scores can be accomplished by using the polynomial time polynomial delay listing algorithm in [26]. This version of the algorithm only outputs minimal perfect phylogenies, and so the optimal trees (with respect to compatibility criteria) that result from using this algorithm are necessarily minimal.

4.3 Finding Common Themes

Finding the common themes of a profile of trees each leaf-labelled by the same species set S is a standard problem in evolutionary tree construction. In our case, we will construct the profile by using either the directed parsimony or directed compatibility criteria and selecting all the trees which are sufficiently close to optimal. One way to achieve this is to use the polynomial time polynomial delay perfect phylogeny algorithm in [26] as part of a greedy heuristic to compute all the best minimal trees with respect to compatibility (or directed compatibility). Having gathered these trees, we can then consider the question of inferring from these trees either a single tree (called a *consensus tree*) on the entire set of languages, or a set of trees (each called an *agreement tree*[38, 18, 28]) on subsets of the languages.

There are many models of consensus trees[10, 41, 27], but the one whch seems most relevant to the evolutionary tree problem for languages is the relaxed discord local consensus tree[27]. Another relevant approach is to identify all maximal agreement subsets of the set S.

Since there is little variation in the optimal trees, a third approach is to identify all the common themes (indicated by the maximum agreement trees), and a limited set of options which each tree in the profile selects from. This has been our approach in the analysis of IE.

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[40]. [2]. Each semantic slot was treated as a single character and judgements of cognation were made on the basis of this method. Once the states were encoded for each character, we detected evidence of parallel development. These included:

1. all characters for which two or more lexical roots are reconstructable for the protolanguag. (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- 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.

5.2 Applying the methodology

Our initial analysis of the data determined that the inclusion of Germanic (represented by Old English (OE)) resulted in trees with low compatibility score. We analyzed the data without Old English and found four trees with extremely high compatibility scores, ranging from all but 4 to all but 7 characters convex on the tree.

5.2.1 Comparing the best trees without Germanic

Our analysis indicates that Albanian can be placed anywhere in the tree with equal benefit, provided as it is above the Satem Core and not in the minimal subtree containing both Greek and Armenian; there is

[2] Our list has one more item than Tischler's[40] because we split the item *day* into two items, *period of 24 hours* and *period of daylight*.

not enough data linking Albanian to specific other IE languages. As a consequence we do not indicate the placement of Albanian in any of these trees.

The four best trees have many common features, and can be distinguished only in terms of the following two criteria:

- the placement of Tocharian B (it has however only two possible locations), and
- whether the Italo-Celtic hypothesis is denied or not; note that none of these trees actively supports this hypothesis.

Thus, most of the structure of the IE family tree can be deduced from examining these four trees. In particular, all of our trees support the Indo-Hittite hypothesis, and in fact, all of the trees we examined with parsimony scores anywhere close to the optimal parsimony score supported this hypothesis. This is the first evidence for the Indo-Hittite hypothesis that does not rest on inconclusive traditional arguments.

5.2.2 The problem of Germanic

Our analysis with Germanic indicated a sharp distinction between the lexical data and the morphological data, in that the lexical data supported a placement of Old English much higher in the tree as compared to the morphological information. This dual allegiance of Germanic is unique among the first-order subgroups of IE. It appears to point to a situation in which Germanic began to develop within the Satem Core (as evidenced by its morphology) but moved away before the final satem innovations. It then moved into close contact with the "western" languages (Celtic and Italic) and borrowed much of its distinctive vocabulary from them at a period early enough that these borrowings cannot be distinguished from true cognates. We represent this situation with two graphs: one is the genetic tree given in Figure 5 and the other is the historical tree given in Figure 6 at a time after the migration of Germanic out of the Satem Core. Note that in the historical tree we do not indicate the placement of Germanic with tree edges, but rather with directed edges to indicate historical contact rather than genetic descent.

5.2.3 Conclusions for Indo-European

Our analysis establishes the following conclusions:

- The Indo-Hittite hypothesis is completely supported without question by the data.
- The Italo-Celtic hypothesis is weakly denied by the data. To argue in favor of the this conjecture it is necessary to impugn the two lexical characters, *eye* and *ye*, which are incompatible with the best tree, and at the same time to find at least one reasonable character (morphological or lexical) which forces a grouping of Latin and Old Irish together. Such a character would have the same state for Latin and Old Irish and a different state shared between Hittite or Proto-IE and some other IE language.
- Germanic began to develop in the Satem Core and then migrated out of the core at an early date to join the western languages.

6 Summary

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 ([31]). Subsequent work along traditional lines produced no further positive results[33, 34]; arguments in favor of a more articulated tree structure supporting the Indo-Hittite

hypothesis[39, 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[40, 15], and the best-informed mathematical linguist who has attempted such work makes notably modest and reserved claims for the method[16].

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.

7 Acknowledgements

The first author wishes to thank Paul Angello, the National Science Foundation for a National Young Investigator Award #CCR-9457800, and ARO (Grant #DAAL0389-C-0031) whose financial support has made this research possible.

8 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] M. Bellare and M. Sudan, "Improved non-approximability results", *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing*, (Montreal), ACM, pp. 184-193.

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

[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] Day, W.H.E. 1983: Computationally difficult parsimony problems in phylogenetic systematics, *Journal of Theoretical Biology* 103:429-438.

[10] Day, W.H.E. 1985: Optimal algorithms for comparing trees with labeled leaves, Journal of Classification 2:7-28.

[11] Day, W.H.E., Johnson, D.S. and Sankoff, D. 1986: The computational complexity of inferring rooted phylogenies by parsimony, *Mathematical Biosciences* 81:33-42.

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

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

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

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

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

[17] Farach, M., Kannan, S., and Warnow, T. 1994: A robust model for finding optimal evolutionary trees, appeared, *Algorithmica*, 1995.

[18] Farach M., and Thorup M. 1994: Fast Comparison of Evolutionary Trees, Proceedings of the Fifth Annual ACM-SIAM Symposium on Discrete Algorithms.

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

[20] Fitch, W.M. 1971: Toward defining the course of evolution: minimum change for a specified tree topology, *Systematic Zoology* 20:406-416.

[21] Foulds, L.R. and Graham, R.L. 1982: The steiner problem in phylogeny is NP-Complete, *Advances in Applied Mathematics* 3:43-49.

[22] Gusfield, D. Efficient algorithms for inferring evolutionary trees, *Networks*, Vol. 21, pp. 19-28, 1991.

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

[24] 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.

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

[26] 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.

[27] Kannan, S., Warnow, T., and Yooseph, S. 1995: Computing the local consensus of trees, Proceedings of ACM-SIAM Symposium on Discrete Algorithms, 1995.

[28] Keselman, D. and Amir, A. 1994: Maximum agreement subtree in a set of evolutionary trees - metrics and efficient algorithms, *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, pp. 758-769, 1994.

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

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

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

[32] Phillips, C.A. and Warnow, T.J. *A new model for building consensus trees*, manuscript.

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

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

[35] 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.

[36] 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.

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

[38] Steel, M. and Warnow, T. Kaikoura tree theorems: computing the maximum agreement subtree, Information Processing Letters (48) 1993, pp. 77-82.

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

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

[41] Warnow, T. Tree compatibility and inferring evolutionary history, *J. of Algorithms,* (16), 1994, pp. 388-407.

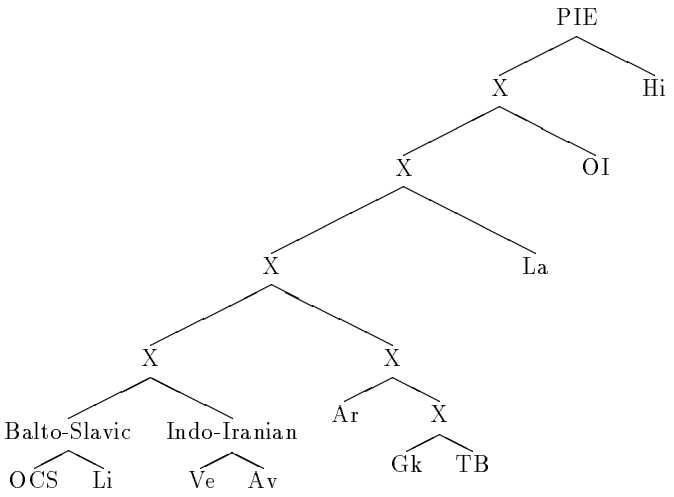

Figure 1: Best tree not including Germanic (non-convex = 2 lex, 2 morph)

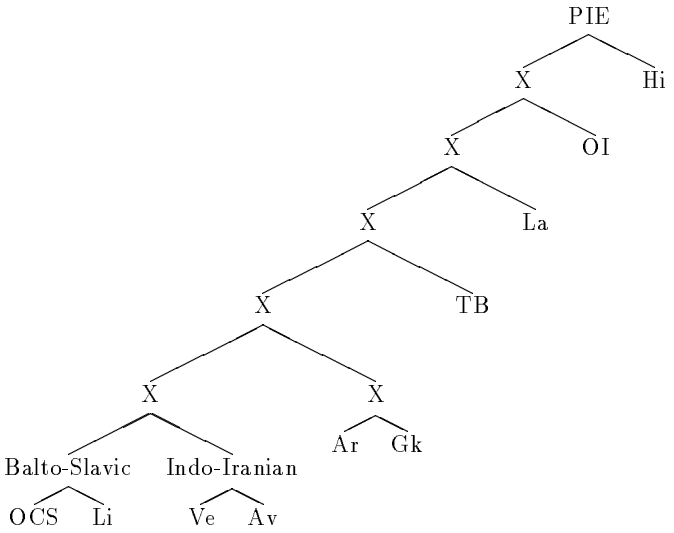

Figure 2: Second best tree not including Germanic (non-convex = 4 lex, 1 morph)

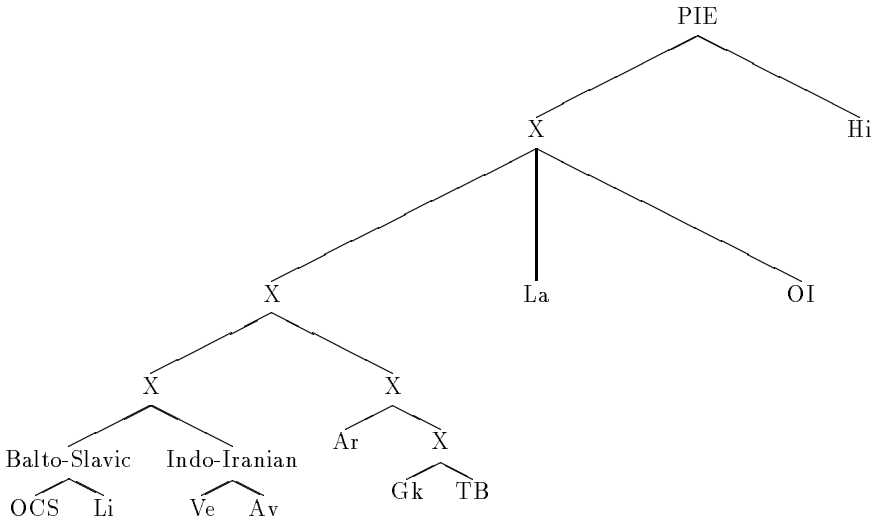

Figure 3: Third best tree not including Germanic (non-convex = 4 lex, 2 morph)

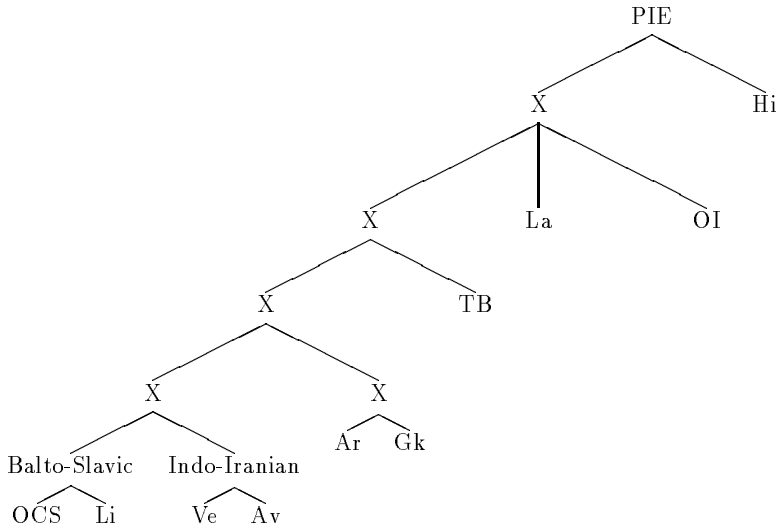

Figure 4: Fourth best tree not including Germanic (non-convex = 6 lex, 1 morph)

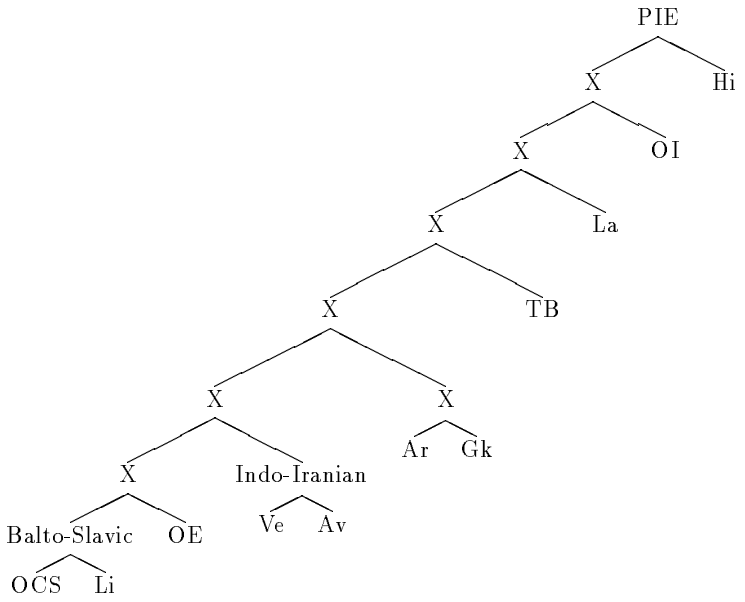

Figure 5: The most likely genetic tree for the entire IE family

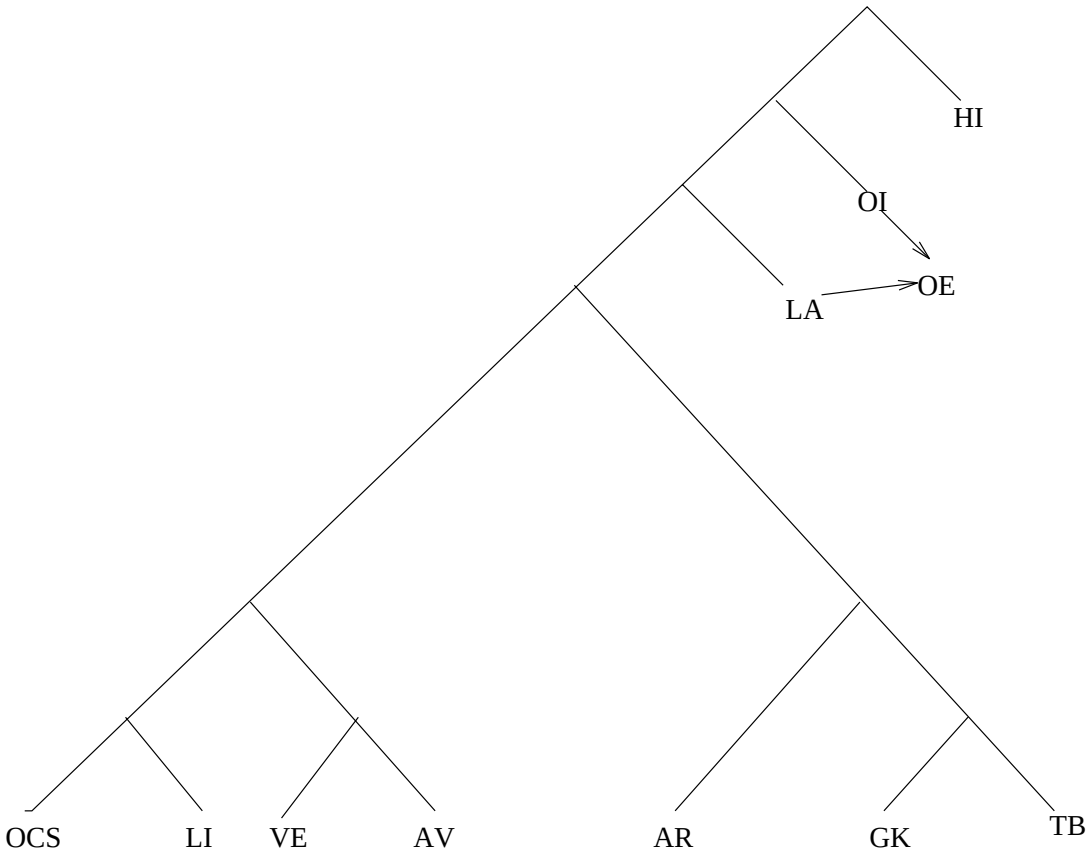

Figure 6: The historical contact tree after the migration of Germanic