

Tutorial on Computational Linguistic Phylogeny

Johanna Nichols

University of California, Berkeley

Tandy Warnow*

University of Texas

Abstract

Over the last 10 or more years, there has been a tremendous increase in the use of computational techniques (many of which come directly from biology) for estimating evolutionary histories (i.e., phylogenies) of languages. This tutorial surveys the different methods and different types of linguistic data that have been used to estimate phylogenies, explains the scientific and mathematical foundations of phylogenetic estimation, and presents methodologies for evaluating a phylogeny estimation method.

1. Introduction

Over the last 10 or more years, there has been a tremendous increase in the use of computational techniques (many of which come directly from biology) for estimating evolutionary histories of languages. Despite the enthusiasm with which the popular press has received these studies, it is probably fair to say that much of the community of historical linguists has been skeptical of the claims made by these studies, and perhaps even dubious of the potential for such approaches to be of use.

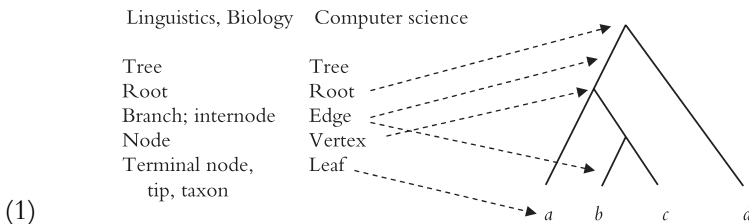
Indeed, while we ourselves are quite positive about the use of computational techniques for phylogenetic estimation in linguistics, it is quite difficult to evaluate phylogeny estimation methods, since the true evolutionary history of any language family is rarely perfectly known. As a result, our position is that any evaluation of computational techniques (or of phylogenetic estimations that are based on such techniques) requires a good understanding of the scientific methodology by which phylogeny estimation methods are evaluated. Explaining this methodology, and showing how it has been applied to phylogeny estimation methods in linguistics, is one of the purposes of this article. In addition, we will survey the attempts to use computational methods in historical linguistics, and explain what can be deduced from these analyses – whether about linguistic evolution or about the phylogeny estimation methods themselves.

A proper evaluation of a phylogenetic analysis also requires a discussion and critique of the data, since these greatly impact the phylogenetic estimation; our discussion will also address these issues.

2. Basics

2.1. TREES

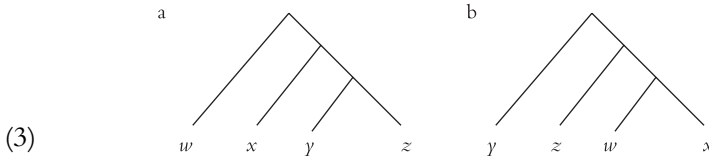
From a graph-theoretic perspective, a ‘tree’ is defined to be a connected acyclic graph consisting of a set of ‘vertices’ (also known as ‘nodes’) and a set of ‘edges’, each of which connects a pair of vertices. The statement that the graph is a connected acyclic graph means that there is exactly one path between every two vertices. Saying that the tree is rooted means that the graph has a distinguished vertex (called the root). In our context, where trees represent the evolutionary history of a set of languages, the root represents the common ancestor of the daughter languages, the languages are represented by the leaves of the tree, the remaining internal nodes of the tree represent intermediate ancestral languages, and the branching order indicated by the tree represents the diversification of the group of languages into subfamilies. A more precise statement is that each language in the input is represented by a path in the tree, representing the language in its different states as it evolved over time. Finally, most phylogenetic trees in biology are assumed to be binary (also called ‘bifurcating’, or ‘fully resolved’), which means that every non-leaf (i.e., non-terminal) node has two daughters (called ‘children’ in the graph-theoretic literature); this assumption is not as strongly held in linguistics, however. Example (1) illustrates these concepts in a diagram of a tree. Note that while linguistics and biology use much the same terms for different parts of trees, these are sometimes different from the terms used in computer science. Also, nodes are also referred to as ‘vertices’, and terminal nodes are also referred to as ‘tips’ or ‘taxa’ (the singular of which is ‘taxon’).



Example (1) is a rooted, or hierarchical, tree in which the top node amounts to the biological or linguistic ancestor (or protolanguage). It shows ancestral nodes for the groups *abcd*, *abc*, and *bc*. An example of an unrooted tree is (2):



Note that this unrooted tree can be rooted on any of the edges in the tree, and hence is compatible with five different rooted trees (one for each edge). In particular, it is compatible with both of the following trees:



Where to root the tree and the ancestral groups this entails are important questions in historical linguistics (the question of whether Indo-European split first into Anatolian vs. the ancestor of the rest is a question of this type). However, most phylogenetic methods do not impose rootings; see Section 2.6.2 below.

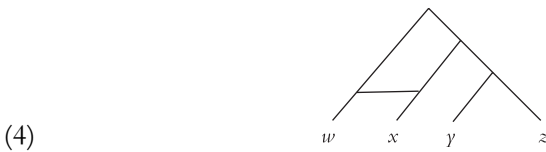
The trees we have drawn here are called ‘binary trees’, since, as rooted trees, each node has two daughters (so the evolutionary history is bifurcating). However, estimated phylogenetic trees are not always binary trees for a number of reasons. For example, phylogenies based on phonological innovations may produce clear subgroups, but be unable to resolve the relationships between the subgroups or the further subgroupings within the subgroups. In these cases, the tree that is proposed for the family may have one or more nodes with many children (which we express by saying the node has ‘high degree’, or more generally by saying the tree is ‘not resolved’). As an example, the root of a widely accepted tree for Indo-European has 10 children, one for each of the established subgroups (Germanic, Italic, Celtic, Balto-Slavic, Indo-Iranian, Albanian, Greek, Armenian, Tocharian, and Anatolian). This tree is almost certainly incorrect if interpreted strictly as indicating a precisely simultaneous radiation, but rather is supposed to be interpreted as being consistent with any of its binary resolutions.

2.2. NETWORKS

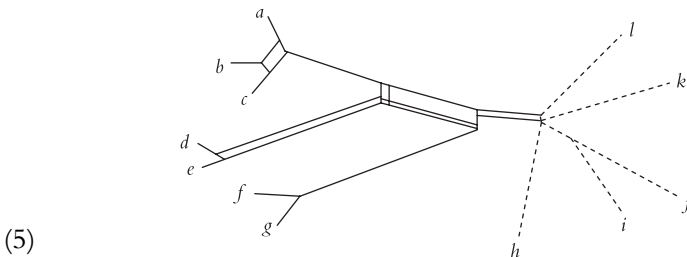
Trees are often reasonable models of evolution, but sometimes a network model is more appropriate. For example, when creolization or language mixture occurs the correct graphical model will contain additional edges between branches in order to indicate the dual parentage. Similarly, when there is borrowing between languages, the proper graphical model will

reflect that borrowing through the addition of contact edges. Such graphical models are called ‘explicit phylogenetic networks’ since they represent an explicit evolutionary scenario. Networks may be drawn as rooted directed graphs, so that each edge has a direction, generally reflecting elapsed time, so that the edge is directed from the parent language to the daughter language. Contact edges may be bidirectional, if both lineages borrowed from each other, or unidirectional if all the borrowing went from one lineage into the other. Note that if these graphs are considered as undirected graphs, then they contain cycles; however, as directed graphs these graphs are considered to be acyclic (since every undirected cycle will contain at least one node with two incoming edges). Thus, while a tree is a very simple kind of phylogenetic network, in general phylogenetic networks are not trees.

Explicit phylogenetic networks have appeared in phylogenetic analyses of Indo-European (Nakhleh et al. 2005a), and have been used in simulation studies (McMahon and McMahon 2005: 111–18; also Barbançon et al. forthcoming). An example of an explicit phylogenetic network is (4), which represents contact between the branches leading to leaves *w* and *x* via the horizontal line between the branches. Note that this graph clearly defines an underlying genetic tree as well as the distinct contact event. Furthermore, we could have directed the contact edge to indicate the directionality of the borrowing.



Other types of phylogenetic network are based on representations of ‘splits’, and produce graphs like (5) (redrawn from the graph for Celtic in McMahon and McMahon 2005: 196; dotted lines *h–l* are other Indo-European branches):



This graph is not acyclic (in that there is more than one way to get from *a* to *c* or *f*, or from *e* to *g*, etc.), and so is not a tree. There are distinct differences between this type of phylogenetic network and the explicit phylogenetic networks given above. This graph does not explicitly indicate

any evolutionary scenario, and instead represents graphically how the input data (distances or characters) do not fit a tree exactly. Thus, the graph represents a combination of tree-like signal and the noise in the data. In particular, the internal nodes of this graph do not represent ancestors of the given languages, but are introduced in order to make possible the representation of the conflict between the different splits that are produced in the data analysis. This makes the interpretation of the graph somewhat difficult – do the parallel lines represent contact events, homoplasy (either independent parallel development or *backmutation*, the reappearance of a state that occurred at an ancestor – see pages 18–20 of <http://www.cs.utexas.edu/users/tandy/newton-linguistics.ppt>), or just deviation from tree-like data due to having an insufficient number of characters? For these reasons, such a graph is known as an implicit network. The difference between implicit and explicit networks is important [and almost all network estimation methods, for example, NeighborNet (Bryant and Moulton 2002, 2004), Network, Splits Tree, etc., produce implicit phylogenetic networks], but most users of phylogenetic networks seem to be unaware of these differences. See, however, Huson and Bryant (2006) for a discussion of these differences, and McMahon and McMahon (2005: Chapter 6) for a comparison between different implicit phylogenetic network methods.

Careful and appropriate modeling of dialect chains is yet another matter, since instead of a few discrete contact events there is an essentially continuous exchange between dialects governed primarily by geographic and temporal proximity. Such scenarios require yet more complicated graphs, which in no real sense resemble phylogenetic trees. The estimation of such graphs is beyond the scope of this article.

2.3. CHARACTERS AND CODING OF CHARACTERS

The input to a phylogenetic analysis is generally a data matrix, where the rows represent the given taxa (in this case languages), and the columns represent different features by which the languages are described. These comparanda are known as characters in biology (so that the data matrix is also called a ‘character matrix’), and variously as features, traits, properties, variables, and probably other terms in linguistics. Linguistic characters can be divided into two kinds. Cognates are individual morphemes consisting of a form and a function or meaning and demonstrably inherited from a unique ancestor, as proven by regular sound correspondences. Examples are English *two* and German *zwei* from Proto-Germanic, or these plus Latin *duo*, Polish *dwa*, Armenian *erku*, etc. from Proto-Indo-European. Note that each cognate set or its ancestral form is a unique or particular individual. It can be inherited, lost, changed, or borrowed, but by definition it cannot recur independently. (An accidentally homophonous word may occur independently, but it is a historically unrelated word and

not a cognate.) Slightly broader but similar to cognates are phyletic characters such as the change of $\star s$ to $\star \zeta$ or the like after $\star r$, $\star k$, $\star u$, or $\star i$ in early Balto-Slavic and Indo-Iranian (branches of Indo-European), known as the *ruki* rule. This is a type of assimilation and might recur in principle, but it is sufficiently specific, even quirky, that it can be presumed to be non-homoplastic, that is, unique in the family.

Very different are typological (in biology, phenetic) characters. These are types and are expected to recur by definition (in biological terms, to be homoplastic). Examples might include glottalized consonants, tone systems, accusative alignment in nouns, dual number, case-number coexistence, OV word order, first person singular pronoun in *m-*. These are all drawn from Haspelmath et al. 2005, where each is attested in languages of more than one family. Typological characters are homoplastic by definition. Cognates are non-homoplastic by definition, but they may undergo phonological and/or morphological changes and/or semantic shifts that recur, that is, are homoplastic. Cognates are morphemes (lexical items, roots, and grammatical morphemes) or larger constructions; typological characters can come from any and all parts of the grammar and lexicon. Biological characters have long been exclusively typological (though this is not a biological term), but molecular genetics makes it possible to define characters that are sufficiently rare and specific as to be unique for all practical purposes.

Lexicostatistics is the cover term for phylogenetic techniques that use the words of languages as characters. Lexicostatistics is usually spoken of as using wordlists (e.g., Kessler 2001), but in fact what it actually uses is lists of glosses, that is, overlaying on each language an etic semantic grid for which the investigator finds the word that surfaces in each semantic cell. The different character states are then the different forms that occur in the same gloss slot in different languages. Commonly used gloss lists are the 100-word and 200-word lists of Swadesh (1952, 1955, see also Lees 1953), but others exist that are adapted to particular language areas [e.g., Matisoff 2000 for Southeast Asia, Alpher and Nash 1999 for Australia, Blust for Oceania (references in Blust 2000)]. An example of a gloss list approach is these Swadesh list fragments from English, French, Russian, and (unrelated) Ingush. Character states, the different values for each feature or variable, are numbered for each gloss. Known cognates are coded as the same character state.

Gloss	English	French	Russian	Ingush
BLOOD	blood (1)	sang (2)	krov' (3)	c'ii (4)
BONE	bone (1)	os (2)	kost' (2?)	t'exk (3)
DOG	dog (1)	chien (2)	sobaka (3)	zhwalii (4)
HEART	heart (1)	coeur (1)	serdce (1)	dog (2)
SUN	sun (1)	soleil (1)	solnce (1)	maax (2)

Some lexicostatistical approaches track not just the character states for glosses but also, and chiefly, which states are cognate and inherited from the protolanguage. (In contrast, what the classic comparative method usually tracks is not glosses and not cognacy of glosses but cognate forms and their development regardless of meaning shifts undergone.)

As examples of types of linguistic characters, a series of studies by Ringe and Warnow and their colleagues (Ringe et al. 2002; Nakhleh et al. 2005a) seek out cognates and diagnostic phyletic characters that are non-homoplastic. Dunn et al. survey exclusively typological characters that are quite homoplastic. Saunders (2005) and Wichmann and Saunders (2007) combine typological and lexical characters.

2.4. CHANGE

Language change is ongoing constantly in all languages. Its cumulative effects over time produce dialect splits, then non-intelligibility, then distant relatedness, and ultimately loss of any evidence of relatedness by descent. Change can affect characters in various ways; for example, the following changes can affect words:

- loss with replacement (by a different word or morpheme);
- partial change (e.g., a word is suffixed, derived, or reanalyzed);
- change at one level that does not affect change at another (e.g., sound change that does not affect morphological structure or the identity of cognates);
- splitting of one word into two separate derivatives; and
- addition of a new word in the same sense or a similar sense to an existing one, creating homophony (a form of polymorphy)

In lexicostatistics, it is normally only loss and replacement that are modeled, that is, the character states are cognates and the various borrowings or innovations that replace them in individual languages. The database for Dyen et al. (1992) codes each lexeme as cognate or not for each pair of languages. Strictly speaking, an unchanged cognate and one changed by derivation probably should be represented as different character changes: for example, English *five* and Latin *quinque* are one character state, while Russian *pjat'*, reflecting the same root plus an originally ordinal suffix **-t-*, is another. Alternatively, this derivation could be entered as a phyletic character in its own right, reflecting an innovation in one branch. Serva and Petroni (2008) take the unusual tack of representing lexical substitutions not as holistic character state changes but as distances measured by sound changes, phoneme by phoneme (or letter by letter) (see Section 5.1.3 below).

2.5. STATISTICAL MODELS OF LANGUAGE EVOLUTION

Statistical models of language evolution are critical to understanding phylogeny estimation methods for many reasons. The most obvious reason is

that some of the newer methods (e.g., the Bayesian methods of Nicholls and Gray 2008 or Gray and Atkinson 2003) explicitly reference a statistical model of language evolution, and use the properties of the model to estimate the evolutionary history. However, even methods that are not explicitly statistical can be studied using simulations, and these simulations have to depend on some explicit statistical model. We therefore begin with a discussion of stochastic processes.

A stochastic process for phylogenetic purposes describes how a set of traits (i.e., characters) evolves within a family of languages. Each trait can assume several states, and changes between the states will occur with some probability on each of the branches of the tree. The probability with which the trait will change its state can depend on the branch, and it can also depend on the character; thus, stochastic processes need not assume that all characters evolve identically, nor that a given character evolves identically on all branches of the tree. Furthermore, although it is generally assumed that characters will evolve independently (so that changes in one character will not impact the probability of change for another character), this is not always the case; the stochastic process will thus make an explicit statement about the independence or lack thereof between characters. Character states can also be borrowed (i.e., transferred from one language to another), and creolization or language mixture can occur, both conditions thus requiring that network models of evolution (rather than tree models) be used. The incidence of homoplasy must also be modeled. Some stochastic models forbid homoplasy (so that all state changes produce new states), while some allow it. Finally, polymorphism (meaning that a character exhibits two or more states in one language; examples are two words, like *rock* and *stone*, for a given meaning) must also be addressed. Thus, in each model, the process by which character states change, the degree to which characters evolve homoplastically, the amount of polymorphism (both in terms of percentage of characters that exhibit it and the number of states that are permitted to simultaneously be present in one language for one character), the amount of borrowing (both in terms of number of contact events and the percentage of characters that evolve with borrowing), and the degree to which different characters are allowed to evolve differently, must be addressed.

We now describe some of the different models that have been proposed in linguistics. The simplest example of a stochastic model is for binary (i.e., two-state) characters (with State 0 representing the absence, and State 1 representing the presence, of some feature). A parametric model of how such characters might evolve would consist of a tree with the leaves labeled by the set of languages, would include a probability for the root to exhibit the cognate (i.e., to have the character state at the root be 1), and with substitution probabilities given for each edge of the tree indicating the probability that the character will change its state on the edge. Thus, the character would evolve down the tree, and would assign

states (either 0 or 1) to each node of the tree. Note therefore that the change on an edge from State 0 to State 1 indicates the appearance of the cognate (but not whether it is the first appearance, or a reappearance), and that a change from State 1 to State 0 indicates the loss of the cognate. A model tree thus consists of a rooted tree (which is generally assumed to be a binary tree), and the numeric parameters (specifically, the probability for each character of being in State 1 at the root, and the substitution probabilities on the edges). Note that in this model, if a character changes twice on some path, then it will begin and end in the same state; thus, this model allows homoplastic events. Constraints on the model so as to prohibit homoplastic changes can be made, however. Note that in this model, the number of parameters does not grow with the number of characters; thus, all characters are supposed to evolve under the same process. However, in some cases, some additional variation between characters is enabled – but is constrained so as to be able to be represented by just one additional parameter for the entire set of characters (i.e., the characters draw a rate of evolution from a distribution which can be described by that single parameter).

Multistate characters can also be modeled in essentially the same way – a single rooted tree, along with substitution probabilities on the edges, and the probability for each state at the root. Here, however, the model needs to specify if the number of states is bounded (e.g., by saying that there are only four possible states, and all substitutions are between these four states) or an unbounded number of states. Here, too, homoplasy can be allowed or prohibited.

These models can be used for simulation purposes (and so used to produce synthetic data that can be given as input to phylogeny estimation methods), and can also be used for phylogeny estimation methods as we now show. Note that the tree and the values for the various parameters define the probability of any given set of sequences at the leaves of the tree. For example, the maximum likelihood estimation method would try to find the tree and parameter values that are most likely to produce the input sequences.

Stochastic models like these have been used in linguistic phylogenetics in several ways. First, Gray and Atkinson estimated the Indo-European history by treating lexical characters as presence/absence of cognates, and performed a Bayesian analysis under the assumption that these binary characters (indicating the presence or absence of a cognate in a language) evolve under the simple two-state model we described above. The homoplasy-free version of the multistate character model has been used for simulation purposes in several studies (see, for example, Minett and Wang 2003; McMahan and McMahan 2005; Wang and Minett 2005). Extensions of these models to allow for borrowing have also been considered, with homoplasy-free versions studied by McMahan and McMahan (2005), Minett and Wang (2003), Wang and Minett (2005), and more

generally by Warnow et al. (2006) (who proposed a parametric model allowing homoplasy) and Barbañon et al. (forthcoming) (who performed a simulation study under the Warnow et al. 2006 model).

Phylogeny estimation methods based on unrealistic models of language evolution are unlikely to produce accurate estimations of evolutionary history, and simulation studies based on unrealistic models of language evolution are unlikely to be informative of performance on real data. Thus, a critical understanding of the models underlying estimation methods and simulation studies is essential to a proper interpretation of phylogenetic analyses and simulation studies.

2.6. PHYLOGENY ESTIMATION METHODS

Phylogeny estimation methods use information about languages to produce an estimate of the evolutionary history of the languages. This information generally takes one of two forms: a character matrix (where the languages are described by a set of characters), or a distance matrix.

As we have discussed above, characters in linguistic analysis are of different types: cognates (which are words or morphemes), phyletic characters, and typological characters; the latter two potentially include characters from all domains of grammar (phonology, morphology, syntax, etc.). The choice of characters for use in a phylogenetic analysis is of great importance, and has often been one of the main issues involved in critiquing a phylogenetic analysis: which characters did the authors use, and what are the consequences of that choice? In addition, the characters must then be coded for each language – a step that is itself critically important, as it is often here where mistakes are made, even by trained linguists (see Eska and Ringe 2004). Indeed, while there can be many problematic issues in a phylogenetic analysis, it is often these two issues – the choice of characters on which to base the phylogenetic analysis, and the coding of these characters for the languages in question – that are the focus of the linguist's attention.

As these issues have been rather thoroughly discussed elsewhere, we will focus our attention here on understanding other aspects of a phylogenetic analysis, namely, the choice of phylogenetic estimation method.

In order to be able to evaluate how well these methods work, we will describe how each method operates, the conditions under which each method is guaranteed to perform well, and the basic assumptions under which each method operates. Finally, in Section 3, we will survey the studies that have used these methods, and the implications of these studies with respect to the reliability of these methods for use in linguistic phylogenetics.

2.6.1. Computational Complexity and Running Time Analyses

One of the important aspects of a computational method is the time it takes to run the method on inputs of different sizes. The usual way of describing this running time is in terms of its worst-case analysis (i.e., the

slowest possible case). For example, the agglomerative clustering technique used in glottochronology has running time described as ‘ $O(n^2)$ ’ (read ‘order n squared’) time, for inputs with n languages. This means that the running time is never more than some constant times n^2 , no matter how big n is. Algorithms that have worst-case running times that are bounded from above by some polynomial in the input size (i.e., that are never larger than some polynomial in their input size) are said to be ‘polynomial time’, while algorithms that have worst-case running times that can be exponential in the input size are said to be ‘exponential time’. Fundamentally, an algorithm that runs in exponential time is likely to take a long time on all but very small inputs (as functions like 10^n grow very quickly), while algorithms that run in polynomial time will still be reasonably fast even on relatively large inputs. Thus, algorithms that are polynomial time are considered fast, and algorithms that are exponential time are considered slow. Note that these running times are based on the worst-case performance, and so do not predict the running time of the method for typical dataset sizes. While the running time analysis of the agglomerative clustering technique is straightforward, some algorithms are difficult to analyze. For example, some methods attempt to solve computational problems, such as finding the tree with the minimum total number of evolutionary changes. For these problems, strategies that examine every possible tree and then return the best one (in this case, the one that has the minimum total number of changes) will be provably optimal, but will take a very long time. Alternative techniques are based on what is called ‘hill-climbing’. These techniques start with an initial tree, then look at small changes to the tree (typically obtained by detaching one part of the tree and reattaching it elsewhere) to see if a better tree can be found. The process terminates when a tree is found that is better than all the previous trees (meaning it has a better score), but where none of the ‘neighboring’ trees are better. Hill-climbing strategies have the attractive property that they can sometimes find good solutions on big datasets, sometimes even relatively quickly, but they are not guaranteed to find the globally optimal solution. Instead, they can get ‘stuck in local optima’ – trees that are better than all the neighboring trees, but not as good as the best tree for the dataset. Techniques for escaping local optima are also part of these heuristics, and generally use randomness to move. Running time analyses of such hill-climbing heuristics are difficult to obtain, and are usually given in terms of empirical experience. To cite one concrete example, Wichmann and Saunders (2007: 391) report a running time of 41 hours on an iMac G4 to apply the Gray and Atkinson Bayesian analysis to a dataset consisting of 12 pairs of sister languages and 17 characters. The samples of the other work cited here are considerably larger in both languages and characters, and would require much longer running times.

Another issue, related to this issue of analyzing the running time of a heuristic, is the computational difficulty of the optimization problem

itself. Some optimization problems are easy to solve efficiently, so that polynomial time algorithms have been developed to answer the optimization problem exactly. An example of such an optimization problem is to find the largest element in a list, a problem that can be trivially solved in polynomial time. Other optimization problems are quite difficult to solve, to the point that no one has yet been able to devise a polynomial time algorithm that will solve the problem correctly. Examples of such problems include the famous ‘Traveling Salesman’ problem – given a map of cities, and the distances between each pair of cities, find the shortest tour that visits every city in the list. The class of NP-hard problems consist of problems of this sort: ones that are provably ‘hard’ in the sense that if any of them could be solved in polynomial time, then all of them could be – and none of them seem to be able to be solved efficiently. Thus, although this is an informal statement, if a problem is proven to be NP-hard, it means that it is unlikely to be solvable in polynomial time. Instead, heuristics (often based on hill-climbing, but frequently including randomization to get out of local optima) are designed to ‘solve’ these problems. It is important, however, to realize that the heuristics may themselves only find local optima rather than global optima, and may also take a very long time to reach good solutions.

2.6.2. Maximum Parsimony and Maximum Compatibility

Maximum parsimony (MP) and maximum compatibility (MC) are related optimization problems.

Maximum Parsimony. Maximum parsimony is an optimization problem for phylogeny estimation, where the objective is to find a tree on which the minimum number of evolutionary changes occurs. Putting this more precisely, the input is a set of n languages described by a set of k characters, so that each language can be represented by its k -tuple of character states. The output is a tree T in which the leaves represent the n languages, and where each internal node is assigned a k -tuple of character states, but the objective is to do this while minimizing the total number of changes across all edges of the tree (where the number of changes on an edge is the number of characters for which the k -tuples at the endpoints are different). While finding the best assignment of k -tuples to the internal nodes of a given fixed tree is computationally easy, finding the best tree in the set of possible trees is a computationally hard problem (technically it is NP-hard: Foulds and Graham 1982). In some analyses, the different characters are assigned positive weights, and the objective is then to find the tree with the minimum total weighted sum of the number of changes over all its edges; this weighted version is again NP-hard, and hard to solve in practice.

Maximum parsimony analyses can be modified to allow the inclusion of characters with missing entries, for example, semantic slots for which

a language has no word. Since MP produces unrooted trees, if the location of the root is also desired, then additional steps must be taken. The usual way to root the tree is to use an outgroup (a language that is related to the rest of the languages in the set, but clearly is not as closely related to any language in the set as they are to each other), which allows the tree to be rooted on the edge leading to the outgroup. For example, including a Germanic language with a group of Balto-Slavic languages allows the tree to be rooted on the edge leading to Germanic. Alternatively, including directed characters (i.e., ones where you know the ancestral state and therefore the direction of the change) in the data matrix can be used to restrict the possible locations of the root to a subset of the tree.

Maximum parsimony analyses are performed using heuristics (most often hill-climbing heuristics combined with some randomization techniques to escape local optima), which are not guaranteed to find optimal solutions, but which seem to perform reasonably well in practice. Software products that perform effective MP analyses include PAUP* (Swofford 1996) and TNT (Goloboff et al. 2003).

Maximum Compatibility. The input to MC is the same as for MP, so that each language is described by a set of k characters and identified with its k -tuple of character states. The objective for MC is to find a tree on which a maximum number of characters evolve without any homoplasy (back mutation or parallel evolution); these characters are said to be compatible on the output tree, and hence the number of such characters is the compatibility score of the tree. When the characters are weighted, then the objective is to find a tree with the total weighted compatibility score (defined to be the sum of the weights of the compatible characters). Like MP, MC is NP-hard (Steel 1992). Also like MP, MC does not return rooted trees, and so outgroups or directed characters are used to help locate the root. Unfortunately, unlike MP, there are no readily available heuristics that are highly accurate in practice.

Consensus Trees: Processing the Output of Maximum Compatibility and Maximum Parsimony Analyses. In the case of both MP and MC, it is frequently the case that several trees are found with the same best score. In this case, the usual practice is to return a consensus tree of some sort. To do this, the trees with the best scores are passed to an appropriate algorithm, which examines the different trees to determine the features that they either all share, or which a majority share. The usual features that are considered in this construction are the splits, or bipartitions, on the leaf set induced by the edges of each tree. As an example, a common consensus tree is the strict consensus tree that has exactly those splits that are present in every one of the input trees. Other consensus trees used commonly in practice include the majority consensus and the greedy consensus, both of which

can be characterized by their splits (the majority consensus contains those edges whose corresponding bipartitions appear in strictly more than half of the input trees, and the greedy consensus is formed by computing the majority consensus and then refining the tree by adding bipartitions from the input trees). By construction, the strict consensus tree is the least resolved, the greedy consensus tree is the most resolved, and the majority consensus is in between these two trees with respect to resolution. Also, however, the greedy consensus tree refines the majority consensus tree, and the majority consensus tree refines the strict consensus tree. All these consensus trees can be computed in polynomial time.

Frequently, the edges of a majority or greedy consensus tree are annotated with the percentage of the input trees containing that bipartition. These edge annotations can be interpreted as the support for each edge (and hence suggest the level of confidence the analysis provides for that particular edge).

Bootstrapping: Another Technique for Estimating Support Values. Sometimes bootstrapping (a statistical resampling technique) is used to produce these edge annotations. Bootstrapping is a technique whereby random datasets are created by randomly picking characters from the input data matrix and these random datasets are then analyzed using the same method as the original phylogenetic analysis. Sometimes bootstrapping is used just to estimate support values for the edges in a phylogenetic tree, where the support for an edge is the fraction of times that bipartition shows up in the bootstrap analysis trees. At other times, the different bootstrap trees are used as input to a consensus method (like the strict consensus, majority consensus, or greedy consensus), which is then annotated with the support estimates.

The way the random matrices are produced is by what is called ‘sampling with replacement’, in which each sample character is drawn randomly from the entire dataset (not the dataset minus characters already drawn). Recall that the rows of the data matrix correspond to languages and the columns represent characters. Sampling with replacement produces a new data matrix with the same number of rows and columns, but where each column is obtained by picking one of the columns in the original matrix at random. Thus, some columns from the original matrix will appear once, some two or more times, and some not at all, in the newly created random ‘bootstrapped’ matrix. Note that in general any two random matrices obtained in this way will be a little different from each other, and hence phylogenies produced using these different bootstrapped matrices will also potentially be a little different. The basic idea behind this technique is that if a bipartition in the estimated tree has high support, then it is likely to be true of the true tree. However, even high bootstrap support on a bipartition does not really indicate the probability that the bipartition is true.

2.6.3. Maximum Likelihood and Bayesian analyses

Two very common methods for phylogenetic analysis in biology are maximum likelihood and Bayesian analyses. These methods are based on explicit parametric models of evolution, and simulation studies of biomolecular evolution have shown that these methods produce very good estimates of the true evolutionary history – provided that the sequences evolve under the same model as is used to estimate the history. An additional advantage that these methods have over others is that they make explicit statements about the assumptions regarding the evolutionary process operating on the data. Within this category, maximum likelihood methods explicitly attempt to find the tree and the associated model parameters so as to maximize the probability of producing the observed data. Bayesian methods, in contrast, do not try to find the specific model parameters, or even the tree. Instead, they try to estimate the probability that each tree is the true tree (and hence they produce not a single tree, but a probability distribution on the set of trees). Bayesian methods used in linguistic phylogenetics include Gray and Atkinson (2003) and Nicholls and Gray (2007).

Bayesian methods used in phylogenetics utilize the following basic algorithmic strategy. The algorithm begins with an initial model tree (i.e., a rooted tree with initial values for each of the associated parameters of evolution). Then, the algorithm follows a ‘random walk’ through ‘model tree space’, at each point computing the probability of the observed sequences being produced by the given model tree. If this probability is higher than the previously computed probability, the move to the new model tree is accepted; if it is lower, the move is accepted with some lower probability. After a ‘burn-in’ period, the random walk is supposed to be in the stationary distribution, and then the algorithm randomly samples from the model tree space that it visits. This collection of model trees is then used to produce a probability distribution on the space of model trees. A standard output of a Bayesian analysis is a consensus tree (usually the majority consensus tree) of the sampled trees. Sometimes, however, the tree appearing the most frequently (called the ‘maximum posterior probability tree’) is returned. The edges of these trees are then often annotated, using the proportion of the sampled trees in which that edge appears. These annotations are interpreted as the support for the edge.

Bayesian analyses can take a long time; indeed, it is not clear how long they need to run in order to reach the stationary distribution (this is an open research question). Furthermore, Bayesian analyses clearly depend very closely on the parametric model of evolution, and their accuracy will depend on the fit between the parametric model used to estimate the evolutionary history and the data. The main advantage that Bayesian analyses have over other statistical methods is that they automatically provide support estimations, whereas maximum likelihood or MP would normally use bootstrapping (which is computationally intensive) to provide support estimates.

2.6.4. Distance-Based Methods

While character-based methods like MP and MC are the most frequently used techniques in linguistic phylogenetics, distance-based methods are also used. These operate in two steps. First, the character matrix is used to compute a distance matrix, for example, by defining the distance between two languages to be the proportion of the character matrix in which the two languages are different, or the Levenshtein distance (defined in Section 5.1.3 below: a distance based on defining costs for each of the possible operations – substitutions, insertions, or deletions of letters – used to transform one word into the other). This distance matrix is then used to construct a tree (see page 18 of the slide presentation at <http://www.cs.utexas.edu/users/tandy/nov26.ppt>).

Methods like these are generally very fast, and can also be quite accurate – but the accuracy of the method depends both on how the distances are computed and on how they are used to obtain the tree.

The issues involved in computing the distance matrix are subtle and mathematically somewhat complicated. The fundamental objective in computing distances is to produce a matrix of pairwise distances that is close to being additive on the true tree, where being ‘additive on the true tree’ means that it is possible to assign lengths to the edges of the true tree so that the distance between any two leaves in the tree exactly equals the sum of the lengths of the edges in the path between the two leaves. Additive distances fit trees perfectly and enable exactly accurate trees to be reconstructed using very simple (and fast) algorithms (Waterman et al. 1977, see also Kim and Warnow 1999; page 18 of the slide presentation at <http://www.cs.utexas.edu/users/tandy/nov26.ppt> shows how an additive matrix defines the tree). The purpose of correcting distances is to account for hidden character state changes (e.g., where a character changes state twice) and so better approximate the actual number of changes that occurred in the evolutionary pathway relating two languages. More generally, simulation studies in biology have suggested that correcting distances to account for hidden character state changes produces improved phylogenetic estimations. An alternative approach is to assume that evolution is clocklike (i.e., the number of character state changes is proportional to time), so that even if the distances are not additive, they will be reasonable estimates of the time that has elapsed since the common ancestor. Under these conditions, even the simple distances used in glottochronology will be sufficient to ensure an accurate tree.

Once the distance matrix is computed, methods like UPGMA (unweighted pair group method with arithmetic mean; Michener and Sokal 1957), the algorithmic technique used in glottochronology to produce a tree from the distance matrix, or neighbor joining (NJ) (Saitou and Nei 1987) are used to produce a tree. Both UPGMA and NJ are agglomerative clustering techniques, but they use different strategies to produce the tree. UPGMA operates by repeatedly joining (as sister

languages) the two languages in the group that have the smallest distance. UPGMA is guaranteed to perform correctly when the input distances are produced from a dataset that has evolved with clocklike evolution; otherwise, it can make a mistake and infer that two languages are siblings when they are not. NJ, however, operates by computing a transformation of the given distance matrix, and selecting as a pair the languages that minimize that transformed distance. Interestingly, this transformation allows NJ to be correct even when the languages do not evolve under a lexical clock!

2.6.5. Phylogenetic Network Methods

Several methods have been developed that produce networks instead of trees. In the context of linguistic phylogeny, NeighborNet (Bryant and Moulton 2004) is the most popular method, but Network (Bandelt et al. 1999), SplitsTree (Huson 1998), and Perfect Phylogenetic Networks (Erdem et al. 2003; Nakhleh et al. 2005a) have also been used. Of these, Perfect Phylogenetic Networks are explicit phylogenetic networks, and the others are implicit phylogenetic networks. We will say more about this subject below.

3. *Evaluating Phylogenetic Estimation Methods*

3.1 OVERVIEW

Many evolutionary trees (phylogenies) have been constructed for Indo-European languages, Bantu languages, Sino-Tibetan, Austronesian, and other families, and often these constructions conflict with each other and/or with traditionally accepted family trees. How is the interested linguist to evaluate the reliability of these trees? [Note that biological and computational terminology refer to hypothesized phylogenetic trees (or networks) as reconstructed or estimated. We use exclusively the latter term here, since in linguistics ‘reconstruction’ is not phylogeny but determination of the ancestral state.]

The accuracy of a phylogeny produced by some estimation method depends on many factors: the method used to analyze the data, the particular choice of languages used (and how sparsely or densely sampled that set is), the number and type of characters used and how these characters are encoded. If all of these aspects are done well (which is itself a non-trivial issue), then it is likely that the phylogeny that is estimated will be meaningful. However, more frequently some aspect of the analysis is handled poorly, and this can call the estimation into question. Thus, faults in a phylogenetic estimation may be due to the estimation method, or perhaps the data, or the encoding of the data; determining where the fault lies is sometimes a difficult task.

But evaluation of phylogeny estimation methods is inherently difficult for another reason: we cannot, in most cases, know the true evolutionary history of any language family beyond some basic features. All of these

issues make the evaluation of phylogeny estimation methods quite difficult. In this section, however, we will examine the various ways that phylogeny estimation methods can be studied.

Phylogeny estimation methods are tools that produce estimations of evolutionary history from input data. Like all tools, their performance needs to be evaluated so that the user can have a sense of the accuracy that they can expect from the tool. Tools for simple problems can generally be evaluated easily; for example, cheap bathroom scales are not generally exactly accurate, but the errors they produce can be estimated (by comparing their estimations to those of better scales). Tools for more complex problems require more sophisticated testing: consider, for example, political polls, or estimates of population sizes. Understanding the errors in these techniques requires rather careful modeling both of the sampling process and of the population, and neither is particularly easy to do correctly. Thus, the estimated error in such techniques itself exhibits error, which is a result of any errors in the modeling. Despite these challenges, because of the fundamental importance of tools to both the sciences and social sciences, significant efforts are made to develop statistical techniques for understanding the error rates and the conditions under which tools produce sufficiently accurate measurements.

Similarly, phylogeny estimation methods are tools that need to be evaluated. Such evaluations need to inform the user about the conditions under which the tools will produce highly accurate estimations, and also to quantify the error rates. However, unlike many other tools (for which the ‘true’ answer can be determined using other techniques and compared to the result obtained by the tool), phylogeny estimation methods are attempts to answer questions about history, and this is something that generally is either very difficult or impossible to know with a high level of accuracy and certitude.

3.2. BENCHMARKS AND MINIMAL EVALUATION STANDARDS

3.2.1. Overview

One approach to evaluating phylogeny estimation is to use well-understood datasets as benchmarks. For instance, the Germanic, Romance, and Slavic branches of Indo-European are well-understood as to subgrouping and dating and have copious and reliable lexical and grammatical data available. Some benchmarks may be of limited utility. For example, Indo-European is very well-understood as to its basic subgroups (Anatolian, Tocharian, Celtic, Italic, Greek, Armenian, Germanic, Indic, Iranian, Baltic, and Slavic), the two higher groupings Balto-Slavic and Indo-Iranian, and its age (dispersal about 5500 years ago), but there is no consensus on its higher structure. For other families, status as a family is well-understood but subgrouping is not understood or is debated. Thus, specialized knowledge is required to decide which datasets can be benchmarks and in what respects. Below in Section 3.2.3 we list some clear cases for convenience.

3.2.2. Criteria

Although even the best datasets often provide only partial information about history, that partial information can be quite adequate for benchmarking, and indeed we posit that the following should be a minimum and essential requirement of any phylogenetic estimation method using data from a well-studied language family:

- **Compatible resolution;** no conflict. The estimated tree may differ from the benchmark subgroupings, but the estimation and benchmark tree must be compatible (in technical terms they should share a common refinement). Importantly, the constructed tree should not mix established subgroups, e.g., placing one Italic language in Germanic.

Note that this criterion only requires that the constructed tree be compatible with the established subgroupings, and that this criterion can be met even if the constructed tree fails to return the established groupings. For example, a tree that has a Germanic clade, but which is completely unresolved within Germanic (and so does not subdivide Germanic into North Germanic and West Germanic) is compatible with West Germanic and North Germanic. That is, undifferentiated groupings are compatible with differentiated ones.

Key to understanding this criterion is the concept of resolution in a tree. A rooted phylogenetic tree in which every node has two children is said to be binary (or fully resolved), and represents a completely determined evolutionary scenario. When this fails, so that some node has three or more children (in which case it is called a polytomy or high-degree node), the interpretation of the tree depends on what is meant by the high degree node. In some cases, the node represents a true radiation, an exception to the usual assumption that diversification is bifurcating. Such nodes are called hard polytomies. In most cases, however, the high degree node represents instead a lack of information, so that the true history (which is possibly bifurcating) is unclear; in this case, the high degree node is called a soft polytomy. Most phylogenetic estimations produce binary trees, and the only time trees with high degree nodes are returned when a consensus tree is used, or when the data are insufficient to provide complete resolution. In particular, polytomies in phylogenetic trees estimated using the comparative method (using, for example, phonological innovations) are all soft. The reason for this is that individual innovations provide information about splits in the tree, but do not forbid other compatible splits, so that the lack of resolution is only an indication of the inability of the data to fully resolve the history. Note therefore that a tree with soft polytomies is asserting that all refinements of the tree (i.e., fully resolved trees that contain all the subgroupings present in the initial tree) are viable candidates for the true evolutionary history. A soft polytomy is consistent with any more resolved analysis of its sub-branches, while a hard one is not. Thus, our minimum requirement for a phylogenetic

estimation method amounts to the requirement that its reconstructed trees must be compatible with the established benchmark trees.

Note that the default polytomy, in this article and in general, is soft. Below we identify some hard ones, and some cases where interpretations as soft and hard have been in conflict.

A second property, desirable but not essential, is:

- **No missing subgroups.** The reconstructed tree should include all established subgroups.

For example, a good estimation method for Indo-European should not fail to return Germanic as a family, and should not fail to separate Baltic from Slavic. (It may, of course, fail to return more controversial subgroupings, such as Italic-Celtic or Greek-Armenian.) Note also that any tree satisfying the second property automatically satisfies the first property, thus, the second property is strictly stronger than the first.

It is important to emphasize that compatibility and recognition of all groups are in some ways independent of decisions as to whether and where to root trees, which is a matter partly of method and partly of interpretation (see Section 2.6.2 above). For instance, many phonological changes have a natural directionality, and these help narrow down the possibilities of rooting.

A third criterial property, for models involving dating, is again essential:

- **Calibration.** A method for dating should be tried out on one or more benchmark datasets, and the reconstructed or computed date must be reasonably close to the established date. If it is not, the established date should be used to calibrate the method (rather than the method's output being announced as a new date for the family).

We close this section with some comments about these three criteria. While it is certainly desirable that a method never fails any of these criteria, any comparison between two methods may have to take the degree of failure into account in order to make a recommendation between two methods. Thus, a method that is very conservative (and hence produces trees that are relatively unresolved) will be likely to satisfy the first criterion of compatibility but fail to recognize established subgroups; on the other hand, a method that is less conservative may succeed in recognizing many established subgroups, but be incompatible with some established subgroup. Therefore, comparisons between methods will be somewhat complicated, and judgments about the relative importance of the criteria will have to be taken into account. These judgments, however, may depend on the linguist's assessment of each violation.

3.2.3. Recommended Benchmark Datasets

For tests of phylogenetic methods, including calibration of dating, the following are young families for which both dating and subgrouping are firm because of written evidence and/or historical records:

- Germanic (about 2000 years for the modern languages, perhaps 2500 including extinct Gothic), Romance (about 2000), and Slavic (about 1500). Three branches of Indo-European, very well-dated on historical grounds and with well-understood subgrouping. The Germanic and Romance families both have grammatically and lexically divergent but recently formed creoles (i.e., creoles with Germanic or Romance lexifier languages and, usually, West African or Austronesian substrate languages) that provide well-known acid tests for methods of subgrouping and (especially) dating. Germanic additionally offers the lexically and grammatically extremely conservative Icelandic and extremely innovative (Romance-influenced) English as milder acid tests. Wikipedia entries for Germanic, Romance, and Slavic languages as accessed March 2008 appear to be in good accord with the received view. The database to Dyen et al. (1992) includes lexical data for several languages of each of these branches.
- Common Turkic (i.e., the Turkic languages minus Chuvash, the sole survivor of the Bulgar branch that is sister to Common Turkic). Well dated on historical grounds; spread across the Eurasian steppe and nearby beginning about 1500 years ago. The daughter languages are usually classified geographically, as northeastern, southeastern, etc., but these are not considered innovation-defined subgroups.
- Chinese (lexical data in the appendix to Minett and Wang 2003).
- Mixe-Zoque (see Cysouw et al. 2006, with bibliography on the history of subgrouping; lexical data in the appendix).

Older groups for which the subgrouping is well-understood and the dating clear from archeological evidence include:

- Austronesian (AN). The family is about 6000 years old on archeological and linguistic grounds (e.g., Blust 1985, 1995; Pawley and Ross 1993). The highest-level and lowest-level groupings are fairly well-understood (including modern and subgroup ancestral dialect continua, for example, Ross 1988, 1996, 1997), but intermediate levels are not always well-understood. The received view of the higher-level structure of the tree is (e.g., Blust 1985, 1995):

AN Formosan subgroups

Malayo-Polynesian:

Western MP: e.g., Philippine and western Indonesian languages;
Chamorro and Palauan are isolate branches at some high level

Central & Eastern MP:

Central

Eastern:

S. Halmahera – W. New Guinea

Oceanic (including Micronesian and Polynesian as subsubgroups)

Whether Malayo-Polynesian is really a discrete stage vs. on a par with the first-order Formosan subgroups is debated (see Ross 2002: 19–20 for a brief review). Zobel (2002) proposes a high-level reorganization of Malayo-Polynesian based on potentially good phyletic characters, but these need more analysis and testing (Ross 2002). (The Wikipedia entry for Malayo-Polynesian as accessed March 2008 incorrectly reports Zobel's analysis as the received view.) Fuller listings of AN languages with the received subgrouping include Ross (1995); for Oceanic, Ross (1988); Haspelmath et al. (2005) lists a number of the languages by intermediate or lower subgroup; Ross and Naess (2007) resolve the question of whether the languages of the Reef and Santa Cruz Islands are Oceanic (deciding in favor) and revise the Oceanic tree accordingly (this revision is correctly reflected as the received view in the Wikipedia entry for Oceanic Languages as accessed March 2008).

- Oceanic (subgroup of Austronesian). This group is very firmly dated at about 4000 years on strong archaeological and linguistic evidence (e.g., Kirch 1997, 2000), and the wide distribution of a number of its discrete subgroups is due to fairly well-understood episodes of migration and colonization. For subgrouping, see Ross (1988, 1996, 1997). That Austronesian, and especially its Oceanic subgroup, presents linguists, geneticists, and archeologists with a near-ideal natural laboratory because of the well-understood age and subgrouping, large number of daughter languages, and good number of well-understood complex sociolinguistic situations, has been mentioned a number of times, since the beginning of serious comparative work on the family; a recent such mention is Gray et al. (2007).
- Indo-European. The date and place of origin are very clear on archeological grounds (e.g., Mallory 1989; Darden 2001; Anthony 2007): an ultimately very wide dispersal from the western steppe not before ca. 5500 years ago, except that the extinct Anatolian group (Hittite, etc.) is likely to have separated slightly earlier.¹ The major subgroups are very well-understood: Celtic, Germanic, Italic/Romance, Balto-Slavic (subgroups: Baltic, Slavic), Indo-Iranian (subgroups: Indic, Iranian), extinct Tocharian, extinct Anatolian; isolate branches Greek, Armenian, and Albanian. Higher-level subgrouping is uncertain and usually left unresolved, except that an initial split into Anatolian vs. all others is linguistically well-supported. The database to Dyen et al. (1992) provides lexical data (Indo-European lexical data arranged by Swadesh gloss and coded for cognacy vs. non-cognacy with each other language. The larger set of characters, including lexical as well as phonological and morphological ones, assembled by Don Ringe and Ann Taylor for 24 Indo-European languages for the papers by Ringe, Warnow, and their colleagues, is available at <http://www.cs.rice.edu/~nakhleh/CPHL/#datasets>. These two databases for Indo-European differ in several ways: the Dyen et al. data are based upon modern languages, while the

Ringe–Taylor data are based upon the earliest well-attested languages in each branch; furthermore, the Dyen et al. data have not been checked carefully with respect to cognate judgments, and it is likely that there are false positives (words noted as cognates that are not true cognates) in that database.

Two other families with some uncertainties in their subgrouping but high-quality datasets available on-line are: (i) Bantu (sub-subgroup of Benue-Kwa, a subgroup of the large and very old Niger-Congo, or Niger-Kordofanian, family); The Comparative Bantu On-line Dictionary has extensive lexical data (www.cbold.ddl.ish-lyon.cnrs.fr/); and (ii) Sino-Tibetan – a large and old family of which Chinese is one branch; The Sino-Tibetan Etymological Dictionary and Thesaurus has extensive lexical data (<http://stedt.berkeley.edu>).

3.3. IMPLICIT OR EXPLICIT ASSUMPTIONS ABOUT THE EVOLUTIONARY PROCESS

Evaluating phylogeny reconstruction methods using our criteria has the benefit of being acceptable to linguists (provided that the established subgroups really do reflect the accepted views, and these views are based on careful linguistic scholarship). However, a drawback of using just the established subgroups within well-studied language families is that most of the time, most (or perhaps all) the methods that are being compared will correctly reproduce the established subgroups. For example, in Nakhleh et al. (2005b), MP, MC (weighted or unweighted), NJ, and the Gray-Atkinson method each produced trees that had all the established subgroups and were thus also compatible with the constraint tree. The only method that was incompatible with the constraint tree was UPGMA – which split Italic and Iranian incorrectly (by putting Latin and Persian respectively into different groups). This study showed that the ability to match established subgroups is a relatively basic capability that can be accomplished by many methods, which can otherwise produce very different solutions. Thus, methods will differ in terms of their phylogenetic estimations, but these will be typically in the first-order subgrouping and within subgroups, rather than in the identification of the subgroups; these differences, while quite important to linguists, will not be reflected in the results of the testing.

Given these limitations, how are we to evaluate phylogeny estimation methods, and – equally importantly – determine whether a tree constructed for a family is accurate?

One very natural way to understand a method, and to help form a judgment about the likelihood that the method will provide accurate solutions, is to try to understand the assumptions made by the method. But this is not always straightforward. Many methods (such as Bayesian methods, or maximum likelihood) rely on explicit parametric models of

language evolution and may have provable performance guarantees under their models. Other methods (such as NJ, UPGMA, MP, or MC) do not rely on an explicit parametric model, but may depend on some implicit assumptions about language evolution. Understanding a method can often require some investigation into the assumptions – whether explicitly stated or not – that underlie the algorithmic design.

3.4. THEORETICAL GUARANTEES

The first, and most obvious, way to evaluate a phylogeny reconstruction method is to consider the conditions under which a method is guaranteed to produce an accurate estimation (i.e., to reproduce the true evolutionary tree, modulo the edges on which no characters change), and then see if it seems reasonable to assume that the data being analyzed will have the required properties. Note that these sufficient conditions may not be necessary in the sense that the method may produce the true tree even when the conditions do not hold. The following are examples of conditions that suffice to guarantee accurate phylogeny estimations for different methods.

- Methods for solving MP and MC (weighted or unweighted) are guaranteed to produce the true tree when characters evolve in such a way that the tree with the minimum number of events is the true tree. One condition under which this second statement holds is when the characters evolve without any homoplasy at all (meaning no back mutation or parallel evolution) and without any borrowing. This condition is highly unrealistic, even for very well-understood languages analyzed by very knowledgeable linguists. However, all guarantees regarding methods for MP and MC apply only if the optimization problems are solved optimally – a significant constraint, as algorithms that are guaranteed to produce optimal solutions are limited to small datasets (generally of about 30 languages). In particular, exhaustive search techniques that evaluate every tree will have to explore $(2n - 5) \times (2n - 7) \times \dots \times 3$ trees, where n is the number of languages, which limits the languages that can be handled to a much smaller number.
- For most models that are considered in phylogenetics, maximum likelihood and Bayesian methods will produce the true tree provided that the true evolutionary process matches the assumed model, the dataset contains enough characters, and the methods are run exactly instead of heuristically (equivalently, run long enough).
- Neighbor joining is guaranteed to produce the true tree when the pairwise distance matrix is sufficiently close to an additive matrix (see the definition and discussion in Section 2.6.4) defining the true tree. Most typically, additive matrices in linguistic phylogenetics define the pairwise distance between two languages to be equal to the number of

character state changes (e.g., word replacements) that occurred in the evolutionary history between the two languages.

- UPGMA, the algorithm used in glottochronology, is guaranteed to produce the true tree when evolution is sufficiently close to clocklike and there are enough characters. In particular, UPGMA is not guaranteed to reconstruct the correct tree on additive matrices defining the true tree unless the true tree branch lengths obey the lexical clock.
- Gray and Atkinson's Bayesian method is guaranteed to be correct (with respect to estimations of the tree topology) when all the binary characters (each based on a single cognate class) evolve identically and independently of each other. The dating aspect of their method is guaranteed to be correct when the data evolve under a lexical clock.

Several points should be made here. First, for most of these statements, one of the required conditions is that there should be enough characters. How many characters are enough? This is complicated, and very little is really known analytically (though see Erdős et al. 1999a,b; Warnow et al. 2001 for some preliminary information). However, experimental results (based on simulation studies) show that the number of characters that are sufficient for the condition to hold is often very large (in the tens of thousands or more; Nakhleh et al. 2001; Moret et al. 2002). Thus, for many datasets it is likely that there are not enough characters to guarantee complete accuracy with high probability.

A second issue is the assumption that the phylogeny estimation method (whether a Bayesian analysis or a technique for maximum likelihood, MP, or MC) be run long enough. How long is 'long enough'? Computational problems such as MP, MC, or maximum likelihood, are all NP-hard, which – as we discussed earlier – means that in practice heuristics (typically based on hill-climbing techniques combined with special techniques to move out of local optima) are used. Unfortunately, there is no reliable way of determining whether a heuristic has been run 'long enough'. In practice, therefore, the methods are run until improvements have not been obtained for some period of time. Thus, the need to run heuristics for NP-hard optimization problems 'long enough' makes sense. But why do we require that Bayesian analyses be run 'long enough', since they are not trying to solve hard optimization problems? The answer is subtle. Bayesian techniques, such as used in phylogenetics, employ Markov Chain Monte Carlo (MCMC) techniques (which are based on randomization) to move through 'model tree space'. These techniques have the theoretical guarantee that if run long enough, the random walk is supposed to have arrived at a stationary distribution, in which further random walking produces only temporary fluctuation, and when this happens the output from the Bayesian analysis will have the correct theoretical properties. When this happens, the maximum posterior probability tree produced by the analysis will be guaranteed to be the true tree. Thus, in theory the MCMC analysis, if

run long enough, will produce the correct tree. How long is long enough? Unfortunately, we do not know. In practice, there are techniques that can be applied to detect that the method has not been run long enough, but these techniques do not allow the user to determine that the method has for sure been run long enough. The problem is a serious one, since it is easy to be mistaken about whether the analysis has run long enough (see, for example, the discussion about this issue on the tutorial page for the MrBayes software at http://mrbayes.csit.fsu.edu/wiki/index.php/Tutorial_3.2). Thus, some methods (Bayesian analyses and heuristics for NP-hard problems like maximum likelihood and MP) cannot be guaranteed to produce correct answers unless run for what may be much longer than any practitioner is willing to perform the analysis. For the Bayesian analyses used in linguistics, it is quite possible that runs of a few days are sufficient (for small enough datasets), but it is also possible that the phylogenetic analysis would be different in interesting and important ways if the analyses were allowed to run for longer. Until these questions are studied carefully, however, caution seems the best policy.

In conclusion, therefore, while there are theoretical guarantees for many methods, often the conditions sufficient for correct estimations do not hold. Then the issue becomes one of ‘given that exact accuracy cannot be guaranteed, how can we predict the performance of a method?’ Or, more simply, if the data do not satisfy the required conditions for any of the available methods, what method should be used?

3.5. SIMULATION STUDIES

A simulation study has the following basic structure:

- Step 1. A model phylogeny T (tree or network) is produced, along with the associated parameters of evolution (e.g., probabilities of substitution, homoplasy, polymorphism, and borrowing).
- Step 2. Characters are evolved from the root of the phylogeny to the leaves, producing a set S of languages at the leaves of the phylogeny.
- Step 3. The set S is given to a phylogeny estimation procedure, and a tree or network T' is produced by the analysis.
- Step 4. The estimated phylogeny T' is compared to the true phylogeny T , and the error is computed.

Key to this procedure is the fact that the true history is known by the person performing the simulation, but not known by the phylogeny estimation procedure (obviously). It is this knowledge that makes it possible to evaluate the accuracy of the phylogeny estimation procedure. However, the simulation study will be relevant to phylogeny estimation only to the extent that the simulation study is based on a sufficiently realistic model of language evolution. Thus, the simulation should reflect, as much as possible, the properties that are observed to be true of real languages.

We begin by discussing the evaluation of phylogenetic tree estimation methods, where the true history is a phylogenetic tree (as opposed to a phylogenetic network). These evaluations are based on comparing the underlying unrooted trees (not the location of the root, and not the associated parameters such as branch lengths) for topological accuracy. Of the various ways of measuring topological accuracy, the Robinson–Foulds (RF) metric (Robinson and Foulds 1981), is the standard measure when both trees are binary (i.e., fully resolved). Here we consider each tree to be represented by the set of edges it contains, and each edge is defined by the way it splits the set of leaves into two parts. The RF metric measures the difference between these two sets, normalized to produce a number between 0 and 1. Thus, if the normalized RF distance between trees T_1 and T_2 is 10%, then this asserts that 90% of the edges of the two trees define the same splits, and they differ only in 10% of their edges (see page 15 of <http://www.cs.utexas.edu/users/tandy/newton-linguistics.ppt> for an example of a calculation of the RF distance between two trees). RF rates above 10% are generally considered poor, but sometimes the data are such that better estimates are not really possible. Some studies have separated out the RF error into two types of errors: false positives (edges that appear in the estimated tree but not in the true tree) and false negatives (edges that are in the true tree but not the estimated tree). These studies allow the two types of error to be distinguished. When both the estimated tree and the true tree are bifurcating, there will be equal numbers of both types of errors; however, as one or both of these trees can be incompletely resolved, distinguishing between the two types of errors can be quite informative. These error rates too are normalized by the number of edges in the respective trees, and error rates above 10% are considered poor.

When either the estimated or true phylogenies have reticulations, and thus are phylogenetic networks instead of trees, the problem of evaluating the accuracy of the estimated evolutionary history is more complicated. However, when the model phylogeny is a network representing borrowing between languages but no creolization or koine formation and the estimated phylogenies are all trees, then it is possible to extend the standard approach for evaluating phylogenies as we now show. In this case, the model phylogenetic network has an underlying genetic tree on top of which there are borrowing edges. In such a case, the estimated trees can be compared to the genetic tree using the standard RF criterion given above. However, evaluating the accuracy of an estimated phylogenetic history under more general conditions (e.g., when the network represents creolization or koine formation) is still a matter of mathematical research, and there are not yet any accepted metrics for comparing two networks.

Some studies have considered repeatability of their results to be evidence of the accuracy of their findings. However, repeatability is not the same thing as accuracy. For example, a method that always produces the

same outcome (independent of input) will be repeatable without being accurate. Thus, while repeatability is desirable to some extent, the accuracy of the method must be established using either theoretical guarantees, simulations, or benchmark datasets.

Finally, we make some comments about the use of statistics to provide evidence in favor of a phylogenetic estimation. Many studies use support values estimated using the bootstrap, or those returned in a Bayesian analysis, or the percentage of best MP trees containing an edge, to suggest that a particular feature of the phylogeny is likely to hold in the true tree. This practice is common in molecular systematics, but the statistical basis of this approach is somewhat questionable. The most appropriate interpretation of support values of this sort is that they indicate the repeatability of the feature, under the hypothetical scenario that there might be a larger family of characters from which to draw from. MP trees are also often evaluated using criteria such as the CI, rescaled consistency index (RCI), or other such statistic. These statistics (which range from 0.0 to 1.0) measure the amount of homoplasy in the data, with 1.0 indicating no homoplasy at all (Naylor and Kraus 1995: 559). Here, too, the interpretation of these statistics is a subtle issue. What these statistics indicate is the quality of the data, in terms of how much homoplasy they reveal. Since MP performs well (in that it produces relatively accurate estimates of the true tree) under conditions of low homoplasy, a MP analysis of a dataset with very little homoplasy (RCI close to 1.0) will potentially be highly accurate, and conversely MP analyses of datasets with a lot of homoplasy (RCI much smaller) may not be particularly accurate.

These statistics are used to validate the trees, but also to lend support to the assertion that the data are evolving in a tree-like fashion. Since the CI and its RCI are the most frequently used statistics, we focus on the use of these two statistics here. Both the CI and RCI measure homoplasy. Note that if there is no homoplasy at all, then no borrowing needs to be posited to explain the data. Thus, when either the CI or RCI is very close to 1.0, the argument that the data are evolving in a tree-like fashion is probably reasonable. However, CIs and RCIs vary significantly between different datasets, even for datasets obtained for the same language family (since this will depend on the encoding and the cognacy judgments).

3.6. DISCUSSION

Phylogeny estimation methods are studied in terms of their accuracy in simulation, on benchmark datasets, or with respect to their theoretical guarantees. However, each of these approaches has its strengths as well as its drawbacks. While theoretical guarantees are appealing, they only apply to rather unrealistic conditions: languages evolve exactly according to the assumed model, and there are a sufficiently large number of characters

available for the analysis. Performance under simulation too has its appeal, because in this case you know the ‘truth’, but is limited as well since simulations depend on models, and the models are always simplifications of reality. Performance on benchmark datasets might seem to be the only real test, but here, unfortunately, the utility of these benchmarks has been limited by their paucity and by the relative lack of resolution available for even the best understood language families.

4. *Special and Advanced Topics*

4.1. DESCENT AND CONTACT; SPLITS AND RETICULATION

All linguists are taught in introductory courses that family trees are an idealization: splits of ancestral languages into descendants are rarely abrupt and clean. Canonical linguistic family trees reflect only descent, that is, transmission of most morphology and basic lexicon from a single ancestor. No language, however, is without borrowed words and grammar from neighboring or influential languages. (The long-established linguistic terms are loanword and borrowing, but copying is becoming more frequent and is more transparent; adoption is occasionally used and is even more satisfactory. The biological equivalent is lateral transfer. The broadest linguistic term for interaction of various kinds is contact or contact-based change.) A descent history or phylogeny with notable amounts of copying is known as reticulate. Traditional linguistic family trees explicitly ignore reticulation, but there is much awareness that there has been contact between branches. Areal linguistics is the study of extensive reticulation between languages and between families in limited geographical areas. Infrequently, under the right sociolinguistic conditions, a language may be so influenced by another that it cannot be described as having a single ancestor. This happens in the case of creoles, which arise in drastic contact situations where a majority shifts rapidly to the language of a minority without adequate access to speech in the latter (the typical outcome is words from one language and grammar and phonology from another). The Caribbean creoles have, broadly speaking, European lexifiers and West African grammar and phonology. Long-term bilingualism together with the right combination of sociolinguistic conditions can produce mixed languages in which different parts of the grammar and/or vocabulary descend from different languages (e.g., Michif, spoken in North Dakota and Canada, in which the verbs, verbal inflection, and verb phrases are Algonquian, while the nouns and noun phrases are French). These distinctive situations, somewhat analogous to hybridization, are hard to capture well in trees or other graphs. (These and similar situations are discussed in Thomason 2001; Thomason and Kaufman 1988.)

A language family can be called a ‘stock’, if it is a maximal clade and both demonstrable and reconstructible (this term from Nichols 1992, 1997).

The oldest well-dated stocks are about 6,000 years old, on joint linguistic and archeological evidence, for example, Semitic, Indo-European, Austro-nesian, and Uto-Aztecan (inscriptional evidence of ancient languages also helps date the first two). A few are likely to be somewhat older: Uralic (Janhunen 2001), Nakh-Daghestanian (Nichols 2007), and Austroasiatic (Blust 1994). The only clear case known so far of a family so old that it can be demonstrated but not reconstructed is Afroasiatic, in which lexical cognates cannot easily be identified, though relatedness can be proven on some highly specific and quirky morphological characters. Given that cognates cannot be firmly identified, subgrouping and ancestral states of words and sounds cannot be known. For these reasons, a family at a higher level than the stock can never serve as a benchmark.

In lexical studies, contact between unrelated or distantly related languages is easily determined when the families are at all well-understood, as non-cognate vocabulary can easily be identified; for example, English *face* is a patent loan of French *face*. Contact between more closely related languages is less immediately apparent; expertise in Germanic philology is required to determine that English *egg* is a borrowing from Norse and not native West Germanic. The major hazard for phylogeny is undetected borrowings. These can arise when borrowing occurs between close sister languages before sound changes occur that identify the clades later descending from those sisters; or when borrowing postdates diagnostic sound changes but the borrowed word happens not to contain any sounds that undergo the diagnostic changes; or when a borrowing is reshaped to 'nativize' it, undoing or adding diagnostic changes (e.g., Grace 1996). Undetected borrowing also occurs when comparative work on languages is in its early stages and the diagnostic sound changes are not well-understood yet. In all of these cases, linguists estimate the likelihood and magnitude of undetected borrowing using various techniques.

For example, Embleton (1986) estimates the extent of undetected borrowing from such things as the number of sisters of each language, the number of neighbors, and indices of similarity (based on typological properties) between the compared languages. Minett and Wang (2003) assume that recurrent non-cognate lexemes are due to borrowing and use statistics to evaluate the probability that characters evolve through borrowing by examining incompatibility patterns on MP trees. In a somewhat similar fashion, Nakhleh et al. (2005a) proposed a model of language evolution whereby all characters evolve without any homoplasy but potentially with borrowing and showed how to use it to analyze the Indo-European family. A mathematically formal parametric model of language evolution with these properties was proposed in Warnow et al. (2006). Hinnebusch's method (1996) of detecting borrowing through lexical skewing was studied in Wang and Minett (2005) and shown to be a reliable indicator of borrowing. McMahan et al. (2005) compare numbers of resemblances in genealogically stable vs. unstable vocabulary to propose that certain

resemblances are better explained by borrowing than by descent. (More detailed discussions of these articles are below.)

The graphical models best suited to represent linguistic evolution in the presence of borrowing are phylogenetic networks rather than trees, but there are no standards for measuring amounts of reticulation and no averages for comparison. As a consequence, commentators differ in whether they consider a given phylogenetic network to be basically reticulate or basically tree-like. A number of studies draw networks (splits graphs) showing considerable reticulation in the central part of the graph (i.e., in the early stages of differentiation) and yet assert that the graph is tree-like: Gray and Atkinson (2003); Forster and Toth (2003); Bryant et al. (2005); McMahon and McMahon (2005: 163) (discussed above); to some extent Greenhill and Gray (2005). Appropriate interpretation of phylogenetic networks is thus still an open problem at this point.

4.2. DISTANCE MEASURES; DATING (ABSOLUTE CHRONOLOGY)

Some phylogenetic methods (largely distance-based methods like NJ and UPGMA, but also maximum likelihood) produce trees that indicate not only subgrouping but also branch lengths. The interpretation of these branch lengths depends on the model assumptions, but in most cases these lengths correspond to the expected number of changes that should occur for a randomly selected character. When evolution is clocklike, however, these also correspond to the elapsed time for that edge. Thus, the algorithmic techniques involved in glottochronology produce estimates of branch length and family ages, because they make the assumption that the evolution is clocklike. All these techniques fundamentally depend on the original character data, and use their various assumptions to produce estimates of elapsed time from character state changes.

Differential Rates of Change. It is a truism of comparative-historical linguistics that high-frequency words are more stable (resisting phonological, morphological, and lexical change) than low-frequency words. Pagel et al. (2007) test lexical replacement in Indo-European and find that the effect is consistent across branches. McMahon and McMahon (2005: 103–11) and Lohr (1999) identify more and less stable vocabulary items. Field (2002) and Haspelmath and Tadmor (n.d.) find cross-linguistic semantic regularities in the propensity of words to be borrowed. Nichols (2006) and Nichols and Nichols (2007) show that the overall lexical type of a language imposes differential propensities to change independent of lexical frequency (e.g., stance verbs are more stable in base-intransitive languages than in base-transitive languages). Apart from these frequency and structural effects, languages differ in overall lexical stability. Lexical replacement often proceeds more rapidly under intense contact (the permeation of English by Romance vocabulary is a classic example). Dialect mixture can produce

a vocabulary containing ultimately cognate vocabulary from different dialect sources, confounding the task of distinguishing cognates from borrowings (Lynch 1999 for southern Vanuatu). Social change can sometimes produce rapid language change, as when the Norman conquest of England disrupted social institutions and unseated West Saxon from its position as standard and interethnic language, allowing dialect variety to emerge, undermining existing stabilizing attitudes, and making possible the rapid transition to the grammatically and lexically innovative Middle English. The movement of Blackfoot, Arapahoan, and Cheyenne away from the eastern mass of Algonquian languages to the Great Plains and into contact with Plains culture seems to have coincided with rapid phonological change in these languages; on the other hand, no similar acceleration of change occurred when some Apachean (Athabaskan) languages and Kiowa (Kiowa-Tanoan) similarly moved onto the plains and/or into the penumbra of Plains culture (for some of these developments, see Goddard 1996; Ives 2003).

Contact and dialect mixture are readily identifiable, and outliers like English are disregarded or downplayed in estimating glottochronological dates. It is more problematic if an entire branch or family shows unusually fast or slow change. The oft-repeated opinion that vocabulary changes rapidly in Australian languages, because a taboo on pronouncing the name of a deceased person drives lexemes out of use is shown to be false by Alpher and Nash (1999). Foley (2007) presents documentary evidence of unusually rapid grammatical and lexical change in the Lower Sepik-Ramu family of Papua New Guinea. Lynch (1999) shows that the languages of southern Vanuatu have undergone much more phonological change than most Oceanic languages; Grace (1996) traces part of the aberrancy to the fact that they make up a language area. It is often said that languages of very small speech communities should be expected to change rapidly, partly because the impact of any idiolect is greater in a smaller society and partly because small populations usually have a good deal of intermarriage and multilingualism. However, there are many languages and language families of small communities whose rates of change since first European contact or since major archeological episodes can be measured well, and whose rates of change are unremarkable compared with those of languages with long written histories: examples include the Algonquian, Athabaskan, Iroquoian, and Kiowa-Tanoan families of North America (some of these discussed just above).

Finally, we note that the mathematical conditions required in order to be able to accurately estimate dates include significant restrictions on how much variability there can be between the different characters' evolutionary processes. In particular, different characters need to speed up or slow down relative to each other in such a manner that if one character evolves twice as quickly as another character on one edge in the tree, then it evolves twice as quickly as that other character on every edge in the tree (Evans et al. 2006). This assumption, called the 'rates-across-sites' assumption,

may not be realistic. In addition, even when the rates-across-sites assumption holds, if the characters draw their rates from a distribution that is not sufficiently simple, then it may still be impossible to estimate dates correctly (Evans and Warnow 2005).

5. *Review of Literature*

In this section, we review a number of recent contributions, attempting to answer the following questions:

- What is the evidence for or against each phylogeny estimation method?
- What are the implications of these studies for taxon sampling and character selection and encoding? and
- What new insights do we have about language evolution, whether of particular families or more generally, as a result of these studies and the new methods they introduce?

5.1. STUDIES EMPLOYING NEW COMPUTATIONAL METHODS

In this section, we discuss studies that employ new methods for phylogenetic analysis of languages, restricting our attention to those methods that produce trees instead of networks and that use ‘traditional’ data (e.g., lexical, phonological, or morphological characters). These methods include the Gray-Atkinson method, the Nicholls and Gray method, methods based on edit distances between words, and Holm’s separation group method.

5.1.1. The Gray-Atkinson Method

This method was introduced in Gray and Atkinson (2003), where it was used on Indo-European languages, and subsequently used to study the Bantu language family (Holden 2002; Holden and Gray 2006; Rexová et al. 2006).

The Gray-Atkinson method uses lexical characters (though it could be used on any type of character) encoded as binary characters. The lexical characters are created using cognate judgments, but then each cognate class for each semantic slot produces one binary character, with ‘present’ indicating that the language contains that cognate, and ‘absent’ indicating that the language does not contain that cognate. These binary characters are then analyzed using a Bayesian MCMC analysis under a model that assumes that the individual binary characters evolve independently and identically under a Markov model. The analysis is performed using the MrBayes software (Huelsenbeck and Ronquist 2001).

This model thus allows characters to be homoplastic (i.e., to evolve several times on the tree) and incorporates polymorphism: because the binary characters for different cognates evolve independently, each

language can exhibit two or more words for any basic semantic slot. As in all Bayesian MCMC analyses, the analysis produces a greedy consensus tree for the set of trees sampled during a random walk through model tree space. The edges of this greedy consensus tree are each annotated with the percentage of the sampled trees that contain the bipartition of taxa represented by the edge.

Evaluating the performance of the Gray-Atkinson method is challenging, since in some cases the analyses were run in such a way as to ensure that the result satisfies all established subgroups for the language family (and hence will satisfy our first two criteria). For example, many of the Bayesian analyses of Indo-European by Gray and Atkinson (see Gray and Atkinson 2003; Atkinson and Gray 2006) had an intermediate step in which the sampled trees were filtered through the use of constraint trees before being used to compute a consensus tree: 'To ensure that the sample was consistent with well-established linguistic relationships, we filtered the 10,000-tree sample using a constraint tree' (Gray and Atkinson 2003: 436). Although Gray and Atkinson explored different constraint trees, even their 'minimum' consensus tree enforced all the established subgroups, as well as a split separating Anatolian and Tocharian from the rest of Indo-European (the difference between the two constraint trees was only in whether they forced Germanic to group with Italic and Celtic). Thus, we cannot use these specific analyses to determine whether their method passes our required compatibility test or our full recognition desideratum for phylogenetic estimation methods.

On the other hand, Atkinson and Gray (2006) also included an analysis of the Swadesh 100-word list in which no filtering of the sampled trees was performed. This analysis succeeds in reproducing all the established subgroups, and thus passes our first two criteria. (It also finds a split between Anatolian and Tocharian and the remainder of Indo-European (IE), and provides moderate support (46%) for Italic-Celtic and the *ruki* group.) Furthermore, Nakhleh et al. (2005b) tested the Gray-Atkinson method on four versions of the IE dataset compiled by Ringe and Taylor (some containing only lexical characters, but others containing lexical, phonological, and morphological characters), and found that the results did satisfy the first two of our criteria. Thus, the Gray-Atkinson method satisfies our first two criteria even without filtering.

However, the main objective of Gray and Atkinson (2003) is the dating of the root of Indo-European, so they are not so concerned with the estimation of the topology (their date is considerably earlier than the established date, so their phylogeny fails our calibration criterion). Furthermore, they assert that reanalyses of the data in which certain aspects of the analysis are varied produce dates in the same range, and from that they infer that their dating estimates are reliable. But it is important to understand that they have established only that their analysis is repeatable, which is not the same thing as accurate. To use an extreme

case as illustration, a method that always produces the same outcome (independent of input) will be repeatable without being accurate.

5.1.2. The Nicholls-Gray Method

Quite similar to Gray and Atkinson's method is the one developed by Geoff Nicholls and Russell Gray (2008). This method uses the same data, that is, binary encodings of the multistate lexical characters, and also subjects the data to a Bayesian MCMC analysis. However, here the statistical model that is used is somewhat different. This analysis uses a single character for each basic meaning, but the state of each character on a given language is the entire set of words for that meaning in that language. This set evolves through the appearance and disappearance of words, and hence explicitly prohibits homoplasy, and permits polymorphism (two or more words for the same basic meaning). Their method, however, cannot handle missing data, and so is constrained to work with characters for which all languages have attestations. They applied this method to the Indo-European family using various subsets of two basic datasets: the lexical dataset of Dyen et al. and the lexical dataset of Ringe et al. (2002). As in Gray and Atkinson (2003), they screened their sampled trees using a constraint tree to ensure that the Indo-European subgroups were reproduced. Hence, we cannot evaluate the performance of this method. However, we note that despite the constraints the relationships between the different subgroups of IE changed according to the dataset used, with some aspects of IE history strongly supported by most analyses, while the support for other aspects changed dramatically depending on the dataset. For example (see Figure 1 in Nicholls and Gray 2008: 13), the rooting of the tree at Anatolian is supported strongly for the Ringe et al. datasets, and not at all for the other datasets; Germanic-Italic is highly supported for some datasets of Dyen et al., but not supported on the other datasets; and Greek-Armenian is supported strongly by some subsets of Ringe et al., but not supported at all by the other analyses. They performed a simulation study (based on the same binary-state model, but enlarged to allow for borrowing between languages), and found that similar differences in topology occurred when analyzing different subsets of synthetic datasets. Interestingly, however, they found that estimates of the date of the root were close across all analyses, and inferred from this that while topological estimations may be faulty, estimating the age of a family could be performed with greater accuracy and reliability. This conjecture, however, needs to be tested.

5.1.3. Analyses Based on Edit Distances between Words

Some of the new methods for phylogenetic analysis use traditional distance-based methods (e.g., UPGMA and NJ) on edit distances between words for the same basic meaning in different languages. Edit distances are costs for replacements of letters by other letters (i.e., sounds by other

sounds), as well as insertions and deletions of letters. Thus, in a sense these distances are an attempt to model the transformation of one word into another by sound changes, without being restricted to uncontroversially regular sound changes. Note also that words that are cognate but have undergone different sound changes will have positive distances between them. Thus, these distances are quite different from the usual application of distance-based methods to the traditional lexical characters, where two words that are cognate will not contribute to the distance between two languages. Edit distances are defined by the cost of each transformation (i.e., the cost of each replacement, as well as the cost of an insertion or a deletion of a letter); thus phylogenetic analyses based on different edit distances may produce different trees.

The most extensive use of edit distances appears in Brown et al. (2008). Here the authors used NJ [as implemented in SplitsTree4; Huson and Bryant 2006) analyses of distances based on the ASJP (Automated Similarity Judgment Program)] to produce trees on Huitoto-Ocaina, Taconan, Chocoran, Mixe-Zoquean, Mayan, Muskogean, Indo-European, and Austro-Asiatic.

The ASJP method had mixed performance in this article, passing both tests for some families, while failing to pass even our minimum compatibility requirement on others. For example, the analysis for Muskogean is incompatible with the Central Muskogean subgroup (based on the expert classification by Haas 1949, 1979), and the analysis for Austro-Asiatic is incompatible with Katuic-Bahnaric (established by Bradley 1994). It is also questionable whether the ASJP analysis of Mixe-Zoquean actually meets our minimum criterion, since the subgrouping of Lowland Mixe, South Highland Mixe, and North Highland Mixe contradicts the subgrouping established by Wichmann as reported in Cysouw et al. (2006). Furthermore, for those families on which the method passes both tests, the sparse sampling of languages for these groups is problematic in that it makes it easier for the analysis to reproduce the subtrees. For example, the Mixe-Zoquean analysis omits many of the languages included in Cysouw et al. (2006), such as Midland Mixe, Tapachultec, and Ayapa Zoque, and the Indo-European analysis omits Greek, Tocharian, Anatolian, and Albanian. In our opinion, therefore, the method is insufficiently accurate to provide confidence in the phylogenetic estimations it might produce.

Serva and Petroni (2008) used these distances with UPGMA on 50 Indo-European languages (excluding Anatolian and Tocharian). Because UPGMA was used, the output is a rooted tree. The tree obtained satisfies our desideratum of no missing groups in that it captures the major daughter branches (Indo-Iranian, Balto-Slavic, Celtic, Germanic, Romance, etc.), but it has some questionable higher-level structure. Technically, the compatibility criterion is inapplicable at higher levels of IE, as there is no consensus on higher-level grouping, but this does not mean that all possible

higher-level structures are equally plausible or would find equal support among Indo-Europeanists. The two subgroupings that would find the most support (recall that Anatolian and Tocharian are not included) would be a radial phylogeny and one in which Celtic (or Italic-Celtic for proponents of that group) branches off very early. Also plausible would be a grouping consisting of Balto-Slavic and Indo-Iranian, or an Armenian-Greek one, or Italic-Celtic. Serva and Petroni's tree contains none of these. Its basic structure is the following (note that 'Other' means a node ancestral to the rest of the languages at that level of structure, and that some lower nodes have been omitted):

- Armenian
- Other: Greek
 - Other: Indo-Iranian
 - Other: Balto-Slavic
 - Other: Albanian
 - Other: Celtic
 - Other: Germanic
 - Other: Romance

The initial subdivision into Armenian vs. all others (i.e., the rooting of this portion of the IE tree) is not advocated in any well-supported claim we are aware of, nor are we aware of support for the other subgroupings. Thus, the higher-level subgrouping is problematic even for a family for which this is not yet agreed. There are also some problematic low-level groupings: the internal subgrouping both of Slavic and Romance violates our compatibility requirement.

Serva and Petroni describe their method as designed to address the surreal concern that 'Cognates are words inferred to have a common historical origin, their identification is often a matter of sensibility and personal knowledge. Therefore, subjectivity plays a relevant role. Furthermore, results are often biased since it is easier for European or American scholars to find out those cognates belonging to western languages. . . .' They seem to believe that each phylogeneticist determines cognacy independently by eyeballing forms or perhaps consulting ordinary defining dictionaries that sometimes give information on etymology. In fact, cognacy judgments are expert decisions based on certain requisite kinds of data and criteria and published in the scientific literature. Phylogenies are normally drawn up with an expert on the language family among the co-authors (e.g., Ringe in Nakhleh et al. 2005a,b, Wichmann in Cysouw et al. 2006) or acknowledged as a consultant (e.g., Robert Blust in Gray and Jordan 2000).

5.1.4. Holm's Separation Group Method

Holm (2000, 2003, 2007) introduced a new method, called the 'separation group method', to deal with some known problems of lexicostatistical analysis. Holm addresses the 'symplesiomorphy trap' (shared archaisms are

hard to distinguish from shared innovations) and the ‘proportionality trap’ (the assumption that a higher proportion of shared cognates indicates a closer genealogical relationship, when in fact later changes can obscure earlier relationships). Holm uses shared retentions rather than shared innovations to do subgrouping, inferring subgroups from the particular etyma retained (and not assuming a constant rate of change). The description of this method, as given in his articles, is difficult to understand, but Cysouw et al. (2006) provide another description that is somewhat more accessible. Cysouw also tested the method on lexical data for Mixe–Zoquean and found that it fails our minimum compatibility requirement (see Section 5.4.3).

5.2. STUDIES EMPLOYING TRADITIONAL COMPUTATIONAL METHODS

Here we discuss the articles that have used traditional methods, that is, weighted or unweighted MP and MC, NJ, and UPGMA, on traditional data (characters based on lexical, phonological, or morphological features) to construct trees on languages.

5.2.1. Maximum Parsimony

Many studies have used MP to perform phylogenetic analyses of linguistic data. For example, several groups have analyzed the Bantu language family using MP on a variety of datasets, with Bastin (1983), Holden (2002), Holden et al. (2005) using only lexical data, and Rexová et al. (2006) also using grammatical data. Indo-European has been studied using lexical data (Nakhleh et al. 2005b; Rexová et al. 2003) and using a combination of types of data (Nakhleh et al. 2005b). Maximum parsimony was also used by Cysouw et al. (2006) to analyze Mixe–Zoquean. In each of these studies, MP passed both of our topological conditions, thus showing that MP is ‘reasonable’.

It is worth noting that though Cysouw et al. stated that the MP analysis of Mixe–Zoquean was incompatible with the subgroupings established by Wichmann (1995), the parsimony analysis actually produced a tree that was a proper refinement of Wichmann’s unresolved constraint tree and so was compatible with it. The performance of the method on the Bantu family is not particularly informative, since very little is established about the subgrouping of Bantu.

Gray and Jordan (2000) used MP to analyze binary-encoded versions of lexical data for Austronesian and produced a unique MP tree with a low consistency index (CI) of 0.25. They claim considerable similarity to the traditional subgrouping (e.g., Blust 1995; Ross 1995). In fact, though, their tree shows several violations of our minimum compatibility constraint: Malay is removed from its close sisters (in the traditional Malayic subgroup; they include Indonesian, Javanese, etc.) and incorrectly placed with languages of southern Borneo, such as Ngaju Dayak and Malagasy (this

latter is correctly placed with its geographically distant close sister); Chamorro and Palauan, probable high-level daughters of Western Malayo-Polynesian on the traditional tree and high-level daughters of Nuclear Malayo-Polynesian on a recent proposed reworking (Zobel 2002), are incorrectly placed in the Oceanic sub-branch; and Fijian, a close sister of the Polynesian subgroup, is removed from Polynesian and incorrectly placed with the Micronesian languages. These are violations of both the traditional classification they cite and its recent proposed reworking (e.g., Ross and Naess 2007; Zobel 2002, discussion by Ross 2002; these recent contributions are reflected in the classifications of Austronesian and its subgroups as accessed on Wikipedia in March 2008).

5.2.2. Maximum Compatibility

The only group to use MC is the Ringe and Warnow team (Ringe 2002; Nakhleh et al. 2005a,b), who use it in a weighted version where the different characters have weights to indicate the relative propensity to homoplasy and/or borrowing. This method depends on the linguist to assign weights to characters, and the output of the method will depend on these weights as well as on the input matrix (whereas other methods will depend only on the input matrix). The method was used to analyze Indo-European in Nakhleh et al. (2005b); Ringe et al. 2002, and passed our first two criteria.

5.2.3. Neighbor Joining

Neighbor joining is a technique that is frequently used in molecular phylogenetics, but rather infrequently in linguistics. However, McMahon and McMahon (2005) and Nakhleh et al. (2005b) included NJ in analyses of Indo-European, and Cysouw et al. (2006) used it on Mixe-Zoquean. The only one of these analyses that failed our first two criteria was the NJ analysis of Holm's distance data, which suggests that the problem might have been the data rather than the analysis. Thus, the method seems to pass our first two criteria. The tree produced by a NJ analysis will also depend upon the technique used to estimate pairwise distances between languages, with theoretical guarantees only when these estimated distances are close to 'additive' (see Section 3.4). Therefore, it is noteworthy that generally no attempt has been made to use distances that are approximately additive – instead, Hamming distances (i.e., the number of characters that differ between two languages) were used in almost all of these NJ analyses (the only exception being Nakhleh et al. 2005b, who used the distance correction provided in Warnow et al. 2006). In conclusion, it seems that NJ performs well, but additional testing is called for.

5.2.4. UPGMA

Unweighted pair group method with arithmetic mean (defined in Section 2.6.4) is the algorithmic component of glottochronology, and although it

was used frequently in the past, it has largely been dropped from the linguistic toolkit due to its reliance on a lexical clock. An evaluation of UPGMA on Indo-European data, and a comparison to other methods, was performed by Nakhleh et al. (2005b), who found that the trees that were obtained were incompatible with two established subgroups (Iranian and Italic). However, the data on which they applied UPGMA included languages selected at different time depths, a condition that is unfavorable to UPGMA. Finally, as discussed earlier, Serva and Petroni also used UPGMA on Levenshtein distances.

5.3. STUDIES USING TYPOLOGICAL CHARACTERS

5.3.1. Dunn et al. 2005

Most studies have used traditional cognate and phyletic data, for which the encoding uses cognate judgments. Dunn et al. (2005) is the first attempt at computational phylogeny using typological characters. They surveyed 125 grammatical characters (phonological, morphological, and syntactic; all were binary yes/no characters) across 16 languages from the Western Oceanic subgroup of Austronesian spoken in the part of Island Melanesia to the northeast of New Guinea (henceforth 'AN languages'), and 15 non-Austronesian (or Papuan) languages (henceforth 'NAN'), not known to be related to each other, from the same general area. In a MP tree, the languages fell into two clusters corresponding to AN and NAN. As there have been cross-family contacts since the first settlement of AN speakers in the area ca. 3500 years ago (the islands have been settled for some 35,000 years, and there is no way to know whether a given NAN language goes back to the first settlers or represents later immigrants; all of them are surely pre-AN in any case), this is a successful demonstration that typological characters can preserve family lines even under much contact.

Dunn et al. assert that their Western Oceanic tree is similar to received trees, but in fact the tree is incompatible with the established subgrouping as determined by Ross (this violation was pointed out by Donohue and Wichmann (2008), as we discuss below). While one interpretation of the study would be to reject MP as a phylogenetic estimation method, the more appropriate interpretation would be to see the data as problematic. There are 125 binary characters, but the best tree has parsimony length that is almost twice as long, indicating that on average each character changes close to twice on the tree, that is, there is a great deal of homoplasy. The analysis of the NAN dataset demonstrates even more homoplasy (on that dataset, characters change on average almost three times). For this kind of data, parsimony analyses may not provide sufficient levels of phylogenetic accuracy. However, this may be inevitable with typological characters whose stability and worldwide frequency are not known.

Further interpretation of this pattern of results has been controversial. Dunn et al. say that their ability to detect relatedness is proved by the success of the AN tree, and that therefore the rather similarly compact

NAN tree is also likely to reflect a ‘phylogenetic signal’, this one from ca. 10,000 years ago. The two trees are similar not only in compactness but in branch length, although the AN languages of their sample represent a transparently related, 3,500-year-old sub-branch, while the NAN languages have not been shown to be related and are not believed to be. Dunn et al. are cautious in their wording (‘If grammatical structures can retain a phylogenetic signal beyond the current temporal ceiling on the reconstruction of language history, then the possibility is opened up of finding relationships between others of the world’s 300 or so existing language families and isolates’ [2075]), but the case for familyhood of the NAN languages in their sample seems too slim for even this claim.

Donohue and Musgrave (2007) note data gaps, language gaps, and non-independent features in Dunn et al.’s survey. Using Dunn et al.’s data, they extract 22 of the features that are significantly different in frequency between the NAN and AN languages of the survey and show that there is considerably more structural dispersal among the NAN languages (13 of the 22 features have 0 or 100% frequency among the AN sample languages, but none of the features do for the NAN languages). They conclude that Dunn et al. should have used some NAN languages from outside the area to try to root the tree, and some more distant AN (but not Western Oceanic) languages to increase the time depth in the AN languages. They believe that the similarities in the NAN languages reflect areal factors rather than descent.

Dunn et al. (2007), in a reply to Donohue and Musgrave (2007), remove 10 of the 27 features identified as problematic by Donohue and Musgrave and use Bayesian analysis to draw trees of their sample languages on this smaller set of features and on the entire set. The resultant trees are very similar (only one subgroup missed), indicating to Dunn et al. that the data problems do not vitiate the calculation of phylogeny (we do not know which 10 characters they chose, or on what grounds). They perform an autocorrelation test to see whether there is appreciable isolation by distance in either set of languages (i.e., the greater the geographical distance between the two members of the pair the fewer their shared features), and find that pairs of AN languages and pairs of NAN languages both show strong isolation by distance, while mixed pairs do not. This supports internal relatedness of the two sets. They add to their sample Äiwoo of the formerly unclassified Reefs-Santa Cruz group east of their survey area. The Reefs-Santa Cruz group has just been shown to be Austronesian (though not Western Oceanic) (Ross and Naess 2007), and Äiwoo patterns strongly with the AN languages of the study.

Dunn et al. do not address Donohue and Musgrave’s question of whether the NAN languages of the survey exhaust some clade or are simply among the members of some larger clade, and they do not address their question of how the time depth between the NAN and AN sets of languages can be so different.

Donohue and Wichmann (2008) also critique Dunn et al., noting that the AN tree found by Dunn et al. actually violates the Ross subgrouping (and so fails our compatibility criterion). In addition, they divide the features into nine that differentiate strongly between NAN and AN languages and 103 that do not, and they perform NeighborNet analyses of the two character sets. They find very striking differences between the phylogenetic networks they obtained: the first does suggest support of the monophyly of the AN languages of the study, while the second intermingles the two sets almost totally (i.e., shows numerous compatibility violations). They show that the topological distance (measured using the RF distance as defined above) between the Dunn et al. tree and the benchmark tree is 7, which is what one would expect between two random trees on the same 16-taxon dataset.

Finally, in an attempt to further determine the ability of this method applied to these data to distinguish between related and (probably) unrelated languages, they add English to the languages. This results in a tree in which English is not clearly distant from the rest of the languages. They add two more distant AN languages and observe a similar response. Adding English to the dataset allows the tree to be rooted at the branch at which the added language attaches. This rooting, however, is clearly incorrect. At a minimum, this demonstrates that the method is unable to accurately analyze datasets with distantly related (or unrelated) languages. [This analysis of the trees obtained by adding the outgroups is our own; the analysis presented in Donohue and Wichmann (2008) attempts to interpret the trees without first rooting the trees where the outgroup attaches.]

5.3.2. Wichmann and Saunders 2007

This study used a smaller set of typological data to compare methods. The comparison is discussed below in Section 5.4.3. Although the study is a test of methods, their results do also show that typological data can produce a recognizable to fairly good approximation of the known phylogenetic relationships and (in the case of NeighborNet) can appropriately reflect the non-tree-like nature of sets of non-sisters. They also tested the phylogenetic signal of typological data by producing a NJ tree using 96 characters from Haspelmath et al. (2005) and 63 native American languages containing a few pairs and small sets of known sisters but overall not known to be related. This analysis produced a tree (their Figure 8) that they observed to be ‘mostly wrong’. They performed a second NJ analysis by adding additional ‘ghost’ characters designed to enforce the known subgroups. This second analysis produced a tree (their Figure 9) that, as intended, reproduced the known subgroups. Wichmann and Saunders then examined those subgroupings that appeared in both trees. This examination produced a few pairings of languages that might be worth further investigation as possible distant sisters. Using expert knowledge to examine one of the pairs (Chiapas Zoque of the Mixe-Zoque family of Mexico and

Mapudungun, an isolate of Chile and Argentina), they found 10 words on a Swadesh 100 word list that they considered ‘similar to the results one would find by comparing, say, English and Farsi’ (399). In fact, though, what they have done is analogous to comparing either English or Farsi to a Proto-Indo-European (PIE) wordlist, which would yield more resemblances than a comparison of two modern languages would. Up to four of their 10 resemblant words involved free parsing or extraction of one or the other piece from a compound to compare to the other language, which would not be done in comparing Indo-European languages. Although we are less sanguine than they are about this particular trial, we agree that the technique is promising in the abstract and very much worth applying to larger datasets with characters better suited to the task.

5.3.3. Saunders 2005

This work applies various methods to Austronesian data, examining the accuracy of the resultant trees by comparison to the established tree for Austronesian, and considering the impact of data choice (lexical alone, or a mixture of lexical and structural) on the resultant trees. We discuss this work in Section 5.4.4.

5.4. STUDIES COMPARING DIFFERENT PHYLOGENETIC ANALYSES

5.4.1. Indo-European Studies

Nakhleh et al. (2005b) compared UPGMA (the agglomerative clustering algorithm involved in glottochronology), NJ, MP and MC (these two in both weighted and unweighted forms), and the Gray and Atkinson method on four different subsets of the Indo-European dataset of 336 characters compiled by Donald Ringe and Ann Taylor. The four datasets differ according to whether they are screened or unscreened and based on lexical characters alone or all the characters (lexical, phonological, and morphological). Nakhleh et al. found that UPGMA did the worst, violating known subgroups (e.g., producing incompatibilities for Italic and Iranian), no matter which dataset was analyzed; the poor performance of UPGMA is, however, perhaps to be expected, since the languages were not sampled at the same time depth (a condition that would be needed for UPGMA to perform well). However, for every other method, and for each of the four datasets, the trees that were produced included all the established groups. The analyses also all shared certain additional subgroupings, such as Greek-Armenian, and internal subgroupings within Germanic, Italic, and Indo-Iranian, and each produced a split between Tocharian and Anatolian and the remainder of Indo-European (i.e., if the tree is rooted with Anatolian as the first daughter, then Tocharian is the second daughter). The placement of Albanian varied widely in the different analyses and so was considered too variable to be used for benchmarking purposes. However, otherwise the trees were quite different; in particular,

where Italic and Celtic were located in the tree changed according to the choice of data and the method, forming a clade with Indo-Iranian and Balto-Slavic for all analyses except when using weighted parsimony or weighted compatibility with one of the morphological characters (the character encoding the shape of the mediopassive primary person-and-number endings) given sufficiently high weight.

Rexová et al. (2003) also explored the impact of character encoding, namely, whether to use the original multistate character encoding, a modified version of the multistate encoding, or a binary encoding of these characters. They used the 200 lexical characters in the Dyen, Kruskal, and Black database. They analyzed each dataset using MP and then computed greedy consensus trees, annotating each edge with the percentage of the MP trees exhibiting the same bipartition (edges without such annotations presumably appear in fewer than 50% of the trees). Their study shows that trees estimated on multistate and binary-encoded versions of the multistate characters differ substantially, and in important ways. For example, the analysis of the multistate character dataset finds the *ruki* group (the branches that have undergone the *ruki* rule, defined in Section 2.3), but the analysis of the binary-encoded version of that data is incompatible with the *ruki* group. The main conclusion of their study is that the binary-encoded data matrix produces trees that are less reliable than analyses based on the original multistate data.

There are some problems with this finding, however. They comment that ‘all the unrooted trees agree that there are four supergroups of IE languages (Balto-Slavonic, Romano-Germano-Celtic, Armenian-Greek, and Indo-Iranian).’ However, these supergroups do not appear in all the trees, but only in some. What is true is that if the majority consensus instead of the greedy consensus had been used to represent the results of each parsimony analysis, then these supergroups would be compatible with all the trees (because the majority consensus trees would be unresolved, and these supergroups could be obtained by refining the majority consensus trees in some way). To see this, it is only necessary to contract all the edges in the trees in their Figure 1 that are not annotated with support values; thus, keeping only the edges with support values that are greater than 50%.

Holm (2000, 2003, 2007) makes good points regarding the use of methods that require a lexical clock in order to be reliable, and also points out the difficulties in obtaining globally optimal solutions to maximum likelihood and MP. Holm emphasizes the need to use characters for which one knows which state is innovative, and dismisses analyses based on methods that fail to have provable performance guarantees. Thus, he rejects both MP and maximum likelihood analyses of datasets beyond 20 taxa, because global optima cannot be provably found, and he rejects any distance-based method that relies on a lexical clock. While we find his desire for rigor admirable, we think that it is too rigid a position;

understanding the performance of an algorithm is a complicated issue, as we show in this article. Furthermore, the assumption that methods will fail absolutely unless they are perfectly designed for the problem is also misleading. As we have seen, even mediocre methods can reproduce much of the true tree (see also our discussion below of Cysouw et al. 2006, applying Holm's method to other data).

5.4.2. Bantu

Marten (2006) compared several analyses of the Bantu family, a group for which there is much dispute about the subgrouping and which is likely to exhibit a fair amount of dialect interaction and hence be not very tree-like. He focused on two recent studies (Holden 2002; Holden and Gray 2006) and compared them to earlier studies (Guthrie 1967–1971; Bastin et al. 1999), finding significant differences between the phylogenetic trees, and observing that the authors of the studies draw very different conclusions about the feasibility of inferring a tree for the family. Marten comments on several methodological issues, pointing out the potential for language selection (called 'taxonomic sampling' in the molecular systematics field) to impact the phylogenetic analysis, especially with respect to dialect chains and other dialect interaction. His general conclusion is that linguistic phylogenetics is problematic and that the interpretation of an analysis is not at all straightforward. To quote Marten,

Bastin et al. deny the possibility of constructing any one family tree for Bantu in principle, while Holden (2002) proposes something a version of a tree which is at least definite enough to be used to support historical hypotheses. The results reported in Holden and Gray (this volume) seem to come closer to Bastin et al.'s results, in that several networks are presented. Yet Holden and Gray's conclusions with respect to the possibility of developing a sub-classification of Bantu with the aid of statistical methods are still more optimistic than those of Bastin et al. Wherever the truth of the matter lies, this difference, to me, indicates some caution about the use of quantitative methods in linguistics . . . It seems, thus, that quantitative methods are a highly useful and welcome addition for the study of linguistic relationship, and improve our methodological toolkit, but that the application of the method, and the interpretation of its results remains messy and subject to a certain degree of subjectivity. (2006: 50–51)

5.4.3. Mixe-Zoque

Cysouw et al. (2006) apply the method of Holm 2000 (see also Holm 2003, 2007) to the Mixe-Zoque family of Central America. This is a fairly shallow family whose subgrouping is well-understood, having been carefully done independently and on independently collected data by Wichmann (1995) and Kaufman and Justeson (2004) (whose trees differ only minimally). Holm's method has two steps: in the first step, it produces a distance matrix (called the 'N values'), and then from this an evolutionary

scenario can be obtained. Cysouw et al. apply various tree-building methods as well as Holm's narrative interpretation, and all give what they consider similar results that are different from the received view; this means that Holm's N values and not the interpretation of them is what is at fault. In particular, the application of Holm's method to this dataset yields a tree very different from the received one (with violations of the compatibility criterion, one missing subgroup, and formation of one sub-subgroup that is not in the established tree at all; Cysouw et al. consider the latter error minor, p. 240). Apparently, this is due to two problems. First, use of protoforms 'creates problems of circularity that cannot be overcome'. Wichmann was careful not to reconstruct as Proto-Mixe-Zoquean (PMZ) words found in one branch plus only one language of the other branch, especially when that language was in a contact situation; Holm's reconstruction of IE protoforms does not take this precaution. In Wichmann's approach, every protoroot survives in both M and Z by definition, so Holm's method applied to Wichmann's data cannot possibly detect the M and Z major nodes. Cysouw et al. use only PMZ roots; if they used lower-level protoroots (PM roots, PZ roots), they would be building in tree structure by definition, and this would amount to classifying based on shared innovations. Second, the method is sensitive to the amount of lexical documentation available for the daughter languages.

Cysouw et al. find that other methods (NJ, MP, and Split Decomposition) return more accurate estimations of the history of Mixe-Zoquean. They used NJ (called the Fitch algorithm in the article) and MP for tree estimation, and Split Decomposition to produce a network. The NJ and MP trees are both compatible with the constraint tree and reproduce all the established groups, but they also introduce additional subgroups (thus passing both our tests). Cysouw et al. interpret the two trees differently, finding the additional subgroup produced by the Fitch algorithm acceptable but the additional subgroup produced by MP an error; we do not understand why this additional subgroup is considered an error. Their evaluation of the Split Decomposition network is that it is largely in agreement with the constraint tree.

Cysouw et al. conclude that Holm's method would work if etymological dictionaries and information about protolanguages were perfect, but in practice it does not work. They also conclude that although Holm's criticisms of lexicostatistics are correct, in practice lexicostatistics yielded 'surprisingly good' results (254).

5.4.4. Austronesian

Saunders (2005) studied several methods (NJ, MP, the Gray-Atkinson method, and NeighborNet) on a combination of lexical and typological data for Austronesian. He used the Austronesian lexical database of Greenhill et al. (2003–2008), which contains word list items and expert cognate judgments. [Saunders needed to make a few of the judgments himself

when they were lacking in the 2005 version he used. He says cognates were mostly ‘fairly obvious’, which is correct in the sense that it is not difficult to see whether a word in one language is the same Austronesian word as those in the others. There is always the possibility of some words like English *egg*, mentioned above (Section 4.1), where the word in some language is clearly Austronesian but only an expert can tell that it is not directly inherited but borrowed from a sister language. These should not be counted as cognates.] These lexical characters were then binary-encoded and combined with structural characters drawn from the World Atlas of Language Structures database. There were two versions of the structural character matrix: one multistate and one binary.

Saunders performed a statistical test to see if the phylogenetic signals in the structural and lexical datasets are compatible (meaning that they could reflect the same tree), and he found that they did not (2005: 50). Despite this incompatibility, Saunders constructed trees using the various phylogenetic estimation methods on the five possible versions of the data: lexical alone, structural alone (in its two forms), or lexical plus structural (in the two forms). The trees he computed were then compared to the traditional tree (Blust 1978; the same as our benchmark tree, Section 3.2.3 above). To evaluate accuracy, he computed the RF distance between the estimated trees and the Blust tree. He found that the Gray-Atkinson tree obtained by analyzing the lexical data (with or without any structural data added) and the NJ tree on the lexical data (or lexical plus binary encodings of the structural data) had the smallest distance to the Blust tree, and that MP was somewhat worse. On the basis of this (and some other observations), he recommended the use of the Gray-Atkinson method.

This is an interesting study, but we have several concerns. First, and importantly, the use of lexical data encoded as binary characters, and not in its multistate form, is problematic and makes it difficult to assess the performance of the phylogeny estimation methods alone. For example, trees computed using NJ on multistate character matrices might be different from trees computed using NJ on the derived binary-encoded character matrices, and potentially more accurate (see, for example, Rexová et al. 2003, where it is shown that trees based on binary encoded characters were in general less accurate than trees based on the original multistate characters). Second, the use of the RF distance to evaluate trees is also problematic: as was pointed out by Rannala et al. (1998), the RF distance is biased in favor of unresolved trees, which is what the Gray and Atkinson method produced. More significantly, the RF metric is inappropriate when the true tree (in this case, the benchmark tree produced by Blust’s analysis) is not binary. Third, we would consider any tree that resolves the benchmark tree to be accurate, but the RF distance counts any additional resolutions beyond what are in the benchmark tree. This has the unfortunate consequence that it is possible for an estimated tree that we would consider to have passed our first two tests to have a high RF error, while

another tree that we would consider to violate the compatibility constraint might have a lower RF error.

Most importantly, one of Saunders's main conclusions is that Bayesian analysis of combined lexical and typological data adds resolution compared with a tree drawn on only lexical data. We find that his trees for lexical and combined data are nearly identical in their ability to capture the low-level groupings, but the resolution in the combined tree is incompatible with the established tree as regards the placement of the Fijian-Hawaiian-Maori subgroup, the Malagasy-Maanyan-Yakan-Toba Batak subgroup, and *Tukang Besi*, and in its failure to detect the *Kapampangan-Bontok* subgroup (compare his Fig. 12, p. 39 with his Fig. 8, p. 32).

5.4.5. Native American Languages

Wichmann and Saunders (2007) report on a study in which they used four methods (NeighborNet, MP, Gray and Atkinson's Bayesian Method, and NJ) on a small sample dataset of six pairs of known sister languages with no known higher relatedness, and the 12 best-represented and genealogically most stable typological characters in Haspelmath et al. (2005). The article contains a lucid, brief, non-technical discussion of each method and how it works, and gives URLs and/or bibliographical information for each. It includes a brief comparison of biological and linguistic methods.

We begin with a discussion of the NJ, NeighborNet, and Bayesian analyses. Their NJ analysis of the six pairs of languages produces five of the six pairings, and is incompatible (though with low bootstrap support) with *Rama+Ika*; thus, NJ fails each of our two topological tests, but perhaps not badly. NeighborNet's output (Figure 3) can be interpreted as strongly supporting only one pairing (*Slave + Navajo*), showing somewhat weak support for *Lealao Chinantec + Chalcatongo Mixtec*, but no real support for the other pairings (and also representing the relationships between the six pairs of languages as highly reticulate). Thus, NeighborNet's output can be considered as (possibly) meeting our minimum criterion once it is processed, and recovering fewer of the pairings than any of the other methods. The Bayesian analysis using the Gray and Atkinson method gives a tree that is fairly unresolved, but gives three of the pairings and is compatible with the remaining three. Thus, Gray and Atkinson's method passes our first criterion, but fails the second – and fails it more clearly than NJ.

The MP study is more complicated. They give a discussion of MP in two places and provide two figures for the MP tree: one in Figure 4 (page 26) and one in Figure 6 (page 34). (They also varied the MP analysis to allow for different weights on different characters and different costs for changing between states; neither technique made any difference in MP tree topologies.) These two trees are very different: the tree in their Figure 4 is clearly poor, as it contains splits that are incompatible with the

established pairings, while the tree in their Figure 6 not only is compatible with the established pairings but in fact recognizes three of the six groupings. It turns out (Wichmann, personal communication) that their first tree is just one of the many MP trees, and the second tree is the majority consensus tree of the set of MP trees. By convention, the majority consensus is the correct tree to use, but the authors use the tree in Figure 4 instead in order to evaluate MP. As a result, they conclude that MP performs unacceptably. However, if the majority tree had been used to evaluate MP, then the opposite conclusion would have been reached.

A comparison between the various methods is interesting. MP and Gray-Atkinson were the only methods that were compatible with the established groupings, and each recovered the same number of pairings (in fact, they each recovered the same three pairings). NeighborNet's output can be considered as (possibly) meeting the minimum criterion once it is processed, but it recovers fewer of the pairings than any of the other methods. NJ's performance is more complicated: it is incompatible with one of the established groupings, but with low bootstrap support. Indeed, it seems that if the majority consensus tree of the bootstrapped NJ analyses had been used to represent the result, it would have been much less resolved – and would have passed the compatibility test, but lost one of the correct pairings.

The authors conclude that more precise testing will require larger and better-filled datasets, and that multistate characters are more informative than binary ones. They also conclude that typological characters may be useful for detecting unsuspected relatedness; see Section 5.3.2.

5.4.6. Simulated Data

Barbançon et al. (forthcoming) explored the performance of MP and MC (weighted and unweighted), NJ, UPGMA, and the Gray-Atkinson method on simulated data. In this study, they generated model phylogenetic networks (i.e., genetic trees with borrowing between lineages) and evolved characters down the networks, thus producing synthetic character matrices that could be analyzed using the different phylogeny estimation methods. By varying the number of borrowing edges (contact events), the number of characters, and the amount of homoplasy, they were able to determine the conditions under which the different methods performed well or badly. Since each method produces trees, they compared the estimated trees to the genetic tree contained inside each network for topological accuracy. They found that the two distance-based methods (NJ and UPGMA) did worse than the character-based methods, with UPGMA not as accurate as NJ, and that using weighted maximum with higher weight given to characters that are more resistant to borrowing and homoplasy improves the phylogenetic accuracy in some circumstances. The performance of the Gray-Atkinson method was in between that of the two distance-based methods and MP, and generally quite close to MP.

5.5. STUDIES ON DETECTING BORROWING

Phylogenetic analyses are increasingly being performed through the use of algorithms that produce networks rather than trees, undoubtedly because of the realization that linguistic evolution is rarely ‘clean’, and that correct detection (and representation) of contact between lineages is desirable. In addition, methods designed just to detect borrowing, without also estimating the genetic tree, have also been developed. In this section, we describe studies that have used these methods on linguistic datasets or studied them in simulation.

In earlier sections, we have discussed studies that used phylogenetic network methods (most often NeighborNet) to estimate phylogenetic histories on different linguistic groups. In this section, we focus our attention on those studies that were not discussed in earlier sections. We discuss the analysis by Forster and Toth of Indo-European languages, the analysis by Holden and Gray of Bantu languages, several studies by McMahon and colleagues, two papers by Minett and Wang, and a paper by Nakleh et al.

5.5.1. Forster and Toth

Forster and Toth (2003) survey 39 lexical and typological characters across 14 languages and use network to estimate a phylogenetic network. The network they construct is in fact unimportant to their paper, whose main point is to estimate dates, and the network is used only to enable Forster and Toth to produce a tree on which to estimate the dates. This tree has an initial split into Continental vs. Insular Celtic, as well as egregiously early dates for the Celtic split and for Proto-Indo-European differentiation in Europe. The findings are unreliable because of problems with language classification, data, and dating method (see Eska and Ringe 2004).

5.5.2. Holden and Gray

Holden and Gray (2006) analyzed lexical data from 93 Bantu languages (including two non-Bantu Bantoid languages in order to root the trees) using the Gray-Atkinson method and also using NeighborNet. Their analysis using the Gray-Atkinson method produces a Bayesian sample of 200 trees, which they represent in two forms: a greedy consensus tree (with the percentage of the trees exhibiting each edge annotating the edges of the tree) and also a consensus network (as computed using the method of Holland and Moulton 2003). While Southeast Bantu appears in all the sampled trees, and a group containing Central Bantu and East Bantu also appears in all the sampled trees, many groups appear in fewer than half the sampled trees: Central Bantu appears in only 27% of the trees, East Bantu appears in only 18% of the sampled trees, while Southeast Africa, East Africa, and West Bantu appear as a group in somewhat less than half the sampled trees. Holden and Gray analyze the lexical data again using NeighborNet. This analysis produces a highly reticulate network

(their Figure 2.6) that is then modified by elimination of 'short edges' to produce a more tree-like network (their Figure 2.8). Their discussion of the two networks indicates that they find the result to be clearly indicative of support for several groupings, including Central Bantu, West Bantu, Southwest Bantu, East Africa, Southeast Africa, and East Bantu. However, none of these groupings are unconflicted, and very few seem well-supported. An examination of their two networks suggests that East Africa and Southwest Bantu are somewhat well-supported, and perhaps also Southeast Africa (though this is less certain), but West Bantu seems only weakly supported and Central Bantu is not so clear. Interpretation of these networks is, clearly, difficult.

5.5.3. Minett and Wang

Minett and Wang (2003) and Wang and Minett (2005) propose two methods for detecting borrowing. The first method (Minett and Wang 2003) is based on the assumption that phylogenies estimated using distances would reflect borrowing by producing negative branch lengths (so as to reflect the fact that two languages that have evolved in contact will be closer together than would be predicted by purely genetic descent). They studied this method in simulation under a model in which characters evolve without any homoplasy, but with borrowing. Their study showed that the method was able to detect borrowing, and so had a low false-negative rate, but that it also had a high false-positive rate – and so detected borrowing under conditions in which no borrowing occurred. On the basis of the high false-positive rate, they do not recommend the method. They briefly discussed the lexical skewing method of Hinnebusch (1996) and argued that because of its similarity to the method they studied, it would likely also have a high false-positive rate. This assumption, however, turned out to be false, as they show in Wang and Minett (2005), where Hinnebusch's lexical skewing method provided reasonably good results in simulation. They then proposed a new character-based method, as follows. First, they analyze their linguistic dataset using MP, thus producing the set of all the most parsimonious trees for the dataset. These trees are then treated as candidates for the genetic tree for the language family. For each of these candidates, they identify the characters that are incompatible (meaning that they cannot evolve down the tree without homoplasy). This enables them to compute, for each of the characters in the dataset, the percentage of the candidate trees on which the character is incompatible. By their assumption, all incompatibilities are due to borrowing. Therefore, they also attempt to add contact edges to each candidate tree in order to derive the character on the resultant network without homoplasy (i.e., each character evolves down the network via a combination of genetic descent and borrowing, but without homoplasy). They then attempt to determine which of the characters are most likely truly borrowed by using a statistical technique to see if the null hypothesis that the MP trees are

random and that a given character is incompatible on a randomly selected MP tree can be ruled out. They apply this method to a collection of Chinese dialects and are able to rule out the null hypothesis for certain lexical characters ('feather', 'small', and 'what'); on the basis of this, they infer that these lexical characters probably did evolve with borrowing. They conclude that this new technique is promising for detecting borrowing. The method identifies which dialects are involved in borrowing but cannot determine which of them is source and which is borrower.

These two papers make good contributions to understanding the performance of methods for detecting borrowing. The 2003 paper definitively shows the weakness of the distance-based method for detecting borrowing based on negative branch lengths, while the 2005 paper shows some initial promise for Hinnebusch's method of lexical skewing. This result would need to be followed up with a more extensive study to see how well Hinnebusch's method works when the datasets are larger, have greater deviations from a lexical clock, or evolve with homoplasy, since any of these conditions is likely to increase the false-positive error. However, it is very good initial work, and there is very little work in this area. Furthermore, Wang and Minett have also introduced a very natural character-based approach for detecting borrowing. However, there is a problem with the interpretation of the statistical test they applied. While they show clearly that the null hypothesis for the characters 'feather', 'small', and 'what' can be rejected, the conclusion that these characters evolved with borrowing does not follow. We can only conclude that the null hypothesis is false. But the null hypothesis has two parts, the first being that the MP trees are random, and we know that that is false. Therefore, the null hypothesis is easy to reject, and unfortunately no inferences can be made about whether the characters are borrowed.

Interestingly, a different argument can be made that would suggest that these characters are borrowed. Suppose, for example, that we are willing to believe that the true tree is one of the MP trees that was obtained. In this case, if a character is incompatible on every one of the MP trees, it will in particular be incompatible with the true tree. Then, if we also assume that all incompatibilities are due to borrowing, we can infer that any character that is incompatible on every MP tree must have been borrowed. These two assumptions (that the true tree is one of the MP trees, and that all incompatibilities are due to borrowing) would still need to be carefully considered.

5.5.4. McMahon et al.

McMahon et al. (2005) shed light on a long-standing question of language classification with a lexicostatistical study of the Quechuan and Aymaran families of Andean South America. Using Lohr (1999) and additional Andean data collected by Heggarty, they compile word lists of the 30 words shown to be genealogically most stable and the 30 least stable across

language families. (Lohr's glosses were taken from Dyen et al. 1992 and tested on Indo-European, Austronesian, Sino-Tibetan, and Afroasiatic. High-level Afroasiatic comparisons are problematic as mentioned above in Section 4.1; within subgroups, however, cognacy can be established, so Lohr's findings can be regarded as reliable though in need of further testing.) NeighborNet analyses of these two sets differ greatly (p. 165): the one on the more stable words groups the Quechuan and Aymaran languages as quite discrete and quite distant from each other with almost no reticulation between the two groups, while the unstable words show considerable reticulation between the two groups and less distance between the two than within one of them. This shows that the resemblances between the two families are probably due to contact (specifically borrowing in the case of resemblant words). Since it is not known in advance whether Quechuan and Aymaran are related, words shared between the two cannot be considered cognates, and the coding standards term them correlates for purposes of assessing character states.

McMahon and McMahon (2005) report this study and others in their Chapter 6 (139–75). They produce a phylogenetic network using NeighborNet on the full data of Dyen et al. (1992: 164) and pronounce it 'clearly a tree, though there are some obvious reticulations too' (163). Since a tree, by definition, cannot contain cycles (as are produced by the reticulations in the graph), this statement is not quite true, and potentially confusing. However, the graph is close to a tree in the following sense: if all the low-width bands of parallel edges are collapsed into a single edge, the graph becomes a tree. This statement will not hold for all networks, obviously; see, for example, their Fig. 6.15 (p. 172), where turning the network into a tree will require collapsing even the relatively wide bands of parallel edges.

When we collapse the narrow bands of parallel edges in the NeighborNet network (their Fig. 6.12, p. 164), we do obtain a graph that is tree-like at its center, but still reticulate within the subgroups. This graph reproduces all the established subgroups of Indo-European, but provides little resolution between the subgroups. Thus, it passes our criteria of compatibility and our desideratum of recognizing all groups.

A similarly constructed network based on data from 26 areally clustered languages of Australia (from Nash 2002) displays much more reticulation (p. 165) and is certainly much less tree-like than the IE graph. These are languages whose genealogical interrelationships are not well known, though all ultimately belong to the Pama-Nyungan family that spread across much of Australia in a time frame comparable to the age of Indo-European. Reticulation is expected given the small geographical area surveyed. Another graph (p. 166) shows that even the highly reticulate Australian data will produce a tree when an NJ method is used.

McMahon and McMahon also studied how different phylogeny estimation methods performed in simulation, when the characters evolve with borrowing but without homoplasy. Their study used a 12-leaf tree with

borrowing from one language into another and compared estimated trees to the true tree. They showed that small amounts of borrowing would not introduce error into an NJ analysis, but that as the amount of borrowing increased the estimated tree would deviate from the true tree. They then examined how different phylogenetic network methods performed on the same data. Their study showed that NeighborNet was the most sensitive, in that it produced reticulations (suggestive of borrowing) more readily than SplitsTree. They find that each method has points in its favor, but they prefer NeighborNet, since ‘Splitstree . . . has the problematic tendency to generate progressively more tree-like graphs . . . when confronted with large and complex data sets’ (McMahon and McMahon 2005: 160).

These applications show that the degree of reticulation varies as expected depending on the linguistic situation and the data. Given the few contrastive applications such as these in the literature, the field has no standards for whether a graph should be considered tree-like or what is an absolutely high, low, or average degree of reticulation. The relative assessments of McMahon, McMahon and their colleagues are clear, however, and the graphic presentations in McMahon and McMahon 2005: 164–72 are striking and should be pedagogically effective.

5.5.5. Perfect Phylogenetic Network analyses

Nakhleh et al. (2005a) proposed the Perfect Phylogenetic Network model of language evolution, in which characters can be borrowed but evolve without any homoplasy. The graphical model that is produced in a perfect phylogenetic network is an explicit phylogenetic network, so that there is an underlying tree with added contact edges. Two different computational methods have been developed to construct these networks from input data: a graph-theoretic algorithm (described in Nakhleh 2005 and used in Nakhleh et al. 2005a) and an algorithm using Artificial Intelligence techniques (given in Erdem et al. 2003). The analysis of Indo-European (Nakhleh et al. 2005a) used the Ringe and Taylor dataset of lexical, phonological, and morphological characters (with detected loans not coded as cognates) and produced a network with three additional contact edges, which was optimal with respect to each of a number of mathematical and linguistic criteria.

6. Discussion

One of the main observations in the studies reviewed here is that trees obtained for the same language family but using different datasets and/or different methods can differ in substantial ways – and yet can all pass the two topological tests we described earlier. Hence, it is clear that even the stronger of our two topological conditions is insufficient to imply accuracy with respect to the remaining features of interest (whether first-order subgrouping or internal subgrouping within a subfamily). Furthermore,

the critical examination we have made of the numerous studies reveals that even the best methods can have poor accuracy if used on poor-quality data (e.g., lexical data for which cognate judgments are questionable, or binary-encoded versions of good-quality multistate data). Therefore, our examination reveals that data selection (both of characters and languages) and the encoding of the character data have the potential to significantly impact the resultant phylogenetic estimation. Finally, these issues taken together show that while the development of methods for phylogenetic estimation in linguistics is exciting, we still do not have evidence that any of these methods is capable of accurate estimation of linguistic phylogenies. To establish whether any of these methods do, in fact, have good promise will require additional study – in the development of good benchmark datasets, in the statistical modeling of language evolution, and in the development of (probably new) methods.

Acknowledgments

The authors wish to thank the referees for their very helpful comments, and the National Science Foundation for its support to Tandy Warnow (grants ITR-BCS 0312830 and ITR 0331453).

Short Biographies

Johanna Nichols is Professor of Slavic Languages and Literatures and Affiliate Professor of Linguistics at the University of California, Berkeley. Her research interests include Russian and Slavic languages, languages of the Caucasus, typology, language spreads, and reconstructing ancient linguistic prehistory. She is co-director with Balthasar Bickel of the AUTOTYP typology database project and director of Itt Ezar Sahwat, a project aimed at compiling a very large database of spoken Ingush and Chechen. She has written an Ingush-English dictionary and co-authored a Chechen-English dictionary and is completing a reference grammar of Ingush. She was awarded the Bloomfield Book Award of the Linguistic Society of America in 1994 and the Order of Merit from the Government of Ingushetia in 2006 and was elected an AAAS Fellow in 2007.

Tandy Warnow is Professor of Computer Sciences at the University of Texas at Austin. Her research combines mathematics, computer science, and statistics to develop improved models and algorithms for reconstructing complex and large-scale evolutionary histories in both biology and historical linguistics. Tandy received her PhD in Mathematics at University of California, Berkeley, under the direction of Gene Lawler, and did postdoctoral training with Simon Tavaré and Michael Waterman at the University of Southern California. She received the National Science Foundation Young Investigator Award in 1994, and the David and Lucile Packard Foundation Award in Science and Engineering in 1996. Tandy is

a member of five graduate programs at the University of Texas, including Computer Science; Ecology, Evolution, and Behavior; Molecular and Cellular Biology; Mathematics; and Computational and Applied Mathematics. She is also the director for the multidisciplinary Cyber-Infrastructure for Phylogenetic Research Project, currently funded by the National Science Foundation under their Information Technology Program.

Notes

* Correspondence address: Tandy Warnow, Department of Computational Science, University of Texas at Austin, One University Station, Austin, TX 78712, USA. E-mail: tandy@cs.utexas.edu.

¹ Renfrew (1987) put forth the view that the spread of Indo-European was a first farming spread and originated in Anatolia some 10,000 years ago. Renfrew's view has had wide publicity, which has led some geneticists, phylogeneticists, and others to believe that it has appreciable support among linguists and archaeologists. We emphasize that growth of knowledge in linguistics, as in other sciences, is a matter of evidence and criteria, not opinion polls; there is a clear received view in linguistics, and it is that Indo-European spread some 5500 years ago from the western steppe. Some facts about Indo-European bear Renfrew's interpretation (e.g., the presence of terminology for domesticates; IE daughter languages now occupy much of the range where domestication was early to spread); some do not (e.g., the existence of PIE terms for wheeled transport and wool, proving that the dispersal of PIE did not occur before the development of these technologies); and we know of no linguistic facts that demand his interpretation and many that demand the received interpretation.

Works Cited

- Alpher, Barry, and David Nash. 1999. Lexical replacement and cognate equilibrium in Australia. *Australian Journal of Linguistics* 19.4–56.
- Anthony, David W. 2007. *The horse, the wheel, and language: how Bronze age riders from the Eurasian steppes shaped the modern world*. Princeton, NJ: Princeton University Press.
- Atkinson, Quentin, and Russell Gray. 2006. How old is the Indo-European language family? Illumination or more moths to the flame? *Phylogenetic methods and the prehistory of languages*, ed. by Peter Forster and Colin Renfrew, 91–110. Cambridge, UK: MacDonal Institute Press, University of Cambridge.
- Bandelt, Hans-Jurgen, Peter Forster, and A. Röhl. 1999. Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution* 16.37–48.
- Barbançon, François, Tandy Warnow, Steven N. Evans, Donald A. Ringe, Jr, and Luay Nakhleh. forthcoming. An experimental study comparing linguistic phylogenetic reconstruction methods. *Proceedings of a conference on Language and Genes*, University of California, Santa Barbara, September 2006. <http://www.cs.utexas.edu/users/tandy/nanterre-talk.ppt>.
- Bastin, Yvonne. 1983. Classification lexicostatistique des langues bantoues (214 releves). *Bulletin des Séances de l'Académie royale des sciences d'outre-mer* 27.173–99.
- Bastin, Yvonne, A. Coupeze and M. Mann. 1999. *Continuity and divergence in the Bantu languages: perspectives from a lexicostatistical study*. Tervuren, Belgium: Musée royal d'Afrique Centrale.
- Blust, Robert A. 1978. Eastern Malayo-Polynesian: a subgrouping argument. *Second international conference on Austronesian linguistics: proceedings (Pacific Linguistics C-61)*, 181–234. Canberra, Australia: Research School of Pacific Studies, Australian National University.
- . 1985. The Austronesian homeland: a linguistic perspective. *Asian Perspectives* 26.45–67.
- . 1994. The Austronesian settlement of mainland Southeast Asia. *Papers from the Second Annual Meeting of the Southeast Asian Linguistics Society*, ed. by Karen L. Adams and

- Thomas John Hudak, 25–83. Tempe, AZ: Program for Southeast Asian Studies, Arizona State University.
- . 1995. The prehistory of the Austronesian speaking people: a view from language. *Journal of World Prehistory* 9.453–510.
- . 2000. Why lexicostatistics doesn't work: the 'universal constant' hypothesis and the Austronesian languages. *Time depth in historical linguistics*, Vol. 2, ed. by Colin Renfrew, April McMahon and Larry Trask, 311–31. Cambridge, UK: The McDonald Institute for Archaeological Research.
- Bradley, David. 1994. East and Southeast Asia. *Atlas of the world's language*, ed. by Christopher Moseley and R. E. Asher, 159–92. London, UK: Routledge.
- Brown, Cecil H., Eric W. Holman, Søren Wichmann, and Viveka Vilupillai. forthcoming 2008. Automated classification of the world's languages: a description of the method and preliminary results. *Sprachtypologie und Universalienforschung*.
- Bryant, David, and Vincent Moulton. 2002. NeighborNet: An agglomerative method for the construction of planar phylogenetic networks. *Proceedings of the 2nd International Workshop on Algorithms in Bioinformatics (WABI'02)*, Rome (2002), *Lecture Notes in Computer Science (LNCS #2452)*, ed. by Roderic Guigó and Daniel Gusfield, 375–91. London, UK: Springer-Verlag.
- . 2004. NeighborNet: an agglomerative algorithm for the construction of phylogenetic networks. *Molecular Biology and Evolution* 21(2).255–65.
- Bryant, David, Flavia Filimon, and Russell Gray. 2005. Untangling our past: Languages, trees, splits, and networks. *The evolution of cultural diversity: a phylogenetic approach*, ed. by Ruth Mace, Clare J. Holden and Stephen Shennan, 67–83. London, UK: University College London Press.
- Cysouw, Michael, Søren Wichmann, and David Kamholz. 2006. A critique of the separation base method for genealogical subgrouping, with data from Mixe-Zoquean. *Journal of Quantitative Linguistics* 13.225–64.
- Darden, Bill J. 2001. On the question of the Anatolian origin of Indo-Hittite. *Greater Anatolia and The Indo-Hittite language family*, ed. by Robert Drews, 184–228. Washington, DC: Institute for the Study of Man.
- Donohue, Mark, and Simon Musgrave. 2007. Typology and the linguistic macrohistory of Island Melanesia. *Oceanic Linguistics* 46.348–87.
- Donohue, Mark, and Søren Wichmann. 2008. Typology, areality, and relatedness. *Oceanic Linguistics* 48.223–32.
- Dunn, Michael, Angela Terrill, Ger P. Reesink, Robert A. Foley, and Stephen C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science* 309.2072–5.
- Dunn, Michael, Robert A. Foley, Stephen C. Levinson, Ger P. Reesink, and Angela Terrill. 2007. Statistical reasoning in the evaluation of typological diversity in Island Melanesia. *Oceanic Linguistics* 46.388–403.
- Dyen, Isidore, Joseph B. Kruskal and Paul Black. 1992. An Indoeuropean classification: a lexicostatistical experiment. *Transactions of the American Philosophical Society* 82, pt. 5. Philadelphia, PA: American Philosophical Society.
- Embleton, Sheila. 1986. *Statistics in historical linguistics*. Bochum, Germany: Brockmeyer.
- Erdem, Esra, Vladimir Lifschitz, Luay Nakhleh, and Donald A. Ringe, Jr. 2003. Reconstructing the evolutionary history of Indo-European languages using answer set programming. *Proceedings of the 5th International Symposium on Practical Aspects of Declarative Languages (PADL 03)*, 13–14 January 2003, pp. 160–76.
- Erdős, Peter L., Michael A. Steel, Laszlo Székely, and Tandy Warnow. 1999a. A few logs suffice to build almost all trees – I. *Random Structures and Algorithms*, 14.153–84. (Also appears as DIMACS Technical Report 97–71.)
- . 1999b. A few logs suffice to build almost all trees – II. *Theoretical Computer Science* 221(1–2).77–118.
- Eska, Joseph F., and Donald A. Ringe, Jr. 2004. Recent work in computational linguistic phylogeny. *Language* 80.569–82.
- Evans, Steven N., Donald A. Ringe, Jr, and Tandy Warnow. 2006. Inference of divergence times as a statistical inverse problem. *Phylogenetic methods and the prehistory of languages*,

- ed. by Peter Forster, and Colin Renfrew, 119–29. Cambridge, UK: McDonald Institute for Archaeological Research, University of Cambridge.
- Evans, Steven N., and Tandy Warnow. 2005. Unidentifiable divergence times in rates-across-sites models. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 1.130–4.
- Field, Fredric W. 2002. Linguistic borrowing in bilingual contexts. Amsterdam, The Netherlands/Philadelphia, PA: Benjamins.
- Foley, William A. 2007. Language change and diffusion in the Sepik region, PNG. MS, University of Sydney.
- Forster, Peter, and Alfred Toth. 2003. Toward a phylogenetic chronology of ancient Gaulish, Celtic, and Indo-European. *Proceedings of the National Academy of Sciences, USA* 100.9079–84.
- Foulds, L. R., and Ronald L. Graham. 1982. The Steiner problem in phylogeny is NP-complete. *Advances in Applied Mathematics* 3.43–9.
- Goddard, Ives. 1996. The west-to-east cline in Algonquian dialectology. *Actes du Vingt-cinquième Congrès des algonquinistes*, ed. by William Cowan, 187–211. Ottawa, Canada: Carleton University.
- Goloboff, Pablo, Steven Farris, and Kevin Nixon. 2003. T.N.T: tree analysis using new technology, version 1.0. Program and documentation. <<http://www.zmuc.dk/public/Phylogeny/TNT>>.
- Grace, George W. 1996. Regularity of change in what? The comparative method reviewed, ed. by Mark Durie and Malcolm Ross, 156–79. New York, NY/Oxford, UK: Oxford University Press.
- Gray, Russell D., and Fiona M. Jordan. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature* 405.1052–5.
- Gray, Russell D., and Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origins. *Nature* 426.435–9.
- Gray, Russell D., Simon Greenhill, and Robert Blust. 2007. Computational phylogenetic methods and Austronesian subgrouping. Presented at International Congress of Historical Linguists, Montreal.
- Greenhill, S. J., and R. D., Gray 2005. Testing population dispersal hypotheses: Pacific settlement, phylogenetic trees, and Austronesian languages. The evolution of cultural diversity: phylogenetic approaches, ed. by R. Mace, C. Holden and S. Shennan, 31–52. London, UK: UCL Press.
- Greenhill, Simon J., Robert A. Blust, and Russell D. Gray. 2003–2008. Austronesian Basic Vocabulary Database <<http://language.psy.auckland.ac.nz/austronesian>>.
- Guthrie, Malcolm. 1967–71. *Comparative Bantu* (4 vols.). Farnborough, UK: Gregg.
- Haas, Mary R. 1949. The position of Apalachee in the Muskogean family. *International Journal of American Linguistics* 15.121–7.
- . 1979. Southeastern languages. The languages of native America, ed. by Lyle Campbell and Marianne Mithun, 299–326. Austin, TX: University of Texas Press.
- Haspelmath, Martin, and Uri Tadmor. n.d. Loanword Typology <<http://www.eva.mpg.de/lingua/files/lwt.html>>
- Haspelmath, Martin, Matthew S. Dryer, David Gil, and Bernard Comrie (eds) 2005. *World Atlas of Language Structures*. Oxford, UK: Oxford University Press.
- Hinnebusch, Thomas J. 1996. Skewing in lexicostatistic tables as an indicator of contact. Paper presented at the Round Table on Bantu Historical Linguistics, Université Lumière 2, Lyon, France, 30 May–1 June 1996.
- Holden, Clare J. 2002. Bantu language trees reflect the spread of farming across Sub-Saharan Africa: a maximum-parsimony analysis. *Proceedings of the Royal Society of London Series B* 269.793–9.
- Holden, Clare J., A. Meade, and M. Pagel. 2005. Comparison of maximum parsimony and Bayesian Bantu language trees. The evolution of cultural diversity: a phylogenetic approach, ed. by R. Mace, C. J. Holden, and S. Shennan, 53–65. London, UK: University College London Press.
- Holden, Clare J., and Russell Gray. 2006. Rapid radiation, borrowing, and dialect continua in the Bantu languages. *Phylogenetic methods and the prehistory of languages*, ed. by Peter Forster and Colin Renfrew, 19–31. Cambridge, UK: MacDonal Institute Press, University of Cambridge.

- Holland, B., and V. Moulton 2003. Consensus networks: a method for visualizing incompatibilities in collections of trees. Workshop on algorithms for bioinformatics, ed. by G. Benson and R. Page. Lecture Notes in Computer Science, Vol. 2812, 165–76. Berlin, Germany: Springer.
- Holm, Hans J. 2000. Genealogy of the main Indo-European branches applying the separation base method. *Journal of Quantitative Linguistics* 7.73–95.
- . 2003. The proportionality trap. Or: What is wrong with lexicostatistical subgrouping? *Indogermanische Forschungen* 108.38–46.
- . 2007. The new arboretum of Indo-European ‘Trees’. Can new algorithms reveal the phylogeny and even prehistory of Indo-European? *Journal of Quantitative Linguistics* 14.167–214.
- Huelsenbeck, John and Fredrik Ronquist, 2001. MrBayes: Bayesian inference of phylogeny. *Bioinformatics* 17.754–5.
- Huson, Daniel. 1998. SplitsTree: a program for analyzing and visualizing evolutionary data. *Bioinformatics* 14(1).68–73.
- Huson, D., and D. Bryant 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 32.254–67.
- Ives, John W. 2003. Alberta, Athapaskans, and Apachean origins. *Archaeology in Alberta: a view from the new millennium*, ed. by J. W. Brink and J. F. Dormaar. Medicine Hat, Alberta: Archaeological Society of Alberta.
- Janhunen, Juha. 2001. Indo-Uralic and Ural-Altai: on the diachronic implications of areal typology. Early contacts between Uralic and Indo-European: linguistic and archaeological considerations, ed. by Christian Carpelan, Asko Parpola and Petteri Koskikallio, 207–20. Helsinki, Finland: Suomalais-Ugrilainen Seura.
- Kaufman, Terrence, and John Justeson. 2004. Epi-Olmec. *The encyclopedia of the world's ancient languages*, ed. by R. D. Woodard, 1071–111. Cambridge, UK: Cambridge University Press.
- Kessler, Brett. 2001. *The significance of word lists*. Stanford, CA: CSLI.
- Kim, Junhyong, and Tandy Warnow. 1999. Tutorial on Phylogenetic Tree Estimation. Intelligent Systems for Molecular Biology, Heidelberg 1999 <<http://www.cs.utexas.edu/users/tandy/tutorial.ps>>.
- Kirch, Patrick V. 1997. *The Lapita peoples: ancestors of the oceanic world*. Cambridge, MA: Blackwell.
- . 2000. *On the road of the winds: an archaeological history of the Pacific islands before European contact*. Berkeley, CA: University of California Press.
- Lees, Robert B. 1953. The basis of glottochronology. *Language* 29.113–27.
- Lohr, Marisa. 1999. *Methods for the genetic classification of languages*. PhD dissertation, University of Cambridge.
- Lynch, John. 1999. Linguistic change in southern Melanesia: linguistic aberrancy and genetic distance. *Archaeology and languages*, ed. by Roger Blench and Matthew Spriggs, 149–59. London, UK: Routledge.
- Mallory, J. P. 1989. *In search of the Indo-Europeans: language, archaeology, and myth*. New York, NY: Thames & Hudson.
- Marten, Lutz. 2006. Bantu classification, Bantu trees, and phylogenetic methods. *Phylogenetic methods and the prehistory of languages*, ed. by Peter Forster and Colin Renfrew, 43–56. Cambridge, UK: MacDonald Institute Press, University of Cambridge.
- Matisoff, James. 2000. On the uselessness of glottochronology for the subgrouping of Tibeto-Burman. *Time depth in historical linguistics*, ed. by Colin Renfrew, April McMahon and Larry Trask, 333–72. Cambridge, UK: The McDonald Institute for Archaeological Research.
- McMahon, April, and Robert McMahon. 2005. *Language classification by numbers*. Oxford, UK: Oxford University Press.
- McMahon, April, Paul Heggarty, Robert McMahon, and Natalia Slaska. 2005. Swadesh sublists and the benefits of borrowing: an Andean case study. *Transactions of the Philological Society* 103.147–70.
- Michener, C. D., and Robert R. Sokal. 1957. A quantitative approach to a problem in classification. *Evolution* 11.130–62.

- Minett, James W., and William S.-Y. Wang. 2003. On detecting borrowing: distance-based and character-based approaches. *Diachronica* 20.289–330.
- Moret, B. M. E., U. Roshan, and T. Warnow. 2002. Sequence length requirements for phylogenetic methods. Proceedings of the 2nd International Workshop on Algorithms in Bioinformatics (WABI'02), Rome (2002), Lecture Notes in Computer Science (LNCS #2452), ed. by Roderic Guigó and Daniel Gusfield, 343–56. London, UK: Springer-Verlag.
- Nakhleh, Luay. 2005. Phylogenetic networks, PhD dissertation. The University of Texas at Austin.
- Nakhleh, Luay, Katherine St. John, Usman Roshan, Jerry Sun, and Tandy Warnow. 2001. Designing fast converging phylogenetic methods. 9th Int'l Conf. on Intelligent Systems for Molecular Biology (ISMB 2001), Copenhagen, in *Bioinformatics* 17(suppl. 1).S190–8.
- Nakhleh, Luay, Donald A. Ringe, Jr, and Tandy Warnow. 2005a. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language* 81.382–420.
- Nakhleh, Luay, Tandy, Warnow, Donald A., Ringe, Jr, and Steven N. Evans. 2005b. A comparison of phylogenetic reconstruction methods on an IE dataset. *Transactions of the Philological Society* 103.171–92.
- Nash, David. 2002. Historical linguistic geography of southeast Western Australia. *Language in native title*, ed. by John Henderson and David Nash, 205–30. Canberra, Australia: AIATSIS Native Title Research Unit, Aboriginal Studies Press.
- Naylor, Gavin, and Fred Kraus. 1995. The relationship between s and m and the retention index. *Systematic Biology* 44(4).559–62.
- Nicholls, Geoff, and Russell Gray. 2008. Dated ancestral trees from binary trait data and their application to the diversification of languages. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 70.545–66.
- Nichols, Johanna. 1992. *Linguistic diversity in space and time*. Chicago, IL: University of Chicago Press.
- . 1997. Modeling ancient population structures and movement in linguistics. *Annual Review of Anthropology* 26.359–84.
- . 2006. Quasi-cognates and lexical type shifts: rigorous distance measures for long-range comparison. *Phylogenetic methods and the prehistory of languages*, ed. by Peter Forster and Colin Renfrew, 57–65. Cambridge, UK: MacDonald Institute Press, University of Cambridge.
- . 2007. Language dispersal from the Black Sea region. *The black sea flood question: changes in coastline, climate, and human settlement*, ed. by Valentina Yanko-Holmbach, Allan S. Gilbert, Nicolae Panin and Pavel M. Dolukhanov, 775–96. Dordrecht, The Netherlands: Kluwer.
- Nichols, Lynn, and Johanna Nichols. 2007. Lexical derivational properties resist diffusion. Presented at Workshop on Language contact and morphosyntactic variation and change, 7th biennial meeting, Association for Linguistic Typology, Paris.
- Pagel, Mark, Quentin Atkinson, and Andrew Meade. 2007. Frequency of word use predicts rates of lexical evolution throughout Indo-European history. *Nature* 449.717–21.
- Pawley, Andrew K., and Malcolm D. Ross. 1993. Austronesian historical linguistics and culture history. *Annual Review of Anthropology* 22.425–59.
- Rannala, Bruce, John P. Huelsenbeck, Ziheng Yang, and Rasmus Nielsen. 1998. Taxon sampling and the accuracy of large phylogenies. *Systematic Biology* 47.702–10.
- Renfrew, Colin. 1987. *Archaeology and language: the puzzle of Indo-European origins*. Cambridge, UK: Cambridge University Press.
- Rexová, Kateřina, Daniel Frynta, and J. Zrzavý. 2003. Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. *Cladistics* 19.120–7.
- Rexová, Kateřina, Yvonne Bastin, and Daniel Frynta. 2006. Cladistic analysis of Bantu languages: a new tree based on combined lexical and grammatical data. *Naturwissenschaften* 93(4).189–94.
- Ringe, Donald A., Jr, Tandy Warnow and Ann Taylor. 2002. Indo-European and computational cladistics. *Transactions of the Philological Society* 100.59–129.
- Robinson, D. F and L. R. Foulds. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences* 53.131–47.
- Ross, Malcolm D. 1988. *Proto Oceanic and the Austronesian languages of western Melanesia*. Canberra, Australia: Australian National University.

- . 1995. Comparative Austronesian dictionary: an introduction to Austronesian studies. *Trends in linguistics*, 10, ed. by Darrell Tryon. Vol. 1, 27–42. Berlin, Germany/New York, NY: Mouton de Gruyter.
- . 1996. Contact-induced change and the comparative method: cases from Papua New Guinea. *The comparative method reviewed*, ed. by Mark Durie and Malcolm D. Ross, 180–217. New York, NY: Oxford University Press.
- . 1997. Social networks and kinds of speech-community event. *Archaeology and language I: theoretical and methodological orientations*, ed. by Roger Blench and Matthew Spriggs, 209–61. London, UK/New York, NY: Routledge.
- . 2002. Final words: Research themes in the history and typology of western Austronesian languages. *The history and typology of Western Austronesian voice systems*, ed. by Fay Wouk and Malcolm Ross, 451–74. Canberra, Australia: Pacific Linguistics.
- Ross, Malcolm D., and Áshild Naess. 2007. An Oceanic origin for Āiwoo, the language of the Reef Islands? *Oceanic Linguistics* 46(2).456–98.
- Saitou, Naryua, and Masatoshi Nei. 1987. The neighbor-joining method: a new method, for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4.406–25.
- Saunders, Arpiar. 2005. Linguistic phylogenetics of the Austronesian family: a performance review of methods adapted from biology. BA thesis, Swarthmore College. <http://arpiarsaunders.com/documents_download/Saunders_AustroThesis2005.pdf>.
- Serva, Maurizio, and Filippo Petroni. 2008. Indo-European languages tree by Levenshtein distance. *Europhysics Letters* 81.68005–9.
- Steel, Michael A. 1992. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification* 9.91–116.
- Swadesh, Morris. 1952. Lexicostatistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* 96.452–63.
- . 1955. Toward greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21.121–37.
- Swofford, David, 1996. PAUP*: Phylogenetic analysis using parsimony (and other methods), version 4.0. Sunderland, MA: Sinauer Assoc.
- Thomason, Sarah G. 2001. *Language contact*. Washington, DC: Georgetown University Press.
- Thomason, Sarah G., and Terrence Kaufman. 1988. *Language contact, creolization, and genetic linguistics*. Berkeley, CA: University of California Press.
- Wang, William S-Y, and James W. Minett. 2005. Vertical and horizontal transmission in language evolution. *Transactions of the Philological Society* 103(2).121–46.
- Warnow, Tandy, Bernard M. E. Moret, and Katherine St. John. 2001. Absolute convergence: true trees from short sequences. *Proceedings of 12th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 01)*, Washington, DC, USA, 186–95. Philadelphia, PA: Society for Industrial and Applied Mathematics.
- Warnow, Tandy, Steven N. Evans, Donald A. Ringe, Jr, and Luay Nakhleh. 2006. A stochastic model of language evolution that incorporates homoplasy and borrowing. *Phylogenetic methods and the prehistory of languages*, ed. by Peter Forster and Colin Renfrew, 75–87. Cambridge, UK: MacDonald Institute Press, University of Cambridge.
- Waterman, Michael S., Temple Smith, M. Singh, and W. A. Beyer. 1977. Additive evolutionary trees. *J. Theoretical Biology* 64, 199–213.
- Wichmann, Søren. 1995. *The relationship among the Mixe-Zoquean languages of Mexico*. Salt Lake City, UT: University of Utah Press.
- Wichmann, Søren, and Arpiar Saunders. 2007. How to use typological databases in historical linguistic research. *Diachronica* 24.373–404.
- Zobel, Erik. 2002. The position of Chamorro and Palauan in the Austronesian family tree: evidence from verb morphosyntax. *The history and typology of Western Austronesian voice systems (Pacific Linguistics; 518)*, ed. by Fay Wouk and Malcolm Ross, 405–34. Canberra, Australia: Australian National University.