A Comparison of Phylogenetic Reconstruction Methods on an IE Dataset

Luay Nakhleh
Dept. of Computer Science
Rice University
nakhleh@cs.rice.edu

Tandy Warnow
Dept. of Computer Sciences
University of Texas
tandy@cs.utexas.edu

Don Ringe
Dept. of Linguistics
University of Pennsylvania
dringe@unagi.cis.upenn.edu

Steven N. Evans
Dept. of Statistics
University of California
evans@stat.berkeley.edu

Abstract

Researchers interested in the history of the Indo-European family of languages have used a variety of methods to estimate the phylogeny of the family, and have obtained widely differing results. In this paper we explore the reconstructions of the Indo-European phylogeny obtained by using the major phylogeny estimation procedures on an existing database of 336 characters (including lexical, phonological, and morphological characters) for 24 Indo-European languages. Our study finds that the different methods agree in part, but that there are also several striking differences. We discuss the reasons for these differences, and make proposals with respect to phylogenetic reconstruction in historical linguistics.

1 Introduction

Reconstruction of the phylogenies of language families is a part of historical linguistics which has recently received significant attention from the non-linguistic scientific research community, some of whom are interested in seeing if phylogenetic reconstruction methods originally designed for biological data can be used on linguistic data to good effect. In this paper we examine the results of using phylogenetic reconstruction methods from both biology and linguistics on the character database we have used over the last decade to analyze the diversification of the Indo-European family. In addition to varying the methods we use to analyze the dataset, we study the consequences for phylogenetic reconstruction of restricting the data to lexical characters alone, and of screening the data to eliminate characters that might have evolved with borrowing or have undergone parallel evolution. Our study shows that the differences in the phylogenies obtained by different reconstruction methods are due at least in part to data selection, with analyses based upon datasets that use only lexical characters being probably less accurate than analyses based upon datasets that include morphological and phonological characters and that give these additional characters extra weight. We also find significant differences between methods, even on the same dataset. Finally, we find that equal treatment of characters is probably unwise, with improved results obtained by recognizing that some characters (notably characters derived from inflectional morphology and complex phonological characters) are less likely to evolve in parallel or with back mutation.

Our paper is organized as follows. We begin by defining the concepts and terminology in Section 2. The methods we use to analyze linguistic datasets are described in Section 3. In Section 4 we discuss the dataset we use to compare reconstruction methods, briefly discussing how the characters were selected and coded. The results of our phylogenetic analyses are presented in Section 5. We summarize our results and make recommendations about phylogenetic reconstruction in Section 6.

2 Basics

2.1 Characters

A (linguistic) character is any feature of languages that can take one or more forms; these different forms are called the "states" of the character. Our characters are of three types. For lexical characters the different states are cognate classes, so that two languages exhibit the same state for the lexical character if and only if they have cognates for the meaning associated with the lexical character. Phonological characters record the occurrence of sound changes within the (pre)history of the language; thus a typical phonological character has two states, depending of whether or not the sound change (or, more often, constellation of changes) has occurred in the development of each language. Most of our morphological characters represent inflectional markers; like lexical characters, they are coded by cognation. Thus each character defines an equivalence relation on the language family, such that two languages are equivalent if they exhibit the same state for the character. Given a partition of a set into disjoint subsets, we can define an equivalence relation by making two languages equivalent if and only if they are in the same subset; thus, a partition of a set into disjoint subsets defines an equivalence relation (and the converse holds as well).

For each character, we can assign numbers to the states of the character so that the character is defined to be a function that assigns every language in a set $\mathcal{L}$ of languages a real number; the number assigned to the language is called the "state" of the character for that language. Thus, the states of all our characters are real numbers, and when we write $c(L)$ for a language L and a character c, we mean the state of the character c exhibited by the language L. However, the particular real number used to label a state is irrelevant, and all that matters is whether two states are equal or different.

2.2 Homoplasy, Character Compatibility, and Perfect Phylogenies

The phenomenon of back–mutation and/or parallel evolution is called "homoplasy". When there is no homoplasy in a character, then all changes of state for that character result in new states. When all the characters evolve without homoplasy down a tree, then the tree is called a "perfect phylogeny", and each of the characters is said to be "compatible" on the tree.

For example, the characters c_1 and c_2 in Figure 1(b) are compatible with the tree T in Figure 2, whereas character c_3 is not.

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

(a)

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

(b)

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

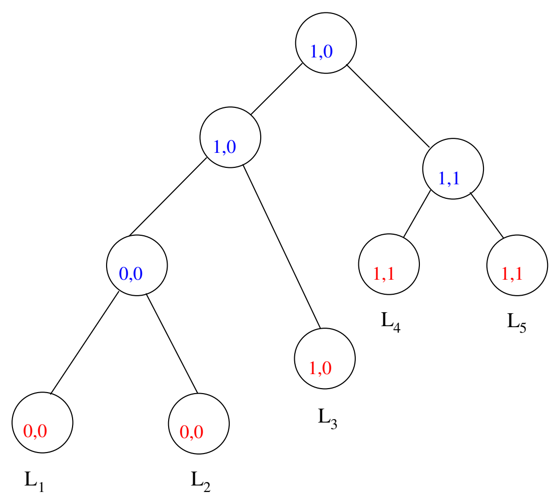


Figure 2: A perfect phylogeny T for the languages and character states of Figure 1(a).

2.3 Study design

We examine the performance of six phylogeny reconstruction methods (two distance-based methods and four character-based methods) on four versions of an IE database. We evaluate the accuracy of these methods with respect to established aspects of the Indo-European history, and also with respect to the number and type of characters that are incompatible with each of the trees returned. We use the IE dataset we have developed over the last decade as the basic dataset. This dataset contains lexical, phonological, and morphological characters, but with polymorphic characters (those which exhibit two or more states on a given language) removed; thus, we had 336 characters to work with, of which 297 are lexical, 17 are morphological, and 22 phonological. We also look at three subsets of this dataset: a screened version of the full dataset, the lexical characters alone, and a screened version of the lexical dataset. The screened versions of the full and lexical datasets are obtained by removing all characters which show strong evidence of homoplasy. Thus, the study is designed so that we can study the impact of restricting the characters to different subsets, as well as so that we can evaluate and compare different methods on the same dataset.

3 Phylogeny reconstruction

The phylogeny reconstruction methods we study in this paper include most of the standard methods used in molecular phylogenetics as well as two newer methods proposed explicitly for reconstructing phylogenies on languages. The methods studied include four character-based methods and two distance-based methods. The four character based methods each explicitly uses a consensus method called the "majority consensus tree" (defined below) in order to return a single estimate of the tree. (See (Felsenstein, 2003; Felsenstein, 1982; Swofford *et al.* , 1996) for a discussion of phylogenetic reconstruction methods used in biology, including many of the methods studied here, including maximum parsimony, maximum compatibility, UPGMA, neighbor joining, and the majority consensus tree.)

3.1 The methods

We begin our discussion of the methods we study with a description of the majority consensus tree, since it is fundamental to four of the methods (all the character-based methods, in fact).

Consensus methods Whenever a method returns a set of trees, rather than one single tree, a *consensus tree* is typically returned. Of the various different possible consensus methods, the *majority consensus* is most typically returned. This is the tree which contains exactly those edges which exist in the strict majority of the trees in the set (where we identify an edge in a tree with the bipartition it induces on the set of leaves). In our study, the MP, weighted and unweighted MC, and Gray & Atkinson methods all have the potential to return several trees (the Gray & Atkinson method returns a random sample of trees visited after burn-in, while the other methods return the trees that optimize their respective criteria). Therefore, for each of these four methods we return the majority consensus of the set of trees returned.

UPGMA The UPGMA algorithm is a distance-based method which is designed to work well when the evolutionary processes obeys the *lexical clock* assumption, and it is the algorithm used in lexicostatistical analyses. UPGMA finds the pair of languages (say L_1 and L_2) which have the smallest distance, and makes them siblings. It then removes one of the two languages (L_1) from the set, recursively constructs the tree on the remaining languages, and then inserts the removed language (L_1) back into the tree by making it a sibling to L_2.

Neighbor joining NJ, or *Neighbor Joining* (Saitou & Nei, 1987), is a standard distance-based reconstruction method used in molecular phylogenetics. Unlike UPGMA, it is able to reconstruct accurate phylogenies even if the evolutionary process does not obey the lexical clock assumption. A key aspect to making the method have good performance is the ability to estimate with a high degree of accuracy the number of evolutionary events (changes in character state) on every leaf-to-leaf path in the tree. Given a distance matrix which is exactly correct with respect to these leaf-to-leaf distances, NJ will return the correct tree, albeit not the location of the root. One of the differences between NJ and UPGMA is that when the set of languages does not evolve under a lexical clock, the pair of languages which have the smallest distance may not be siblings in the true tree. In this case, UPGMA will incorrectly join the nearest pair, but NJ will not (provided, of course, that the distance matrix given to NJ is sufficiently accurate).

Maximum Parsimony *Maximum Parsimony*, or MP, is an optimization problem which seeks a tree on which a minimum number of character state changes occurs. This is a hard optimization problem which can be solved exactly only for small datasets of up to about 20 languages. Beyond that size, heuristics are used to "solve" MP, which may or may not find optimal solutions. Because several equally good trees can be returned in an MP analysis, we report results for the majority consensus tree of the best trees found. MP is one of the most important and frequently used methods in molecular phylogenetics, and heuristics for MP exist in a variety of software products.

Maximum Compatibility Maximum Compatibility, or MC, is an optimization problem which seeks a tree on which the maximum number of characters are compatible. Like MP, it is hard to solve exactly, and so heuristics (which are not guaranteed to solve the problem) are used to analyze datasets above 20 or so languages.

Weighted Maximum Compatibility In a weighted maximum compatibility analysis we are given weights for each character, so that characters that are considered to be more resistant to homoplasy are given higher weight. In this case, rather than seeking a tree on which the smallest number of characters are incompatible, we seek a tree which has the smallest total weight of incompatible characters (so that we sum up the weights of the incompatible characters). As an example, if each lexical character has weight one, and each phonological character has weight three, and a morphological character has weight five, then a tree with four incompatible lexical characters would be preferable to a tree with one morphological character incompatible, but a tree with just one phonological character incompatible would be better than both of these.

Gray & Atkinson The method (originally presented in (Gray & Atkinson, 2003)) designed by Russell Gray and Quentin Atkinson operates as follows. First, each multistate characters is replaced by a binary encoded version of the character, and these binary characters are then interpreted as restriction sites and analyzed under a rates-across-sites model in the MrBayes software (Huelsenbeck & Ronquist, n.d.). After an initial burn-in period, a majority consensus of a random sample of trees is returned.

3.2 Software

PAUP* for basic methods We also used PAUP* (Swofford, n.d.) for NJ, UPGMA, heuristic maximum parsimony, and for computing the majority tree. The PAUP* block we used for heuristic search MP is:

```
set criterion=parsimony maxtrees=100 increase=no;
hsearch start=stepwise addseq=random nreps=25 swap=tbr;
filter best=yes;
set maxtrees=100 increase=no;
hsearch start=current swap=tbr hold=1 nbest=1000;
```

We also use PAUP* to compute the majority consensus tree, our second phase in MP, MC, weighted MC, and the Gray & Atkinson method.

Weighted and unweighted maximum compatibility In order to attempt to solve either weighted or unweighted maximum compatibility, we did the following. First, we used our MP heuristic search to search for good MP trees, maintaining a list of the best trees we see. Once we have completed our search, we scan the trees we explored and compute the (weighted or unweighted) compatibility score of each tree. We compare the best-scoring trees we find this way to other trees we have found using other techniques, and return as our (weighted or unweighted) maximum compatibility trees those that have the optimal score. Thus, the MP heuristic serves to produce a set of potential candidates, but these are not the only ones that are evaluated.

The weighting function we used for our weighted maximum compatibility criterion had only two values – infinite (thus requiring the character to be compatible on the optimal trees) or 1. We assigned weights as follows: No lexical character is required to be compatible, and so all lexical characters that were not omitted due to polymorphism are given the same weight. Of the morphological characters, one (M7) was omitted from both datasets due to polymorphism, and three (M9, M10, and M11) were considered sufficiently likely to evolve either in parallel or to spread through contact that they were also treated equally with lexical characters. Of the phonological characters, two (P2 and P3) were considered potentially able to spread through contact and so were assigned weight one. (These two characters define the "satem" subgroup and might reflect either shared descent in the strictest sense or have spread through a dialect continuum; see e.g. (Hock, 1986):442-4). Downweighting these two characters to be equivalent in importance with lexical characters thus enables us to reconsider the Satem Core.

Gray & Atkinson The Gray & Atkinson approach uses the MrBayes software package (Huelsenbeck & Ronquist, n.d.), and we duplicated the approach they used with respect to the number of iterations, sampling frequency, and reported the majority consensus. MrBayes was run using the following block:

```
lset rates=gamma;
mcmc ngen=10000000 burnin=300000 printfreq=10000 samplefreq=10000
      nchains=4 savebrlens=yes;
```

Distances For UPGMA we used Hamming distances (the number of characters in which two languages are different). For NJ we compute distances under the formula given in (Warnow *et al.* , 2004) (this is a statistically consistent distance estimation technique for homoplasy-free evolution - no statistically consistent distance estimator yet exists for a model with homoplasy but an infinite number of states).

4 The IE Dataset

Our basic dataset consists of 336 characters for 24 IE languages (that is, we removed 40 characters from a larger dataset of 376 characters, because these were clearly polymorphic). We will first describe and explain our choice of languages and characters, then describe our coding of the characters.

4.1 Selection of languages

The languages are listed in Table 1.

As can be seen, they represent all ten well-attested subgroups of the IE family (namely Anatolian, Tocharian, Celtic, Italic, Germanic, Albanian, Greek, Armenian, Balto-Slavic, and Indo-Iranian). To represent each subgroup we have chosen a language or languages that are attested relatively fully at as early a date as possible. For instance, Indic is represented by early Vedic, since the Rigveda and other very early texts are extensive enough to provide us with data for most of our characters; but we have used "younger" Avestan rather than the earlier Gatha-Avestan to represent eastern Iranian, since the Gathas are too restricted for our purposes. Greek is represented by Classical Attic rather than Homeric, both because our attestation of Attic is far more extensive and because the Homeric language is known to be an artificial literary dialect. Similar decisions have been made in the other cases. We have used modern data for Welsh, Lithuanian, Latvian, and Albanian because earlier data are much less accessible and because we judged that it would make little difference in those cases.

Table 1: The 24 IE languages analyzed.

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

The fact that the languages of our database are not contemporaneous has a possible negative impact on the UPGMA method, since this method operates best when the evolutionary process is clock-like, and all the leaves are at the same time depth. However, this selection of our data will not necessarily negatively impact the performance of any of our other methods. (In fact, it is advantageous to character-based methods to use the earliest attested languages, since these are more likely to have retained character states that are informative of the underlying evolutionary history.)

In order to represent as many of the major subgroups as was practicable we were obliged to use some fragmentarily attested ancient languages for which only a minority of the lexical characters could be filled with actual data. The languages in question are Lycian (for which we have only about 15% of the wordlist), Oscan (ca. 20%), Umbrian (ca. 25%), Old Persian (ca. 30%), and Luvian (ca. 40%). At the other extreme we have complete or virtually complete ($\geq$ 99%) wordlists not only for the modern languages but also for Ancient Greek, Latin, Old Norse, Old English, and Old High German; we also have nearly complete ($\geq$ 95%) wordlists for Vedic, Classical Armenian, Old Irish, and Old Church Slavonic. Coverage of the remaining wordlists ranges from about 70% to about 85%.

The inclusion of three Baltic languages and of four Germanic languages introduces parallel development in a considerable number of lexical characters, thus decreasing the amount of usable evidence. We have retained the full set of languages in the database because the internal subgrouping of Balto-Slavic and of Germanic are matters of ongoing debate in the specialist community.[1] On the other hand, the inclusion of only two West Germanic languages–Old English and Old High German, the northernmost and southernmost respectively–potentially

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

avoids much greater character incompatibilities, since the internal diversification of West Germanic is known to have been radically non-treelike (cf. (Ringe *et al.* , 2002):110).

4.2 Character selection

The original database, which included polymorphic characters, had a total of 376 characters; 40 (one morphological, and 39 lexical) of these were removed due to evidence of polymorphism, leaving us with 336 characters. This unscreened database includes 22 phonological characters encoding regular sound changes (or, more often, sets of sound changes) that have occurred in the prehistory of various languages, 17 morphological characters encoding details of inflection (or, in one case, word formation), and 297 lexical characters defined by meanings on a basic wordlist. (A modern English example of a polymorphic character would be the meaning 'small', for which English contains at least two basic equivalents, *small* and *little*. Polymorphic characters were omitted from the dataset, both because no approved methodology exists for analyzing polymorphic characters and because we have not yet evolved a proposal for analyzing such data.) The data were assembled by Don Ringe and Ann Taylor with the advice of other specialist colleagues. Details of the character selection can be obtained from (Ringe *et al.* , 2002). The database just described was the basis of the analysis reported in (Ringe *et al.* , 2002).

Gaps in the data are coded with unique states, which are compatible with any tree. Therefore, though gaps do not cause problems for the maximum parsimony or the weighted or unweighted maximum compatibility methods, they do decrease the robustness of certain subgroups under these analyses–which is, of course, realistic. The impact of this encoding on distance-based methods or on the Gray and Atkinson method is currently unknown.

We then produced a *screened* dataset, excluding all characters that clearly exhibit parallel development (whether or not they are compatible with any plausible tree). The result of this screening eliminated 38 lexical characters, four (M9a, M9b, M10a, and M10b) of the morphological characters, and none of the phonological characters. Some discussion of the characters and of the rationale for eliminating particular characters is provided in (Ringe *et al.* , 2002); a detailed discussion of the elimination of specific characters will be made available on our project website (Nakhleh *et al.* , 2004).

4.3 Summary of the four datasets

Recall that we do not include any polymorphic characters in any dataset, and so we draw our subsets from a full dataset of 336 monomorphic characters. Because we consider lexical alone and all the character types, and screened as well as unscreened, we have four datasets:

- Full, unscreened. This dataset is the complete set of characters; thus, no character was eliminated due to evidence of homoplasy, even if that evidence was uncontroversial. This dataset has a total of 336 characters, 297 of which are lexical, 17 are morphological, and 22 phonological.

- Full, screened. In this dataset we eliminate only those characters with strong evidence of homoplasy. The total number of characters in this dataset is 294, of which 259 are lexical, 13 morphological, and 22 phonological.
- Lexical, unscreened. Here we include every lexical character, whether likely or not to evolve with homoplasy. The total number of characters in this dataset is 297.
- Lexical, screened. We remove all characters from the lexical unscreened dataset that have clear evidence of homoplasy. The total number of characters in this dataset is 259.

5 Results

5.1 Evaluation criteria

We evaluate trees (and therefore the methods used to infer the trees) according to two basic criteria: first, the number and type of characters that are incompatible with the tree, and second, agreement with established aspects of IE history. These established aspects are: Indo-Iranian, Balto-Slavic, and the remaining eight (8) families - Italic, Celtic, Greek, Armenian, Germanic, Albanian, Anatolian, and Tocharian.

5.2 Initial observations

The most striking observation about the different methods we examined, and their inferred trees, is that UPGMA did clearly the worst with respect to both criteria. In particular, UPGMA failed to find the Iranian clade, as it separates Persian from Avestan and Vedic, and it also failed to find Italic, as it split Latin off from Oscan and Umbrian. Furthermore, UPGMA had the most incompatible characters, including a large number of both phonological and morphological characters. Consequently, UPGMA is clearly inferior on these datasets. However, UPGMA's poor performance may be a consequence of the process we used to select our languages, as we discussed earlier.

Because of UPGMA's poor performance, we will focus our attention primary on the other methods, which are maximum parsimony (MP), weighted and unweighted maximum compatibility (WMC and MC), neighbor joining (NJ), and the technique of Gray and Atkinson (GA).

The most striking observations about the methods other than UPGMA are as follows:

- These five methods each recovered all the established subgroups of Indo-European, as well as also constructing Greco-Armenian (that Greek and Armenian are sister subgroups). They also agree about the internal subgrouping within Germanic, Italic, and Indo-Iranian, but not always within Anatolian. However, the different methods posit *very different* relationships between these major subgroups. With the exception of maximum parsimony on the unscreened lexical character dataset, all methods reconstructed Anatolian-Tocharian (that under the assumption that Anatolian is the first subgroup off the root of IE, Tocharian is the second subgroup off).

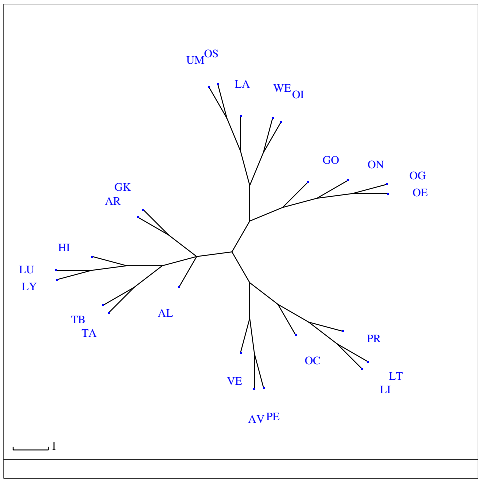
- Albanian is found in varied positions within the trees, so that consensus about its relative placement is unlikely; for this reason we ignore Albanian in evaluating these methods on these datasets.
- On most datasets, Maximum Parsimony and unweighted Maximum Compatibility return extremely similar trees (identical or compatible much of the time), modulo the position of Albanian. (The single exception is the unscreened lexical dataset, in which MP and MC are quite different.)
- Most methods (other than weighted MC) return different trees on screened and unscreened datasets, though for most methods the changes in the tree are relatively local.
- Certain posited relationships *only* show up if morphological and phonological characters are included in the analysis.
- The only trees in which Italic and Celtic are not placed within the "core" are those based upon Weighted Maximum Compatibility, with morphological character M5 receiving significant weight.
- All methods other than Weighted MC return trees with incompatible morphological and phonological characters, suggesting that weighting of characters is an important aspect of a phylogenetic analysis.

We now turn to a careful discussion of the trees and their incompatible characters, on each dataset in turn from (in our opinion) most reliable (screened full dataset) to least reliable (unscreened lexical dataset).

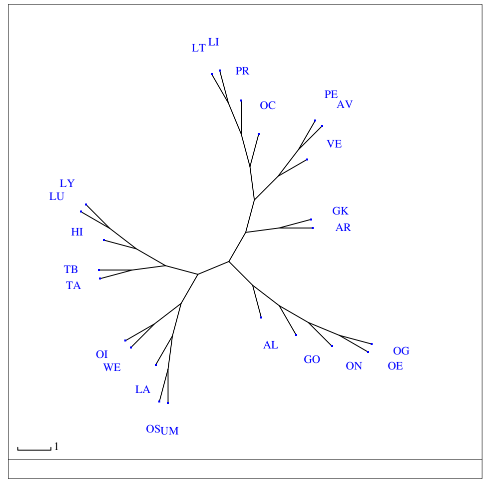
5.3 Screened Full Dataset

We constructed six trees on this dataset, one for each method we studied. Since MP and (unweighted) MC return identical trees, we show figures only for trees on five of these methods: weighted MC (given as T_{WMC} in Figure 3(a)), Gray & Atkinson (given as T_{GA} in Figure 3(b)), MP=MC (given as T_{MC} in Figure 3(c)), UPGMA (given as T_{UPGMA} in Figure 3(d)), and NJ (given as T_{NJ} in Figure 3(f)).
Lists of incompatible characters for each tree

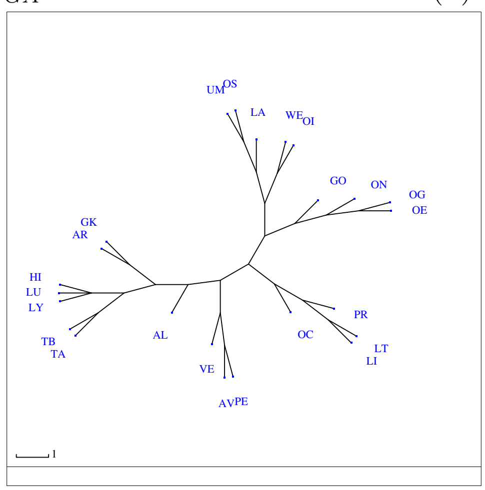
Characters incompatible on the Gray & Atkinson tree T_{GA} (15):
M5 float2 ice straight suck2 arm beard break1 free leave1 thousand1 young2 tear head nine



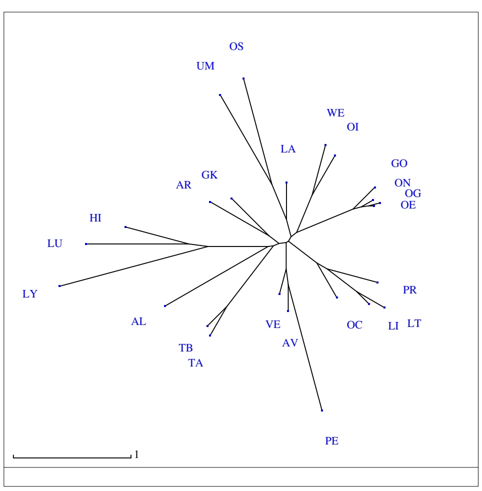
(a) T_{GA}



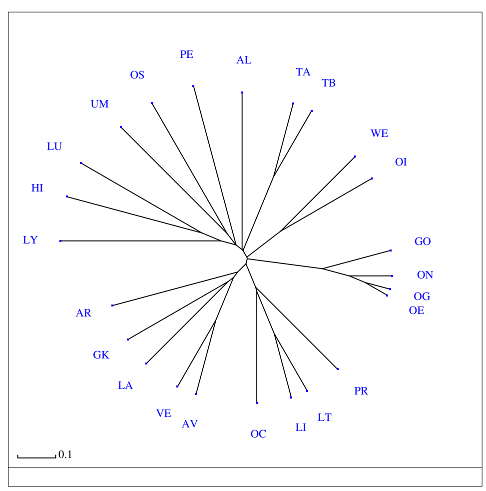
(b) T_{WMC}



(c) T_{MC}



(e) T_{NJ}



(f) T_{UPGMA}

Figure 3: Five trees inferred on the screened full dataset.

Characters incompatible on the weighted MC tree T_{WMC} (15):
all1 breast1 float2 one straight arm beard break1 free pour thousand1 young2 tear head nine

Characters incompatible on the MC tree T_{MC} (14):
P2 P3 M5 float2 ice straight suck2 break1 free leave1 young2 tear head nine

Characters incompatible on the neighbor joining tree T_{NJ}**:** (17)
P1 P2 P3 M5 M6 M8 M11 drink float2 head ice straight suck2 break1 leave1 nine tear

Characters incompatible on the UPGMA tree T_{UPGMA}**:** (75)
P1 P2 P3 P12 P14 M1a M1b M5 M6 M8 M11 M12 M13 M14 all1 all2 and bad breast1 day1 day2 dig drink ear1 ear2 earth1 earth2 eye1 eye2 few1 fish1 float2 full2 give1 hand1 hand2 head long1 long2 mountain1 not one other righthand river see sing stand2 stone straight suck2 there tongue2 tongue3 us wind1 wind2 yellow2 arm beard bee2 break1 buy daughter-in-law free goat gold grind king now1 now2 pour young1 young2 tear

Note that other than UPGMA's extremely large number of incompatible characters, the remaining methods are all able to reconstruct trees with a small number of incompatible characters. However, the particular characters that are incompatible on the trees differ in some important ways, especially with respect to particular phonological and morphological characters. We will return to a discussion of these characters later.

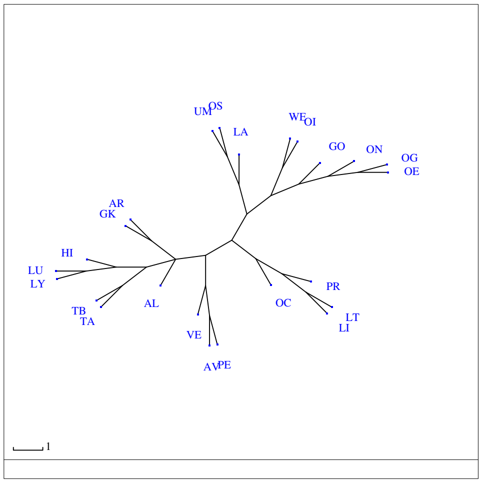
5.4 Unscreened Full Dataset

Of the six trees we constructed on this dataset, we will show only those obtained from maximum parsimony, unweighted maximum compatibility, weighted maximum compatibility, Gray & Atkinson, and NJ. UPGMA's performance, as before, is very poor. See Figure 4 for these trees.
Lists of incompatible characters for the trees

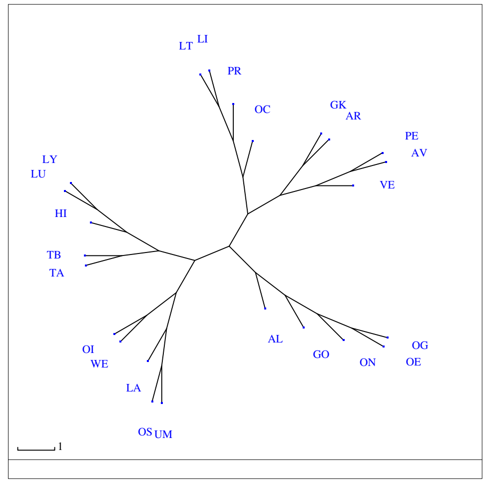
Characters incompatible on T_{GA}**:** (51)
P1 P2 P3 M5 M6 M8 M9a M11 black blood1 blood2 fall father fire float2 fog1 fog2 head hear1 hear2 here hold horn1 horn2 husband I me ice if lie sleep straight suck2 swim thee tooth we where ye arrow break1 honey house1 house2 leave1 nine ox sweat weave wolf tear

Characters incompatible on T_{WMC}**:** (53)
P2 P3 M9a all1 black blood1 blood2 breast1 fall father fire float2 fog1 fog2 head hear1 hear2 here hold horn1 horn2 husband I me if lie one sleep straight suck2 swim thee tooth we where ye arm arrow beard break1 duck free honey house1 house2 nine ox pour sweat weave wolf young2 tear

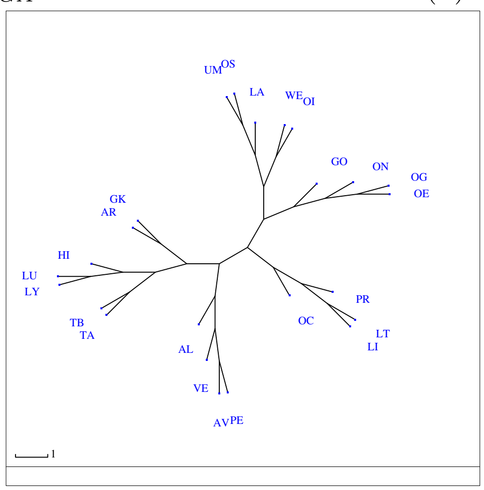
Characters incompatible on T_{MC}**:** (48)
P2 P3 M5 M9a black blood1 blood2 fall father fire float2 fog1 fog2 head hear1 hear2 here



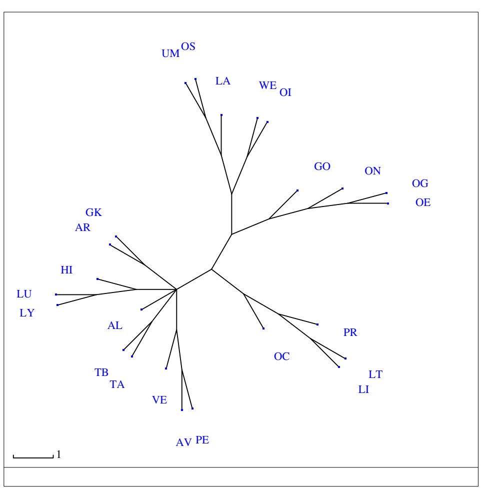
(a) T_{GA}



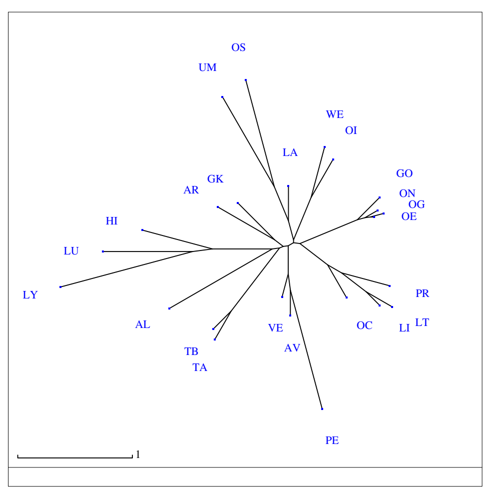
(b) T_{WMC}



(c) T_{MC}



(d) T_{MP}



(e) T_{NJ}

Figure 4: Five trees inferred on the unscreened full dataset.

hold horn1 horn2 husband I me ice if lie sleep straight suck2 swim thee tooth we where ye arrow break1 free honey house1 house2 leave1 nine ox sweat wolf young2 tear

Characters incompatible on T_{MP}: (52)
P2 P3 M5 M9a all1 black blood1 blood2 drink fall father fire float2 fog1 fog2 give1 head hear1 hear2 here hold horn1 horn2 husband I me ice if lie sleep straight suck2 swim thee tooth we where ye arrow break1 free honey house1 house2 leave1 nine ox sweat weave wolf young2 tear

Characters incompatible on T_{NJ}: (53)
P2 P3 M5 all1 black blood1 blood2 drink fall father fire fish1 float2 fog1 fog2 head hear1 hear2 here hold horn1 horn2 husband I me ice if lie long1 long2 sleep straight suck2 swim thee tooth we where ye arrow break1 free honey house1 house2 leave1 nine ox sweat weave wolf young2 tear

Characters incompatible on T_{UPGMA}: (115)
P1 P2 P3 P12 P14 M1a M1b M5 M6 M8 M9a M10a M10b M11 M12 M13 M14 all1 all2 and bad black blood1 blood2 breast1 day1 day2 dig drink ear1 ear2 earth1 earth2 eye1 eye2 fall father few1 fire fish1 float2 fog1 fog2 full2 give1 green2 hand1 hand2 head hear1 hear2 here hold horn1 horn2 husband I me if lie long1 long2 mountain1 not one other righthand river see sing sleep snow1 stand2 stone straight suck2 swim there thee to ngue2 tongue3 tooth we us where wind1 wind2 ye yellow2 arm arrow beard bee2 break1 buy daughter-in-law duck free goat gold grind honey house1 house2 king lamb now1 now2 pour sweat weave wolf young1 young2 tear

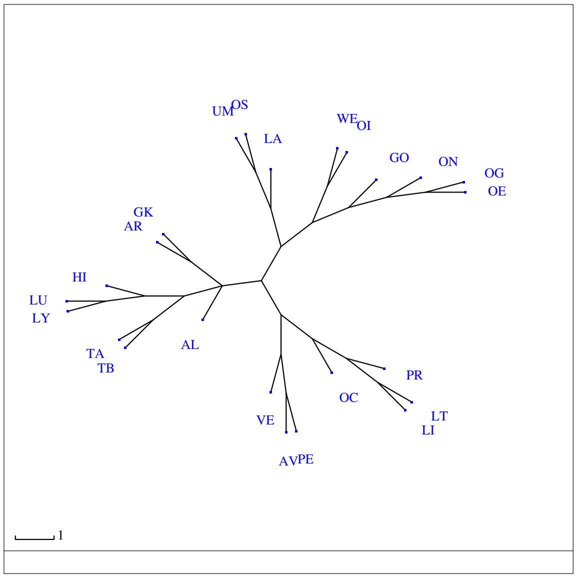
5.5 Screened Lexical Dataset

Under the weighting we use, all lexical characters have the same weights, and hence weighted MC and unweighted MC are identical on any lexical (screened or unscreened) dataset. For the screened lexical dataset, MC and MP differ only with respect to the placement of Albanian. Thus, for this dataset, we will report results for only four methods: MC, Gray & Atkinson, UPGMA, and NJ. We present these trees in Figure 5.
Lists of incompatible characters for the trees

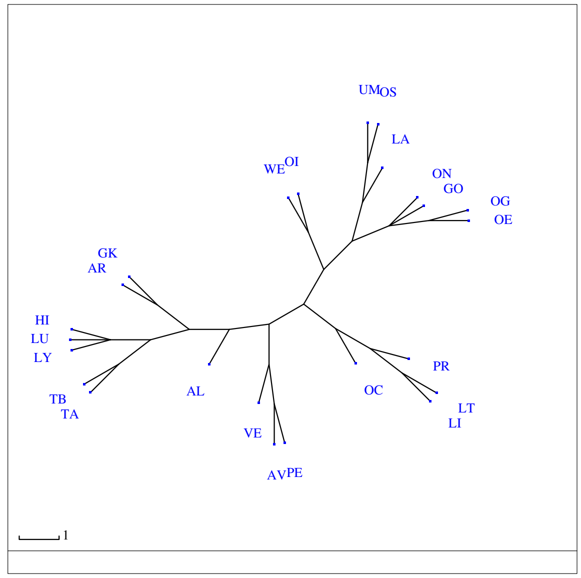
Characters incompatible on tree LT_{GA}: (12)
float2 head ice straight suck2 arm beard break1 leave1 nine thousand1 tear

Characters incompatible on tree LT_{MC}: (9)
float2 head ice free horse leave1 nine young2 tear

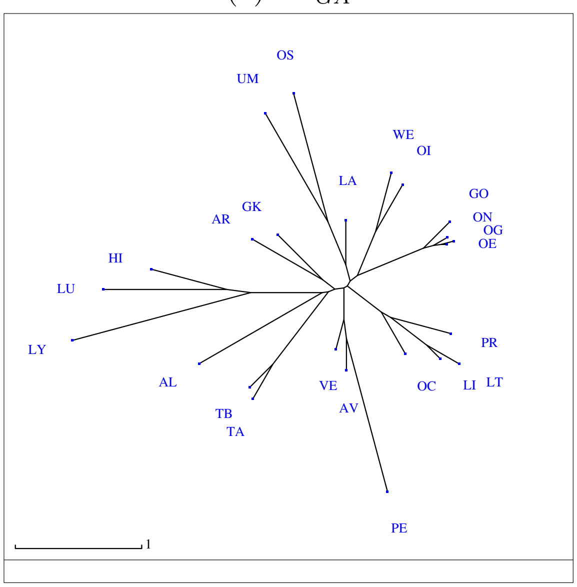
Characters incompatible on tree LT_{NJ}: (10)
drink float2 head ice straight suck2 break1 leave1 nine tear



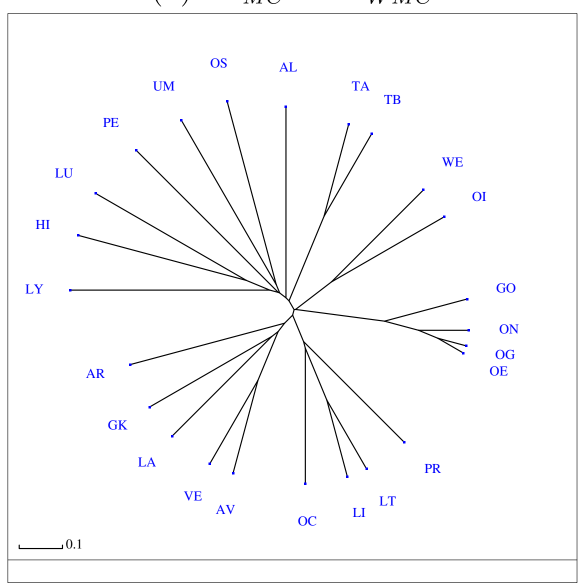
(a) LT_{GA}



(b) $LT_{MC} = LT_{WMC}$



(c) LT_{NJ}



(d) LT_{UPGMA}

Figure 5: Four trees inferred on the screened lexical dataset.

Characters incompatible on tree LT_{UPGMA}: (61)
all1 all2 and bad day1 day2 dig drink ear1 ear2 earth1 earth2 eye1 eye2 few1 fish1 float2 full2 give1 green2 hand1 hand2 head ice long1 long2 mountain1 not one other righthand river see sing stand2 stone straight suck2 there tongue2 tongue3 us wind1 wind2 yellow2 arm beard bee2 break1 buy daughter-in-law goat gold grind king now1 now2 pour young1 young2 tear

As before, UPGMA is much worse than the others, but there is little difference with respect to compatibility scores for the other methods. More striking, however, is the differences between these trees and some of those obtained using morphological and phonological characters, at least when the morphological and phonological characters are weighted.

5.6 Unscreened Lexical Dataset

As explained before, weighted and unweighted compatibility methods do not differ on this dataset. Interestingly, and in contrast to the other datasets, MP's reconstruction is quite different from MC's. We present only four of these methods (see Figure 6), omitting UPGMA since its performance is so poor.

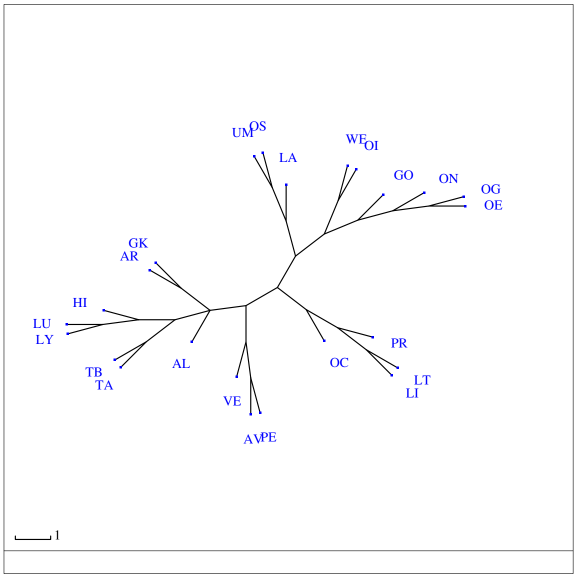
There were 45 characters incompatible on the maximum parsimony tree, 44 on the maximum compatibility tree, 43 on the Gray & Atkinson tree, 98 on the UPGMA tree, and 44 on the NJ tree. (The maximum compatibility tree was computed by taking the majority consensus of all trees with 43 incompatible characters; this included some trees explored during the maximum parsimony search, as well as the NJ tree.) Note the difference between the MP and the MC trees.

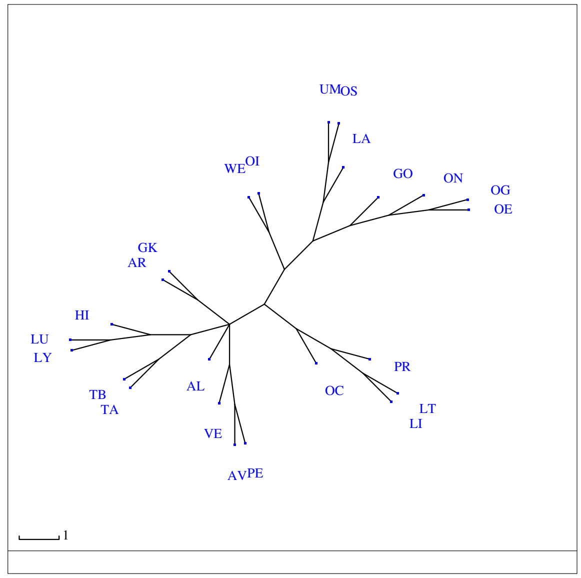
Characters incompatible on LT_{GA}: (43)
black blood1 blood2 fall father fire float2 fog1 fog2 head hear1 hear2 here hold horn1 horn2 husband I me ice if lie sleep straight suck2 swim thee tooth we where ye arrow break1 honey house1 house2 leave1 nine ox sweat weave wolf tear

Characters incompatible on $LT_{MC} = LT_{WMC}$: (44)
all1 black blood1 blood2 fall father fire float2 fog1 fog2 head hear1 hear2 here hold horn1 horn2 husband I me ice if lie sleep swim thee tooth we where ye arrow free honey horse house1 house2 leave1 nine ox sweat weave wolf young2 tear

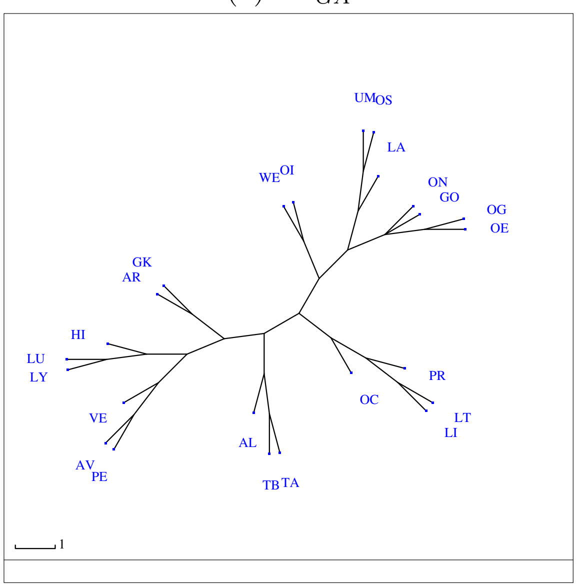
Characters incompatible on LT_{MP}: (45)
all1 black blood1 blood2 drink fall father fire float2 fog1 fog2 give1 head hear1 hear2 here hold horn1 horn2 husband I me ice if sleep swim thee tooth we where ye arrow free honey horse house1 house2 leave1 nine ox sweat weave wolf young2 tear

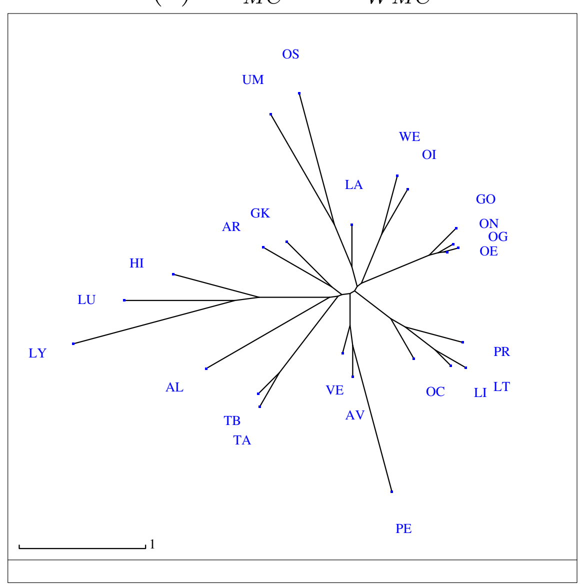
Characters incompatible on LT_{NJ}: (44)
black blood1 blood2 drink fall father fire float2 fog1 fog2 head hear1 hear2 here hold horn1 horn2 husband I me ice if lie sleep straight suck2 swim thee tooth we where ye arrow break1 honey house1 house2 leave1 nine ox sweat weave wolf tear





(a) LT_{GA} (b) $LT_{MC} = LT_{WMC}$





(a) LT_{MP} (b) LT_{NJ}

Figure 6: Four trees inferred on the unscreened lexical dataset.

Characters incompatible on LT_{UPGMA}: (98)
all1 all2 and bad black blood1 blood2 breast1 day1 day2 dig drink ear1 ear2 earth1 earth2 eye1 eye2 fall father few1 fire fish1 float2 fog1 fog2 full2 give1 green2 hand1 hand2 head hear1 hear2 here hold horn1 horn2 husband I me if lie long1 long2 mountain1 not one other righthand river see sing sleep snow1 stand2 stone straight suck2 swim there thee tongue2 tongue3 tooth we us where wind1 wind2 ye yellow2 arm arrow beard bee2 break1 buy daughter-in-law duck free goat gold grind honey house1 house2 king lamb now1 now2 pour sweat weave wolf young1 young2 tear

5.7 Character incompatibility patterns

The trees we have seen have differed topologically in interesting and significant ways, but also with respect to the specific characters on which they are incompatible. Since not all characters are equally readily borrowed, nor equally likely to evolve in parallel or with back mutation, it makes sense to consider the incompatibility patterns underlying the different trees.

A full examination of the lexical characters is beyond the scope of this paper, but in any event lexical characters are the most readily borrowed. For this reason, we will focus on the morphological and phonological characters that were incompatible with a phylogeny reconstructed using one of the methods other than UPGMA (since UPGMA's results are so poor). In what follows we begin with the least strongly supported such characters, and move through the list of such characters that are incompatible on at least one of the trees in our study.

M9a is one alternative coding of the athematic dative plural ending. It forces three subgroups: (1) Anatolian; (2) Indo-Iranian plus Italo-Celtic; (3) Germanic plus Balto-Slavic. The actual difference between (2) and (3) is whether the ending *-os is preceded by *-bh- or *-m-. Since it is not inconceivable that one replaced the other, the members of these subgroups need not be nearest sisters; the subgroups can instead be nested. However, they cannot overlap topologically. This character appears only in the unscreened full dataset because we are fairly confident that it reflects parallel development; there is a growing consensus to that effect among specialists (see (Beekes, 1985):143-4, (Beekes, 1995):115-8, (Hajnal, 1995):327-37, and (Katz, 1998):248-51). Therefore the fact that it is incompatible with all the trees constructed from the unscreened data except that found by NJ is not significant; it is not even a weak endorsement of NJ.

P2 (the "satem" development of dorsals) and P3 (the "ruki"-rule) require Balto-Slavic and Indo-Iranian to be nearest sisters. They therefore impugn almost all the trees constructed from the unscreened full dataset (excluding only WMC), as well as the trees constructed from screened data by MP (= MC) and NJ–but not the trees constructed from screened data by G&A's method and by WMC. If we could be certain that the phonological developments represented by these characters necessarily occurred during a period of shared genetic descent, they would impugn most of the trees and most of the methods. However, there is at least some possibility that these sound changes spread from one diversifying dialect of PIE to another that was already significantly different, though not so different as to impede communication (see e.g. (Hock, 1986):442-4).

Such a scenario would be considerably more plausible to the extent that the two subgroups have a recent (as measured in terms of linguistic divergence) common ancestor; the further back in time that common ancestor is, the less likely the scenario will seem on linguistic grounds.

Because P2 and P3 are less secure than the other phonological characters and than most morphological characters, one cannot easily judge the performance of any given method by how it treats these two characters.

M11 (representing the extension of the abstract noun suffix *-ti- with a further suffix *-Hen-) requires Italic and Celtic to be nearest sisters. It therefore impugns the tree constructed by G&A's method from unscreened data, and that constructed by NJ from screened data. Unfortunately this character is not as reliable as most of the morphological characters. For one thing, it is the only morphological character which encodes an aspect of word- formation rather than inflection, and it appears that such "derivational" morphology is not as resistant to borrowing as inflectional morphology is. Moreover, there has long been some question whether the same change might not have occurred independently in Armenian, which no method finds to be a near sister of Italic or Celtic (see (Olsen, 1999) with references); more recently Craig Melchert has suggested that the same development might have occurred in Anatolian as well, and that too must have been an independent event (see (Melchert, 2003)).

Consequently we also cannot use this character to judge the performance of different methods with any confidence.

The four characters discussed from this point forward (P1, M5, M6, and M8) are, in our opinion, completely reliable indicators of shared genetic descent. They therefore impugn not only trees with which they are not compatible, but also the methods by which those trees were constructed. In all cases the shortcoming of the methods is the same: they treat all characters alike, with no weighting. These characters thus amount to four strong arguments in favor of the weighting of characters.

P1 (the sound change $*\mathrm{p}\ldots\mathrm{k}^{\mathrm{W}} > *\mathrm{k}^{\mathrm{W}}\ldots\mathrm{k}^{\mathrm{W}}$) requires Italic and Celtic to be nearest sisters. It therefore impugns the tree constructed by G&A's method from unscreened data, and that constructed by NJ from screened data. In our judgment this sound change is odd enough to guarantee the Italo-Celtic clade (though not quite all colleagues would agree). It thus impugns not only those trees, but also the methods that found them. However, it is only a single sound change affecting three lexemes (!); thus it would not be completely unreasonable to argue that incompatibility with this character is a relatively minor matter.

M6, which encodes the thematic optative suffix, and M8, which encodes the (most archaic) superlative suffix, require that the portion of the tree including Italic and Celtic not overlap with the portion of the tree including Germanic, Greek, Indo-Iranian, and (in the case of M6) Balto-Slavic–though the clades can be nested, and if they are not nested they do not need to be nearest sisters. Thus these characters impugn the tree found by NJ from the screened data and that constructed by G&A's method from unscreened data. Unlike P1, these characters are clearly nontrivial markers of genetic descent. The fact that NJ found a tree incompatible with them even using screened data is enough to eliminate NJ as a viable method. The fact that G&A's method found such a tree from unscreened data is at least a strong argument that data should be screened when some methods are used.

M5, which encodes the shape of the mediopassive primary person-and-number endings, divides the tree into a portion containing Anatolian, Tocharian, Italic, and Celtic and a portion containing Germanic, Greek, and Indo-Iranian, which must not overlap (though the clades can be nested, and most specialists think they are). There is very wide, though not quite universal, consensus on both the coding of this character and its importance (though each of these aspects has had critics with alternative viewpoints). It impugns all trees found on both datasets except those found by weighted maximum compatibility. In our view this is a clinching argument for the weighting of linguistic characters.

6 Discussion

The main observations we can make on the basis of this study are that these methods *do* differ in ways that are significant to historical linguists and Indo-Europeanists in particular, and that these differences seem to point substantially to the significance (and probable importance) of assigning appropriate weights to different characters. Exactly how to do this is clearly a matter which should be addressed in the historical linguistics research community. Furthermore, since different linguists are likely to assign different weights to different characters, and hence potentially obtain trees that differ in significant ways, these observations also point to the difficulty inherent in recovering the diversification of IE with precision. In particular, while we believe that higher weighting of most of the morphological and phonological characters reflects a general consensus among IEists, this in itself will not resolve all of the remaining disputes related to the history of the IE family. Resolving questions such as whether Greco-Armenian and the Satem Core are true genetic clades of Indo-European history will require additional research.

Our future research will explore the consequences of using a different weighting scheme from the extremely simple (two-valued) weighting scheme we used in this paper. In particular, we will investigate weighting schemes that ensure that characters that survive the screening process have higher weights than characters that are eliminated during that screening. Such weighting schemes are clearly suggested by linguistic scholarship, and using them in a weighted maximum compatibility analysis would likely result in different estimations of evolutionary history than we have obtained using our weighting scheme. In particular, the trees that we obtained using weighted compatibility on the screened and unscreened datasets differ in terms of the location of Greek and Armenian, and consequently differ with respect to the incompatible characters (with the difference primarily being a choice between M10a and P2 and P3). Thus, the selection of one tree over another depends very significantly on the relative confidence one has in the different characters.

Our observations thus strongly support the need for linguists to incorporate into cladistic analyses their own judgments about the relative reliability of different characters. It seems possible that phylogenetic reconstruction methods are best suited to working out, in a maximally rigorous fashion, the consequences of linguists' judgments. Whether they can recover the actual history of a language family's diversification is a separate question. Of course it does not follow that rigorous phylogenetic reconstruction is unimportant; finding the tree(s) which best reflect the judgments of qualified specialists is a computationally difficult problem, so computational techniques are needed. However, it does mean that more work will be needed on the part of linguists to formalize their scholarship so that it becomes amenable to use in rigorous phylogenetic reconstruction. Furthermore, a comparative analysis of trees obtained using datasets (with appropriate weights) constructed by different specialists will help us to determine those aspects of IE history that are reliably reconstructible.

We close with a comment about statistical methods for inferring phylogenies on languages. While all methods can be explored with respect to performance on established datasets, a greater understanding of methods can be obtained if they are also explored on synthetic data generated under sufficiently realistic models of evolution. Our group is working on developing these models (see (Warnow *et al.* , 2004) in particular). One of the benefits of developing good models of evolution is that they may allow us to move beyond the current paradigm inherent in these methods – whereby we either allow characters to evolve under different processes but require the user to specify parameters for the ways in which they vary as input to the program (e.g., Weighted MC), or explicitly assume that all the characters evolve under the same model (e.g., UPGMA, NJ, and the Gray & Atkinson method). An important long term objective would be to develop statistical estimation techniques which can estimate the parameters for each of the different characters from the data. Provided that these techniques are based upon models of language evolution that make sense to historical linguists, they would potentially be able to greatly improve our understanding of the evolutionary processes underlying language evolution, and also allow us to recover the true genetic histories of IE and other language families with greater accuracy than we can currently.

References

Beekes, R.S.P. 1985. *The origins of the Indo-European nominal inflection*. Innsbruck: Innsbrucker Beitrage zur Sprachwissenschaft.

Beekes, R.S.P. 1995. *Comparative Indo-European linguistics: an introduction*. Amsterdam: Benjamins.

Felsenstein, J. 1982. Numerical methods for inferring evolutionary trees. *Quarterly Review of Biology*, **57**, 379–404.

Felsenstein, J. 2003. *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates, Inc.

Gray, Russell D., & Atkinson, Quentin D. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, **426**, 435–439.

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

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

Huelsenbeck, J., & Ronquist, F. *MrBayes: Bayesian inference of phylogeny*. http://morphbank.ebc.uu.se/mrbayes/.

Katz, J.T. 1998. *Topics in Indo-European personal pronouns*. Ph.D. thesis, Harvard University.

Melchert, H.C. 2003. Hittite Nominal Stems in -anzan-. *Pages 129–139 of:* et al., E. Tichy (ed), *Indogermanisches Nomen*. Bremen: Hempen.

Nakhleh, L., Warnow, T., Ringe, D., & Evans, S.N. 2004. http://www.cs.rice.edu/ nakhleh/CPHL/.

Olsen, B.A. 1999. *The noun in Biblical Armenian*. Berlin: Mouton de Gruyter.

Ringe, D., Warnow, T., & Taylor, A. 2002. Indo-European and computational cladistics. *Transactions of the Philological Society*, **100**(1), 59–129.

Saitou, N., & Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.

Swofford, D.L. *PAUP*: Phylogenetic Analysis under Parsimony (and Other Methods)*. version 4.0. Sinauer Associates, Sunderland, Mass.

Swofford, D.L., Olsen, G.J., Waddell, P.J., & Hillis, D.M. 1996. Phylogenetic inference. *Pages 407–514 of:* Hillis, D.M., Mable, B.K., & Moritz, C. (eds), *Molecular Systematics*. Sunderland, Mass.: Sinauer Assoc.

Warnow, T., Evans, S.N., Ringe, D., & Nakhleh, L. 2004. A Stochastic model of language evolution that incorporates homoplasy and borrowing. *In: Phylogenetic Methods and the Prehistory of Languages*.