How to use typological databases in historical linguistic research*

Søren Wichmann and Arpiar Saunders
Max Planck Institute for Evolutionary Anthropology & Leiden University /
Swarthmore College

Several databases have been compiled with the aim of documenting the distribution of typological features across the world's languages. This paper looks at ways of utilizing this type of data for making inferences concerning genealogical relationships by using phylogenetic algorithms originally developed for biologists. The focus is on methodology, including how to assess the stability of individual typological features and the suitability of different phylogenetic algorithms, as well as ways to enhance phylogenetic signals and heuristic procedures for identifying genealogical relationships. The various issues are illustrated by a small sample of empirical data from a set of Native American languages.

Keywords: Diachronic typology, stability of linguistic features, Native American languages, phylogenetic algorithms, Neighbor-joining, Neighbor Net, Maximum parsimony, Bayesian analysis

0. Introduction

This paper will focus on methodological issues involved in drawing inferences concerning genetic relatedness from typological data. We mainly look at ways of detecting vertical inheritance, although we do not wish thereby to imply that lateral diffusion via language contact is not an equally important issue or that it could not be addressed through related procedures. We do not attempt to derive empirical conclusions concerning actual historical relations, but simply wish to

* We would like to acknowledge comments from Bernard Comrie, Joseph Felsenstein, Cecil H. Brown, and Sean Crist on an earlier version of this paper, as well as comments from Marianne Mithun on an oral presentation at the 17th International Conference on Historical Linguistics, Madison, August 2005. Special thanks go to Melanie Keledjian for help with translations of the summary.

Diachronica 24:2 (2007), 373–404.
ISSN 0176–4225 / E-ISSN 1569–9714 © John Benjamins Publishing Company

summarize the results of an initial exploration of the application of phylogenetic methods and techniques to a typological dataset. The dataset used is *The World Atlas of Language Structures* (Haspelmath et al., eds., henceforth *WALS*). For illustrative purposes we draw upon a small subset of *WALS* data from the native languages of the Americas.[1]

We first discuss some similarities between our approach to phylogenetics and approaches in biology and briefly comment on the currently available phylogenetic programs that seem most useful for historical linguistic purposes. We then go on to discuss one of the most pressing problems encountered in using typological databases to infer genealogies, namely, which typological features to use. Then we discuss the use of four different types of algorithms for inferring phylogenies as well as methods for enhancing phylogenetic signals. Finally, we demonstrate how lexical data reinforcing known, shallow genetic relations may be used to enhance the heuristic value of a phylogeny generated from typological data.

Only during the last decade have serious attempts been made to study evolutionary relationships of languages using computationally driven phylogenetic approaches. The pioneering (and controversial) works of Russell D. Gray and colleagues differ from the present contribution by drawing uniquely upon lexical data (Gray & Fiona 2000, Gray & Atkinson 2003), as do other recent contributions (Forster et al. 1998, Forster & Toth 2003, Holden 2002, McMahon & McMahon 2003, McMahon et al. 2005, Atkinson et al. 2005, Cysouw et al. 2006). Don Ringe, Tandy Warnow and associates have assembled sets of not only lexical characters but also phonological ones (for Indo-European) and have subjected them to rigorous phylogenetic analyses using computational algorithms (Warnow 1997, Nakleh, Ringe & Warnow 2005, Nakleh, Warnow, Ringe & Evans 2005). Much of this work has focused on the development of the model of so-called 'perfect phylogenies'. In such phylogenies, all — or at least the great majority — of the linguistic characters selected are compatible with a single evolutionary tree. That is, cases of borrowing, parallel changes, and changes back to an earlier state (backmutation) are excluded (see Ringe, Warnow & Taylor 2002 for data). In the present paper, however, we shall address problems arising from using data which, because of their generic, typological nature, are not expected to yield perfect phylogenetic networks. The type of data which we consider, have not gone through prior analysis and sifting by scholars working within the framework of the traditional comparative method.

[1] In future research we hope to expand our agenda to making actual inferences regarding the classification of these languages, since they present a special challenge to historical linguistics given the genetically fragmented picture according to the consensus view coupled with the widely accepted hypothesis that the Americas were one of the last areas in the world to be populated by humans. The present paper, however, is purely methodological.

Thus, we use typological features, several of which may have suffered diffusion, parallel changes and backmutation, but we try to identify and use features that are least amenable to this type of behavior and search for algorithms that minimize the phylogenetic noise produced by this type of data. Unlike that of Ringe, Warnow and colleagues, our approach is not so much a refinement of a particular aspect of traditional comparative linguistics — i.e. the phylogenetic aspect — but is better described as a new alternative based on the employment of methods adapted from biology. As the present paper was submitted, another relevant work, namely Dunn et al. (2005), appeared. This is the first attempt to draw phylogenetic inferences from typological data and, as such, is congenial to our approach. Our paper, however, complements rather than overlaps with that of Dunn et al. While the latter authors immediately attempt to draw empirical conclusions and largely exclude the methodological considerations that have let them to make choices such as using parsimony methods or organizing their data as binary-valued typological features, we directly address such methodological issues. Several of our observations can be brought to directly bear on Dunn et al. (2005), but we defer direct discussion of their contribution to another context.

In the following section we briefly review some of the productive parallels of phylogenies based on generic typological data vs. biological data.

1. Phylogeny in biology and in linguistics

The wealth of new data resulting from the molecular biological revolution did not miraculously resolve the ancestral relationships of related organisms. Initially, biologists were baffled at contradictions between molecular data and organismal phylogenies established through comparative morphology and behavior. Even more troubling, the molecular data (i.e. DNA, RNA and Protein) were often internally inconsistent; the natural histories of genes and gene products differed within an organism. Biological phylogenetics had to confront the philosophical and methodological problem of how to resolve the evolutions of characters from different types of data with the evolution of the organism as a whole. The debate and subsequent computational solutions to this dilemma have productive applications for historical linguistics.

Building phylogenies for both organisms and languages depends on comparing the states for a set of shared characters. The hope is that by analyzing these differences with an appropriate algorithm a phylogenetic signal can be recovered with a quantifiable degree of certainty. In sequence based phylogeny (DNA, for example), each character is a linear position along the gene and the character state is the nucleotide building block occupying that place. Since homologous sequence characters are easily defined, comparing them depends only on proper alignment. Morphological and behavioral characters must be defined more subjectively since each character and its representative character states are constructed by the researcher. In defining characters of this type, the researcher must use their best judgment to assure that characters are both truly homologous (i.e. share the same developmental mechanism) and behave independently during evolutionary change. When these conditions are not met, skewed phylogenies will ensue. A final difference between molecular and morphological data involves the nature of the character state labels used for coding. With molecular sequence data, character states are shared across characters. For example, with DNA data, if Adenine ("A") is coded as "1," all "1's" can be treated as a group across characters. With morphological and behavioral data, a "1" may mean "50–75mm beak depth" for the first character and "brown/black coloration" for the second. This latter coding scheme, where character state labels are arbitrary, is appropriate for typological and lexical data. Once coded, language data can be analyzed by a wide variety of algorithms built for biological morphology.

Phylogenies are the result of the interaction between coded data and a specific algorithm. While the advantages of different encoding schemes and algorithms are hotly debated among biologists, linguists have done little to systematically evaluate which methods, if any, are appropriate for inferring language evolution. The results presented here represent an initial attempt to narrow down the methodological choices faced by linguists for determining which algorithms are most appropriate for typological data. Often in biological contexts, multiple algorithms are employed on the same dataset to demonstrate the robustness of the phylogenetic signal (e.g., Als et al. 2004). Here we shall employ a similar strategy in order to investigate the utility of *WALS*-type data for phylogenetic research.

Many of the most basic confounds faced by historical linguists when constructing evolutionary trees, including the possibility of horizontal transfer of linguistic features and differing rates of change for different linguistic components, have biological counterparts. The sophisticated methods biologists have developed to empirically approach these issues offer powerful new perspectives for diachronic linguistics. Two languages may share the same state for a given typological feature through common ancestry, contact or random chance. Determining which of these possibilities is being instantiated requires a comparison with the behavior of other typological features. Is only a single state shared between the two languages or are there groups of shared states? And do these similarities conflict phylogenetically with shared states among other languages in the sample? Biologists are confronted with similar problems. For example, while ancestrally related genes do accumulate tractable changes, genes can also be transferred horizontally between unrelated species. Especially rampant in the bacterial kingdom, lateral gene transfer is considered to be a major player in evolution (Ochman, Lawrence & Groisman 2000). While identifying areal typological features using current biological software is outside the scope of this study, we hope that a brief description of strategies developed by biologists may inspire historical linguists in the future.

One method compares phylogenies produced by different genes of the genome. The underlying logic is that a consensus of phylogenetic signals across genes under different evolutionary constraints most accurately represents the natural history of the whole organism (Griffiths 1999, Nylander et al. 2004). A second strategy, amenable to organisms with sequenced genomes, involves genetic comparisons between distantly related species; high similarity between genes not shared by 'intermediate' organisms suggests the candidate genes may have been acquired through horizontal transfer (cf. Lawrence & Ochman 2002). A third strategy, similarly non-phylogenetic but intra-species, identifies genes that appear outside of their genomic context (i.e. patterns of nucleotides), the logic here being that long-term evolution in a separate species might leave an identifiable 'fingerprint' (cf. Lawrence & Ochman 2002). Perhaps the software implementations of these approaches can be tuned to assess horizontal transfer of typological and lexical data in the future.

In addition to detecting similarities due to contact, linguists also must contend with heterogeneity in the rates of evolutionary change for different components of language, including different typological features. Although in practice little is known empirically about the rates of change for specific typological features, those with rapidly changing states may be used to infer shallow evolutionary signals, while features with slower rates of change may inform deeper relationships. Large differences in rates can similarly occur between different regions of the genome. For example, DNA coding for highly conserved, essential proteins is in general less tolerant of mutations than non-coding, non-regulatory DNA. When character sets are diverse with respect to rate, one hopes for the 'shallow' and 'deep' phylogenetic signals to be clearly represented by the appropriate character type and for the uninformative type to introduce only negligible noise.

Some of the more parameter-rich biological algorithms allow evolutionary rates to differ between characters based on a variety of distributions. Simpler algorithms do not take rate heterogeneity into account. Thus, the issue of rate heterogeneity exemplifies an important general principle in phylogenetics: researchers must understand the assumptions and mechanics of their algorithms to properly interpret the results. Models operating on a large number of parameters provide the opportunity for heightened realism, while simpler models have the advantage of being more transparent making the resulting trees easier to interpret.

One widespread parameter in models of biological evolution is the use of forced directional changes between characters states. Just as many types of phonological changes are directional, the probability of different amino acid substitutions can be quantified and applied to models of evolution. Incorporating directional information into character state transitions may be one simple and productive way in which biological parameters can be successfully adapted to language data.

The work of Russell Gray and colleagues represents an admirable initial effort to ask historical linguistic questions based on lexical data and biological algorithms (Gray & Atkinson 2003, Gray & Jordan 2000). Gray's technique involves coding lexical items as cognate classes. Such a strategy produces a huge number of characters for comparison, but is still limited by the need for reliable cognacy judgments. If one goal of linguistic phylogenetics is to infer more ancient relationships than those distinguishable by words alone, typological data may be the only choice. Some of our unpublished work (Saunders 2006) suggests that introducing even a small number of typological characters into a predominantly lexical dataset can dramatically increase the accuracy of the phylogeny. Presumably, the data types are complementary; the inclusion of stable typological features may resolve higher order relationships relatively uninformed by lexical data. Similar improvements have been noted in biological analyses, where combining molecular and morphological data to infer organismal phylogenies can increase certainty and improve the accuracy of the resulting trees (Baker & Gatesy 2002).

Despite the striking theoretical and methodological parallels between biological and linguistic evolution, the transfer of methods must be done with caution. Efforts must be made to evaluate both the diversity of methods available and the appropriateness of different options for analysis within each method. This is a short term goal. In the long run, the best performing algorithms and evolutionary models can be incorporated into software designed specifically for linguistic data. Towards the short term goal, linguists may currently benefit from the achievements of biological phylogenetics through the use of the available software.

The expansive collection of computational tools developed to pursue a variety of questions relating to phylogenetics should inspire linguists. A good place to start is to go to Joseph Felsenstein's webpage, which contains the most up-to-date collection.[2] While the number of tools available may be staggering, many operate on the same basic principles. To help linguists efficiently apply the methods evaluated in this paper and others found on the Felsenstein webpage, we provide a brief introduction to the discipline in the following.

Phylogenetics software can be roughly divided into two categories, generalist and specialist. Two of the simplest and most widely-used generalist packages are

[2] The URL for Felsenstein's webpage is http://evolution.genetics.washington.edu/phylip/software.html.

PAUP* (Swofford 2002) and PHYLIP (Felsenstein 2005),[3] which allow for a large variety of Maximum Likelihood, Maximum Parsimony and Distance Methods. In addition to raw output, these programs have many post-analysis features enabling the user to measure, compare and visualize trees. Both PAUP* and PHYLIP have been around for a long time, and are constantly updated. The specialist programs are usually designed to complement the generalist programs, often representing methodological extensions or improvements in efficiency.

Most phylogenetics programs require that input data be organized in a format called 'NEXUS' (Maddison et al. 1997). The NEXUS format consists of organizational divisions called 'blocks,' each of which contains information concerning the nature of the data or the type of analysis. The most basic information (how the data are coded and how many characters and taxa are present) is listed in blocks used by all of the programs; the more program-specific information is listed in specialized blocks that can easily be added or removed. The NEXUS format allows a single data set to be interchanged between programs with minimal effort.

In §§4–5 below, after having discussed criteria for sampling the most reliable typological data and presenting the languages in the sample, we introduce four of the basic phylogenetic methods and performance-evaluate their results on our dataset.

2. A method for evaluating the strength of typological features for phylogenetic analyses

In the preceding section we mentioned a phenomenon shared between biological phenomena and languages, namely the differential rates of change for different characters. Our main motivation for using typological features for building genealogies is the hope that we may be able to reconstruct language history at time depths that are not within reach of the traditional comparative method. To reach the maximally possible depth of time one will need to choose the features which tend to be most stable diachronically and which are least amenable to change as the result of areal convergence. In this section we briefly summarize an approach to the problem of determining the relative stabilities of different typological features (the approach derives from Wichmann & Kamholz forthcoming, and is discussed in more detail there).

In order to make meaningful statements about the stability of linguistic features it is necessary to study their behavior within genetic units where there are no controversies over genetic relatedness and where time depths are roughly equal.

[3] PAUP* is downloadable through www.sinauer.com and PHYLIP through the URL cited in the previous note.

A convenient list of such uncontroversial genetic units is provided by the 'genera' of *WALS*, as defined by Matthew S. Dryer. A genus is a group of languages which are fairly obviously related and whose time depth, whenever this is known, does not exceed 4000 years. The level of classification corresponding to a genus is intended to be comparable across the world. It is not an entirely objective notion but relies to some extent on the informed intuitions of Dryer (cf. Dryer 1989 for more discussion and Dryer 2005a for a complete list of genera).

A starting assumption for our approach is that the feature value which is most favored in a given genus is the one that should be reconstructed for the proto-language of the genus and that languages exhibiting other values will have undergone changes. There would obviously be counterexamples to this, but we think that at a shallow time depth the assumption would hold true in the great majority of cases. Moreover, we often simply do not know what the true history was, so we need an assumption like the one just stated. On the basis of these considerations we can now make the crucial inference that *the more representative a feature value is within a given genus, the more stable that feature may be assumed to be within the genus*. In other words, the more widespread the feature value, the more stability is inherent in the feature. Following this logic we may study the distribution of values of a given feature for each genus and then calculate an average of how well represented the best represented value is throughout all genera in the *WALS* sample. It is irrelevant to what degree 'the best represented value' varies across genera — when each genus has a high degree of consistency of one particular value then the feature as a whole should count as highly stable.

In order to implement this kind of evaluation strategy we need to tackle the problem of how to compare features that have different numbers of values or for which the number of languages attested vary. For an illustration of the question, suppose that we are comparing the stability of the following two features: 'Order of Subject, Object, and Verb (*WALS* Ch. 81 = Dryer 2005b) and 'Associative Plural' (*WALS* Ch. 36 = Daniel & Moravcsik 2005).[4] Even if we limit the test to one genus, for instance Germanic, we still have the problem that the numbers of languages sampled for the two features are different, and the number of possible values is also different. Thus, the best represented value ('SVO') out of 7 possible ones occurs 5 times in a sample of 8 Germanic languages. How may we now compare this to the Associative Plural? Here the best represented feature value ('unique periphrastic associative plural') is one out of 4 possible values and occurs 5 times in a sample of

[4] For the present discussion it is irrelevant exactly what precisely is meant by 'Associative plural' or how the word order feature is defined. Moreover, it would take up too much space to present feature definitions here. For such information, the interested reader is referred to the relevant *WALS* chapters.

7 Germanic languages. Let us translate the problem into a situation which might be more familiar.

In the case of the Associative Plural, we have to decide what the probability is of drawing the same card 5 times when there are 7 cards to draw from and each may have 4 different values. This can only be decided if we either know what the probability of each of the 4 different values is or if we simply assume that they are equal. For the present, let us make the latter assumption.[5] With a bit of mathematical intuition it is easy to see that the probability of the situation represented by the Associate Plural is higher than that of drawing the same card 5 times when there are 8 to draw from and the number of (equally) possible values is 7. Obviously, all else being equal, it is a bit harder to draw the same card 5 times of out 7 than 5 times out of 8. But when there are several more equally possible values involved, as in the second situation, it is going to get *very* hard to draw 5 that are the same out of 8. Since the probability that the number of occurrences of the best represented feature value could be due to sheer chance is lower for 'Order of Subject, Object, and Verb' than for 'Associate Plural', the Germanic evidence suggests that Order of Subject, Object, and Verb is the more stable feature of the two.

When calculating the probabilities (henceforth p-values), the variables involved are the different possible values of a feature (we label these *a*, *b*, *c*, …), the number of possible values, *k*, and the number of languages in the set, *n*. The number of times that the best represented feature value occurs is labelled *r*.

As an illustration, we provide the example in Table 1. Here we have a feature with 2 possible values (*a* or *b*). Thus $k=2$. There is a set of 4 languages ($n=4$). Table 1 provides all the logically possible distributions and the corresponding value of the best represented feature values, *r*.

Table 1. An example of distributional possibilities when k = 2 and n = 4.

Distribution	r	Distribution	r
aaaa	4	bbab	3
bbbb	4	bbba	3
aaab	3	aabb	2
aaba	3	abab	2
abaa	3	abba	2
baaa	3	baab	2
abbb	3	baba	2
babb	3	bbaa	2

[5] This assumption begs the questions of areal convergence and universal preferences for particular typological features, but toward the end of this section we discuss why, for the present purposes, it may nevertheless be sustained.

Since there are 16 logically possible distributions, the different values of *r* have the following probabilities attached to them:

r	probability
4	2/16
3	8/16
2	6/16

There are two ways that the p-value could be calculated. One is to generate a table like Table 1 for the different values of *k* and *n* and then go through the table for given values of *r* and calculate the p-value. We might call this the 'brute force' approach. Another way would be to derive a general mathematical formula for calculating these values. We have chosen the former, less elegant approach. Thus, we profited by a computer program written by David Kamholz in Perl which simply calculated the p-values for each feature and genus in *WALS*. By averaging the p-values found for each feature a ranked list of features were generated where the (averaged) p-value is inversely proportional to the rank-order of the corresponding feature in terms of its usefulness for genealogical analyses.[6]

It is important to keep in mind that we are dealing with features strictly and only as defined and attested in *WALS*. If a particular feature were defined differently or were documented differently it would also have a different p-value. While the p-values are useful for deciding which features to select for a phylogenetic investigation we should also emphasize that one should be careful not to blindly select a set of features to use on the basis of p-values alone. The features of *WALS* are structured in such a way that some are problematic even if they may be high-ranking in terms of feature values. The two major problems with the ways that data are encoded in *WALS* are what we might term 'the interdependency problem' and 'the wastebasket problem'.

The interdependency problem crops up with some features that refer to one another. For instance, 'Relationship between the Order of Object and Verb and the Order of Relative Clause and Noun' (*WALS* Ch. 96 = Dryer 2005c) combines data from 'Order of Object and Verb' (*WALS* Ch. 83 = Dryer 2005d) and 'Order of Relative Clause and Noun' (*WALS* Ch. 90 = Dryer 2005e). If both feature no. 96 and one of the features no. 83 or no. 90 were drawn upon to make inferences

[6] The approach whereby we simply average the p-values was licensed by a one-way ANOVA. This statistical test was used to estimate whether there are significant overall differences in variance between p-values within a given feature as opposed to between features. In order to avoid the controversial issue of how to deal with empty cells when doing this test we reduced the dataset to the maximal possible size not containing empty cells. In this set there were p-values for 65 features and 7 genera. The result was highly statistically significant. More detail, as well as results of other statistical test are given in Wichmann & Kamholz (forthcoming).

concerning relationships among languages, then the values for the features object-verb word order or relative clause-noun word order would be represented twice in the data matrix, giving extra weight to these features — something which is not necessarily warranted. In such cases one must exclude one or more features such that overlap does not occur.

The wastebasket problem involves features that contain an 'other'-value, which will group languages together that do not share a positively defined trait; such a negatively defined feature-value will obscure relationships. An example would be the value 'marked by mixed or other strategies' of the feature 'Distributive Numerals' (*WALS* Ch. 54 = Gil 2005). In this case, one has to encode scores for the 'other'-value as gaps in the data rather than as true character states.

As mentioned earlier in this section, the logic by which the p-values are derived hinges upon the assumption that the probability of finding one particular character state out of x possible ones manifested in a given language is $1/x$. If this were a statement about how languages actually behave, it would be highly problematic. We know that some character states are heavily geographically skewed (a phenomenon which we might label 'the areal factor'), and we also know that certain character states are more widespread in human languages at large than others ('the universals factor'). We believe, however, that the areal factor in the end may not constitute as great a problem as one might think, and that there are ways of compensating for the universals factor or perhaps even reasons to ignore it.

As regards the areal factor it is instructive to leaf through *WALS*. It quickly becomes clear that no two distributional maps are quite the same. This is expected since it is well known that even the best established 'linguistic areas' such as the Balkans (Sandfeld 1930), India (Emeneau 1956), Arnhem Land (Heath 1978), Mesoamerica (Campbell et al. 1986), the Circum-Baltic area (Koptjevskaja-Tamm & Wälchli 2001), Amazonia (Aikhenvald & Dixon 1998), and Europe (Dahl 1990, Haspelmath 1998), are defined on the basis of just a small handful of typological character states, and that even such areas have fuzzy boundaries rarely exhibiting neat alignments of the isoglosses that define them. If genealogies were to be established on the basis of a small handful of typological characters, areal effects could skew the results heavily, but the greater the number of characters used, the more such effects should be expected to neutralize one another because of the differential areal distributions of the states of different characters. This is why we think that in the context of the way that the p-values are meant to be applied it is viable to operate with the $1/x$-probability assumption.

As for the universals factor, this may be compensated for in a parsimony analysis by introducing step matrices specifying that a change to a given character state should be penalized in inverse proportion to how often it occurs among human languages at large, cf. explanations in §5.2 below. We are not sure, however, that we want to subscribe to the necessity of such a compensatory measure. In the end, the universals factor may just be a special instance of the areal factor. If we view the entire world as just another large area — the largest, it so happens — it becomes clear that there is no a priori reason to assume that the languages which it contains should exhaust the possibilities of human language or that particularly widespread typological character states — so-called 'statistical universals' — will necessarily be more preferred biologically than others. To really be able to make claims about universally preferred character states we would need evidence independent of world-wide distributions. In the absence of such evidence we may consider the universals factor as part and parcel of the areal factor.

3. A test sample

The *WALS* data on languages of the Americas are far from complete enough to make any actual inferences concerning historical relationships among these languages. They do, however, constitute a good sample for the purpose of testing various phylogenetic methods and for gauging to what extent an amplification of the dataset would be needed for making interesting empirical inferences in the future. The database contains data for 621 Native American languages (not including Eskimo-Aleut). For these languages there are a total of 15,046 data points, which means an average of 24.2 data points per language. The representation of the various languages and the extent to which different features have been investigated varies greatly. There is only a relatively small set of languages for which enough features have been investigated that comparisons may be expected to yield interesting results. For instance, there are only 70 languages that have been investigated for 60 or more features, and these languages are widely dispersed over the continent(s).

Given the limitations of the dataset, we have selected a small sample from the languages of the Americas. Since a major aim of this paper is to investigate the utility of typological features for detecting genetic relatedness, our sample includes languages that are known to be related. We have sampled pairs of languages whose status as members of one and the same family is undisputed among specialists. Another criterion was that the languages selected be well attested. By these two criteria, we ended up selecting six pairs of related languages, all of which were among the top fifth in the Americas in terms of the number of features for which they were investigated. The languages in question are the Athapaskan languages Slave and Navajo, the Chibchan languages Ika and Rama, the Aymaran languages Aymara and Jaqaru, the Uto-Aztecan languages Yaqui and Comanche, the Oto-manguean languages Chalcatongo Mixtec and Lealao Chinantec, and the Carib languages Hixkaryana and Carib.

In the following section we simultaneously exercise our dataset and discuss some features of different algorithms exemplifying the range of algorithms developed to date. All of the trees shown in this section, except the one in Figure 1, are based on the same selection of just 17 *WALS* features.[7]

4. Introducing four different phylogenetic algorithms

4.1 Neighbor-joining

Neighbor-joining, an algorithm developed by Saitou & Nei (1987), is fast and practical for a high number of taxa. Like other clustering algorithms, it produces a distance matrix from the data and builds up the tree starting by uniting the two closest taxa under a node. It then computes new distances where the node just added is treated as a single taxon replacing the two original taxa. This process is repeated until a whole tree is produced. The resultant tree has branch lengths indicating relative distances. In Figure 1 we have provided so-called bootstrap values. These values give a statistical measurement of the amount of support for each node (Felsenstein 1985). A single bootstrapping procedure consists in making *n* random samples with replacement of whole characters from the set of *n* characters in a given data matrix. This means that one and the same character may be sampled more than once and others left out. On the basis of this sample a tree is constructed. The procedure is repeated many times — say 10,000 times. That produces a collection of 10,000 trees. One can now count how often a given node in the original tree recurs among the trees in the collection and thereby derive an idea of how well supported each node is.[8] Without such bootstrap values, the

[7] The features selected are the 17 highest ranking in terms of p-value which additionally satisfy the following criteria: at least 10 out of the 12 languages should be attested for the feature selected, the features must not be non-informative (i.e. having the same value for all languages or for all but one language), and the features should not suffer from interdependency in the sense that that their definitions overlap. We have not required the features to be independent in the sense that they are free of mutual statistical implication relations. Notably, in our selection of features there are several features of affix and constituent order which correlate to various degrees. Given the recent availability of an exact method for quantifying implicational relations (Holman, manuscript) it would now be possible select features that are completely free of mutual implicational relations. It is not certain, however, that features which are independent in this sense are necessarily to be preferred since there is a tendency for such features to also be less stable. More testing is required in this area. See Appendix for the data matrix and short feature descriptions.

[8] The term 'bootstrapping' derives from the saying 'to lift oneself by one's own bootstraps' and hints at the fact that the method does not involve an external yardstick but, so to speak, evaluates

tree in Figure 1 would be deceptive, suggesting, as it does, that the relationships among the 12 languages can and have been conclusively resolved. Given the small amount of data (17 features), it is encouraging that with the exception of Ika and Rama all the pairs of languages that are actually related are joined under exterior nodes even if the Otomanguean and Athapaskan branches are the only ones to be supported by bootstrap values in the 95% range, which is usually considered indicative of strong statistical significance. The two nodes exhibiting values in the 80–90% range are moderately well supported. But the extremely low bootstrap values found at all but one of the interior branches of the tree indicate that the data give no strong support for any relatedness among the 6 language pairs, except perhaps between Uto-Aztecan and Aymaran.

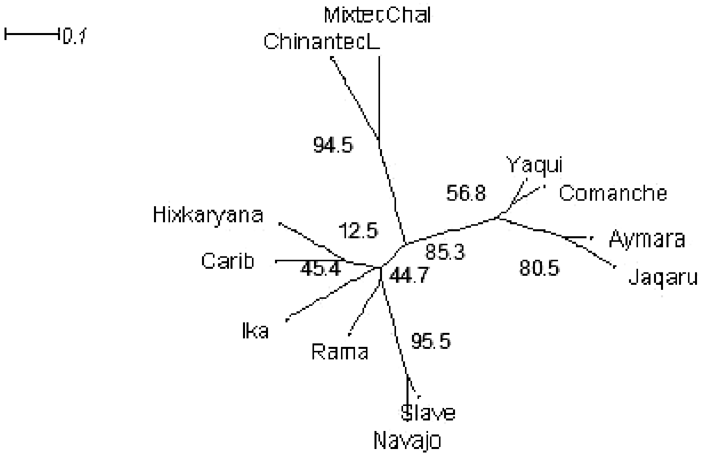

Figure 1. Tree based on 17 highest-ranking features produced by Neighbor-joining with bootstrap values (10,000 runs) using SplitsTree4 (Huson & Bryant 2006).

In Figure 2 we have made a quick test of the validity of our p-values (cf. §2) by selecting the 17 lowest-ranking features.[9] In the resulting phylogeny not only the two Chibchan languages Ika and Rama but also the Athapaskan languages Slave and Navajo as well as the Uto-Aztecan ones Comanche and Yaqui are divorced.

the data by the data themselves. For presentations of this and related statistical procedures see Felsenstein (2004: Ch. 20).

[9] That is, the lowest-ranking features which simultaneously satisfy the same criteria as the ones chosen as the best ones (cf. Appendix). Since one of the criteria is that no more than two languages must be unattested for a given feature we are actually forced to include some relatively high-ranking features. The ones selected are: 1, 2, 3, 4, 35, 49, 71, 73, 77, 91, 100, 102, 103, 107, 113, 114, 131.

This shows the usefulness of the p-values, and at the same time gives an example of the kind of quick test for which Neighbor-joining is very useful.

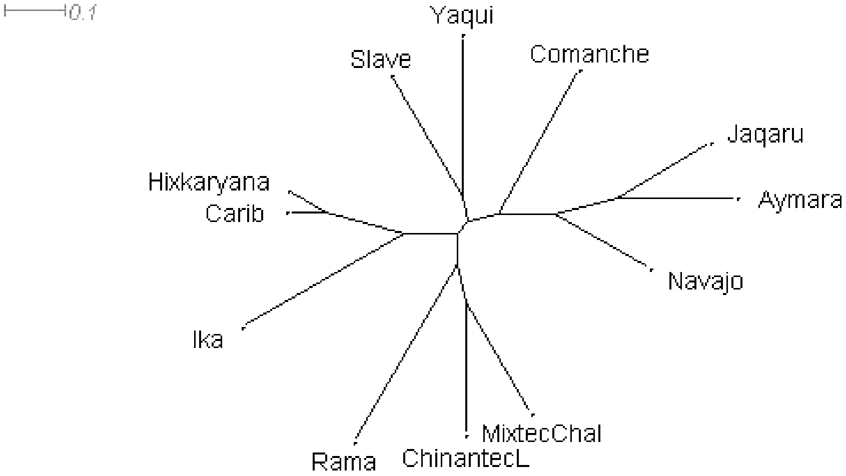

Figure 2. Tree produced by Neighbor-joining based on 'worst' p-valued features and using SplitsTree4.

4.2 Neighbor Net

Whereas most methods will impose a tree on data regardless of the extent to which the data actually map on to a tree-like phylogenetic evolution, the SplitsTree4[10] implementation of the Neighbor Net method allows for an effective visual method of depicting network structures that are truer to the data. In such a network, the problem of finding an optimal tree is left unresolved when it is in fact not resolvable. Instead, alternative trees are suggested. These various trees may be arrived at by collapsing the parallel edges of the structure in all possible ways. As a quick way of getting an overview of the degree to which the data conform to a phylogeny, this method is very effective. Like the bootstrap values in Figure 1, but perhaps somewhat clearer, the network-like structure of the representation in Figure 3 shows that the data cannot tell us much about possible relationships among the six pairs of languages in our sample.

[10] SplitsTree4 may be downloaded from www.splitstree.org.

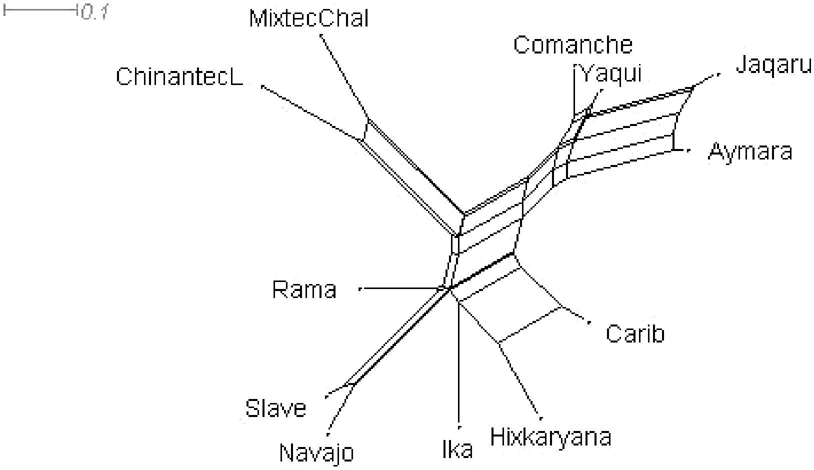

Figure 3. A Neighbor Net representation produced by SplitsTree4.

4.3 Maximum parsimony

Parsimony methods were among the first to be implemented in the early days of modern phylogenetics about half century ago and are still widely used. They are fast and logically quite transparent. The idea behind such methods is to find a tree which accounts for the data using the fewest number of changes. This Occam's Razor principle is familiar to historical linguists who, when they produce phylogenies from phonological evidence, also try to arrive at reconstructions that account for the data with a minimum of changes and then shape their family trees accordingly. It is a question, however, whether the principle is equally adequate for typological data. When using sound changes for building trees we let certain changes unique to subsets of taxa define the various nodes in the tree. But when using typological data, all taxa are typed for some state (value) of all characters (features). For this reason and also because of the somewhat rough and generic character of the data, there will be many independent innovations and a good deal of fluctuation back and forth among character states between nodes. Therefore we need as much data as possible and a method which increases its potential for arriving at an adequate tree the more data we add. But with parsimony methods adding data may actually decrease precision. For instance, branches characterized by many changes will tend to attract one another and appear more closely related than they necessarily are — a phenomenon known as 'long branch attraction'. Felsenstein (2004: 122) concludes a penetrating discussion of the statistical properties of parsimony by stating that "parsimony will work particularly well for recently diverged species whose branch lengths are not long." In spite of the disadvantages of parsimony methods they are difficult to ignore since, unlike other methods, they allow for constraining an algorithm such that it takes into account known tendencies for changes among individual characters. Since parsimony involves a calculation of cost in terms of the 'steps' or changes needed to transverse the tree it is possible to stipulate that certain changes are more costly than others, making the tree-construction process conform to particular hypotheses about changes in linguistic structure. Thus parsimony may be used to test such assumptions, an issue to which we return in §5 below.

In Figure 4 we show a cladogram of our 12 sample languages constructed by a maximum parsimony algorithm implemented in the program PAUP* (and using TreeView[11] for graphic editing).

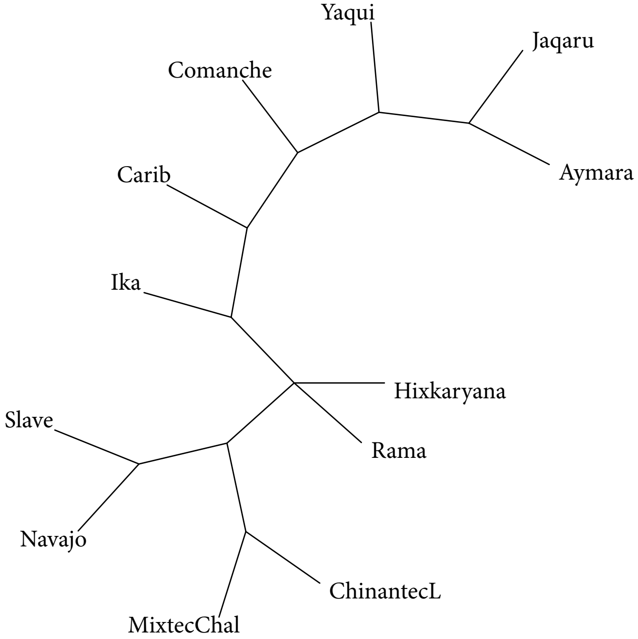

Figure 4. A Maximum Parsimony analysis of the dataset using PAUP*.

The unrooted cladogram in Figure 4 is deficient in that it falsely unites the Carib language Hixkaryana and the Chibchan language Rama. Only three pairs of languages known to be related, i.e. Chinantec-Mixtec (Otomanguean), Navajo-Slave (Athapaskan), and Jaqaru-Ayamara (Aymaran), are correctly joined. It looks like parsimony is not an adequate method for the type of data we are using here. We have obtained similarly poor results by looking at datasets involving other American and Austronesian languages and feel that these explorations are sufficient to

[11] TreeView is downloadable from http://taxonomy.zoology.gla.uk/rod/treeview.html.

cast serious doubt on the validity of parsimony for the purpose of building phylogenies from linguistic data (Cysouw et al. in 2006, Saunders 2006). In §5 we return to parsimony in order to test whether the results improve when we add weighting or step matrices.

4.4 Bayesian analysis

Bayesian inference of phylogeny is executed through an algorithm which searches through a space of possible trees, preferably steering toward those trees which maximize a value called the 'posterior probability'. The 'posterior probability' is a numerical evaluation of the probability that a given tree is the correct one for the data, and is formulated from Bayes' Theorem. Bayes' Theorem allows the probability of one event to be computed from observations of another event and knowledge of their joint distributions (Huelsenbeck & Ronquist 2001). For phylogenetic inference, the event of interest is the probability of the tree given the data, which can be described with the following formulation of Bayes' Theorem:

$$P(\text{tree+model}|\text{data}) = P(\text{data}|\text{tree+model}) \times P(\text{tree+model}) / P(\text{data})$$

Here the probability of the tree (and the model used to generate the tree) given the data is equal to the likelihood of the tree and model, multiplied by the probability of the tree and model itself, and then divided by the probability of the data. The P(data|tree+model) can be calculated through established maximum likelihood methods, while the P(tree+model) is assumed through the prior probabilities defined by the researcher. The values assigned for the prior probabilities determine which trees have a high 'posterior probability'; it is this subjectivity that makes Bayesian methods controversial. The problematic term, P(data), has been dealt with computationally through the use of Markov chain Monte Carlo (MCMC) algorithms.

Starting with an arbitrary tree in the defined space of possible trees, MCMC algorithms select a neighboring tree at random. Instead of calculating the 'posterior probabilities' for each individual tree, the algorithm compares the ratio of likelihood between the current and the neighboring tree. In comparing this ratio, the P(data) term cancels out. If the neighboring tree has a higher likelihood (i.e. if the likelihood ratio of $T_{\text{neighbor}}/T_{\text{current}}$ is >1), then the neighboring tree becomes the current tree. If the likelihood ratio is <1, then the algorithm compares that ratio value to a random number between 0 and 1. If the ratio is higher than this random value, the current tree is kept; if not, another neighbor tree is selected and the process continues.

The ratio comparison and tree selection steps make up one iteration of the algorithm. As the iterations increase, the 'posterior probability' of the sampled trees is also prone to increase. This progression is reported by the software through 'posterior probability' values collected at a consistent rate (i.e. every 100 iterations). As the sampling continues, a point of convergence is reached when the 'posterior probability' of subsequent trees fails to improve. After this point, the algorithm collects trees with near equal probability. Once the desired number of iterations is complete, the pre-convergence trees are thrown out. From the remaining sample of optimal trees, a consensus tree is assembled following a basic rule: only include those groups which exist in a more than a certain threshold percentage (50% and above) of optimal trees. Each expressed node has a 'posterior probability score' to represent its strength, defined as the probability of finding that node in the set of optimal trees. Branch lengths are similarly derived from the set of optimal trees.

Bayesian inference of phylogenies is implemented in the software package MrBayes.[12] Under default conditions, MrBayes uses four tree-searching algorithms simultaneously and in concert. Three of these algorithms are considered 'hot' and allowed to make large jumps in tree space to find neighbors. One chain is always 'cold' and is constrained to make only local comparisons. After every generation of the program, the chain with the highest 'posterior probability' becomes the cold chain. Together these chains avoid locally optimal areas of 'posterior probability' and hone in on those global areas of tree space containing the most probable trees.

Bayesian analysis is at the cutting edge of phylogenetic algorithm development. It neither has the kind of conceptual simplicity that characterizes distance or parsimony analysis, nor is it as speedy. In fact, when the number of taxa and characters run into the dozens it may take days for even a fast computer to perform the analysis. Nevertheless, it may be well worth considering in building linguistic phylogenies, as our initial explorations suggest that it could be superior to other methods.

Figure 5 shows an unrooted phylogram based on our dataset constructed using MrBayes (edited in TreeView). We made the program run 10 million generations, sampling every 100. That produced a total sample of 100,000 trees. Of these, only the last 25,000 were kept, which means a 'burn in' of 75,000 trees. From the 25,000 subset a 50% majority rule consensus tree was made. The analysis took around 41 hours on a G4 iMac. The resulting tree at first glance seems a bit confusing because of its not very tree-like structure. But on further inspection it becomes clear that it represents the relationships among the languages more adequately than any of the other methods tested. First, all pairs of truly related languages are in fact joined as daughters under immediately shared nodes. The tree shows strong support for the pairs Mixtec-Chinantec, Navajo-Slave, and Jaqaru-Aymara.

[12] MrBayes is downloadable through http://mrbayes.csit.fsu.edu/index.php.

For Comanche-Yaqui there is some weak support. For Carib-Hixkaryana and Ika-Rama the tree is neutral: in an unrooted tree a star-like shape (*) reflects absence of specified relationships, so the tree neither supports nor denies the possibility of relationships among these four languages. Thus, for all the relationships mentioned so far the tree is either correct or neutral. Except in one case its topology does not make statements that could be wrong. The one surprise in the topology is the node that attaches the Jaqaru-Ayamara pair to the Uto-Aztecan node of Yaqui and Comanche. We would expect a branch going from directly from the Jaqaru-Aymara node to the central one. But apparently there is support for this relationship in the dataset at hand.

It is very instructive to compare the Bayesian tree in Figure 5 to the one produced by Neighbor-joining in Figure 1. If all the branches of Figure 1 that have bootstrap values below 80 are pruned away one arrives at a tree having the exact same topology as the one in Figure 5 (ignoring branch lengths). This suggests that Neighbor-joining and Bayesian analysis are quite compatible, the major difference being that the latter method is more conservative and does not impose topological structure when support for such structure is weak. Thus, while Neighbor-joining has to be used in conjunction with bootstrapping and its resulting trees revised in light of these values, Bayesian analysis, as implemented in MrBayes, directly shows strongly supported hypotheses reflected in topology and avoids positing structure for which there is little support.

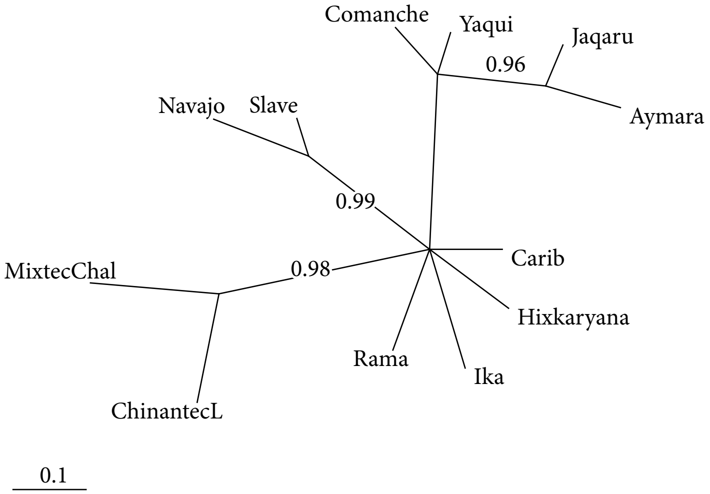

Figure 5. A Bayesian analysis of the dataset using MrBayes.

4.5 Conclusions regarding methods

The four different methods tested are all useful, but for different purposes. Neighbor joining is a quick method for inspecting the structure of a phylogeny for a large dataset, but should be used with caution and be supplemented with bootstrap analysis. Neighbor Net, like similar methods implemented in SplitsTree, is another fast method and the visually most effective, directly showing in the graphic representation how tightly knitted a phylogeny is. Parsimony may be the least adequate method for the kinds of data that we have been looking at in this paper, but could potentially be valuable for the analytical purpose of testing the effects of different methods for enhancing phylogenetic signals (see the next section). Bayesian analysis is probably the type of method which is most adequate for linguistic typological data and the results seem to directly make for strong hypotheses concerning actual genealogical relationships. Often it will be useful to run a dataset through all the different methods (for instance, in the order in which they were presented here), but it seems that the firmest conclusions are reached by means of Bayesian analysis.

5. Methods for enhancing phylogentic signals

5.1 Weighting

If some characters are more suited for establishing genealogies than others it should ideally be possible to enhance phylogenetic signals by giving more weight to more suitable characters than to less suitable ones. PAUP* is among the programs that allows for such weighting schemes, which require the use of parsimony. We have tested the effect of weighting characters according to their p-values (cf. §2 above). Since, as already mentioned, the p-values roughly decrease linearly by each step in the rank-order, we have applied a simple weighting procedure by which the lowest-ranking feature — the one with the highest p-value — carries a numerical weight of 1 and the highest-ranking feature among the 139 relevant ones carries a weight of 139. Testing this method on various sets of languages of the Americas consistently shows little effect on tree topologies. Weighting nevertheless does affect the phylogenetic analysis in subtle ways, as revealed by bootstrapping. Thus, it may be informative to compare bootstrap values for trees produced without as opposed to with weighting. The weights would seem to be correctly set when 'wrong' nodes receive decreased support and 'right' nodes increased support; inversely, when 'right' nodes are weakened and 'wrong' ones strengthened something must be wrong with the weighting scheme. A comparison of Figures 6 and 7 provides an example of the effect of weighting. As can be seen, the differences are quite small.

```
/--------------------------------------------------------------- Aymara
|
|                               /------------------------------- Carib
|                               |
|                               |              /---------------- ChinantecL
|                               +------85------+
|                               |              \---------------- MixtecChal
|                               |
|                               +------------------------------- Hixkaryana
|               /------85-------+
|               |               +------------------------------- Ika
|               |               |
|               |               |              /---------------- Navajo
|               |               +------88------+
|               |               |              \---------------- Slave
+------72-------+               |
|               |               \------------------------------- Rama
|               |
|               +----------------------------------------------- Comanche
|               |
|               \----------------------------------------------- Yaqui
|
\--------------------------------------------------------------- Jaqaru
```

Figure 6. Tree produced by default parsimony method of PAUP* without weights.

```
/--------------------------------------------------------------- Aymara
|
|                               /------------------------------- Carib
|                               |
|                               |              /---------------- ChinantecL
|                               +------85------+
|                               |              \---------------- MixtecChal
|                               |
|                               +------------------------------- Hixkaryana
|               /------79-------+
|               |               +------------------------------- Ika
|               |               |
|               |               |              /---------------- Navajo
|               |               +------92------+
|               |               |              \---------------- Slave
+------72-------+               |
|               |               \------------------------------- Rama
|               |
|               +----------------------------------------------- Comanche
|               |
|               \----------------------------------------------- Yaqui
|
\--------------------------------------------------------------- Jaqaru
```

Figure 7. Tree produced by default parsimony method of PAUP* with weights based on p-values.

The two trees have the same — mostly wrong — topology. In two cases, however, bootstrap values change, and changes in both cases go in the right directions. Thus, the node falsely uniting Carib, Chinantec, Mixtec, Hixkaryana, Ika, Navajo, Slave, and Rama is weakened from 85 to 79 and the correct node united Navajo and Slave is strengthened from 88 to 92.

On the basis of these initial explorations we conclude that weighting is not enough to 'save' a parsimony method. On the other hand, testing the effects of weights may be a valuable analytical tool in its own right. Thus, assumptions about the relative merits of different features as input to genealogical analyses may be tested by translating the assumptions into quantitative weights and then seeing how the results are affected.

5.2 Step matrices

Since maximum parsimony seeks out the tree that overall requires a minimal amount of changes it is possible to give different weights to different changes, 'penalizing' changes that are presumed to be more unexpected. For instance, we might stipulate that the expression of some linguistic category by means of a free particle switching to a proclitic is to be regarded as a single step, while a change involving going from a free particle to a prefix might be regarded as two steps. Similarly, a word order change from VSO to SVO might be one step (by fronting of S) whereas a change from VSO to OVS should be harder and might count as two or more steps. Even if directions of change in typology are often unknown, we can in most cases use simple intuition about the 'costs' of changes and, in so doing, construct sensible 'step matrices.' For each feature, a step matrix should describe the relative number of steps involved in changes among all feature values. We have conducted several experiments constructing step matrices and testing their effects. The tests involved both matrices that simply stipulated steps based on logical reasoning (as described) as well as matrices taking into account the world-wide distribution of different feature values, adding or subtracting fractions of steps in proportion to the percentage of languages in the world which exhibit the given feature value. None of these exercises have conclusively proven step matrices to have either a positive or a negative effect. It is difficult at this point to judge whether the disappointing results relate to the way that we have constructed the matrices or to the datasets. In any case, introducing step matrices into parsimony analyses could be a potentially useful tool for making assumptions about directions of change explicit and testing the effects of such assumptions.

6. Using typological data as a heuristic tool

The results presented so far suggest that modern, computationally-driven phylogenetic methods should have the potential for establishing genealogies based on typological data. It is certainly clear that they may minimally be used as a heuristic tool for quickly disclosing candidates for genealogical relatedness or typologically convergent languages in a large dataset. Although the *WALS* dataset from the Americas is far from extensive enough to allow for a actual empirical hypotheses of this kind, we may nevertheless use the data to illustrate the type of heuristic procedure that might profitably be employed in future research.

In Figure 8 we show a tree based on *WALS* data from the 63 best attested languages[13] of the Americas and the 96 highest-ranking features in terms of p-values.

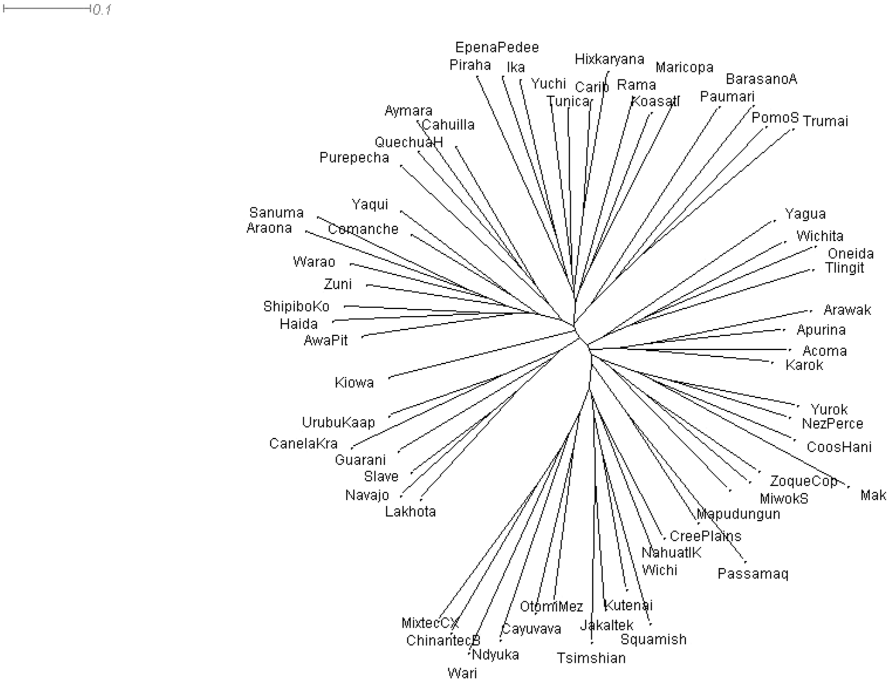

Figure 8. Neighbor-joining tree of 63 languages of the Americas using 96 *WALS* features.

The many nodes on the tree are mostly wrong. So many, in fact, are wrong that it would require too much space to enumerate them all here. A few, however, are correct. For yet others the limitation in our present knowledge of the classification of the languages of the Americas prevents us from making definite judgments. Given the uncertainty about the classification of many of the languages it is impossible to do exact statistics on how well the tree performs. But given 93 taxa it is possible to

[13] The software imposes limitations on the symbols that can be used for naming taxa. Higher ascii characters and spaces are disallowed. Therefore some language names have been changed as follows: Apurina = Apurinã, AwaPit = Awa Pit, BarasanoA = Barasano, CanelaKra = Canela-Krahô, ChinantecB = Lealao Chinantec, CoosHani = Hanis Coos, CreePlains = Plains Cree, EpenaPedee = Epena Pedee, Guarani = Guaraní, MixtecCX = Chalcatongo Mixtec, MiwokS = Southern Sierra Miwok, NahuatlK = Tetelcingo Nahuatl, NezPerce = Nez Perce, OtomiMez = Mezquital Otomi, Passamaq = Passamaquoddy, Paumari = Paumarí, Piraha = Pirahã, PomoS = Southeastern Pomo, Purepecha = Purépecha, QuechuaH = Imbabura Quechua, ShipiboKo = Shipibo-Konibo, UrubuKaap = Urubú-Kaapor, Wari = Wari', Wichi = Wichí, ZoqueCop = Copainalá Zoque.

construct a staggering number of more than 8×10^{100} different unrooted bifurcating trees (see Felsenstein 2004: 25 for how to calculate this), so it would seem that just getting 2–3 correct nodes is more than we would expect by pure chance.

There might be interesting hypotheses scattered throughout the tree, so it would seem too rash to throw out the entire tree just because many of its nodes are problematic. A way of rescuing potentially useful information is to combine the data matrix on which the tree is built with accumulated knowledge concerning genealogical relations, that is, the results obtained by generations of linguists who have employed traditional lexical and morphological comparisons. Some programs, such as PAUP*, allow for directly specifying the nodes that should be found in the output tree regardless of the topology licensed by the data. In Splitstree, the program we have used for producing Figure 8, nodes may be forced by adding a sufficient number of 'ghost' columns to the data matrix. Each column represents some imaginary character and has the same state for the taxa that one wishes to unite under a single node, while other taxa are left unspecified for the character in question. This method actually makes good sense from a linguist's point of view because each 'ghost' column may be imagined to represent some lexical item uniquely present in the set of languages known to be related. In reality it should not be hard to find such items if the languages thus forced together are truly related.

Figure 9 shows a reshaping of the tree of Figure 8, accomplished by forcing eight nodes known to exist, while also marking correct genealogical nodes that emerged from the data (three nodes are controversial and are therefore supplied with question marks).

The utility of a tree with forced nodes is that it allows us zoom in on the nodes that are not either forced or already known or thought to correspond to real relationships; the languages grouped under these nodes may now be investigated for possible relations of either a genealogical or areal nature that have never earlier been considered.

Given the limitations of the data we shall not indulge in discussions of each of the various nodes. For the sake of illustrating the procedure, however, we would like to briefly comment on one of them. This is the node uniting Mapudungun and Copainalá Zoque. The former is a language with many speakers in Chile and Argentina and is usually regarded as a linguistic isolate; the latter is a dialect of Chiapas Zoque, a Mixe-Zoquean language of Chiapas, Mexico. Among the distant relatives involving Mixe-Zoquean which have been proposed, Penutian has figured prominently (beginning with Sapir 1929, cf. Wichmann 1994:238–243 for an overview of related as well as other proposals). In fact, in the unmanipulated tree of Figure 8 a node also includes the Penutian language Southern Sierra Miwok in a group together with Chiapas Zoque and Mapudungun. When the Penutian

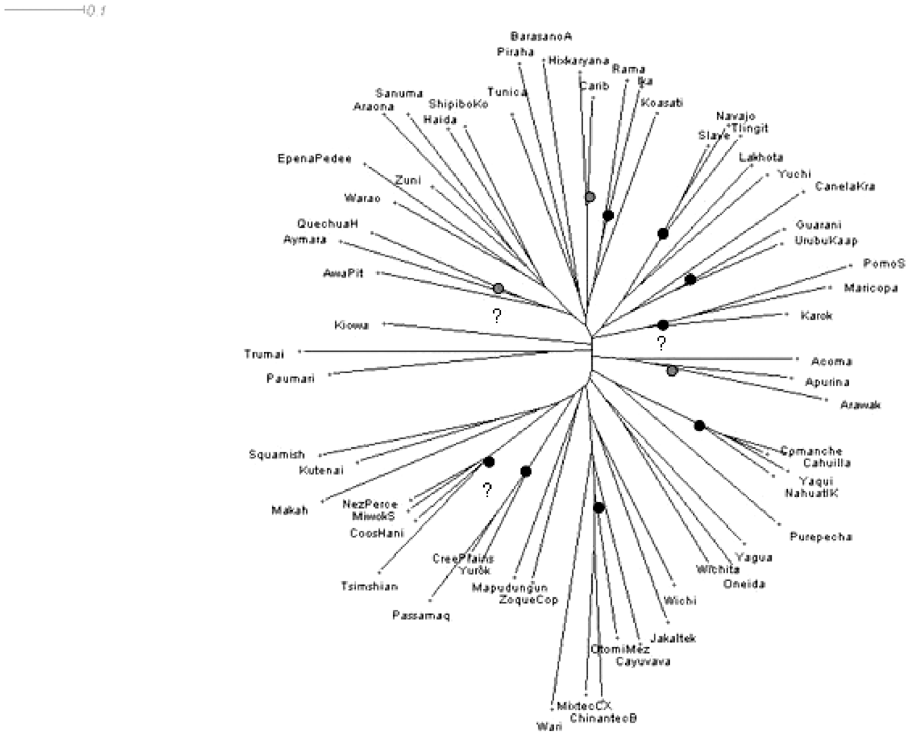

Figure 9. Neighbor-joining tree with forced nodes of 63 languages of the Americas (black dots indicate forced nodes, gray ones unforced genealogical nodes emerging from the data).

languages are forced together, however, Chiapas Zoque remains united with Mapudungun rather than migrating towards the Penutian node. In this way, a Chiapas Zoque-Mapudungun connection is singled out as a hypothesis. One might not have thought of the possibility of a relationship among these two languages — such a relationship has never, to our knowledge, been proposed — but if we follow the prompting of the heuristic tree and take a next step, which might involve searching for lexical parallels, there are some promising further leads. Searching for possible cognates among the items on the 100-word Swadesh list using the dictionary of Mapudungun (Map) of Fernández Garay (2001) and the proto-Mixe-Zoquean (pMZ) lexicon of Wichmann (1995), the following parallels emerge (pZ stands for proto-Zoquean, the ancestral language of one of the two major branches of Mixe-Zoquean; an apostrophe represents a glottal stop):

"I"	Map *inche* : pMZ *'ə:tzi
"who"	Map *iñey ~ iñe ~ iñ ~ ñi* : pZ *'iyə
"what"	Map *chem* : pMZ *ti

“woman”	Map *domo* : pZ *yomo
“man”	Map *wentru* : pMZ *pən
“blood”	Map *mollvün* ~ *mollviñ* ~ *mollvin* : pMZ *nə’pin[14]
“head”	Map *longko* : pMZ *ko-pak[15]
“hand”	Map *küwü* : pMZ *kə’
“to sleep”	Map *umaw-* : pMZ *ma:h’
“name”	Map *üy* : pZ *nəyi

Among the comparisons several may be due to chance, but at least the three items “woman”, “hand”, and “to sleep” seem more than superficially similar. The number of possible cognates, both striking ones and less striking ones, is similar to the results one would find by comparing, say, English and Farsi.

It would be inappropriate to continue the exercise of comparing Mapudungun to Mixe-Zoquean here. We would like to stress that we have not attempted to demonstrate any genetic link. It would be necessary in the first instance to show that the typological parallels and the lexical look-alikes have strong statistical support. And for a conclusive demonstration we would like to be able to find cognate grammatical elements and systematic sound correspondences. Our only point of this little exercise was to describe a heuristic procedure for finding genetic links using an interesting illustration. The procedure may be improved in several respects to produce even better results. The most immediate need is to make the typological database more robust through the addition of data. In the final section we briefly discuss some further items for future research.

7. Conclusions and questions for future research

The present paper has offered the results of explorations of a typological dataset using computationally-driven phylogenetic software developed largely for the use of biologists. First, we have presented a method for quantifying the utility of different typological features for establishing linguistic genealogies. Then we tested four different types of phylogenetic algorithms. To very briefly summarize our findings: Neighbor-joining seems to be most useful for heuristic purposes; Neighbor Net and similar methods are particularly useful for visual analysis; parsimony analyses is a valuable tool for assessing the validity of assumptions regarding the weights of different characters or preferred directions of change among states, but the assumption that the preferred phylogeny is also the most parsimonious one seems

[14] This appears to be an ancient composite, where the word for ‘water’, **nə’*, combines with an element *pin* of unknown etymology.

[15] While *pak* means ‘bone’, *ko-* is the element referring to ‘head’.

to represent a drawback. Our preliminary probing suggests that Bayesian analysis has an advantage over all the other three methods, but this finding of course needs to be tested on other datasets. Finally, have we illustrated a simple, heuristic procedure for identifying hitherto unrecognized cases of genealogical relations.

As mentioned, the utility of typological databases for historical linguistic research cannot be fully assessed until more extensive databases have been constructed. Nor can we hope to bring our results to bear on actual empirical problems before relevant databases have been enlarged. The *WALS* database provides a good beginning, as long as the problems of overlap and 'wastebasket' categories are taken into account. Simply filling out holes in the *WALS* matrix for the set of languages that one would like to compare would already constitute a useful step forward. Eventually, more characters could be added. As should have transpired in our discussion of p-values, a character is likely to be more informative the more states it has. It may not be simple to find characters with several states that can be attested for any language, so it might be tempting to simply use binary ones. Nevertheless, we would recommend avoiding binary characters since their states are expected to be more prone to chance fluctuation than characters with several states.

A crucial issue which we have not directly addressed up to this point is the following: with what degree of confidence can we accept hypotheses concerning genealogical relationships generated by a given algorithm on the basis of typological data? The way to answer this question would be to compare trees constructed from the results of the application of the comparative method with trees constructed from typological data. Variables that should be taken into account are the number of languages involved, the time depth, the number of characters used, and the p-values of the characters. By holding constant some variables and changing others it should be possible to arrive at estimates of confidence. A good deal of theoretical work will need to go into exploring adequate ways of comparing trees; but again the biological literature will be helpful since a by now classical issue for biologists is to assess differences among phylogenies produced by traditional morphological methods (which are comparable, in many respects, to the use of linguistic typological data) and those produced by molecular systematics (comparable, to a certain extent, to lexical comparisons). Before long, we expect that providing estimates of confidence will become an entrenched part of the way that historical linguistics is practiced, just as it is in most other branches of science. Once this happens, emotionally charged arguments from beliefs about what it takes for a given genealogical relationship to be 'proved' may be avoided and replaced by cooler, statistical reasoning.

References

Aikhenvald, Alexandra Y. & Robert M. W. Dixon. 1998. “Evidentials and areal typology: a case study from Amazonia”. *Language Sciences* 20.241–257.

Als, Thomas D., Roger Vila, Nikolai P. Kandul, David R. Nash, Shen-Horn Yen, Yu-Feng Hsu, André A. Mignault, Jacobus J. Boomsma & Naomi E. Pierce. 2004. “The evolution of alternative parasitic life histories in large blue butterflies”. *Nature* 432.386–390.

Atkinson, Quentin, Geoff Nicholls, David Welch & Russell Gray. 2005. “From words to dates: Water into wine, mathemagic or phylogenetic inference?” *Transactions of the Philological Society* 103.193–219.

Baker, Richard H. & Gatesy, John. 2002. “Is morphology still relevant?” *EXS* 92.163–174.

Bickel, Balthasar & Johanna Nichols. 2005. “Inflectional synthesis of the verb”. Haspelmath et al., eds., 94–99.

Campbell, Lyle, Terrence Kaufman & Thomas C. Smith-Stark. 1986. “Mesoamerica as a linguistic area”. *Language* 62.530–570.

Cysouw, Michael, Søren Wichmann & David Kamholz. 2006. “A critique of the separation base method for genealogical subgrouping, with data from Mixe-Zoquean”. *Journal of Quantitative Linguistics* 13.225–264.

Dahl, Östen. 1990. “Standard Average European as an exotic language“. In Bechert, Johannes, Giuliano Bernini & Claude Buridant, eds., *Toward a Typology of European Languages*, 3–8. Berlin: Mouton de Gruyter.

Daniel, Michael & Edith Moravcsik. 2005. “The associative plural”. In Haspelmath et al., eds., 150–153.

Dryer, Matthew S. 1989. “Large linguistic areas and language sampling”. *Studies in Language* 13.257–292.

Dryer, Matthew S. 2005a. “Genealogical language list”. In Haspelmath et al., eds., 584–644.

Dryer, Matthew S. 2005b. “Order of subject, object, and verb”. In Haspelmath et al., eds., 330–333.

Dryer, Matthew S. 2005c. “Relationship between the order of object and verb and the order of relative clause and noun”. In Haspelmath et al., eds., 390–393.

Dryer, Matthew S. 2005d. “Order of object and verb”. In Haspelmath et al., eds., 338–341.

Dryer, Matthew S. 2005e. “Order of relative clause and noun”. In Haspelmath et al., eds., 366–369.

Dunn, Michael, Angela Terrill, Ger Reesink, Robert A. Foley & Stephen C. Levinson. 2005. “Structural phylogenetics and the reconstruction of ancient language history”. *Science* 309.2072–2075.

Emeneau, Murray B. 1956. “India as a linguistic area”. *Language* 32.3–16.

Felsenstein, Joseph. 1985. “Confidence limits on phylogenies: An approach using the bootstrap”. *Evolution* 39.783–791.

Felsenstein, Joseph. 2004. *Inferring Phylogenies*. Sunderland, Mass.: Sinauer Associates.

Felsenstein, Joseph. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Fernández Garay, Ana. 2001. *Ranquel-español/español-ranquel. Diccionario de una variedad mapuche de La Pampa (Argentina)*. Universidad de Leiden: Escuela de Investigación de Estudios Asiáticos, Africanos, y Amerindios (CNWS).

Forster, Peter & Alfred Toth. 2003. “Toward a phylogenetic chronology of ancient Gualish, Celtic, and Indo-European”. *Proceedings of the National Academy of Sciences of the United States of America*100:15.9079–9084.

Forster, Peter, Alfred Toth & Hans-Jürgen Bandelt. 1998. “Evolutionary network analysis of word lists: Visualising the relationships between Alpine Romance languages”. *Journal of Quantitative Linguistics* 5.174–187.

Gil, David. 2005. “Distributive numerals”. In Haspelmath et al., eds., 222–225.

Gray, Russell D. & Quentin D. Atkinson. 2003. “Language-tree divergence times support the Anatolian theory of Indo-European origin”. *Nature* 426.435–439.

Gray, Russell D. & Fiona M. Jordan. 2000. “Language trees support the express-train sequence of Austronesian expansion”. *Nature* 405.1052–1055.

Griffiths, Carole. 1999. “Phylogeny of the Falconidae inferred from molecular and morphological data”. *The Auk* 116:1.116–130.

Haspelmath, Martin. 1998. “How young is Standard Average European?” *Language Sciences* 20:3.271–287.

Haspelmath, Martin. 2001. “The European linguistic area: Standard Average European”. *Language Typology and Language Universals: An International Handbook*, ed. by Martin Haspelmath, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible, 1492–1510. Berlin: Mouton de Gruyter.

Haspelmath, Martin, Matthew S. Dryer, David Gil & Bernard Comrie, eds. 2005. *The World Atlas of Language Structures*. Oxford: Oxford University Press.

Heath, Jeffrey. 1978. *Linguistic Diffusion in Arnhem Land*. Canberra: Australian Institute of Aboriginal Studies.

Holden, Clare Janaki. 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, 269.793–799.

Holman, Eric W. Manuscript. “Approximately independent typological features of languages”.

Huelsenbeck, John P. & Fredrik Ronquist. 2001. “MRBAYES: Bayesian inference of phylogenetic trees”. *Bioinformatics* 17.754–755.

Huson, Daniel H. & David Bryant. 2006. “Application of phylogenetic networks in evolutionary studies”. *Molecular Biology and Evolution* 23.254–267.

Koptjevskaja-Tamm, Maria & Bernhard Wälchli. 2001. “The Circum-Baltic languages. An areal-typological approach”. *Circum-Baltic Languages*. Vol 2: *Grammar and Typology*, ed. by Östen Dahl & Maria Koptjevskaja-Tamm, 615–750. Amsterdam & Philadelphia: John Benjamins.

Lawrence, Jeffrey G. & Howard Ochman. 2002. “Reconciling the many faces of lateral gene transfer”. *Trends in Microbiology* 10:1.1–4.

Maddison, David R., David L. Swofford & Wayne P. Maddison. 1997. “NEXUS: An extensible file format for systematic information”. *Systematic Biology* 46:4.590–621.

McMahon, April, Paul Heggarty, Robert McMahon & Natalia Slaska. 2005. “Swadesh sublists and the benefits of borrowing: An Andean case study”. *Transactions of the Philological Society* 103.147–170.

McMahon, April & Robert McMahon. 2003. “Finding families: Quantitative methods in language classification”. *Transactions of the Philological Society* 101.7–55.

Nakhleh, Luay, Don Ringe & Tandy Warnow. 2005. “Perfect phylogenetic networks: a new methodology for reconstructing the evolutionary history of natural languages”. *Language* 81.382–420.

Nakhleh, Luay, Tandy Warnow, Don Ringe & Steven N. Evans. 2005. "A comparison of phylogenetic reconstruction methods on an Indo-European dataset". *Transactions of the Philological Society* 103.171–192.

Nylander, Johan A. A., Fredrik Ronquist, John P. Huelsenbeck & José Luis Nieves-Aldrey. 2004. "Bayesian phylogenetic analysis of combined data". *Systematic Biology* 53:1.47–67.

Ochman, Howard, Jeffrey G. Lawrence & Eduardo A. Groisman. 2000. "Lateral gene transfer among genomes". *Nature* 405.299–304.

Saitou, Naruya & Masatoshi Nei. 1987. "The neighbor-joining method: A new method for reconstructing phylogenetic trees". *Molecular Biology and Evolution* 4.406–425.

Sandfeld, Kristian. 1930. *Linguistique balkanique: Problèmes et résultats.* Paris: Honoré Champion.

Sapir, Edward. 1929. "Central and North American languages". *Encyclopaedia Britannica*, 14th ed. 5.138–141.

Saunders, Arpiar. 2006. *Linguistic phylogenetics of the Austronesian family: A performance review of methods adapted from biology.* B.A. Thesis, Swarthmore College.

Swofford, David L. 2002. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods).* Sunderland, Mass.: Sinauer Associates.

Warnow, Tandy. 1997. "Mathematical approaches to comparative linguistics". *Proceedings of the National Academy of Sciences of the USA* 94.6585–6590.

Wichmann, Søren. 1994. "Mixe-Zoquean linguistics: A status report". *Panorama de los estudios de las lenguas indígenas de México* (=*Biblioteca Abya-Yala*, 16), ed. by Doris Bartholomew, Yolanda Lastra & Leonardo Manrique, vol. 2, 193–267. Quito: Abya-Yala.

Wichmann, Søren. 1995. *The Relationship among the Mixe-Zoquean Languages of Mexico.* Salt Lake City: University of Utah Press.

Wichmann, Søren & David Kamholz. Forthcoming. "A stability metric for typological features". *Sprachtypologie und Universalienforschung.*

Résumé

Plusieurs bases de données ont été compilées afin de documenter la distribution de traits typologiques à travers les langues du monde. Ce travail cherche des façons d'utiliser ce type de données dans le but de trouver des relations généalogiques en utilisant des algorithmes phylogéniques initialement développés pour des biologistes. Les points forts en sont la méthodologie, y inclus l'évaluation de la stabilité des traits typologiques individuels, la pertinence de différents algorithmes ainsi que la mise en valeur des signaux phylogéniques et les procédures heuristiques afin d'identifier des relations généalogiques. Les différents points sont illustrés par un petit échantillon de données empiriques d'un certain nombre de langues amérindiennes.

Zusammenfassung

Mehrere Datenbanken sind zusammengetragen worden, um die Verteilung typologischer Merkmale auf die Sprachen der Welt aufzuzeigen. Diese Arbeit bemüht sich diese Art von Daten zu benutzen, um Rückschlüsse auf genealogische Beziehungen zu ziehen, indem sie ursprünglich für Biologen entwickelte phylogenetische Algorithmen benutzt. Ihr zentrales Anliegen ist dabei die Methodologie, die Evaluation der Stabilität individueller typologischer Merkmale und die Eignung verschiedener phylogenetischer Algorithmen, ebenso wie Wege phylogenetische Signale und heuristische Verfahren zu verbessern, um genealogische Beziehungen zu identifizieren. Die unterschiedlichen Punkte werden anhand ausgewählter empirischer Daten einiger Sprachen der Ureinwohner Amerikas illustriert.

Authors' addresses

Søren Wichmann
Department of Linguistics
Max Planck Institute for Evolutionary
Anthropology
Deutscher Platz 6
D-04103 Leipzig, Germany

E-mail: wichmann@eva.mpg.de

Arpiar Saunders
Linguistics Department
Swarthmore College
500 College Avenue
Swarthmore, PA 19081–1397, U.S.A.

E-mail: arpiar.saunders@gmail.com

Appendix

Data matrix showing the values of each feature for the selection of 12 languages (rows) and 17 *WALS* features (columns).

	6	7	8	26	27	33	44	51	57	69	81	85	86	87	89	101	111
Aymara	4	2	2	2	1	2	6	6	2	2	1	1	1	1	1	2	2
Carib	1	1	1	4	3	2	2	9	1	2	1	1	1	1	1	?	2
Chalcatongo Mixtec	1	1	2	6	?	7	6	9	4	1	3	2	2	2	1	6	2
Comanche	1	1	1	2	1	2	6	1	4	2	1	1	1	1	1	3	2
Hixkaryana	1	1	3	4	3	7	2	9	1	2	5	1	1	2	1	2	2
Ika	1	1	1	3	3	7	6	6	1	2	1	1	1	2	3	2	2
Jaqaru	2	2	2	2	1	2	6	1	2	2	?	?	1	1	1	?	2
Lealao Chinantec	1	1	2	5	2	7	6	9	2	4	4	2	2	2	1	2	2
Navajo	1	2	4	6	3	2	3	9	1	1	1	1	1	2	2	2	1
Rama	1	1	2	4	?	8	6	9	1	2	1	1	1	2	2	2	4
Slave	1	6	4	6	3	2	6	9	1	1	1	1	1	2	2	2	2
Yaqui	1	1	2	2	1	2	6	1	4	2	1	1	1	1	1	6	2

Short feature descriptions: 6: Uvular Consonants, 7: Glottalized Consonants, 8: Lateral Consonants, 26: Prefixing versus Suffixing in Inflectional Morphology, 27: Reduplication, 33: Coding of Nominal Plurality, 44: Gender Distinctions in Independent Personal Pronouns, 51: Position of Case Affixes, 57: Position of Pronominal Possessive Affixes, 69: Position of Tense-Aspect Affixes, 81: Order of Subject, Object, and Verb, 85: Order of Adposition and Noun Phrase, 86: Order of Genitive and Noun, 87: Order of Adjective and Noun, 89: Order of Numeral and Noun, 101: Expression of Pronominal Subjects, 111: Nonperiphrastic Causative Constructions.