A Pipeline for Computational Historical Linguistics

Lydia Steiner
Bioinformatics Group, Interdisciplinary Center for Bioinformatics, University of Leipzig *

Peter F. Stadler
Bioinformatics Group, University of Leipzig, Interdisciplinary Center for Bioinformatics, University of Leipzig*; Max-Planck-Institute for Mathematics in the Science; Fraunhofer Institute for Cell Therapy and Immunology; Center for non-coding RNA in Technology and Health, University of Copenhagen; Institute for Theoretical Chemistry, University of Vienna; Santa Fe Institute

Michael Cysouw
Research unit *Quantitative Language Comparison*, LMU München **

There are many parallels between historical linguistics and molecular phylogenetics. In this paper we describe an algorithmic pipeline that mimics, as closely as possible, the traditional workflow of language reconstruction known as the comparative method. *The pipeline consists of suitably modified algorithms based on recent research in bioinformatics, that are adapted to the specifics of linguistic data. This approach can alleviate much of the laborious research needed to establish proof of historical relationships between languages. Equally important to our proposal is that each step in the workflow of the comparative method is implemented independently, so language specialists have the possibility to scrutinize intermediate results. We have used our pipeline to investigate two groups of languages, the Tsezic languages from the Caucasus and the Mataco-Guaicuruan languages from South America, based on the lexical data from the* Intercontinental Dictionary Series (IDS). *The results of these tests show that the current approach is a viable and useful extension to historical linguistic research.*

1. Introduction

Molecular phylogenetics and historical linguistics are both concerned with the reconstruction of evolutionary histories, the former of biological organisms, the latter of human languages. Even the underlying data structure — sequences of characters — are very similar, and the evolutionary process can in both cases be modeled as a change

* Bioinformatics Group, Department of Computer Science, and IZBI, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany
** Research unit Quantitative Language Comparison, Faculty of Languages and Literatures, Geschwister Scholl Platz 1, D-80539 München, Germany

of characters influenced by environmental factors. Both fields eventually reconstruct histories by comparing related sequences and tracking their changes. The apparent similarity of these fields has prompted various methodological comparisons, see (Stevick 1963; Platnick and Cameron 1977) for some early visions, (Whitfield 2008) for a recent review, and (Atkinson and Gray 2005) for a historical overview.

A major distinction between research in historical linguistics and evolutionary biology concerns the application of computational methods. In the biological context, computational methods play a dominating role due to the availability of extensive sequence databases and due to sequences that are much too long to be handled by manual inspection. In contrast, computational approaches have not taken an equally strong hold in historical linguistics, presumably because cross-language comparisons of small sets of words can be handled by human experts and because only a tiny fraction of linguistic data is readily available in machine-readable form.

There is a recent surge in computational studies in historical linguistics, which tackle the reconstruction of language phylogenies using typological characteristics and wordlists. These studies employ computational methods from molecular phylogenetics, for example utilizing Maximum Parsimony algorithms (Gray and Jordan 2000; Holden 2002; Rexová, Frynta, and Zrzavý 2002; Dunn et al. 2005; Rexová, Bastin, and Frynta 2006) or Maximum Compatibility algorithms (Warnow 1997; Ringe, Warnow, and Taylor 2002; Nakhleh, Ringe, and Warnow 2005). A distance-based study of the Indo-European languages can be found in (Serva and Petroni 2008). Further, stochastic models of language evolution are required to employ Maximum Likelihood and Bayesian approaches. Models for lexical replacement (Gray and Atkinson 2003; Gray, Drummond, and Greenhill 2009; Kitchen et al. 2009), typological features (Dediu 2010), and combinations thereof (Greenhill et al. 2010) have been developed in recent years.

This line of research presents a long overdue innovation of the study of language history. However, we believe that the traditional approach to historical linguistics, known as the *comparative method* (Campbell 2004), can likewise profit from the computational toolkit as developed in the field of bioinformatics in the last few decades. Even more so as the basic evidence used for language reconstruction in the comparative method (sound change) is in essence very similar to the underlying process of biological evolution (nucleotide mutation and recombination) as both concern changes in character sequences.

Despite the fact that the workflow in the comparative method is quite well standardized, and many of the necessary computational tools are available, there is only a very limited, and not very active, literature on computational approaches to the comparative method (Hewson 1973, 1993; Hartman 1981; Muzaffar 1997; Lowe and Mazaudon 1994; Lowe 1995; Covington 1996; Boitet and Whitelock 1998). Recently, two important steps in the workflow of the comparative method (*viz.* cognate identification and character alignment) have received renewed computational interest. Cognate identification has been automatized in several ways. In (Kondrak 2002) pairwise alignments similarity is used, combined with semantic similarity extracted from WordNet (Fellbaum 1998), while (Kondrak 2005) employs n-gram similarity. The two approaches are combined in (Cysouw and Jung 2007). Similarities based on the consonant skeleton of the words is exploited in (Starostin 2008). Alignments of characters was approached qualitatively by (Heggarty 2000) and quantitatively by (Kondrak 2000). Alignments within words are computed with a scoring model based on *a priori* defined characters features in (Kondrak 2003; Heeringa 2004). Using the same scoring scheme, multiple alignments can be created by a profile Hidden Markov Model (Kondrak 2009) or using iterative strategies (Prokić, Wieling, and Nerbonne 2009).

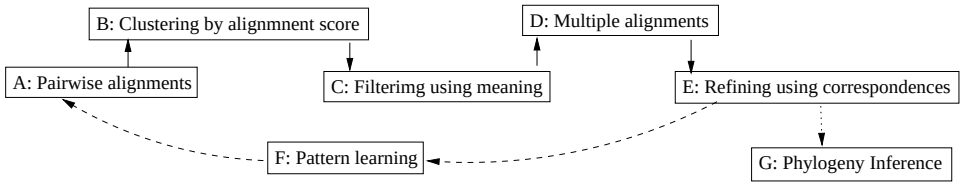

Figure 1
Overview of the steps in the pipeline.

Although the application of computational methods is by no means pervasive in historical linguistics, their usefulness has been amply demonstrated. The next step to be taken is to combine the various approaches into pipelines that can assist historical linguistics in the search for a better understanding of language (pre-)history. The computational pipeline described in this paper was explicitly designed to implement the workflow of the comparative method, much in the spirit of (Lowe and Mazaudon 1994), using suitably modified and adapted algorithms from the toolkit of bioinformatics. The different steps of the pipeline will be first described in Section 2 without delving into practical details, but rather focussing on general principles. The details of the implementation are presented in Section 3. Two case studies are presented, one on the Tsezic languages from the Caucasus in Section 4, and one on the Mataco-Guaicuruan languages from South America in Section 5, using data from the *Intercontinental Dictionary Series* (Key and Comrie 2007). We choose these two illustrative test cases because they are of a rather different kind: one example (Tsezic) should be relatively straightforward, and the other case (Mataco-Guiacuruan) should be difficult, or even impossible because the groups might not even be related. Further, we decided to aim for two groups from completely different parts of the world, and there should be lexical data available in larger amounts (at least about 1000 words per language) for more than five languages for each case (so we could test multi-way alignments). Finally, the data should have a comparable orthography for all languages in each group. Given these restrictions, it is actually not easy to find suitable groups at the current stage of digitization of lexical data, so we were happy to at least have the two current test cases for our research.

2. Organization of the pipeline

2.1 Input

Our algorithmic pipeline is subdivided into various separate steps, as detailed in Figure 1. In this section we outline the organization of the pipeline. Details on the implementation of the individual components can be found in the subsequent *Methods* section 3.

The required input data for our pipeline are word lists alike to the well-known Swadesh-lists (Swadesh 1950). Basically, such lists consist of a set of meanings, and for each of these meanings translations into the object languages are collected. In practice, this means that our pipeline starts from a set of parallel dictionaries in the different languages to be compared. In the Swadesh tradition there is normally an attempt to restrict the set of meanings to such ones that are supposedly slowly changing, but our approach does not apply such a restriction. We search for regular sounds correspondences, and not for lexical replacements. To be able to find regular correspondences, we will need

all data that we can possible get. Further, loanwords should in principle be detectable in the same fashion as they are detectable in the comparative method, namely by using characteristic sub-regularities in the sound correspondences as indications for different strata in the lexicon.

In addition to the preparation of parallel dictionaries, the orthographies have to be harmonized across the languages under investigation. In practice, our pipeline assumes that the input data uses the same orthographic representation throughout. Preferably, the orthography uses some kind of phonological or even phonetic representation, though this is not necessary as long as the orthography remains constant. Further, a pre-classification of characters is required such that at least vowels and consonants are distinguished. Our workflow in itself can accept a more fine-grained pre-classification of the difference characters, up to the point of a detailed pre-determined metric of sounds, as used for example by (Kondrak 2003; Heeringa 2004). But minimally a separation of vowels and consonants is used, which already gives reasonable results, arguing that it might not be necessary to provide more detailed knowledge about the precise meaning of the orthographic symbols. The current attempt to *not* use more fine-grained information about the similarity between characters (e.g. directional preferences of sound change) can be seen as a necessary first round of analysis using minimal assumptions. Only if such minimal assumptions turn out to be insufficient, more information should be added. We do not think it to be methodologically proper to add all possibly relevant information from the outset, because then it becomes impossible to distinguish necessary from non-necessary prior knowledge.

2.2 Cognates

The first processing step (Figure 1, step A, *cf.* Section 3.2) consists of a pairwise comparison of words by means of a dynamic programming alignment algorithm (Needleman and Wunsch 1970; Sellers 1974). Algorithmically, the problem is essentially the same as in DNA or protein sequence comparison. In contrast to bioinformatics, however, the linguistic setting requires a more elaborate scoring function that reflects rules of sound changes and their context dependence, see e.g. (Kondrak 2003, 2009). Since this scoring is not known from the outset, we resort to an iterative approach to learning the scoring function from the data, *cf.* Section 3.7. We start from a very simple scoring scheme that only considers matches between identical characters and distinguishes mismatches only between vowels and consonants (*cf.* Section 3.1). It is plausible to assume that vowels are more variable diachronically than consonants, thus mismatches between vowels are penalized less than mismatches between consonants. Since consonants rarely evolve into vowels and *vice versa*, mismatches between vowels and consonants carry the largest penalty. In the second step of our pipeline preliminary cognate sets are identified by using a clustering approach on the pairwise alignment scores (Figure 1, step B, *cf.* Section 3.3).

In these first two steps of the pipeline only the form of the words influenced their grouping into cognate sets. This results in many superficial lookalikes being grouped together, simply because there is a reasonable chance of similar words arising independently in different languages. To filter out such cases, the comparative method in linguistics enforces the additional constraint that the meanings of the words in a cognate set should also be similar. Likewise, we employ a filter to remove unlikely cognates based on an approximation to the meaning of the words (Figure 1, step C, *cf.* Section 3.4). We approach similarity in meaning using a method inspired by semantic maps as used in linguistic typology (Haspelmath 2003).

Our cognate sets will include various kinds of related words that would not have been called cognates in historical linguistics. There is a strong tradition in linguistics to reserve the term "cognacy" for words that are related by descent (i.e. vertical transfer) in opposition to "loanwords" for words that are related through borrowing (i.e. horizontal transfer). Further, related words within one language would either be called "lectal variants" or be analyzed as synchronic "derivations". In this paper we consider all these kinds of relations between words to be part of one superordinate kind of cognacy (*cf.* the term "allofamy" in (Matisoff 1978)). Basically, in all these situations the words themselves are reflexes of the same object, though the underlying processes that lead to the current situation differ. From an empirical perspective, we think it is important to distinguish the first step of the recognition of words belonging together (i.e. cognacy in the wider definition) from the second step of deciding what kind of process has led to this situation (e.g. distinguishing cognates, in the narrow sense, from loanwords).

Because most of the concrete cases in our current research are really cognates (in the narrow sense of being the result of vertical transfer) we will simply use the term "cognacy" here. However, for future research we think it would be highly valuable to make this more all-encompassing perspective explicit by using the term "homology" for the wider definition of related words, encompassing cognates, loanwords, lectal variants and synchronic derivations. This usage of "homology" fits perfectly with the way this term is used in evolutionary biology (Wagner 2007; Scotland 2010).

2.3 Correspondences

In the next step of our pipeline, the cognate sets are combined to extract correspondence sets, i.e., patterns of corresponding characters in equivalent positions of cognates. In bioinformatics, this step corresponds to the construction of multiple alignments from the collection of initial pairwise alignments. Although the multiple sequence alignment problem is NP-complete (Wang and Jiang 1994; Just 2001), many excellent approximation algorithms, such as `Clustal W` (Larkin et al. 2007), `tcoffee` (Wallace et al. 2006), `muscle` (Edgar 2004), or `mafft` (Katoh et al. 2005), are routinely used very successfully in bioinformatics applications. We here use a similar, but specifically tweaked, progressive alignment algorithm that is more geared towards linguistic data by starting with simultaneous exact alignments of up to four words (Figure 1, step D, *cf.* Section 3.5). Using more than two words in a single step increases the accuracy of the resulting alignment by reducing the inconsistencies within sets of pairwise alignments, see (Colbourn and Kumar 2007).

Correspondence sets in historical linguistics are the columns in a multiple alignment of cognates, or more precisely, the distinct patterns of characters that appear in these columns. Correspondence sets that appear multiple times in the data represent regular transformations between the languages, and hence more strongly support the hypothesis of common ancestry. Thus, we filter the cognate sets again, disregarding candidate cognates that are related only by correspondence sets that appear very rarely throughout the entire data set (Figure 1, step E, *cf.* Section 3.6).

From these refined cognate sets we learn highly conserved correspondence sets and their combinations with the help of the LZ78 Algorithm (Ziv and Lempel 1978). The patterns learned are then used to refine and revise the scoring model of both the pairwise and the multiple alignments (Figure 1, step F, *cf.* Section 3.7). The entire procedure is then repeated to produce more accurate alignments, more complete cognate sets, and more refined correspondence sets. It is possible in principle to iterate this process

more often. We found, however, that only the first two iterations led to significant improvements.

2.4 Phylogenetic inference

The resulting cognate sets and their multiple character alignments form the basis for the inference of language phylogeny. Basically, there are two different approaches to use this information for the reconstruction of language history. First, the regular sound changes can be used to infer a tree, parallel to the comparative method in historical linguistics. From this perspective the multiple character alignments are used for the phylogenetic inference. Second, the cognate sets themselves also contain phylogenetic information. Recording the presence/absence of a member of a cognate set within a particular meaning can be used for historical interpretation, parallel to the Swadesh-list approach.

In both these approaches the input for the phylogeny inference will have the form of a so-called "character matrix" (or maybe better "characteristic matrix") with languages as rows and any comparative characteristics as columns. In the first approach (the one alike to the comparative method) the columns are the multiple character alignments themselves. In the second approach, the columns are defined by specific meanings, and the available cognates within the boundary of such a meaning are marked as having identical characteristics. Although these two interpretations of the data are not completely independent of each other, they show rather different aspects of the data. For example, cognates without any change in their sound structure do not contain any phylogenetic information as far as the first approach is concerned (because all characters are identical), but they might be informative as far as the attested distribution over meanings is concerned (an thus be actually quantifying meaning change).

Several distinct approaches are utilized in bioinformatics to infer the most plausible tree from any set of such data. In parsimony methods (Fitch 1971; Sankoff 1975), the task is to find a tree that minimizes the total substitution score required to explain all columns. Alternatively, the character matrix is translated into pairwise distances of the languages by defining a metric on the rows. This is actually the most widespread approach in the linguistic literature when using Swadesh-type lists for historical comparison. Basically, the distance between two languages is defined there through a suitable function on the number of pairwise non-cognates in the columns of the matrix. When using pairwise distances, the Neighbor Joining algorithm (Saitou and Nei 1987) is preferably used to infer the language tree. Finally, when stochastic models are available, Maximum Likelihood methods (Felsenstein 1973) can be employed.

There is a crucial difference between the approach to tree building as sketched out here (and as is commonly found in molecular phylogenetics) and as it is performed by the comparative method in historical linguistics. This difference concerns the way how the reconstruction of historical directionality is handled. In simple terms, the comparative method in linguistics states that first the ancestral sounds and the direction of attested sound changes has to be inferred from the correspondence sets. This inference is performed using phonetic argumentation or experiences from earlier research (thus assuming a universal model of sound change). On the basis of such assumed directional changes a tree can be constructed. Conversely (and again strongly simplified), one of the central innovations from computational phylogenetics in biology is the insight that the structure of the tree (the "topology") can be inferred without assuming directionality. The distribution of characteristics is sufficient to infer an *unrooted tree*. The question of directionality then reduced to the problem of finding the root. The problem of establish-

ing a root in an unrooted tree is far from trivial in practice, and we will not delve into that problem in this paper. In our case studies in Sections 4 and 5 we will simply use a heuristic method called "midpoint rooting" and compare this to linguistic consensus. This method, which places the root in the most central position of tree assuming that all branches evolve with similar rates, is not very satisfactory. The most common procedure in biology — outgroup rooting (Maddison, Donoghue, and Maddison 1984) — is in most cases unusable for linguistics, because it assumes a far relative that still can be compared to the set of languages under investigation. In molecular phylogenetics, several alternative source of information, such as a molecular clock, non-reversible models of DNA substitution, and innovations of complex features, have been investigated for this purpose. More research is needed to find suitable methods for establishing the root of language phylogenies.

Even in an unrooted tree, however, the sound changes will automatically be mapped to the edges of the tree where they occur; we only lack information about the direction in which they have taken place. Now, assuming a suitable root is found (which might also be inferred from non-linguistic information like geography, archeology, or genetics) the directions are automatically given. Using our simplistic rooting method, we obtain linguistically sensible sound changes as reconstruction. This information about preferred *directional* sound changes could in principle be used to again revise the scoring model for the pairwise and multiple word alignments, starting a next iteration of the process of finding cognates and correspondences.

A central aspect of the historical-comparative tradition is to propose actual proto-forms to represent the reconstructed languages. We have not tried to emulate this step of the traditional approach yet. Given the statistical output available from our method, it should be possible to propose an even better approximation of the proto-forms as traditionally possible, because it should be possible to produce probabilities for various possible proto-forms, and not just one single reconstruction. All this, however, will have to be postponed for future research.

2.5 Evaluation

Overall, we observe that the well-established methodology of molecular phylogenetics can be transferred with only relatively minor modifications to implement the comparative method as used traditionally in linguistics to reconstruct language evolution. The modifications have been incorporated into our computational procedure in order to closely mimic the workflow of the comparative method and to provide access to intermediate results, such as cognate sets, rules of sound change and reconstructions of ancestral states, that play a central role in the linguistic discussion while their analogues are usually of little interest in molecular phylogenetics.

Given the state of the art in historical linguistics, no uncontroversial gold standards exist to test any approach against. However, intermediate results, when they are presented in a linguistically sensible way, can be evaluated manually by linguistic experts. For example, automatically produced cognate sets can be provided for inspection to linguists, either for correction, for testing, or for simply convincing specialists investigating specific language groups that automatic approaches might not be foolproof, but can be very helpful.

The pipeline is designed in such a way that intermediate results are available as the output of one module, to be processed further with the next module. Except for allowing evaluation, this also presents the possibility for prior knowledge to be incorporated into the workflow. For example, the pipeline can be used to align characters in known

cognate sets. Also, it can be used to evaluate correspondence sets based on a known language phylogeny or based on prior knowledge about rules of sound changes for subsets of languages. Further, it can learn relations between the orthographies used for different languages or accept character correspondences on input. Finally, the pipeline can determine similarities of meanings based on cognate distributions, but this information can also be supplied as input.

3. Methods

3.1 Alignment model

We use a basic string-edit model for the evaluation of alignments, distinguishing three different terms in the scoring functions of our implementation: σ is the (mis)match scoring function, δ is the deletion/insertion scoring function and κ is the contraction/expansion scoring function. Contractions (and their counterparts, expansions) of sounds are a frequent phenomenon in languages evolution (Campbell 2004). For example, correspondences between the English /sk/ and the German /ʃ/, such as evidenced by cognates like *school−Schule*, *scale−Schale*, or *scarf−scharf*, can of course be modeled as a combination of a change and a deletion/insertion. However, it makes more sense to consider this to be one single change, which formally has to be modeled as a contraction/expansion. This edit operation has no analog in bioinformatics and hence alignment software developed for biological applications does not implement such operations. Fortunately, a pairwise alignment algorithm that includes them is easily constructed (Oommen 1995) and has been successfully applied in a linguistic context (Kondrak 2000). Here we use a version that considers contractions of two sounds to a single one as well as the corresponding expansions.

The dynamic programming recursions to establish the alignment score of two words then reads as follows, where S_{ij} is, as usual, the optimal score of an alignment of the prefixes $x[1,i]$ and $y[1,j]$ of the words x and y. This dynamic programming scheme generalizes directly to three-way and four-way alignments that can deal with expansions and contractions.

$$S_{i,j} = \max \begin{cases} S_{i-1,j-1} + \sigma(i,j) & (mis)match \\ S_{i-1,j} + \delta_1(i) & deletion \\ S_{i,j-1} + \delta_2(j) & insertion \\ S_{i-2,j-1} + \kappa_1(i-1,i,j) & contraction \\ S_{i-1,j-2} + \kappa_2(j-1,j,i) & expansion \end{cases} \tag{1}$$

While alignment scores in a biological context are typically symmetric, and hence most of the commonly used alignment programs implement symmetric scoring models, asymmetry is a crucial feature in language data. For example, when aligning English with German words, one might find regular correspondences in word-initial position between English /ʧ/ and German /k/, as exemplified by cognate pairs like *church−Kirche*, *chalk−Kalk*, and *cheese−Käse*. The reverse situation, however, does not occur, i.e. German /ʧ/ paired with English /k/. (Word-initial /ʧ/ in German is highly unusual, and mainly occurs in names like *Tschechien* or *Tschetschenien*. The only regular English-German cognate pair with German word-initial /ʧ/ known to us is *chirp−tschirpen*.) So, in the course of learning and fine-tuning the alignment scores when moving from English to German, the σ(ʧ,k) will become higher (as it is a good match),

while the σ(k,ʧ) will become lower (as it is a bad match). This asymmetry, as commonly attested in linguistics, is not difficult to implement, but it will not come for free with standard alignment implementations available in the bioinformatics community. Hence the pairwise and multiple alignment tools specifically developed for our pipeline had to be made aware of possible asymmetric scoring functions.

Note that this asymmetry is not making any claims about diachronic directionality. Most linguists would assume that the change /k/ → /ʧ/ is the preferred direction over a change /ʧ/ → /k/. This is an example of diachronic directional asymmetry. However, the point here is that the underlying scoring function to evaluate cognates is asymmetric in its own sense. From a linguistic perspective this observation is completely trivial, but it is important to realize that this is different from common assumptions in molecular phylogenetics.

Sound changes are often influenced by their context, i.e., by the surrounding sounds. Hence, the scoring functions σ, δ and κ in equation (1) are implemented in such a way as to allow for the inclusion of contextual factors In our current case studies (*cf.* Sections 4 and 5) we have not yet used the power of such context dependencies for the specification of the scoring. The difficulty is not so much the principle of learning such contextual dependencies, but to fine-tune the learning in such a way as to produce dependencies that are interesting for historical linguistics. Most contexts that are found by automatic learning will rather refer to frequently occurring phonotactic patterns in the data instead of diachronically significant contexts.

We have not implemented any attempt to handle metathesis, i.e. the exchange of the position of sounds, as attested in the following examples comparing English with Dutch: *breast−borst*, *needle−naald*, or *fresh−vers*.

3.2 Initial scoring model for pairwise alignments

The first step of our pipeline consists of a rough approximation of pairwise alignments between all pairs of words from all pairs of languages. We define a simple initial scoring model for this first step. The character alignment task is simplified substantially when comparable orthographic systems are used. Given such data, we favor matches of identical characters, setting a score $\sigma(x, x) = 4$. Mismatches of characters result in a lower score. This rule cannot be used when the orthographic systems are incomparable, though note that even completely different orthographies (e.g. latin and cyrillic) can be matched approximatively when at least a few cognates or loanwords are available (Cysouw and Jung 2007).

Except for the comparable orthographic representation, we furthermore require a pre-classification of the characters into vowels and consonants for the initial iteration. This initial scoring function then prefers matches of two arbitrary vowels, $\sigma(V, V) = 2$ over matches of consonants, $\sigma(C, C) = 1$, and vowel/consonant mismatches are maximally dispreferred $\sigma(C, V) = 0$. Optionally, a different list of correspondences between characters that should be considered as "identical" can be supplied. For example, the ASJP orthography (Brown et al. 2008) only distinguishes 41 symbols to represent all possible sounds of the world's languages, merging various different sounds into classes. Such an approach implicitly presupposes a different initial σ function, and could easily be included in our pipeline by defining this function accordingly.

3.3 Preliminary clustering of pairwise alignments

The score relative to the optimal alignment of two words is used to preliminarily distinguish cognates from non-cognates. To this end we employ an affine instead of a constant or linear threshold value so that pairs of longer words may have a larger number of mismatches before they are rejected, while short words are accepted as cognates only when they are nearly identical. After some testing, we determined that the threshold function

$$\theta(\ell_1, \ell_2) = 4 \times [2 + 0.2(\ell_1 + \ell_2)] \tag{2}$$

seems to perform sufficiently for the initial cognacy decision. Here, ℓ_1 and ℓ_2 are the length of the two words and the factor 4 is the maximum attainable σ score for identical characters. The underlying idea is that two words should reach a score higher than the equivalent of two identical characters plus 40% of the average word length to pass. We do not think that this definition of the threshold is ideal, but it worked sufficiently well in our testing to proceed with the pipeline.

Preliminary cognate sets are then constructed from pairwise alignments. Considering each word a node in a graph Γ, we draw an edge between two words whenever their pairwise alignment score exceeds the threshold value θ defined in equation (2). The connected components of the resulting graph Γ are the preliminary cognate sets. The well-known breadth-first search algorithm described already in (Hopcroft and Tarjan 1973) is used to determine the connected components in the graph.

3.4 Filtering of cognate sets using meaning

The overwhelming majority of cognates retain similar meanings. Nevertheless, the meanings of many words change in the course of their historical development. The search for cognate sets should thus not be restricted to expressions of identical meanings across languages. Because it is difficult to decide which 'similar meanings' to include in such a search, the approach taken in the first two steps of our pipeline completely ignored meaning, and investigated only similarities in form. This approach strongly overgeneralizes and produces many diachronically spurious sets of superficial look-alikes. To filter out the real cognates from the look-alikes, we used a metric on meaning inspired by the tradition of semantic maps as used in linguistic typology (Cysouw 2010).

In order to quantify meaning, we used two matrices $\mathbf{S}$ and $\mathbf{D}$ with entries constructed as follows: For every pair of meanings i and j, the entry D_{ij} of $\mathbf{D}$ is the average Levenshtein distance (Levenshtein 1966) of words with meanings i and j, averaged over many different languages. Similarly, the entries S_{ij} of $\mathbf{S}$ count how often the meanings i and j are expressed by the same word in many different languages. Empirically, we used the word lists of all Indo-European languages from the *Intercontinental Dictionary Series* (Key and Comrie 2007) as the set of languages to determine these values. In total, data for 29 Indo-European languages was available. Note that the matrices $\mathbf{D}$ and $\mathbf{S}$ could also be computed from the input world lists themselves, provided the data set covers a large enough set of languages. Such an approach would have the additional benefit that locally occurring semantic shifts might be better represented in these approximations of the similarity of meaning.

The idea behind these metrics is that similar meanings have a larger probability to be expressed similarly in human language than different meanings. Individual language might (and will) deviate strongly from general trends, but on average across

many languages the formal similarity in the linguistic expression of meaning will reflect the similarity in meaning itself. Small values of D_{ij} thus indicate that the meanings are often expressed by similar words across all languages, while large values of S_{ij} provide evidence that meanings i and j are likely to be expressed by related words, including cognates, within a given language. This information can be used to identify unlikely candidates for cognates.

The matrices $\mathbf{S}$ and $\mathbf{D}$ are of course strongly (inversely) correlated, because when two meanings are expressed by the same word in a languages, then these two meanings also will have a small Levenshtein distance. Nevertheless it make sense to use both metrics, since they react with different sensitivities to different levels of semantic similarity. For highly similar meanings, $\mathbf{S}$ will give the best results, because only highly similar meanings will sometimes be expressed using the same word in any language. However, this measure will quickly reach saturation, because differences in meaning will quickly become too large to be expressed by the same word. The matrix $\mathbf{D}$ is a more approximate, but more robust, measurement of semantic similarity. It will also produce results over larger semantic ranges, while becoming rather imprecise in regions of highly similar words, where in turn $\mathbf{S}$ is informative.

In the practical application of these approximation to meaning in our pipeline, the averages $\bar{D}$ and $\bar{S}$ serve as thresholds. Two words are accepted as cognates if $D_{ij} < \bar{D}$ and $S_{ij} > \bar{S}$. Using these constraints, we applied the following greedy procedure to compile cognate sets: Each word x is tested against all other words in any already established groups C whether x satisfies the above condition for each $y \in C$. If so, x is retained in C. Otherwise a new cognate group $\{x\}$ is created. At the end, all singletons, i.e., words without cognates, are removed. There was no test data on which to base our threshold decisions, and all thresholds proposed in this paper should thus be taken as just preliminary proposals. The resulting cognate sets are therefore subsets of the connected components of Γ described in Section 3.3.

3.5 Multi-way alignments

Computational efficiency of alignment algorithms is a major issue in computational biology since the strings that need to be handled are usually very long, typically in the range of $n = 10^3$ characters. The multiple sequence alignment problem is NP-hard (Wang and Jiang 1994; Just 2001), hence heuristics are used in most applications that compose multiple alignments from a collection of pairwise alignments. For our problem at hand, however, the sequence lengths are small, because for most words the number of characters $n \leq 10$. This means that we can afford to solve the alignment problem for 3 and 4 words exactly by means of dynamic programming. Exact dynamic programming algorithms for three-way alignments have been used in computational biology (Gotoh 1986; Konagurthu, Whisstock, and Stuckey 2004). Workable approaches that avoid the computation of all entries in the dynamic tables also exist for more than three sequences (Carrillo and Lipman 1988; Lipman, Altschul, and Kececioglu 1989; Stoye 1998) but are rarely used because of their resource consumption. Since existing tools are not applicable to our extended edit problem, we use our own implementation which currently is restricted to simultaneous comparison of $N \leq 4$ words. In order to score alignments of N strings ("words") the sum-of-pair cost model is used, i.e. the score of a multiple alignment is the sum of the costs of the $N(N-1)/2$ pairwise alignments that are contained in it. In extension of previously available software, we use three-way and four-way alignments that directly generalize the recursions in equation (1).

Like pairwise alignments, three-way or four-way alignments can be combined into multiple sequence alignments (Kruspe and Stadler 2007), an approach that leads to significant increase in accuracy. Here we use a simple progressive alignment approach inspired by `clustalw`, one of the most frequently used tools in computational biology (Higgins, Thompson, and Gibson 1996). Instead of a binary guide tree, however, we use a guide tree in which each interior node has up to four children. In each progressive alignment step, the words (or alignments of words) associated with the 2 to 4 child-nodes are aligned by the exact dynamic programming as outlined in the previous paragraph. The guided tree is then constructed by a modified UPGMA clustering (Sokal and Michener 1958). In our modification, we determine the two most similar words x and y in each step, and then the two words u and v that are most similar to x or y, respectively. These words become the children of the newly created node z. The similarity matrix is updated by deleting the rows and columns belonging to u, v, x, and y and inserting a row and column for z, whose distance d_{zq} is the average distance of u, v, x, and y to a word q.

This simple update rule can easily be replaced by a weighted average if more fine-grained control is deemed necessary.

3.6 Analysis of Correspondence Sets

All correspondence sets are extracted from the multiple alignments as distinct columns. Since most alignments do not contain words from all languages and dialects, most correspondence sets are incomplete. In many cases, an incomplete correspondence set Φ is contained in a larger one, Θ, in the sense that the patterns of Θ and Φ agree for all languages that are represented in Φ. Clearly, Φ can be seen as support for every correspondence set of which it is a subset. More generally, we say that two correspondence sets are consistent if they are identical for all languages that are represented in both of them. Consistent correspondence sets can be merged, representing the common practice in historical linguistics to merge matching correspondences even if the evidence is not available for all languages under investigation. However, there is not necessary just one single possibility to merge correspondence sets. If Φ is consistent with two or more mutually inconsistent correspondence sets Θ_i, we decided to merge Φ to the most frequently occurring Θ_i. We use here a simple greedy heuristic to determine the order in which cognate sets are merged. This problem to determine a suitable merging strategy for incomplete correspondence sets is a problem that to our knowledge has never been noted as a possible problem in the linguistic literature.

One reason for merging correspondence sets is that "missing" cognates distort the estimated branch lengths by forcing transitions to $\varnothing$ from the reconstructed internal nodes of the phylogenetic tree. The merged cognate sets make it possible to "reconstruct" hypothetical words that either are missing from the data set, because the words have either been lost completely from a given language, or because its meaning has diverged too far so that it is not included in the data set, or because it has been removed by the filtering steps of the cognate recognition algorithm.

A second filter for cognacy is then derived from these generalized correspondence sets. Since only recurrent correspondences can be used in support for a common origin of words, we retain only those correspondences that appear at least twice in the word alignments. Candidate cognates that are not supported by the remaining correspondences are rejected. This filter represents the constraint in the linguistic comparative method that sound changes have to be regular, and cognates are only accepted when they can be shown to be related by using regular sound changes. In our practice, for

example, two or more cognate sets are sometimes lumped together in the initial alignments. Using the filtering on reliable correspondences allows us to split them into much more plausible units. To this end, we start with an arbitrary word x in the alignment $\mathbb{A}$ and determine for all words $y \in \mathbb{A}$ whether all correspondence sets connecting x and y appear at least twice in the entire data set.

3.7 Learning regular correspondences

Recently, a modification of the classical LZ78 string compression algorithm (Ziv and Lempel 1978) has shown to be highly efficient in learning recurrent patterns (Begleiter, El-Yaniv, and Yona 2004). In the compression step, LZ78 step-by-step creates a dictionary of the most frequently appearing patterns, irrespective of their length. When a pair of aligned strings is fed to LZ78, the dictionary records recurrent local alignments. Since our scoring model handles only patterns with a maximum length of four characters including the context (*cf.* Section 3.1), we need to limit the size of recorded patterns. At the same time, we are also interested in the frequency of the patterns. Both requirements make it necessary to further modify LZ78, as shown in Algorithm 1.

Note that in our case *data*, *pattern*, *suffix*, and *prefix* are pairwise alignments instead of simple strings. Step 2 to 8 are performed for all pairwise alignments referring to the same pair of languages. When the algorithm has finished, the dictionary contains all frequent local alignments and the frequency of their their occurrence across all cognates of two languages.

Algorithm 1 Variant of the LZ78 algorithms to learn a dictionary of aligned patterns of limited length

Input: pairwise_alignment *data*, integer *limit* /* max pattern size */

```
1:  dictionary ← ∅
2:  prefix ← ''
3:  while (data≠ ∅) do
4:     suffix ← firstOf(data)
5:     pattern ← concatenate(prefix,suffix)
6:     if (pattern ∉ dictionary) then
7:        add pattern to dictionary with frequency 1
8:     else
9:        increase frequency of pattern in dictionary by 1
10:    if (length(pattern) ≥ limit) then
11:       prefix ← suffix
12:    else
13:       prefix ← pattern
```

Based on the results of the learning algorithm, the scoring functions in our alignment model are refined. Substitution scores are derived from the patterns frequencies using a logarithmic transformation. As an alternative, one might consider log-odds substitution scores (Altschul et al. 2010), which have a more natural probabilistic interpretation. However, our score for the substitution rule $x \rightarrow y$ from language i to language j is simply set to

$$\sigma_{ij}(x, y) = 4\frac{\log(f_{ij}(x, y))}{\log f_{\max}} \tag{3}$$

where $f_{ij}(x, y)$ is the observed relative frequency and $f_{\max}$ is the maximal relative frequency in the data set. The values are restricted to the interval $[0, 4]$ to be compatible with the initial scores, which are used for all substitutions that have not been sampled in the first iteration. The score values of expansions and contractions are estimated analogously, while gap scores are left unchanged.

3.8 Phylogenetic inference

Several distances matrices are derived from the cognate sets and the alignment data. Neighbor joining (Saitou and Nei 1987), UPGMA clustering (Sokal and Michener 1958), and split decomposition (Bandelt and Dress 1992) are used to construct phylogenetic trees from these distance data.

One possibility to establish distances is to compute the total similarity score of all aligned cognate pairs between two languages. To account for biases in coverage, this value is normalized by the length of extracted pairwise alignments. The normalized similarity scores can then be translated to a distances measure using e.g.

$$d_{\alpha,\beta} = (s_{\alpha,\alpha} + s_{\beta,\beta})/2 - s_{\alpha,\beta} \tag{4}$$

Alternatively, Holm's *Separation Base* method (Holm 2000) can be employed. Here one counts from the multiple alignments the number $c_{\alpha,\beta}$ of alignments containing cognates in languages α and β. In addition, for each language the number c_α of alignments containing a word from α is determined. The parameter

$$h_{\alpha,\beta} = \frac{c_\alpha c_\beta}{c_{\alpha,\beta}} \tag{5}$$

measures a distance between the languages, based on the assumption that the probability of finding cognates is hypergeometrically distributed. This method is used for both data sets, Tsezic and Mataco-Guaicuruan.

A further distance measurement is based on the frequencies of n-grams. For any two languages, their Pearson's correlation r_{ij} coefficient provides a convenient similarity measurement with $r_{ij} \in [0, 1]$ (in the field of information retrieval this measurement is more widely known as the cosine of the angle between the normalized n-gram vectors). The corresponding distance measure is $1 - r_{ij}$. In bioinformatics, 3-grams are commonly used e.g. in the context of gene expression analysis (Eisen et al. 1998), and in linguistics it is known to function well as an approximation of genealogical relationships (Huffman 2003).

Cognate sets can be used to define two different kinds of character tables. First, it is possible to establish a table for all cognate sets in which all languages are coded by "1" that have a representative in the cognate set, while non-represented languages are coded by "0" A phylogenetic analysis of such a table basically quantifies the process of chances in the vocabulary. The problematic assumption of this approach is that it assumes that we have complete knowledge about the presence or absence of words in all languages studied, which mostly we do not have. To remedy this problem, we use a second method to establish a character table. By sampling a set of meanings, alike to the approach used with words lists like the Swadesh list, we can construct a character table for all these meanings by dividing the cognate sets into subgroups according to these meanings. For each meaning as represented in each cognate set, all languages are coded

by "1" that have a representative in this cognate set, while all other languages are coded as "0". Non-informative columns are immediately removed.

The character tables representing the presence or absence of a language in the cognate sets are used to construct trees by Maximum Parsimony with either the standard (Fitch 1971) or the Dollo (Farris 1977) model, which allows a word to be invented only once in the tree. Given the fact that it is highly unlike in linguistics for the same word to arise more than once in the same language, the Dollo parsimony assumption seems to be the most useful one for linguistics data. We use the `phylip` package (Felsenstein 1998) for these calculations.

Finally, partial splits are derived from all alignments containing correspondence set with at least one change. These can be used directly to infer phylogenetic trees (Semple and Steel 2000). We use the implementations provided by `SplitsTree` (Huson and Bryant 2006). `Dendroscope` (Huson et al. 2007) was used to visualize sound changes and related information along the edges of language trees.

3.9 Evaluation of Alignments

In order to evaluate the performance of the alignment procedure we first determined the trivial alignments, i.e., those that contain no or only a single sound change. These are correct by construction. All remaining alignments were manually examined by linguistic experts. Alignments were classified as *incorrect* when there is no correct cognate pair included. Multiple alignments containing both correct and incorrect cognate pairs are classified as *partially correct*. Alignments are deemed *questionable* if they could be correct according to expertise but a definitive classification could not be made without additional investigations into the particular case.

4. Application I: Tsezic

4.1 Data

The Tsezic languages, forming a subgroup of the Nakh-Daghestanian family of languages, are spoken in small mountain villages with about 500-7000 speakers each in southern Daghestan, Russia. The family consists of five rather closely related languages (Hunzib, Bezhta, Tsez, Hinukh, and Khvarshi), about two of which we have data available for two dialects (viz. for Tsez and Khvarshi), so that our test data comprise seven taxonomic units.

Historical-comparative studies on the Tsezic language family can be found in (Bokarev 1959) and (Alekseev 2003). Opinions about the subgrouping of the Tsezic languages diverge. One of the first researchers of the Tsezic languages (Bokarev 1959) using the comparative method, divides Tsezic into East Tsezic (Hunzib and Bezhta) and West Tsezic (Tsez and Khvarshi) with Hinukh in between the two groups. Differently, Van den Berg (van den Berg 1995) maintains that the West Tsezic languages comprise Tsez and Hinukh, the East Tsezic languages Bezhta and Hunzib, while Khvarshi constitutes a separate northern branch. Recent research (Nikolaev and Starostin 1994; Korjakov 2006) proposes the currently favored subgrouping of East Tsezic (comprising Hunzib and Bezhta) and West Tsezic (comprising Khvarshi, Tsez and Hinukh). Nikolaev & Starostin have used the historical-comparative method, whereas Korjakov has applied the lexicostatistical method. Thus, there is a clear consensus that Hunzib and Bezhta form one branch and that Tsez and Hinukh from another branch. There is no clear

consensus on the placement of Khvarshi, though the preference of the current research is that it should be grouped together with Tsez and Hinukh.

Our study is based on the lexical data collected by M. Š. Xalilov, which contains the lexical equivalences of about 1300 meanings and are made available in the *Intercontinental Dictionary Series* (IDS) (Key and Comrie 2007). Specifically, our dataset consists of 12141 words distributed over 1288 meanings in 7 taxonomic units, i.e. 1.35 words per meaning/language pair on average (see Supplemental Material A.1).

The Latinate orthographic representation used by Xalilov in the preparation of the data is phonemic-based and distinguishing 79 characters throughout all languages, of which 41 are consonant and 38 are vowels. Of the 41 consonants, 33 are used in all languages, viz. /b c cʼ č čʼ d g h ħ ʕ k kʼ l ƛ ƛʼ ɬ m n p pʼ q qʼ r s š t tʼ w x y ɣ z ž/, while 8 are only used by some of the languages, *viz.* /f g^w $lʲ$ k^w x̌ x^w y^n ʔ/. The large number of vowels arises because we treated all combination of cardinal vowels with diacritics (representing length and nasalisation) as separate characters. Some preliminary tests ignoring the diacritics did not change the results substantially. Like with the consonants there is variation in the usage of cardinal vowels characters between the languages. All languages use /a e i o u/, while only some languages use any of the remaining characters /ɑ ä ö ü ɨ/. The variation in the usage of these characters is not randomly distributed across the languages. Actually, treating only the occurrence of characters in the orthographic representation as a phylogenetic characteristic is already sufficient to reconstruct the Tsezic family tree (see Supplemental Material A.2). This result is not very surprising, as the phylogenetic analysis that is performed here actually represents a crude attempt to reconstruct the evolution of the phoneme system, which is exactly the underlying model of the *comparative method*. However, the important implication of this quick result is that reconstructing the Tsezic family tree appears to be relatively easy, and is possible with very limited data. The effort of more elaborate phylogenetic analyses should thus be directed towards providing linguistically interpretable intermediate results, like cognate sets, correspondences and sound changes.

4.2 Cognates and correspondences

Already the initial very simple scores produced high quality cognates. Not surprisingly, short words ($n \leq 3$ characters) are aligned with reduced precision, hence they were excluded from pattern learning. With the refined scoring model we observe a significantly improved recognition of cognates in particular for short words. After the second iteration we obtained 6403 pairwise and 1387 multiple alignments (see Supplemental Material A.4). It should be noted, however, that 76% of the pairwise and 87% of the multiple alignments are trivial in the sense that they contain none, or only a single sound change, i.e., only 13% of the multiple alignments contain parsimony-informative columns. Expert evaluation by D. Forker showed that the overwhelming majority of these were true cognate sets, see Table 1.

We observed that compound words can cause problems with the alignment procedure. In those cases in which the components of compound words are separated by white-spaces, it is straightforward to split the data at the white-spaces and to handle the components as separate words. In other cases, however, the split points have to be determined based on the information contained within the multiple alignment. At the moment no such remedy is implemented; *cf.* (Kondrak 2002) for a proposal how to deal with this problem to ignore typically occurring non-aligned parts of the sequence at the start and the end of words.

Table 1
Evaluation of pairwise and multiple word alignments of the Tsezic data set.

Classification	pairwise	multiple
trivial	4904	1208
correct	1360	147
partially correct	–	16
questionable	22	9
incorrect	117	7
total	6403	1387

Table 2
Summary of Tsezic cognate sets. **Left:** Distribution of the language coverage of cognate sets (alignments). **Right:** Sets consisting of differing words from the same language.

Language set	**#align.**
any comb'n of diff. lgs.	1252
only East Tsezic lgs.	77
only West Tsezic lgs.	768
East and West Tsezic lgs.	407
all languages included	56
all West Tsezic lgs. included	96
all East Tsezic lgs. included	250

Language	**#align.**
only Bezhta	38
only Hunzib	21
only Hinukh	11
only Khvarshi Inxokvari	11
only Khvarshi Khvarshi	23
only Tsez Mokok	11
only Tsez Sagadin	20

The left part of Table 2 provides an overview of the distribution of languages over the cognate sets identified. Despite the close relationships of languages and the fairly small set of languages there are only 56 cognate sets covering all Tsezic languages. The comparably large number of cognate sets within the West Tsezic branch is explained by the dialect pairs for Khvarshi and Tsez. When only cognates sets are counted that cover all West Tsezic languages, we retain fewer cognate sets than in the Eastern branch, which is statistically expected because the Eastern branch comprises fewer languages.

Various alignments proposed by our pipeline consists just of words from a single language (see right side of Table 2). Such 'monolingual cognate sets' occur in each of the seven taxonomic units. Such sets are a by-product of our all-inclusive approach of searching for cognates, because we do not restrict our search to just single words per language. All words from all languages are all compared to each other, so we are bound to find examples of cognate sets with highly similar words from the same language, be it because of lectal variation or through derivational processes producing similar lexemes within one language. Now, it also sometimes happens that a cognate set as proposed by our pipeline in the end only includes words from one single language. In such a situation, it is of course illusive to speak of a 'cognate set' in the strict sense. However, there is no way to exclude the occurrence of such 'monolingual cognate sets' from the outset, though it is trivial to separate them *post-hoc*. As discussed in Section 2.2, we propose to use the term *homology* for cognate sets in the wider sense, including these language-specific 'monolingual cognate sets'. Homologue words are simply words that

Hunzib	i	š	l	-	a	p	a
Bezhta	i	š	l	-	a	p’	a
Khvarshi Khvarshi	i	š	l	y	a	p	a
Khvarshi Inxokvari	i	š	l	y	a	p	a
Tsez Mokok	š	i	l	y	a	p	a
Tsez Sagadin	i	š	l	-	a	p	a
Hinukh	š	i	l	-	a	p	a

Figure 2
Metathesis produces erroneous correspondence sets. The highlighted columns in the alignment occur only in this particular example.

are formally related, be it through descent (vertical transfer), borrowing (horizontal transfer), lectal variation, or structural processes within a language (or even, taking this approach to its extreme, orthographic variants from different sources on the same language). Our pipeline finds ‘homologue sets’, which we see as the first step in an empirical approach to language phylogeny. The counts of cognate sets as presented in the leftmost part of Table 2 exclude the monolingual homologue sets, which are listed in the rightmost part of Table 2. However, we have not yet attempted to distinguish between sets caused by horizontal transfer from sets originating from vertical transfer.

The pipeline identified 1902 correspondence sets, of which 1165 appear only once. These were condensed to 251 distinct maximal correspondence sets using the greedy approach, of which 103 appear only once. Such a high number of correspondences is much too high for a reasonable reconstruction according to the *comparative method*. Following that method strictly, each correspondence set represents an ancestral phoneme, and a language distinguishing hundreds of phonemes is highly unlikely. There are various reasons for this high number of correspondence sets. Most importantly, we did not search for contextual factors influencing differential behavior of ancestral phonemes. Also, we have not yet tried to identify historical strata (i.e. horizontal influences), each of which will ideally cause its own specific type of correspondences. Further, there are a few examples of metathesis in the data (*cf.* Figure 2), a phenomenon we did not include in our model yet. Errors in the alignments are of course also a reason for the inflation of the number of correspondences. Finally, it might very well be the case that the actual changes in the complete lexicon are much more variable than assumed in the *comparative method*, an effect that we are faced with immediately when analyzing large amounts of data instead of hand-selecting crucial evidence.

4.3 Phylogenetic inference

Several methods were used to infer phylogenetic trees of the Tsezic language family: (1) Neighbor joining based on Holm’s separation distance, (2) Maximum Parsimony interpreting the characters in the multiple alignments as characteristics, (3) 1-covariance of 3-grams vectors with subsequent neighbor joining. All approaches agree on the same unrooted tree, which we conveniently rooted by midpoint at the edge separating the Eastern and Western branches, *cf.* Figure 3. Additional outgroup languages would be necessary to demonstrate the correctness of the location of the root. (Cysouw and Forker 2009) used other Nakh-Daghestanian languages as outgroup, which resulted in the same root as we use here.

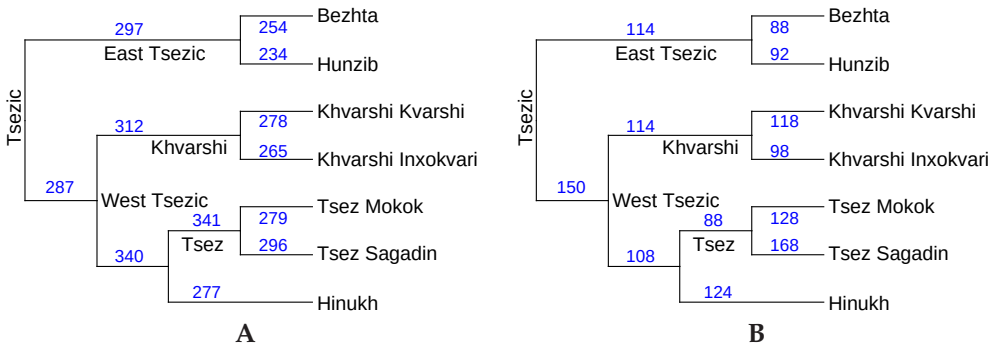

Figure 3
Total number of sound changes per edge inferred from all cognate sets (**A**) and reconstructed directly from the correspondence sets (**B**).

Table 3
Some well-supported sound changes among Tsezic languages.

Languages	sound change	
East Tsezic → Bezhta	ħ → *h*	
East Tsezic → Hunzib	*o* → *u*, *u* → *o*	
West Tsezic → Tsez/Hinukh	*l* → *r*	
Khvarshi → Inxokhvari	*e* → *i*	†
Tsez/Hinukh → Tsez	*e* → *i*	†

† These two sound changes are derived from different correspondence sets, reconstructed as e^A and e^B in (Nikolaev and Starostin 1994).

Maximum parsimony was used to reconstruct both ancestral states of words from the multiple alignments and ancestral characters directly from the correspondence sets. After annotating the internal nodes of the trees, the numbers of changes were counted by comparing the reconstructions at adjacent vertices. The total numbers of inferred changes is shown in Figure 3. A few well-supported sound changes are compiled in Table 3. These changes fit very well in with the developments sketched out in (Nikolaev and Starostin 1994). In general, the number of changes needed on each branch to account for the development of the attested correspondence sets is bewildering from a linguistic point of view. To some extend this is due to the still strongly underanalyzed status of the correspondence sets, as indicated above. However, we would also like to suggest that the sound changes occurring in natural language diachrony might be more variable than often assumed. A less cluttered result might only be achieved through a strong selection of relevant evidence.

5. Application II: Mataco-Guaicuruan

5.1 Data

The second test set consists of seven members of the Mataco-Guaicuruan language family spoken by 1500 to 42000 speakers in Argentina, Paraguay, Brazil, and Bolivia.

Although these languages are are often found grouped together as one large family, the relationship between the Mataco and the Guaicuruan groups is not proven and should be considered at best as a working hypothesis (Campbell 1997; Dryer 2005). Our decision to chose Mataco-Guaicuruan for application with our pipeline was driven by the question whether any interesting lexical similarities between these two groups could be found, possibly enforcing the hypothesis. Further, there has not been in-depth historical-linguistic research into the internal grouping of the Mataco and the Guaicuruan groups, so any results of our pipeline will provide a useful collection of basic evidence for future research.

As in the previous case study, the dataset was retrieved from the *Intercontinental Dictionary Series* (IDS) (Key and Comrie 2007). In this case, the dataset consists of 9665 words distributed over 1266 meanings in seven languages, i.e. on average 1.09 words per language/meaning pair (see Supplemental Material B.1). The data is encoded using a custom-made phoneme-based character set developed by Mary Ritchie Key for the compilation and comparison of various South American word lists. For our current selection of languages the orthography consists of in 41 characters, of which 35 are consonants [č č′ ɸ ɸ′ d f fʷ g h hʷ k kʷ k′ ḳ ḳ′ l lʸ ɬ m n ɲ p p′ r ʁ s š t t′ w x y ǰ ʔ χ] and 6 are vowels [a e i o ɔ u]. For this count we have ignored marginal characters that only occurred once or twice per language. These probably represent errors in the data. The occurrence of such errors does not have any influence on the results as produced by our pipeline, because they occur only extremely infrequently.

We used the differences in occurrence of these characters in the various languages to infer a genealogy, and this very limited evidence immediately gives a strong division between the Mataco and the Guaicuruan languages, though the internal grouping within these two sets of languages is not clearly resolved from just this limited evidence (see Supplemental Material B.2).

5.2 Cognates and correspondences

We found a total of 730 cognate sets; some statistics are shown in Table 4 (see Supplemental Material B.4). Most of the 730 cognate sets are confined to the Guaicuruan subgroup (583), and a smaller number of cognate sets contains only languages the Mataco subgroup (58). While 253 cognate sets contain all four Guaicuruan languages, there is only a single alignment covering all three Mataco languages and just 65 alignments containing at least two of the Mataco languages. Most unfortunate, there is not one cognate set comprising all seven languages, and only 89 cognate sets contain languages of both the Mataco and the Guaicuruan groups. These 89 cognate sets are particularly intriguing because they offer possible evidence for a relationship between the two groups. We have manually inspected this set, and unfortunately not very much of the possible evidence remains after scrutiny.

Most cognate sets linking Mataco and Guaicuruan that were proposed by our pipeline are not convincing because our method tries to find cognate sets by allowing some freedom both on the form and on the meaning level. As an effect, most of the 89 cognate sets are strongly similar regarding one of these two aspects only. Of the 89 possible cognate sets, 63 are very similar in meaning, but the form is too far off to be convincing. Conversely, there are 12 cognate sets for which the form is strikingly similar across the two groups, but the meaning is too diverse to allow for a historical interpretation. The remaining 14 cognate sets (see Figure 4) represent a promising collection of possible cognate sets. This collection is too small for more detailed linguistic interpretation, as there are, for example, no regular correspondences discernible linking

Table 4
Summary of Mataco-Guaicuruan cognate sets (alignments). **Left:** Distribution of the language coverage of cognate sets. **Right:** Cognate sets consisting of different words from the same language..

Language set	#align.	Language	#align.
any comb'n of diff. lgs.	730	only Mocoví	20
only Guaicuruan lgs.	583	only Toba	17
only Mataco lgs.	58	only Pilagá	13
Mataco and Guaicuruan lgs.	89	only Nivaclé	90
all languages included	0	only Maca	65
all Guaicuruan lgs. included	253	only Chorote	192
all Mataco lgs. included	1	only Wichí	42

Mataco and Guaicuruan in these few possible cognate sets. However, these 14 sets can be a basis for future research.

Note that there are still various rather large semantic changes necessary in some of these cognate sets (esp. "scar-footprint", "hit/kill-call by name" and "hawk-frog"). Although we evaluate these changes as highly unlikely, the matches in form are still deemed so strong that we decided to report these 'lookalikes' here for the sake of completeness.

Not only can we produce such lists of promising cognates, we are also rather certain that we have exhausted all evidence that is available in the *IDS* on the Mataco-Guaicuruan languages. Given the approach of our pipeline, we do not think that there are further cognates that we have not yet found. This also implies that when these 14 cases could all be explained by non-phylogenetic arguments (e.g. through borrowing, data errors, or chance) then we would be certain that there is no evidence for a link between the Mataco and the Guaicuruan languages in the *IDS* data.

So, although we have not been able to find substantial evidence for a relationship between Mataco and Guaicuruan, the results indicate how our current approach allows for a fruitful collaboration of computational techniques and manual scrutiny.

5.3 Phylogenetic inference

Although our cognate sets do not provide convincing support for the hypothesis that Mataco and Guaicuruan form branches of a common language family, we attempted to infer phylogenetic trees for the entire data set. Several different methods were employed, including covariance of 3-grams, Maximum Parsimony using the individual alignment columns, and Holm's "Separation Base" method. In contrast to the Tsezic data, where all methods agreed on a single tree, we consistently find two alternatives for the Mataco-Guaicuruan dataset, as shown in Figure 5. Both topologies clearly distinguish the Mataco and Guaicuruan subfamilies, but they disagree on the placement of Nivaclé within the Mataco subfamily. Nivaclé appears either as sister of Maca, or as the most basal branch of the Mataco group. In contrast, the sister-relationship of Pilagá and Toba, as well as the basal position of Mocoví within the Guaicuruan subfamily appear to be stable across methods.

Language	IDS	meaning	alignment							
Nivaclé	5.124	unripe	n	i	y	i	ʔ	y	a	
Pilagá	5.125	rotten	n	i	č	i	ʔ	y	a	
Wichí	10.240	drip	n	i	t	o	n			
Wichí	4.591	dribble	n	i	t	u	n			
Pilagá	5.130	drink	n	i	y	o	m			
Wichí	5.370/3.655	spoon/shell	l	a	n	e	k			
Toba	5.370	spoon	l	e	m	e	k			
Pilagá	5.370	spoon	l	e	m	e	k			
Nivaclé	15.830	wet	w	a	ʔ	a	i			
Pilagá	1.329/1.320	ocean/sea	w	a	ʁ	a	i			
Wichí	7.330/10.710	chimney/road	n	o	y	i	h			
Mocoví	10.710	road	n	a	ʔ	i	k			
Nivaclé	14.332	for a long time	k	a	x	u	ʔ			
Mocoví	14.332	for a long time	k̟	a	w	a	ʔ			
Wichí	1.520/14.530	sun/clock	h^w	a	l	a	ʔ			
Toba	1.520	sun	n	a	l	a	ʔ			
Wichí	9.220	cut	y	i	s	e	t			
Pilagá	9.110	do/make	y	i	ʔ	e	t			
Wichí	9.210/4.760	hit/kill	i	l	o	n				
Toba	18.420	call by name	i	l	o	n				
Wichí	17.172	imitate	i	t	e	n				
Mocoví	16.510	dare	i	t	e	n				
Wichí	4.858	scar	l	a	h	ɲ	i			
Pilagá	4.374	footprint	l	i	i	ɲ	i			
Maca	3.585	hawk	m	i	y	o				
Toba	3.950	frog	m	i	y	o				
Nivaclé	19.590	prevent	f	a	ʔ	m	a	t	a	n
Mocoví	16.670	tell lies	n	a	ʔ	m	a	h	a	n
Chorote	5.123	ripe	y	o	w	e	ʔ			
Toba	5.123	ripe	y	a	m	o	k̟			

Figure 4
Promising candidates for cognate alignments between Mataco and Guaicuruan (*italicized*) languages.

The fact that different inference methods produce different trees implies that the data contains inconsistencies. These inconsistent data points are alignments covering Guaicuruan languages as well as Mataco languages but support different incompatible tree topologies. These inconsistencies suggest that Guaicuruan and Mataco are unrelated. We therefore re-evaluated the word lists separately for both groups.

5.4 Guaicuruan analyzed separately

The Guaicuruan family consists of five languages about which we have information about only three languages, Toba, Pilagá, and Mocoví. We re-ran our pipeline using just the data of these three languages. A summary of the attested cognates sets is given in Table 5. There are 251 cognate sets including all three languages, and another 374

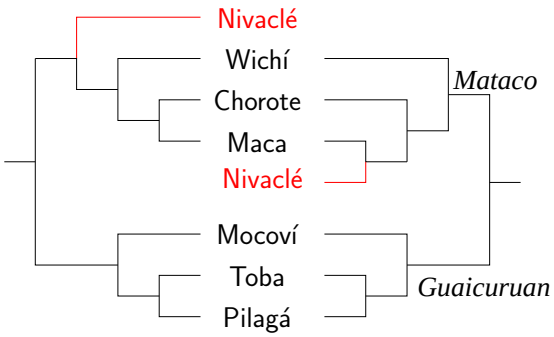

Figure 5
Two alternative tree topologies are obtained for the Mataco-Guaicuruan language family depending on the method employed for phylogeny reconstruction. The left tree is supported e.g. by Holm's distance, while Maximum Parsimony on the presence/absence of cognates supports the tree on the right.

Table 5
Summary of cognate sets in separates analyses of Guaicuruan (left) and Mataco (right). While in Guaicuruan most alignments involve more than language, we observe that the Mataco data are dominated by within-language homologues.

Language set	#align.
all Guaicuruan lgs.	251
some Guaicuruan lgs.	374
only Toba	20
only Pilagá	13
only Mocoí	25

Language set	#align.
all Mataco lgs.	0
some Mataco lgs.	65
only Wichí	52
only Chorote	201
only Maca	67
only Nivaclé	93

comprising two out of the three languages. Only few single-language homologue-sets are attested.

Although the evidence looks good from the number of cognate sets, the alignments of Guaicuruan words are much more variable than those encountered in the investigation of the Tsezic languages. Some examples illustrating a few complications are compiled in Figure 6. The large variation between the three Guaicuruan languages is also reflected by the correspondence sets. Of the 756 correspondence sets contained in the alignments, more than half (viz. 399) appear only once. Even after application of the greedy condensation, 183 of the 416 condensed sets are singletons (44%), a very large number compared to the Tsezic languages. Many of these correspondence sets, however, reflect events of loss. For example, all seven possible combinations of /a/ and the gap character /-/ appear as correspondence sets in the alignments. A few notable recurrent correspondences between Pilagá, Toba, and Mocoví are d-d-r, l-l^{y}-l^{y} and s-š-š, see Figure 7. Also note the unrecognized metathesis in the second example.

Language phylogenies were computed using all methods that were also used in the combined analysis. With the exception of 3-gram covariance, all methods agreed on the same tree, shown in Figure 8A. This topology also agrees with the results of the joint analysis outline above in Figure 5, which places the root of the Guaicuruan subtree between Mocoví and sister group consisting of Pilagá and Toba. This position of the Guaicuruan root is specifically supported when using Dollo parsimony on the cognate distribution. This method is implicitly directed and places the root at the position

A

Language	IDS	meaning	alignment										
Mocoví	18.210/22.220	speak/preach	r	a	ʔ	k̟	a	a	t	a	ʁ	a	n
Toba	18.210	speak	d	a	ʔ	–	a	k̟	t	a	ʁ	a	n
Pilagá	22.220	preach	d	–	ʔ	–	a	k̟	t	a	ʁ	a	n

B

Language	IDS	meaning	alignment								
Mocoví	17.140	think	i	p	e	e	t	e	t	a	ʔ	a
Toba	17.140	think	i	p	–	–	–	e	t	a	ʔ	a
Mocoví	17.440/17.171	guess/suspect	i	p	–	e	–	e	t	a	ʔ	a
Toba	17.440/17.171	guess/suspect	i	p	–	–	–	e	t	a	ʔ	a

C

Language	IDS	meaning	alignment													
Pilagá	9.120	work	d	–	ʔ	–	–	o	n	a	t	a	ʁ	a	n	
Toba	9.120	work	d	o	ʔ	–	–	o	n	a	t	a	ʁ	a	n	
Mocoví	9.120	work	r	o	ʔ	w	e	e	n	a	t	a	ʁ	a	n	
Toba	17.130	think	d	o	–	w	e	n	n	a	t	a	ʁ	a	n	
Mocoví	17.130	think	r	a	ʔ	d	e	e	n	a	t	a	ʁ	a	n	
Mocoví	17.110/17.190	mind/idea	l	a	ʔ	d	e	e	n	a	t	a	ʁ	a	n	a	ʁ	a	k
Toba	17.190	idea	l	–	–	w	e	n	n	a	t	a	ʁ	a	–	–	–	–	–

Figure 6
Some problems encountered with the alignments of Guaicuruan words.
A This alignment exhibits a metathesis /k̟a/ versus /ak̟/ separating Mocoví from its sister group, which is linked to the loss of the 2nd /a/. An additional loss of vowel and /a/ is observed in Pilagá. Note that all changes are concentrated in the first part of the word.
B This alignment is close to the detection limit due to the large number of losses.
C Three alignments with an identical part (shown in read), while the obvious similarities in the first parts are more difficult to interpret algorithmically.

Language	IDS	meaning	alignment										
Pilagá	15.810/15.820	heavy/light	d	e	s	a	l	i					
Toba	15.810	heavy	d	e	s	a	lʸ	i					
Mocoví	15.810/15.820	heavy/light	r	e	s	a	lʸ	i					
Pilagá	9.440	build	n	ʔ	o	ʁ	o	–	s	e	g	e	m
Toba	9.440	build	n	ʔ	o	ʁ	o	o	š	i	g	e	m
Mocoví	9.440	build	n	o	ʔ	ʁ	o	n	š	i	g	i	m

Figure 7
Examples of recurrent correspondences in Guaicuruan.

minimizing the number of independent losses, hence can be expected to be more reliable than e.g. distance-based methods that are intrinsically undirected.

5.5 Mataco analyzed separately

The Mataco languages consists of the four languages Nivaclé, Chorote, Maca, and Wichí. In total, only 65 cognate sets were found linking two or more of these languages. Like-

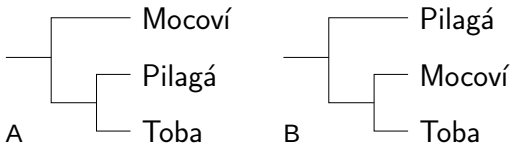

Figure 8
Phylogenetic trees inferred for the Guaicuruan language family. Tree **A** is inferred from all methods with the exception of 3-gram covariance. The alternative tree **B** is supported by the 3-gram method only.

Language	IDS	meaning	alignment							
Nivaclé	8.680	tobacco	f	i	n	ɔ	k			
Maca	8.680	tobacco	f	i	n	a	k			
Nivaclé	6.310	to spin	ɔ	f	t	i	ɬ			
Maca	6.310	to spin	a	f	t	i	ɬ			
Nivaclé	8.690	to smoke	w	a	n	k	a	ɬ	ɔ	n
Maca	8.690	to smoke	w	a	n	k̟	a	ɬ	a	n
Maca	10.613	carry-on-shoulder	t	i	ɬ	o	χ			
Wichí	10.613	carry-on-shoulder	t	i	ɬ	o	h			
Maca	9.220	cut	i	s	a	χ	i			
Wichí	9.222	chop	i	h^w	a	h	i			

Figure 9
Some examples of recurrent correspondences /ɔ/↔/a/ and /h/↔/χ in Mataco.

wise surprisingly, we did not find any alignment containing words of all four Mataco languages. The most frequent combinations of languages attested were 18 alignments with words from Wichí and Chorote, 13 covering Nivaclé and Maca, and 10 alignments with Chorote and Maca words. All other combinations of languages only occurred in a few cognate sets each. This low number of cognates sets between these languages leads to unstable phylogenies. In addition, the cognate sets often contain just a few changes or even no changes at all, further diminishing the probability to obtain a good phylogenetic signal from the data.

From a linguistic point of view, there are very many interesting and highly plausible correspondences in the Mataco cognate sets. However, because there are so few cognate sets, most of these noteworthy correspondences only occur once in the data, and often only in a cognate set with words from just two languages. Two recurrent correspondences are Nivaclé /ɔ/ ↔ Maca /a/ and Wichí /h/ ↔ Maca /χ/, some examples of which are shown in Figure 9.

Actually, most variation between homologous words in Mataco are detected in alignments that compare only words from the same language (*cf.* the rightmost part of Table 5). Most of these intra-language homologues have the same or at least very similar meanings. For example, the insertion of a *y* in Chorote words with the same meaning appears in 49 alignments, a few examples are shown in Figure 10. Another very frequent variation within Chorote is *k*↔*s*, which appears in 19 alignments. These cases represent variability within the language, which is highly important information that could in principle be taken into account to find cognates more easily. However, we have not yet found a suitable technical approach how to deal with this variability in the establishment of correspondences and in the counting of changes, as necessary for our quantitative methods.

Language	IDS	meaning	alignment								
Chorote	4.560	spit	f^{w}	–	a	ɸ	u	x	n	e	n
Chorote	4.560	spit	f^{w}	y	a	ɸ	u	x	n	e	n
Chorote	5.250	oven	k	–	a	l	a	n	a	t	i
Chorote	5.250	oven	k	y	a	l	a	n	a	t	i
Chorote	17.230	insane, crazy	t	–	a	y	e	x	e	e	s
Chorote	17.230	insane, crazy	t	y	a	y	e	x	e	e	s

Figure 10
Examples of variable *y* in Chorote-only homologue sets.

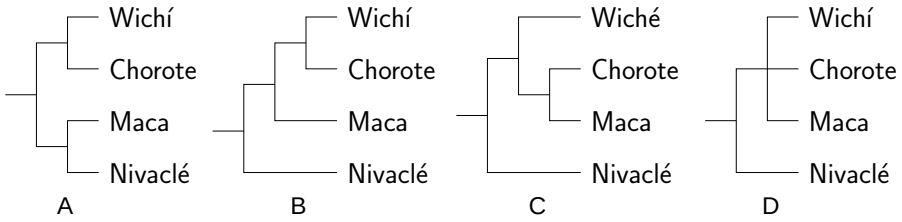

Figure 11
Alternative trees reconstructed for the Mataco languages: **A**: Dollop parsimony method and Holm's method. **B**: Splitstree based on composition of the alignments and Maximum Parsimony based on the presence/absence of cognates. **C**: 3-gram tree. **D**: Maximum Parsimony of composition of alignments.

Because of the limited amount of evidence, all reconstructions based on just these 65 cognate sets becomes unreliable. And indeed, we find different clustering when using different methods to interpret the attested evidence, *cf.* Figure 11. While tree **A** and tree **B** are supported by 2 reconstruction methods, tree **C** and **D** are produced by one method solely. The difference between the first two trees **A** and **B** is only the position of the root, as their unrooted topology is the same. Tree **D** can be seen as consensus of **B** and **C**, placing a multifurcation at the root of a subgroup formed by Wichí, Chorote, and Maca. The only really deviant unrooted topology therefore is **C**, derived from 3-gram covariation. This would suggest that the unrooted version of the trees **A** and **B** is the closest we can currently get to the phylogeny of the Mataco languages. Outgroup-rooting of the Mataco tree using the Guaicuruan data, as shown in in Figure 5, also does not resolve the discrepancies, since those trees favor a grouping of Chorote with Maca like tree **C**, instead of the sister relation of Chorote and Wichí suggested by trees **A** and tree **B**.

6. Conclusion

We have constructed a pipeline implementing the major steps of the workflow of traditional historical linguistics: the recognition of cognates, the determination of correspondence sets, and the construction of phylogenetic trees. The pipeline is constructed in such a way that is can perform both an essentially unsupervised exploratory analysis of the input data and at the same time allows to incorporate a large amount of expert knowledge if such information is available.

We applied the software to two test cases: the Tsezic family consisting of closely related languages and the Mataco-Guaicuruan languages comprising of two groups of more distantly related languages. In both cases we obtain very encouraging results. Most word alignments are correct and hence correctly identify homologous words. In the case of the Tsezic family we obtained an unambiguous phylogeny consistent with previous studies, and we correctly identified several of the previously recognized sound changes. The investigation of the Mataco-Guaicuruan languages was less conclusive, though we have produced numerous cognate sets that can be used for future scrutiny. The fact that only very few cognate sets are attested linking the Mataco and the Guaicuruan groups implies that there is hardly any evidence to link these two families in the data-set that we used. This negative evidence is worthwhile, because we are reasonably certain that we have found all possible evidence in the current data through our all-encompassing search strategy.

The pipeline offers several benefits over the traditional hand-crafted approach of historical linguistics. It can provide a preliminary analysis of language families for which little detailed knowledge is available. Further, the pipeline provides a large amount of intermediate output in the sense of actual proposed cognate sets and sound changes mapped on a phylogenetic tree. This information, while not without errors, provides a suitable set of hypotheses to be investigated further, possibly also without quantitative methods.

For example, the algorithm produced 89 possible cognate sets between Mataco and Guaicuruan from a set almost 10.000 words without any manual input. This number of 89 possible cognate sets is an amount that is easily inspected manually, whereas the manual search for those 89 cases in the 10.000 words is highly laborious. In this way, the current approach can be used to precede and complement manual work, in the sense that it can sort out large amount of data and present a selection that is worthwhile looking at manually. This approach will not replace or verify manual work. In contrast, manual inspection can be used to verify the automatic approach, not vice versa.

The pipeline produces results that are interpretable manually (we have provided all output for manual inspection as Supplemental Material to this paper). We do not test our pipeline against any gold standard or test case, because there is no gold standard, nor any other accepted test case to evaluate linguistic phylogenic methods. Whether the results are convincing or not does not depend on some general test of precision, but on the individual proposals as present in the output (and as summarized in the paper).

Of course, the results of the automatic computational procedures are not perfect. For example, there are false positives in cognate detection, which introduce some noise into the subsequent steps. However, since they form a small minority of cases, they do not compromise frequent correspondence sets and hence have little impact on the outcome of tree reconstruction algorithms. As in molecular phylogenetics, it is possible to extract robust features by focussing on the consistency of results between different variants of data analysis. For instance, there is little doubt about genealogy of Tsezic language as all lines of evidence agree on a single tree.

The approach described in this paper only provides a first step in the direction of the wider application of quantitative methods in the *comparative method*, and many of the details of the pipeline will need to be further investigated and improved before it can be used in the daily practice of historical linguistics. However, the general approach outlined is highly needed, given that only a few dozen of the 460 linguistic genera as listed in (Dryer 2005) are investigated historically in any detail, and even fewer higher order groupings are backed by solid evidence. The problem is not that it is impossible to investigate these questions: the methodology and the data to improve this situation are

available. The problem is that the research is too laborious to be performed by hand on paper, as it has traditionally been done. To speed up the research in historical linguistics, and to finally improve our understanding of the world-wide historical developments of human languages, methods like the ones that we have proposed are duly needed.

Supplemental Data

All word lists and summaries of the individual analysis steps, including alignments and trees, are available for download at `http://www.bioinf.uni-leipzig.de/publications/supplements/10-038`.

References

Alekseev, Mixail E. 2003. *Sravnitel'no-istoričeskaja morfologija avaro-andijskix jazykov*. Nauka, Moscow.

Altschul, Stephen F., John C. Wootton, Elena Zaslavsky, and Yi-Kuo Yu. 2010. The construction and use of Log-Odds substitution scores for multiple sequence alignment. *PLoS Comput Biol*, 6:e1000852.

Atkinson, Quentin D. and Russell D. Gray. 2005. Curious parallels and curious connections — phylogenetic thinking in biology and historical linguistics. *Syst. Biol.*, 54:513–526.

Bandelt, H. J. and A. W. M. Dress. 1992. A canonical decomposition theory for metrics on a finite set. *Adv. Math.*, 92:47.

Begleiter, Ron, Ran El-Yaniv, and Golan Yona. 2004. On prediction using variable order markov models. *J. Artif. Intel. Res.*, 22:385–421.

Boitet, Christian and Pete Whitelock, editors. 1998. *Alignment of multiple languages for historical comparison*, volume 1, Morristown, NJ, USA. Association for Computational Linguistics.

Bokarev, Evgenij A. 1959. *Cezskie (didojskie) jazyki Dagestana*. AN SSSR, Moscow.

Brown, Cecil H., Eric W. Holman, Søren Wichmann, and Viveka Velupillai. 2008. Automated classification of the world's languages: a description of the method and preliminary results. *Sprachtypologie und Universalienforschung*, 61:285–308.

Campbell, Lyle. 1997. *American Indian Languages: The Historical Linguistics of Native America*, volume 4. Oxford University Press, Oxford.

Campbell, Lyle. 2004. *Historical Linguistics*. MIT Press, Cambridge, MA.

Carrillo, H. and D. Lipman. 1988. The multiple sequence alignment problem in biology. *SIAM J. Appl. Math.*, 48:1073–1082.

Colbourn, Charles J and Sudhir Kumar. 2007. Lower bounds on multiple sequence alignment using exact 3-way alignment. *BMC Bioinformatics*, 8:140.

Covington, Michael A. 1996. An algorithm to align words for historical comparison. *Computational Linguistics*, 22:481–496.

Cysouw, Michael. 2010. Semantic maps as metrics on meaning. *Linguistic Discovery*, 8:70–95.

Cysouw, Michael and Diana Forker. 2009. Reconstruction of morphosyntactic function: Nonspatial usage of spatial case marking in tsezic. *Language*, 85:588–617.

Cysouw, Michael and Hagen Jung. 2007. Cognate identification and alignment using practical orthographies. In *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, pages 109–116. Association for Computational Linguistics.

Dediu, Dan. 2010. A bayesian phylogenetic approach to estimating the stability of linguistic features and the genetic biasing of tone. *Proc. Roy. Soc. B*. doi: 10.1098/rspb.2010.1595.

Dryer, Matthew S. 2005. Genealogical language list. In Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie, editors, *World Atlas of Language Structures*. Oxford University Press, Oxford, pages 582–642.

Dunn, M., A. Terrill, G. Reesink, R. A. Foley, and S. C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science*, 309:2072–2075.

Edgar, R C. 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32:1792–1797.

Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, 95:14863–14868.

Farris, J. S. 1977. Phylogenetic analysis under Dollo's law. *Syst. Zool.*, 26:77–88.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. MIT Press, Cambridge.
Felsenstein, J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst. Zool.*, 22:240–249.
Felsenstein, J. 1998. PHYLIP – phylogeny inference package (version 3.2). *Cladistics*, 5:164–166. `http://evolution.genetics.washington.edu/phylip.html`.
Fitch, W. M. 1971. Toward defining the course of evolution: minimum change for a specified tree topology. *Syst. Zool.*, 20:406–416.
Gotoh, O. 1986. Alignment of three biological sequences with an efficient traceback procedure. *J. theor. Biol.*, 121:327–337.
Gray, R. D. and Q. D. Atkinson. 2003. Language tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 425:435–439.
Gray, R. D., A. J. Drummond, and S. J. Greenhill. 2009. Language phylogenies reveal expansion pulses and pauses in pacific settlement. *Science*, 323:479–483.
Gray, R. D. and F. M. Jordan. 2000. Language trees support the support the express-train sequences of austronesian expansion. *Nature*, 405:1052–1055.
Greenhill, S J, Q D Atkinson, A Meade, and R D Gray. 2010. The shape and tempo of language evolution. *Proc Roy. Soc. B*, 277:2443–2450.
Hartman, Steven Lee. 1981. A universal alphabet for experiments in comparative phonology. *Computers and the Humanities*, 15:75–82.
Haspelmath, Martin. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In Michael Tomasello, editor, *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*, volume 2. Erlbaum, Mahwah, NJ, pages 211–242.
Heeringa, Wilbert. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. Ph.D. thesis, Rijksuniversiteit Groningen.
Heggarty, Paul. 2000. Quantifying change over time in phonetics. In Colin Renfrew, April McMahon, and Larry Trask, editors, *Time Depth in Historical Linguistics*. McDonald Institute for Archaeological Research, Oxford, pages 531–562.
Hewson, John. 1973. Reconstructing prehistoric languages on the computer: The triumph of the electronic neogrammarian. In A. Zampolli and N. Calzolari, editors, *Proceedings of the 5th conference on Computational linguistics*, volume 1. Association for Computational Linguistics, Morristown, NJ, pages 263–273.
Hewson, John. 1993. *A computer-generated dictionary of proto-Algonquian*. Canadian Museum of Civilization, Quebec.
Higgins, D G, J D Thompson, and T J Gibson. 1996. Using clustal for multiple sequence alignments. *Methods Enzymol*, 266:383–402.
Holden, Clare Janaki. 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. *Proc. Royal Soc. B: Biol. Sci.*, 269:793–799.
Holm, H. J. 2000. Genealogy of the main indo-european branches applying the separation base method. *J. Quant. Linguistics*, 7:73–95.
Hopcroft, J. and R. Tarjan. 1973. Efficient algorithms for graph manipulation. *Commun. ACM*, 16:372–378.
Huffman, Stephen M. 2003. *The genetic classification of languages by N-Gram analysis*. Ph.D. thesis, Georgetown University, Washington, D.C.
Huson, Daniel H. and D. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.*, 23:254–267.
Huson, Daniel H., Daniel C Richter, Christian Rausch, Tobias Dezulian, Markus Franz, and Regula Rupp. 2007. `Dendroscope`: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics*, 8:460.
Just, W. 2001. Computational complexity of multiple sequence alignment with SP-score. *J Comput Biol*, 8:615–623.
Katoh, Kazutaka, Kei-ichi Kuma, Hiroyuki Toh, and Takashi Miyata. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, 33:511–518.
Key, Mary Ritchie and Bernard Comrie. 2007. Intercontinental dictionary series. `http://lingweb.eva.mpg.de/ids/`.
Kitchen, A., C. Ehret, S. Assefa, and C. J. Mulligan. 2009. Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. *Proc Roy. Soc. B*, 276:2703–2710.

Konagurthu, A. S., J. Whisstock, and P. J. Stuckey. 2004. Progressive multiple alignment using sequence triplet optimization and three-residue exchange costs. *J. Bioinf. Comp. Biol.*, 2:719–745.
Kondrak, Grzegorz. 2000. A new algorithm for the alignment of phonetic sequences. In *Proceedings of NAACL 2000*, pages 288–295, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
Kondrak, Grzegorz. 2002. *Algorithms for language reconstruction*. Ph.D. thesis, University of Toronto, Toronto, Canada.
Kondrak, Grzegorz. 2003. Phonetic alignment and similarity. *Computers & Humanities*, 37:273–291.
Kondrak, Grzegorz. 2005. *N*-gram similarity and distance. In Mariano P. Consens and Gonzalo Navarro, editors, *String Processing and Information Retrieval*, volume 3772 of *Lecture Notes in Computer Science*, pages 115–126, Berlin/Heidelberg. Springer.
Kondrak, Grzegorz. 2009. Identication of cognates and recurrent sound correspondences in word lists. *Traitement automatique des langues et langues anciennes*, 50:201–235.
Korjakov, J. B. 2006. *Atlas kavkazskix jazykov*. Pilgrim, Moscow.
Kruspe, Matthias and Peter F. Stadler. 2007. Progressive multiple sequence alignments from triplets. *BMC Bioinformatics*, 8:254.
Larkin, M A, G Blackshields, N P Brown, R Chenna, P A McGettigan, H McWilliam, F Valentin, I M Wallace, A Wilm, R Lopez, J D Thompson, T J Gibson, and D G Higgins. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23:2947–2948.
Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Phys. Dokl.*, 10:707–710.
Lipman, D J, S F Altschul, and J D Kececioglu. 1989. A tool for multiple sequence alignment. *Proc. Natl. Acad. Sci. USA*, 86:4412–4415.
Lowe, John Brandon. 1995. *Cross-linguistic lexicographic databases for etymological research, with examples from Sino-Tibetan and Bantu languages*. Ph.D. thesis, Univ. California, Berkeley, Berkeley, CA.
Lowe, John Brandon and Martine Mazaudon. 1994. The reconstruction engine: a computer implementation of the comparative method. *Computational Linguistics*, 20:381–417.
Maddison, W P, M J Donoghue, and D R Maddison. 1984. Outgroup analysis and parsimony. *Syst. Zool.*, 33:83–103.
Matisoff, James A. 1978. *Variational Semantics in Tibeto-Burman*. Institute for the Study of Human Issues, Philadelphia.
Muzaffar, Towhid Bin. 1997. *Computer Simulation of Shawnee Historical Phonology*. Ph.D. thesis, Memorial University of Newfoundland, Corner Brook.
Nakhleh, Luay, Don Ringe, and Tandy Warnow. 2005. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language*, 81:382–420.
Needleman, S B and C D Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48:443–53.
Nikolaev, Sergey L. and Sergey A. Starostin. 1994. *A North Caucasian Etymological Dictionary*. Asterisk, Moscow.
Oommen, B. J. 1995. String alignment with substitution, insertion, deletion, squashing, and expansion operations. *Information Sci.*, 83:89–107.
Platnick, Norman I. and H. Don Cameron. 1977. Cladistic methods in textual, linguistic, and phylogenetic analysis. *Syst. Zool.*, 26:380–385.
Prokić, Jelena, Martijn Wieling, and John Nerbonne. 2009. Multiple sequence alignments in linguistics. In Lars Borin and Piroska Lendvai, editors, *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*. Association for Computational Linguistics, Stroudsburg, PA, pages 18–25.
Rexová, K, Y Bastin, and D. Frynta. 2006. Cladistic analysis of Bantu languages: a new tree based on combined lexical and grammatical data. *Naturwissenschaften*, 93:189–194.
Rexová, Kateřina, Daniel Frynta, and Jan Zrzavý. 2002. Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. *Cladistics*, 19:120–127.
Ringe, D., Tandy Warnow, and A. Taylor. 2002. Indo-european and computational cladistics. *Trans. Phil. Soc.*, 100:59–129.
Saitou, N and M Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4:406–425.
Sankoff, D. 1975. Minimal mutation trees of sequences. *SIAM J. Appl. Math.*, 28:35–42.

Scotland, Robert W. 2010. Deep homology: A view from systematics. *BioEssays*, 32:438–449.
Sellers, Peter H. 1974. On the theory and computation of evolutionary distances. *SIAM J. Appl. Math.*, 26:787–793.
Semple, Charles and Mike Steel. 2000. Tree reconstruction via a closure operation on partial splits. In Olivier Gascuel and Marie-France Sagot, editors, *Selected papers from the First International Conference on Computational Biology, Biology, Informatics, and Mathematics*, volume 2066 of *Lect. Notes Comp. Sci.*, pages 126–134.
Serva, Maurizio and Filippo Petroni. 2008. Indo-European languages tree by Levenshtein distance. *Europhys. Lett.*, 81:68005 (5pp).
Sokal, R. R. and C. D. Michener. 1958. A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.*, 38:1409–1438.
Starostin, George. 2008. Making a comparative linguist out of your computer: Problems and achievements. Seminar talk at the Santa Fe Institute in Aug. 2008.
Stevick, Robert D. 1963. The biological model and historical linguistics. *Language*, 39:159–169.
Stoye, J. 1998. Multiple sequence alignment with the Divide-and-Conquer method. *Gene*, 211:GC45–56.
Swadesh, Morris. 1950. Salish internal relationships. *Int. J. Amer. Linguistics*, 16:157–167.
van den Berg, Helma. 1995. *A Grammar of Hunzib (with Texts and Lexicon)*, volume 1. Lincom, Munich.
Wagner, G P. 2007. The developmental genetics of homology. *Nat Rev Genet*, 8:473–479.
Wallace, I M, O O'Sullivan, D G Higgins, and C Notredame. 2006. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res*, 34:1692–1699.
Wang, L and T Jiang. 1994. On the complexity of multiple sequence alignment. *J Comput Biol*, 1:337–348.
Warnow, T. 1997. Mathematical approaches to comparative linguistics. *Proc Natl Acad Sci USA*, 94:6585–6590.
Whitfield, John. 2008. Across the curious parallel of language and species evolution. *PLoS Biology*, 6:e186.
Ziv, Jacob and Abraham Lempel. 1978. Compression of individual sequences via variable-rate coding. *IEEE Trans. Inf. Th.*, 24:530–536.