Computer-Assisted Approaches to Historical Language Comparison

Habilitationsschrift
vorgelegt am 06.07.2020
der Philosophischen Fakultät
der Friedrich-Schiller-Universität Jena

von Dr. Phil. Johann-Mattis List aus Kassel

Gutachter

1. Prof. Dr. Martin Joachim Kümmel (Friedrich-Schiller-Universität Jena, Lehrbefähigung erteilt am 13.06.2005)

2. Prof. Dr. Volker Gast (Friedrich-Schiller-Universität Jena, Lehrbefähigung erteilt am 25.06.2007)

3. Prof. Dr. Gerhard Jäger (Eberhard-Karls-Universität Tübingen, Lehrbefähigung erteilt am 03.07.2002)

Ehrenwörtliche Erklärung

Ich erkläre hiermit eidesstattlich, dass ich die vorliegende Arbeit selbständig angefertigt habe. Die einzelnen Kapitel beinhalten zuweilen Arbeiten, die aus der unmittelbaren Kollaboration mit anderen Forscherinnen und Forschern entstanden sind. In allen Fällen liegen hier detaillierte Angaben zu den beteiligten Personen vor, und wo notwendig wurde auch deutlich gemacht, wer an welchen Aufgaben beteiligt war. Die aus fremden Quellen übernommenen Gedanken sind als solche überall kenntlich gemacht. Daten und Code, welche zum Replizieren der Studien erforderlich sind, wurden auf öffentlichen Repositorien geteilt. Alle Studien, die hier wiederabgedruckt wurden, sind frei verfügbar. Die Arbeit in ihrer Gesamtheit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht. Die einzelnen Studien jedoch wurden ursprünglich in verschiedenen Fachzeitschriften veröffentlicht, die in der Bibliographie allesamt angegeben werden.

Ort, Datum: Jena, 16.06.2021

Unterschrift: Johann-Mattis List

Contents

1 Introduction

The proposal of new quantitative methods supposed to handle problems in historical linguistics has created a gap between what one could call "classical" approaches to historical language comparison and the "new and innovative" automatic approaches. Classical linguists are often skeptical of the new approaches, partly because the results differ from those achieved by classical methods (Anthony and Ringe 2015, Holm 2007), but also because the majority of the new approaches work in a black box fashion and do not allow inspecting the concrete findings in detail. Computational linguists, on the other hand, complain about classical historical linguists' lack of consistency when applying the classical methods.

The use of computer applications in historical linguistics is steadily increasing. With more and more data available, the classical methods reach their practical limits. At the same time, computer applications are not capable of replacing experts' experience and intuition, especially when data are sparse. If computers cannot replace experts and experts do not have enough time to analyse the massive amounts of data, a new framework is needed, neither completely computer-driven, nor ignorant of the assistance computers afford. Such computer-*assisted* frameworks are well-established in biology and translation. Current machine translation systems, for example, are efficient and consistent, but they are by no means accurate, and no one would use them in place of a trained expert. Trained experts, on the other hand, do not necessarily work consistently and efficiently. In order to enhance both the quality of machine translation and the efficiency and consistency of human translation, a new paradigm of computer-assisted translation has emerged (Barrachina et al. 2008: 3).

Following the idea of computer-assisted frameworks in translation and biology, a framework for computer-assisted language comparison (CALC) could be the key to reconcile classical and computational approaches in historical linguistics. Computational approaches may still not be able to compete with human experts, but when used to pre-process the data with human experts systematically correcting the results, they can drastically increase both the efficiency and the consistency of the classical comparative method.

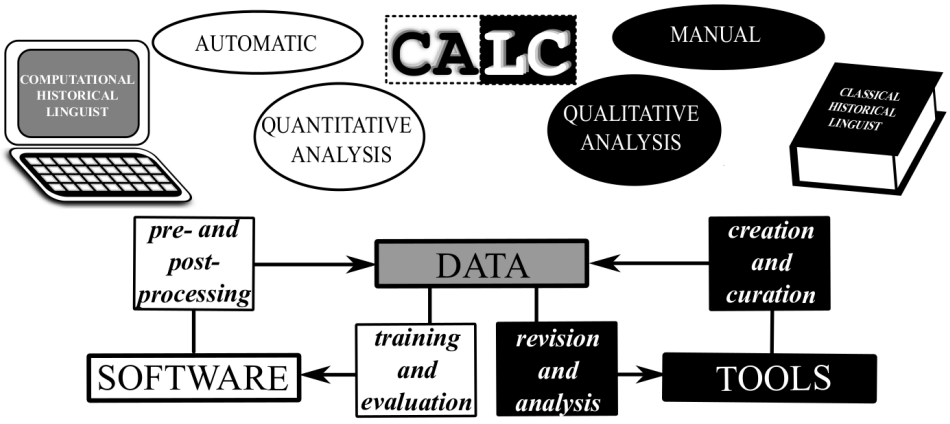


Figure 1.1: Basic idea of data managment within the CALC framework.

This study presents a collection of 12 articles in which several ideas are presented that help to fill the promise of computer-assisted approaches in historical linguistics with life. The articles reflect my past efforts to establish computer-assisted language comparison as an integral part of research in both classical and computational historical linguistics. They are divided into three major building blocks, which are themselves subdivided into two groups, with two thematically close papers exemplifying my work in this area.

The first block, titled "Of trees and webs: phylogenies and networks in historical linguistics" concentrates on phylogenetic reconstruction in a broad sense. After discussion phylogenetic networks in two detailed studies, one exploring an Indo-European dataset (List et al. 2014a), and one a dataset of Chinese dialects (List et al. 2014b), the chapter concentrates on automated approaches to ancestral state reconstruction, as reflected in a theoretical study that emphasizes the importance of specific linguistic models and rejects a naive take-over of models from evolutionary biology (List et al. 2016b), and a practical study which evaluates how well different approaches for ancestral state reconstruction perform across different datasets (Jäger 2018).

The second block, titled "Data formats and annotation frameworks" focuses on enhanced, computer-assisted ways to produce, curate, and analyze linguistic data. The chapter starts from a series of standardization attempts as reflected by the "Cross-Linguistic Data Formats initiative" (`https://cldf.clld.org`), which are introduced in an introductory article (Forkel et al. 2018) and exemplified in form of an article presenting a new linguistic database of phonetic transcription systems (Anderson et al. 2018). Thereafter, the chapter focuses on annotation in historical linguistics, as reflected in two studies, one presenting a new web-based tool which was designed to help linguists to increase the consistency of their data annotations (List 2017), and one discussing theoretical and practical challenges of annotation in Computer-Assisted Language Comparision, taking Burmish languages as an example test case (Hill and List 2017).

The third block deals with automated approaches to sequence comparison in historical linguistics. In a first part, two advanced methods for cognate detection are presented. The first study deals with the detection of partial cognates in multi-lingual wordlists (List et al. 2016b). The second study provides a description of the state of the art in automated cognate detection and introduces a new approach for the partitioning of words into cognate sets (List et al. 2017). The second part concentrates on phonetic alignments and sound correspondences. After a tutorial that introduces the state of the art of sequence comparison methods in historical linguistics (List et al. 2018), a new method for the automated inference of sound correspondence patterns across aligned data from multiple languages is proposed (List 2019b).

Each subsection of each block is accompanied by a short summary of the research and then followed by the original studies as they appeared in the different journals. While I am not always listed as the first author in all of these studies, my contribution to all of them was so substantial that the studies would not have appeared without my assistance. In most of the studies, this can also be seen directly from the author contributions wich are nowadays required by most journals.

In the conclusion, I give a short outlook on computer-assisted approaches in historical linguistics and point to future challenges. While I conclude that the studies illustrated here are some first steps towards the goal of making computer-assisted language comparison an integrative part of historical linguistics, I express optimism that the importance of computer-assisted approaches in historical linguistics will steadily grow and eventually help to bridge the gap between computational and classical approaches to historical language comparison.

2 Of Trees and Webs: Phylogenies and Networks in Historical Linguistics

2.1 Phylogenetic Networks

Phylogenies in the form of phylogenetic trees play an important role in historical linguistics. On the one hand, they capture macro-evolutionary patterns by modeling how language families evolved into their current shape. On the other hand, they can serve as a *backbone* along which evolutionary processes which do not necessarily follow the macro-phylogeny in all its details can be plotted. When comparing how a certain set of character traits evolves along a given *reference phylogeny*, we can gain valuable insights into specific aspects of language evolution and language change. With respect to lexical evolution, for example, we know that the words in a given language variety are not necessarily all inherited from the ancestor language, but may instead also be *borrowed* from neighboring languages. When investigating the evolution of a set of words along a given reference phylogeny, borrowing processes may contradict the macro-phylogeny and lead to conflicting signals in the analysis. These conflicting signals can be represented in form of *lateral edges* drawn on top of the reference phylogeny. As a result, the phylogeny becomes a *phylogenetic network* in which lateral edges represent processes resulting from language contact, while vertical edges represent processes resulting from language change.

The procedure of *character mapping*, which is required to conduct these studies automatically, has been originally developed in evolutionary biology (Dagan and Martin 2009) and later applied to study lexical borrowing in historical linguistics (Nelson-Sathi et al. 2011). The following two studies build on these initial ideas but expand them considerably. While the pilot study by Nelson-Sathi et al. (ibid.) used a simplified technique to map the evolution of words onto a reference phylogeny, the study by List et al. (2014a), title "Networks of lexical borrowing and lateral gene transfer in language and genome evolution" introduces a *weighted parsimony* approach for character mapping and applies this to an improved dataset of 40 Indo-European languages. While software, data, and code, were submitted along with the original study in form of an extended supplementary material that also contained a short tutorial explaining how the code could be used, the software has later been added to the LingPy software package (`http://lingpy.org`, List et al. 2019).

While the first study merely improved the algorithm for character mapping and the Indo-European dataset, the second study by List et al. (2014b), titled "Using Phylogenetic Networks to Model Chinese Dialect History" goes a step further by introducing a new analysis in which lateral connections, i.e., individual scenarios of lexical evolution that are in conflict with the reference phylogeny, are displayed in geographic space. The benefits of these *minimal spatial networks* are illustrated with help of a lexical dataset of 40 Chinese dialect varieties. As in the case of the improved *minimal lateral network* analysis first presented in List et al. (2014a), the code for the calculation of minimal spatial networks has by now been incorporated into the LingPy software package.

Networks of lexical borrowing and lateral gene transfer in language and genome evolution

Johann-Mattis List[1)], Shijulal Nelson-Sathi[2)], Hans Geisler[3)] and William Martin[2)]*

Like biological species, languages change over time. As noted by Darwin, there are many parallels between language evolution and biological evolution. Insights into these parallels have also undergone change in the past 150 years. Just like genes, words change over time, and language evolution can be likened to genome evolution accordingly, but what kind of evolution? There are fundamental differences between eukaryotic and prokaryotic evolution. In the former, natural variation entails the gradual accumulation of minor mutations in alleles. In the latter, lateral gene transfer is an integral mechanism of natural variation. The study of language evolution using biological methods has attracted much interest of late, most approaches focusing on language tree construction. These approaches may underestimate the important role that borrowing plays in language evolution. Network approaches that were originally designed to study lateral gene transfer may provide more realistic insights into the complexities of language evolution.

Keywords:
borrowing; language evolution; lateral transfer; network approaches; prokaryotic evolution

Additional supporting information may be found in the online version of this article at the publisher's web-site.

Introduction

For a long time, both biologists and linguists have been using family trees to model how species and languages evolve. But in contrast to biology – where the tree model is generally accepted to be the most realistic way to model how eukaryotic species (species with nucleated cells, such as animals and plants) evolve – linguists have always treated language trees with a certain suspicion. They have emphasized that – given the important role that horizontal transmission plays in language history – such trees can only capture vertical aspects of language evolution, while horizontal aspects (which linguists traditionally model as "waves" that spread out in circles around a center in geographic space) are ignored.

In the last decade, language trees have experienced a strong revival, especially in the public notion of linguistics as reflected in popular scientific literature and in articles addressed to a not exclusively linguistic readership [1]. Earlier linguistic work on phylogenetic reconstruction was, with a few exceptions [2–8], qualitative in its nature. But starting about 10 years ago, computer methods originally designed to infer trees from molecular sequence data made their way into the analysis of large linguistic datasets, leading to a resurgence of language trees [9–15]. If the reconstruction of trees had only played a minor role in historical linguistics up to that point, it has now become a specific field of interest, and some scholars even go so far as proclaiming tree construction as a priority for historical linguistic endeavor [16].

In traditional historical linguistics, these new approaches are met with a certain amount of reservation, since their results are often not in concordance with those achieved by traditional methods [17–20]. One important reason for such discrepancies is the relatively large number of individual

DOI 10.1002/bies.201300096

[1)] Research Center Deutscher Sprachatlas, Philipps-University Marburg, Marburg, Germany
[2)] Institute of Molecular Evolution, Heinrich-Heine University Düsseldorf, Düsseldorf, Germany
[3)] Institute of Romance Languages and Literature, Heinrich-Heine University Düsseldorf, Düsseldorf, Germany

***Corresponding author:**
Johann-Mattis List
E-mail: mattis.list@uni-marburg.de

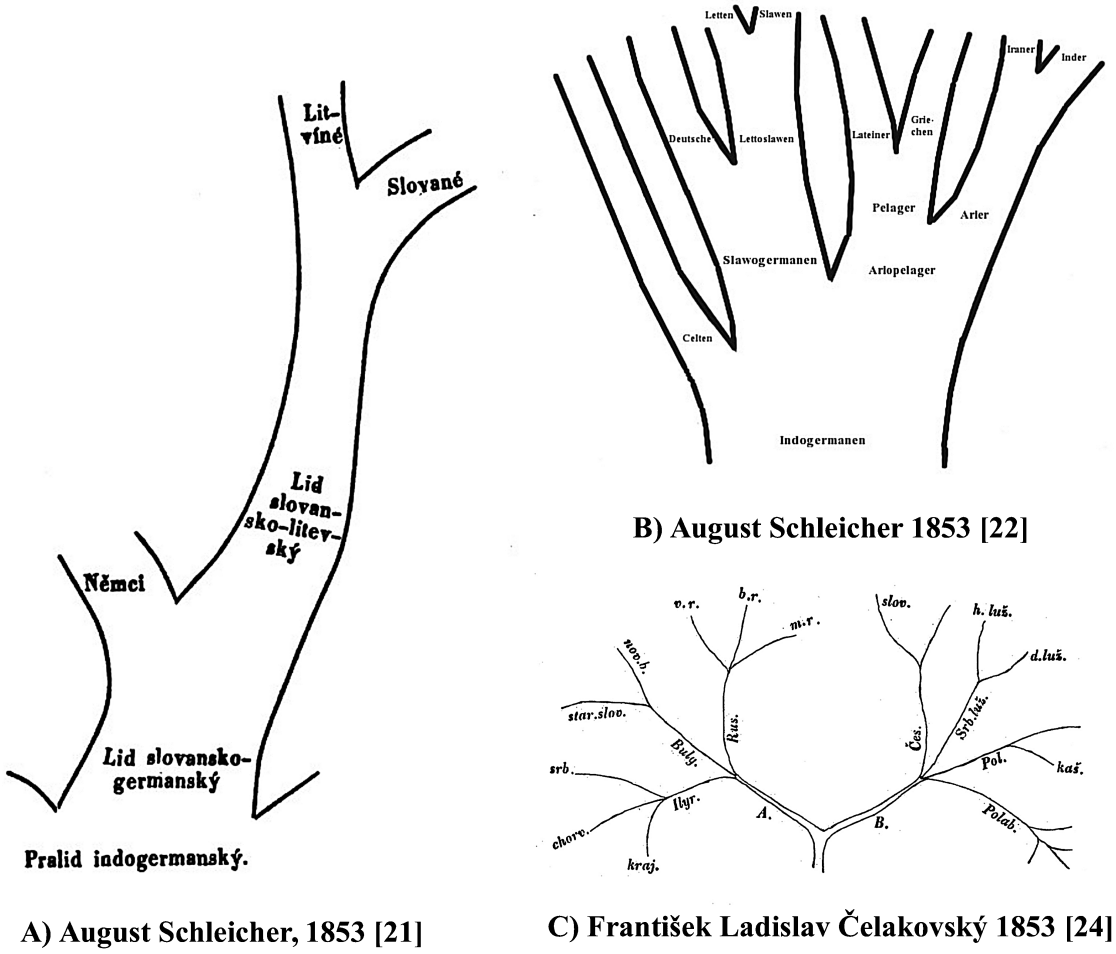


Figure 1. Three early language trees in the history of linguistics. **A:** August Schleicher's first tree of Germanic and Balto-Slavic languages. **B:** Schleicher's first tree of the Indo-European language family. **C:** An early tree of the Slavic languages by František Ladislav Čelakovský.

and methodological errors in linguistic datasets [19]; this is reflected by numerous cases of wrong translations, wrong homology assessments (incorrect identification of cognate words), and undetected cases of lateral transfer (borrowing) [17, 18].

In this paper, we argue that the problem of the new quantitative methods is that they focus too much on the vertical aspects of language evolution, thereby forcing the data into tree-like structures. We show that network approaches that were originally designed to study reticulation and lateral gene transfer in the evolution of prokaryotic species (microbes without cell nuclei, such as bacteria and archaea) can cope with these problems, hence providing a more realistic way to model the complexities of language history by combining both its tree-like (vertical) and its wave-like (horizontal) aspects.

Historical linguists were always skeptical about language trees

In 1853 the German linguist August Schleicher (1821–1868) published two articles [21, 22] (Fig. 1A and B) in which he showed how branching trees can be used to illustrate the historical development of languages (Table 1A). It is possible [23] that Schleicher himself adopted the idea from a colleague, the Czech linguist František Ladislav Čelakovský (1759–1852), whose posthumously published lectures contain an early tree diagram of the Slavic languages [24] (Fig. 1C). Schleicher was very interested in biology, especially botany, and in his work we find many passages where he compares languages with organisms, assuming that they went through stages of birth, youth, middle age, old age, and – finally – death [25]. He emphasized that language classification was quite similar to biological classification of animals or plants [25]. He also mentioned the problem of distinguishing vertically from horizontally transmitted traits, drawing a parallel between "foreign influence" due to language contact in language history, and "crossbreeding" in evolutionary biology [26] (Table 1B).

In biology, the concept of evolutionary trees was not introduced until

Table 1. Early quotes on language history from August Schleicher and Hugo Schuchardt

(A) August Schleicher [26]	
We know both the Old Latin and the Romance languages which demonstrably descended from the former via differentiation and – you would call it crossbreeding – foreign influence	*Wir kennen sowohl das Altlateinische, als auch die durch Differenzierung und durch fremden Einfluss – Ihr würdet sagen durch Kreuzung – nachweislich aus ihm hervorgegangenen romanischen Sprachen*
(B) August Schleicher [22]	
These assumptions which logically follow from the previous research can be best illustrated with the help of a branching tree	*Diese Annahmen, logisch folgend aus den Ergebnissen der bisherigen Forschung, lassen sich am besten unter dem Bilde eines sich verästelnden Baumes anschaulich machen*
(C) Hugo Schuchardt [32]	
We connect the branches and twigs of the family tree with countless horizontal lines and it ceases to be a tree	*Wir verbinden die Äste und Zweige des Stammbaums durch zahllose horizontale Linien, und er hört auf ein Stammbaum zu sein*

Charles Darwin's (1809–1882) mentioning of the "Great Tree of Life" in 1859 [27], but it soon became deeply ingrained in thinking on the topic. Notably, it was later reinforced by many influential drawings from Ernst Haeckel (1837–1919, see [28] for details), culminating in the inference of trees from molecular sequences [29], and the reconstruction of phylogenetic trees for all organisms using ribosomal and informational gene phylogenies [30].

In linguistics the popularity of language trees began to fade soon after it was first proposed [31]. In 1872 Johannes Schmidt (1843–1901) pointed out that linguistic data contradicted the idea of simple, tree-like differentiation [32]. Instead of the family tree theory he proposed the "wave theory" (*Wellentheorie* in German), which states that certain changes spread like waves in concentric circles over neighboring speech communities. And before Schmidt, Hugo Schuchardt (1842–1927) had criticized the idea of split and independent differentiation [33], emphasizing that languages diverge gradually while at the same time mutually influencing each other (Table 1C). Even today, historical linguists continue to hold strong reservations about the tree model. In text books on historical linguistics, both the tree and the wave theory are usually introduced as two complementary models, each of which only depicts one aspect of language history [34, 35]. Thus, if linguists are asked whether language evolves in a tree-like manner, most linguists would probably answer as Hoenigswald did in 1990: "Yes, of course it does, if we so wish; but we had better be very careful" [36].

Borrowing is a constitutive part of language history

If we take the most frequent 1,000 Latin words and look at how they survived in its daughter languages, we will find that 67% of all words were directly inherited in at least one language, yet only 14% were inherited in all Romance languages [37]. However, this drastic loss of Latin words during Romance language history is only part of the story: Since Latin never ceased to serve as a *cultural adstrate language* (a language that co-exists in some form in parallel with another language with which it is in contact), with a particularly great impact on written vernaculars, only 33% of all 1,000 words were completely lost, and about 50% survive as borrowings from the ancestor language in the daughter languages [37]. Moreover, lexical transfer during the history of the Romance languages was not restricted to the influence of Latin alone, and contact among the Romance languages and other neighboring Indo-European languages was very frequent and vivid. According to a recent survey of 2,137 common words in Romanian [38], for example, 894 (41.8%) were classified as loanwords from other languages. The majority of these borrowed words were transferred from Slavic donor languages (about 14%). Only a small number of words were borrowed from Latin (about 3%).

On the "borrowability scale" [39], which ranks the ease with which different elements of language are assimilated by recipient languages, borrowing of *words* ranks highest. Lexical borrowing can affect only small parts of the vocabulary of a given language (such as specific terms for religious concepts, cultural items, or artifacts), or result in a situation where large parts of the language's original lexicon are replaced. This can even result in complete relexification, as in Creole languages. In the World Loanword Database [40] the frequency of direct borrowing events documented for 41 languages varies greatly, ranging from 1% for Mandarin Chinese to 62% for Selice Romani, with an average of 25% and a standard deviation of 13% [41].

Borrowing cannot be ignored in quantitative approaches

With few exceptions [42–44], the majority of the new biological methods for tree construction makes use of lexical language data. This is due to the fact that it is much easier to compile lexical datasets for large numbers of languages: in many cases – especially for less-well studied language families – wordlists are the only things available for study. However, analysis of lexical items also reflects the basic practice of the traditional method for linguistic reconstruction, which starts with the comparison of words and morphemes [35, 45, 46]. Similarly to earlier quantitative approaches in historical linguistics [8], the biological methods require that borrowings be filtered out of the data before the analysis is applied. Since reliable automatic methods are lacking, cognate and borrowing

assignments are usually carried out manually. In order to make this painstaking process easier, scholars revived an old idea proposed in the 1950s [4, 5, 47], and restrict the lexical comparison to words that belong to the realm of the so-called "basic vocabulary" [12]. Basic vocabulary is merely a technical term that refers to a list of about 100–200 basic concepts (such as "hand", "foot", "stone") that are translated into the languages under investigation. These lists are usually called *Swadesh lists*, in acknowledgement of Morris Swadesh (1909–1967), who popularized their use in linguistics. The basic assumption regarding Swadesh lists is that (a) every language has words that express the concepts, (b) the words evolve slowly (enabling us to recognize similarities across languages), and (c) the words are rather resistant to borrowing [16]. Unfortunately, the last assumption, in particular, is highly problematic. Although the use of Swadesh lists may decrease the number of borrowings to a certain degree, it cannot exclude all of them. In a recent survey of 1,504 common words in English, for example, 616 (41%) were judged to be loanwords [48], yet in the traditional English Swadesh list there are still 32 borrowings out of 200 (16.5%), mostly from Old Norse and Old French [18]; and in a recent revision of the Albanian Swadesh list, 34 out of 107 words (31.8%) were identified as possible borrowings [49].

Manual detection of borrowings can range between trivial and impossible, depending on the case in point. Some borrowing processes are very transparent. Neither a linguist nor a German speaker has problems in identifying the word *Job* "job" as a recent borrowing from English, since the initial sound of the word is not yet "integrated" into the German sound system. But the situation is not always that simple. Thus, while no German native speaker would hesitate to assume that *Fett* "grease" is a "normal" German word, the word has in fact been borrowed from Low German dialects [50], as can be proven from its irregular correspondence with English *fat*: If the words were truly cognate, we would expect the German word to end with an [s] (spelled as *ß* in German) instead of a [t], as in German *heiß* "hot," which is truly cognate with English *hot* [50]. Identifying borrowings with help of these techniques requires expert knowledge of the languages under investigation, and the deeper one goes back in time, the harder it becomes even for the experts, since the available phonological information may be lost.

Recent tests on simulated data have shown how crucial it is to screen the linguistic data carefully before applying quantitative analyses [51]. How difficult it is to prepare the data and to filter out all borrowings correctly is reflected by the fact that the most frequently used datasets, the *Comparative Indo-European Database* ([52], http://www.wordgumbo.com/ie/cmp/), and the *Austronesian Basic Vocabulary Database* ([53], http://language.psy.auckland.ac.nz/austronesian/), contain many undetected borrowings and various levels of erroneous cognate judgments [17–19, 49]. But "scrubbing" the data of false cognate assignments does not seem to be feasible for large datasets. Quantitative studies that are based on the *Indo-European Lexical Cognacy Database* (IELex, http://ielex.mpi.nl/), whose goal was to significantly enhance the notoriously flawed database composed by [52], still yield subgroupings that contradict traditional genetic classification (compare, for example, the strange grouping of Polish in [13] and [54]). One reason for these problems is that the database still contains many undetected borrowings and other errors. The other reason is that the exclusion of borrowings necessarily yields a loss of information that can have large impacts on the results [49]. It seems that the a priori exclusion of suspected borrowings from the data is not enough, especially in cases where the history of a language family is not yet well understood. Instead of making tree reconstruction the key objective of historical linguistics, we need quantitative methods that can deal with borrowings and – ideally – handle both vertical and lateral transmission.

Language history bears a close resemblance to prokaryote evolution

If historical linguists want to profit from biological expertise in large-scale analyses of big datasets, they need to make up their mind regarding the methods they need in linguistics, and the methods that biology can provide. That evolutionary biology has developed some sophisticated tools to reconstruct phylogenetic trees, and that these tools can be easily applied to linguistic datasets, has been demonstrated frequently during the last decade. Yet is this really all that biology has to offer?

In several fundamental aspects, the genomes of eukaryotic species – such as animals and plants – and prokaryotic species – such as bacteria and archaea – evolve in very different ways, and lateral gene transfer is generally at the root of those differences. Gene families are one example. Gene families are sets of homologous (cognate) genes that were formed by duplication of an ancestral gene, quite similar to the reflexes of the root of a word in the same or different language. In eukaryotes, gene families arise through duplication: a resident gene duplicates, perhaps several times, and the resulting gene family consists of members that are closely related at the outset and undergo divergence and functional specialization [55]. In prokaryotes, gene families arise via the acquisition of related sequences through lateral gene transfer, not through duplication [56]. As another example, in eukaryotes, meiosis ensures that only members of the same species exchange genes, and recombination is reciprocal. In prokaryotes, there are well-studied mechanisms that mediate gene transfer, both within and across species boundaries [57].

Furthermore, if we sequence 61 human genomes, we will find – to all intents and purposes – the same collection of about 30,000 genes in each individual, with allelic variants at many loci, and the 46 chromosomes will almost always be colinear: the genes appearing at similar positions. If we sequence 61 genomes of *Escherichia coli*, a bacterium usually found in the intestines of warm-blooded species, we will find about 4,500 genes in each individual genome, but only about 1,000 genes that are present in all genomes. Summing up the different genes we find in all individuals, there are about 18,000 different genes distributed among them, and this count will further increase if we add more individual genomes to this calculation,

hence yielding an ever growing pangenome of *Escherichia coli* [58]. These examples underscore fundamental differences in the nature of the processes of evolutionary divergence in prokaryotic and eukaryotic populations: Eukaryotic populations generate tree-like structures of divergence over time [59], while genome evolution in prokaryotes generates both tree-like and net-like components of relatedness over time [60].

Recalling the scores on shared inherited words and borrowings we reported for the Romance languages earlier, it seems obvious that language history shows a much closer resemblance to prokaryotic evolution than to eukaryotic evolution. Thus, if one says that language history and genome evolution have a lot in common, it seems much more appropriate to emphasize that language evolution may resemble prokaryotic evolution much more than it resembles eukaryotic evolution. We do not claim to make a binary distinction here: As the amount of contact-induced change differs from language to language, so do the underlying evolutionary processes, and it is rather a continuum between strictly tree-like and strictly network-like evolution that we are dealing with. Nevertheless, if we want to employ quantitative methods from biology to supplement our research in historical linguistics, it could be much more fruitful to get away from focusing exclusively on those methods that yield simple family trees, and instead look for methods that were designed to handle lateral transfer.

Network approaches offer new possibilities for quantitative analyses in language evolution

Despite the dissatisfaction of many historical linguists with both the tree and the wave model, there are – to our knowledge – only a few attempts to combine both approaches within a new framework [35, 61, 62]; furthermore, unfortunately most of these proposals remain a mere visualization of the scholars' intuitions regarding the data, from which no further insights can be drawn. If one wants to include both the vertical and the horizontal aspects, it seems natural to turn to networks as a format to represent language history.

In evolutionary biology, different network approaches have been developed in order to study reticulation in biological datasets (see the overviews in [63] and [64]). Among the most popular of these methods are those that produce unrooted networks (splits graphs) such as *split decomposition* [65] or *NeighborNet* [66]. These methods enjoy some popularity in recent quantitative studies in historical linguistics, and have been applied to quite a few different datasets [67–71]. In contrast to the popular quantitative methods for tree construction, such as *Neighbor-Joining* [72], or *Bayesian inference* [73], they are unbiased with respect to "tree-likeness", and provide a direct visualization of the degree of conflict in a given dataset [74]. They have proven to be a very useful tool for data exploration, and have even been used to measure reticulation directly from lexical distance matrices across the world's language families [75]. The drawback of these methods is that they are distance-based, hence aggregating lexical information on the taxonomic level. The information on shared cognates in the underlying datasets is converted to distance scores, and the result is an unrooted network that only indicates whether there are conflicting signals in the data, but does not directly point to the cognate sets that are responsible for these conflicts.

A more realistic modeling of language history could be achieved by methods that automatically infer hidden borrowings in the data. While quite common in evolutionary biology [76, 77], these methods are still in their infancy in historical linguistics. Two early approaches [70, 78] are distance-based, and therefore do not allow the direct identification of the characters that conflict in the reference trees. The first character-based approach to this problem [79] uses maximum parsimony to determine the characters that conflict with an inferred family tree. Unfortunately, the method has only been tested on a very small dataset, and no further applications are known to us. An alternative proposal expands the notion of *perfect phylogenetic trees* [10] to the notion of *perfect phylogenetic networks* [80]. The method yields direct statements as to which characters have been inferred as being borrowed in a given dataset. Unfortunately, the algorithm is very time-consuming, and it is thus not feasible to apply it to larger datasets [81].

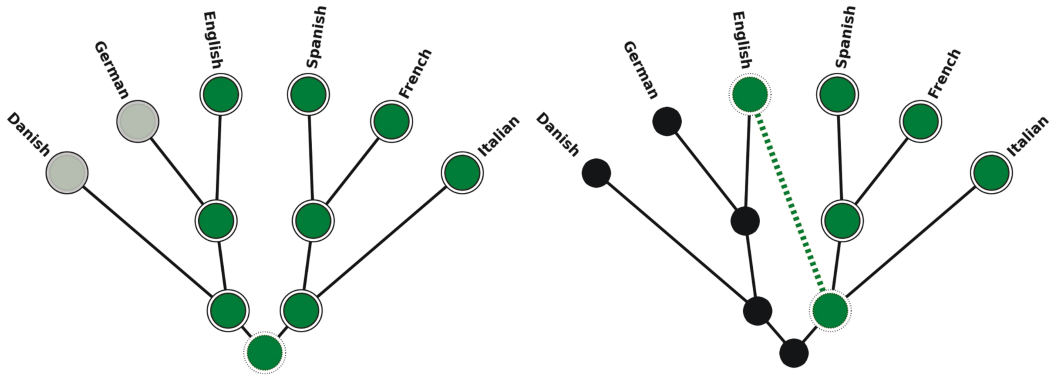
Ancestral genome sizes reveal the minimum amount of lateral transfer in microbial evolution

A more recent method for lateral gene transfer detection in prokaryotic genomes is the so-called *minimal lateral network approach* (MLN, [82]). This method applies the technique of gain-loss mapping [83–85] to presence-absence patterns of gene families in order to infer patterns that are suggestive of lateral transfer. Gain-loss mapping starts from a given reference tree that should reflect the vertical component of evolution as closely as possible. With help of the reference tree, specific gain-loss scenarios for all gene families in the dataset are inferred. A gain-loss scenario provides an explanation of how a given character could have evolved along the reference tree when character evolution is modeled as a simple process of gain and loss events. In order to confirm the assumption that a given character evolves in an exclusively vertical manner, the inferred gain-loss scenario should contain only one gain event. If more than one gain event is inferred, the character is judged to be suggestive of lateral transfer (see Fig. 2 for an example applied to linguistic data).

The crucial point of the MLN method is to select the best gain-loss scenarios out of the multitude of possible ones. The key argument in biology is the notion of ancestral genome size distributions [84]: If, for example, all gene families are assumed to originate only once along the reference tree, this may result in ancestral genomes that contain much more genes than are observed in the contemporary genomes. If, on the other hand, one assumes that all gene families are explained by lateral gene transfer only, then the vertical component of genome evolution disappears, and ancestral genome sizes become too tiny to support life. Between those extremes there are amounts of vertical and lateral inheritance that will

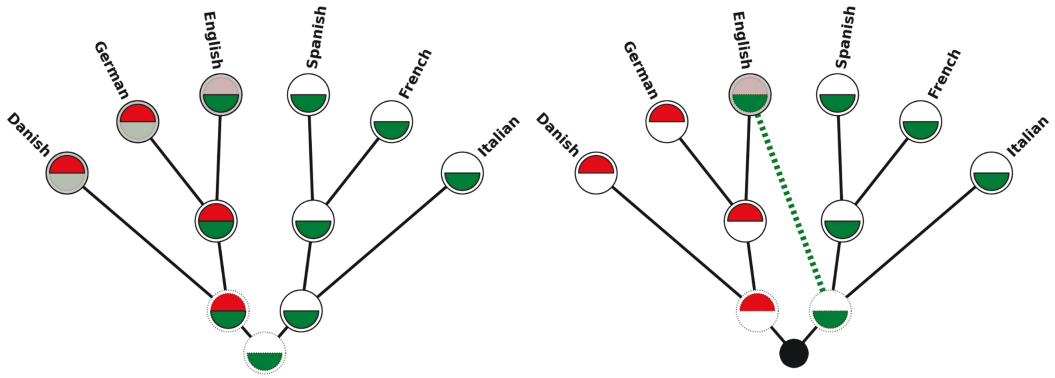
Language Variety	Danish	German	English	Spanish	French	Italian
"to count"	*tælle*	*zählen*	*count*	*contar*	*compter*	*contare*
Latin *computare*	0	0	1	1	1	1
Proto-Germanic **tal-*	1	1	0	0	0	0

A) Presence-absence patterns of two cognates sets in Germanic and Romance languages



B) Loss-only scenario: no lateral transfer event inferred

C) Two-gain scenario: one lateral transfer event inferred



D) Combined loss-only scenario: no lateral transfer event inferred

E) Combined two-gain scenario: one lateral transfer event inferred

Figure 2. Illustration of the MLN method. **A:** Two cognate sets for "to count" in three Germanic and three Romance languages. The English word is a known borrowing from Old French. The original reflex of Proto-Germanic *tal- is still preserved in English "to tell," but its original meaning has shifted under the influence of the borrowing from Old French, and it is thus not listed in this sample. **B:** The loss-only scenario assumes that the cognate set with reflexes of Latin originated in the root and was then lost independently in both German and Danish. **C:** The two-gain scenario infers two separate origins of the cognate sets. The pattern is thus suggestive of lateral transfer, and one lateral transfer event is inferred. This is marked by the link drawn between the two nodes where the characters first originate. **D:** Combination of scenarios for both cognate sets based on the loss-only scenario in B. Note that this scenario forces us to assume that the ancestor of the Germanic languages had two words expressing the concept "to count." While this is not improbable per se, cases of inferred overwhelming amounts of synonymy are suspicious in language history. **E:** Combination of scenarios for both cognate sets based on the two-gain scenario in C. This scenario is preferred by the MLN method, since the number of synonyms in the ancestral languages is in balance with the modern languages. Note that the inference does not tell us which language is the real donor (which is Old French). According to our model, it could be any of the three Romance languages. For this reason, the edge is drawn between the ancestor off all languages.

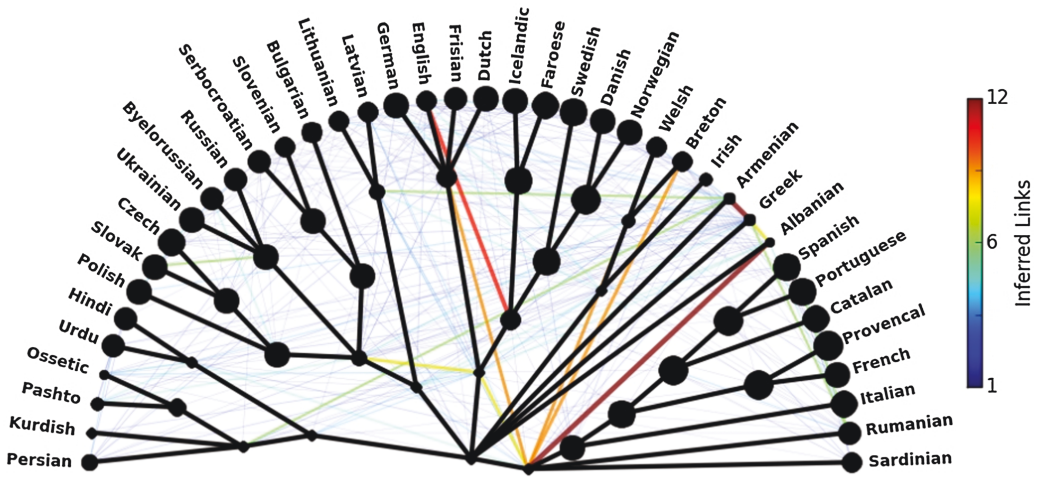


Figure 3. Minimal Lateral Network of 40 Indo-European languages. The size of the nodes reflects the number of cognate sets in each language as inferred by the MLN approach. The links reflect the minimal amount of lateral transfer events that is needed to bring the distributions of synonyms in the contemporary languages (leaves of the tree) and the ancestral languages (internal nodes of the tree) as closely together as possible.

bring the distribution of inferred ancestral genome sizes into agreement with the attested distribution of contemporary genome sizes. Those distributions can be tested statistically, and the gain-loss scenarios with the amount of lateral gene transfer that best fits the data can be determined. Having selected the best scenarios, a rooted phylogenetic network can be reconstructed. Here, multiple origins of the same gene family on different branches of the reference tree are connected by lateral links; edges connecting the same two nodes for different gene families are joined to form weighted edges [82].

How minimal lateral networks can be applied to linguistic data

Technically, the application of the MLN approach to language data can be carried out in a rather straightforward way, by investigating presence-absence patterns of cognate sets instead of presence-absence patterns in gene families. Theoretically, however, the application of the approach requires some caveats: while genomes are physical entities whose size can be directly determined, the linguistic data consist of samples based on meaning lists. We can restate the genome size criterion for scenario selection in such a way that we prefer those scenarios in which the number of words used to express specific meanings does not differ much between ancestral and contemporary languages. However, we need to keep in mind that new words can also shift into the meaning slots from outside the sample. Although parallel semantic shift involving cognate words in different branches of a language family is surely much rarer than borrowing, this has to be considered when applying the method to linguistic data.

The MLN approach was first applied to the well-known Comparative Indo-European Database [52], and revealed a rather high degree of non-tree-like signal: 61% of all 2,346 cognate sets in the data were found to be suggestive of borrowing [86]. Since the study employed a very simple top-down algorithm for gain-loss mapping [84], the inferred amount of cognate sets contradicting the reference tree is surely too high. In order to test whether more refined techniques of gain-loss mapping can yield more realistic results, we applied a refined variant of the MLN approach to a subset of 40 Indo-European languages taken from the IELex (dump from May 2013 kindly provided by M. Dunn). The modified MLN approach is implemented as part of a freely available Python library for quantitative tasks in historical linguistics [87]. It employs weighted parsimony for the task of gain-loss mapping [83] and also allows for a certain proportion of parallel evolution. A Python script along with the data to run all analyses can be downloaded from: https://gist.github.com/LinguList/7475830. The advantage of the IELex is that known borrowings are not only marked as such, but that they are also assigned to the cognate sets to which they would belong, if they were not borrowings. Thus, English *mountain* is clustered with the reflexes of Vulgar Latin **montanea* (derived from Latin *mōns*) in the Romance languages, such as, among others, French *montagne*, Italian *montagna*, and Spanish *montaña*. This gives us the possibility to test the usefulness of the refined MLN approach. We corrected some obvious errors in the data, especially in some of the Slavic languages (the whole dataset is provided in Supplementary Material I). Excluding 1,864 words that could not be shown to be cognate to any other word in the data, this yielded a total of 1,190 cognate sets. As a reference tree, we chose the one provided by Ethnologue [88]. The choice of this tree is for practical reasons, since it was proposed independently of quantitative methods, and reflects an openly available "quasi-standard". This does not mean that we are unaware of the many problems that this tree contains, especially in the classification of the subgroups.

Figure 3 shows the rooted phylogenetic network that the refined MLN

approach reconstructed from the data. As can be seen, the method nicely recovers some well-known cases of contact relations among the languages in the sample. English, for example shows two heavily weighted edges, one with the ancestor of the Scandinavian languages, and one with the ancestor of the Romance languages, nicely reflecting two of its major donors: Scandinavian words made their way into the English lexicon as a result of Danish and northern Scandinavian invasions starting in the 8th–9th century [89], and Old Norman (a northern French dialect) came to England as a result of the Norman conquest in 1066. Old Norman even developed into a distinct variety called Anglo-Norman which was spoken in England by the higher social strata from 12th to 15th century. The ensuing intensive language contact results in a boom of "French" loans, which eventually became a formative element of the English lexicon [89]. Albanian shows also strong connections with the ancestor of the Romance languages, reflecting the large number of Latin loanwords in the language [49].

Of the 105 cognate sets in the data that contain known hidden borrowings, the method identifies 76 correctly (see the specific results in Supplementary Material I). In total, the method identifies 369 out of 1,190 cognate sets (31%) that do not correspond to the reference tree. If the number of known borrowings reflected the true amount of borrowings in the data, and the reference tree displayed the true vertical history of the languages, this would mean that the method largely overstates the amount of lateral transfer. However, given the uncertainty regarding the subgrouping of the Indo-European languages that is also reflected in the reference tree, and the uncertainty of the cognate judgments in the data, we are confident that the results provide a good starting point for further research that may reveal further hidden borrowings and erroneous cognate judgments.

This can be exemplified by an inspection of the specific results that the method yields for English: Of the 32 borrowings into English [18], eight are singletons and five have reflexes in almost all Germanic languages in the sample and can thus technically not be identified by the MLN approach. Of the remaining 19 words, 17 (89%) are correctly identified. 17 further words are found to be not compatible with the reference tree, but three of these words are known borrowings in other languages. Of the remaining 14 words, four words (*belly*, *narrow*, *dull*, *smoke*), are obviously erroneously coded, since they are linked with words outside the Germanic branch, although their deeper etymology or the etymology of their presumed cognates is unclear; and four words (*at*, *leaf*, *small*, *know*) seem to be real cases of parallel semantic development (be it retention or innovation) with other languages (see Supplementary Material II). The remaining six words (*back*, *few*, *many*, *snake*, *tree*, *with*) are exclusively shared with the Scandinavian languages inside the Germanic branch. Whether this pattern results from innovations on the West Germanic mainland, by which the reflexes of the words in Frisian, German, and Dutch were replaced, or from hitherto unnoticed Scandinavian influence requires further investigation. A full list of all words with further comments is supplied in Supplementary Material II.

The modified MLN approach is surely not perfect. It heavily relies on the underlying data, and especially the selection of the reference tree can have a strong influence on the results. Furthermore, it can only recover those cases of borrowing that occur inside a given language family. External influences cannot be recovered. Further research is required in order to assess to which degree it overestimates borrowing rates because of its incapacity of handling independent parallel developments. However, it is a first step en route to more realistic quantitative models of language evolution, and could prove useful for scholars working on quantitative applications in historical linguistics, since it not only tests the tree-likeness of datasets but also provides direct hints as to the characters that cause reticulation. It can help us to improve the quality of our datasets by identifying possible hidden borrowings and erroneous cognate assignments.

Conclusion and outlook

Different metaphors and models have, over the past century or two, been developed to describe the evolution of languages, but realistic quantitative models that can explain horizontal evolutionary processes in addition to genealogical relationships were lacking. Since similar evolutionary processes shaped both genomes and languages into contemporary forms, it is possible to apply methods that are developed to study genome evolution to study language evolution. Since lateral transfer in language evolution constitutes a real form of natural variation, phylogenetic network approaches provide a better means to model language evolution than strictly bifurcating phylogenetic trees. We strongly support the recent attempts to strengthen the quantitative basis of historical linguistics by building large databases and adapting computational methods from biology. Great work has been done in the past 10 years, and we know that errors are unavoidable when building large databases that accumulate historical linguistic knowledge. However, since errors are not only unavoidable, but – in the case of undetected borrowings – also reflect one vivid aspect of language history, we think it is time to rethink claims about the major processes underlying language evolution. Applying network approaches in historical linguistics can provide new insights into both the vertical and the lateral components of language history, and help to bring traditional and more quantitative research closer together.

Acknowledgements

This research was supported by the ERC grants 240816 and F020515005. We thank Søren Wichmann and two anonymous reviewers for constructive comments. We also thank Michael Dunn for providing us with the Indo-European data which we used for the analyses presented in this study.

References

1. **Kiparsky P.** 2014. New perspectives in historical linguistics. In Bowern C, Evans B, eds; *The Routledge Handbook of Historical Linguistics*. London and New York: Routledge.
2. **Kroeber AL**, **Chrétien CD.** 1937. Quantitative classification of Indo-European languages. *Language* **13**: 83–103.
3. **Ross ASC.** 1950. Philological probability problems. *J R Stat Soc* **12**: 19–59.

4. **Swadesh M.** 1952. Lexico-statistic dating of prehistoric ethnic contacts. *Proc Am Philos Soc* **96**: 452–63.
5. **Swadesh M.** 1955. Towards greater accuracy in lexicostatistic dating. *Int J Am Linguist* **21**: 121–37.
6. **Sankoff D.** 1970. On the rate of replacement of word-meaning relationships. *Language* **46**: 564–9.
7. **Embleton SM.** 1986. *Statistics in Historical Linguistics*. Bochum: Studienverlag Brockmeyer.
8. **Starostin SA.** 1989. Sravnitel'no-istoričeskoe jazykoznanie i leksikostatistika [Comparative historical linguistics and lexicostatistics]. In Kullanda SV, Longinov JD, Militarev AJ, Nosenko EJ, et al., eds; *Materialy k Diskussijam na Konferencii [Materials for the Discussions at the Conference]*. Moscow: Institut Vostokovedenija. p. 3–39.
9. **Holden CJ.** 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. *Proc Biol Sci* **269**: 793–9.
10. **Ringe D**, **Warnow T**, **Taylor A.** 2002. Indo-European and computational cladistics. *Trans Philol Soc* **100**: 59–129.
11. **Gray RD**, **Atkinson QD.** 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426**: 435–9.
12. **Atkinson QD**, **Gray RD.** 2006. How old is the Indo-European language family? Illumination or more moths to the flame? In Forster P, Renfrew C, eds; *Phylogenetic Methods and the Prehistory of Languages*. Cambridge: McDonald Institute for Archaeological Research. p. 91–109.
13. **Bouckaert R**, **Lemey P**, **Dunn M**, **Greenhill SJ**, et al**.** 2012. Mapping the origins and expansion of the Indo-European language family. *Science* **337**: 957–60.
14. **Dunn M**, **Levinson SC**, **Lindstroem E**, **Reesink G**, et al**.** 2008. Structural phylogeny in historical linguistics: methodological explorations applied in island Melanesia. *Language* **84**: 710–59.
15. **Gray RD**, **Jordan FM.** 2000. Language trees support the express-train sequences of Austronesian expansion. *Nature* **450**: 1052–5.
16. **Pagel M.** 2009. Human language as a culturally transmitted replicator. *Nat Rev Genet* **10**: 405–15.
17. **Holm HJ.** 2007. The new arboretum of Indo-European "trees". *J Quant Linguist* **14**: 167–214.
18. **Donohue M**, **Denham T**, **Oppenheimer S.** 2012. New methodologies for historical linguistics? Calibrating a lexicon-based methodology for diffusion vs. subgrouping. *Diachronica* **29**: 505–22.
19. **Geisler H**, **List J-M.** 2014. Beautiful trees on unstable ground. Notes on the data problem in lexicostatistics. In Hettrich H, Ziegler S, eds; *Die Ausbreitung des Indogermanischen. Thesen aus Sprachwissenschaft, Archäologie und Genetik [The Spread of Indo-European. Theses From Linguistics, Archaeology, and Genetics]*. Wiesbaden: Reichert.
20. **Häkkinen J.** 2012. Problems in the method and interpretations of the computational phylogenetics based on linguistic data. An example of wishful thinking: Bouckaert et al. 2012. URL: http://www.elisanet.fi/alkupera/Problems_of_phylogenetics.pdf.
21. **Schleicher A.** 1853. O jazyku litevském, zvlástě na slovanský [On the Lithuanian language, and specifically on Slavic]. *Časopis Česekého Museum [J Czech Mus]* **27**: 320–4.
22. **Schleicher A.** 1853. Die ersten Spaltungen des indogermanischen Urvolkes [The first splits of the Proto-Indo-European people]. *Allgemeine Monatsschrift für Wissenschaft und Literatur [Mon J Sci Lit]* **3**: 786–7.
23. **Sutrop U.** 1999. Diskussionsbeiträge zur Stammbaumtheorie [Discussing the theories of family trees]. *Fenno-Ugristica* **22**: 223–51.
24. **Čelakovský FL.** 1853. *Čtení o Srovnavací Mluvnici Slovanské [Readings on the Comparison of Slavic Languages]*. V komisí u F. Řivnáče: Prague.
25. **Schleicher A.** 1848. *Zur Vergleichenden Sprachengeschichte [On Comparative Language History]*. Bonn: König.
26. **Schleicher A.** 1863. *Die Darwinsche Theorie und die Sprachwissenschaft [Darwin's Theory and Linguistics]*. Leipzig: Hermann Böhlau.
27. **Darwin C.** 1859. *On the Origin of Species by Means of Natural Selection, or, the Preservation of Favoured Races in the Struggle for Life*. London: John Murray.
28. **Oppenheimer JM.** 1987. Haeckel's variations on Darwin. In Hoenigswald HM, ed; *Biological Metaphor and Cladistic Classification: An Interdisciplinary Perspective*. Philadelphia: University of Pennsylvania Press. p. 123–35.
29. **Fitch WM**, **Margoliash E.** 1967. Construction of phylogenetic trees. *Science* **155**: 279–84.
30. **Woese CR**, **Kandler O**, **Wheelis ML.** 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA* **87**: 4576–9.
31. **Geisler H**, **List JM.** 2013. Do languages grow on trees? The tree metaphor in the history of linguistics. In Fangerau H, Geisler H, Halling T, Martin W, eds; *Classification and evolution in biology, linguistics and the history of science. Concepts – methods – visualization*. Stuttgart: Steiner. p. 111–24.
32. **Schmidt J.** 1872. *Die Verwantschaftsverhältnisse der Indogermanischen Sprachen [The Relationship of the Indo-European Languages]*. Leipzig: Hermann Böhlau.
33. **Schuchardt H.** 1900. *Über die Klassifikation der Romanischen Mundarten. Probe-Vorlesung, Gehalten zu Leipzig am 30. April 1870* [On the Classification of the Romance Dialects. Test lecture, held in Leipzig 30th of April 1870]. Graz.
34. **Campbell L.** 1999. *Historical Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.
35. **Anttila R.** 1972. *An Introduction to Historical and Comparative Linguistics*. New York: Macmillan.
36. **Hoenigswald HM.** 1990. Does language grow on trees? *Proc Am Philos Soc* **134**: 10–8.
37. **Stefenelli A.** 1992. *Das Schicksal des Lateinischen Wortschatzes in den Romanischen Sprachen [The Fate of the Lexicon in the Romance Languages]*. Rothe.
38. **Schulte K.** 2009. Loanwords in Romanian. In Haspelmath M, Tadmor U, eds; *Loanwords in the World's Languages*. Berlin and New York: de Gruyter. p. 231–59.
39. **Thomason S**, **Kaufman T.** 1988. *Language Contact, Creolization, and Genetic Linguistics*. Berkeley: University of California Press.
40. **Haspelmath M**, **Tadmor U.** 2009. The loanword typology project and the world loanword database. In Haspelmath M, Tadmor U, eds; *Loanwords in the World's Languages*. Berlin and New York: de Gruyter. p. 1–34.
41. **Tadmor U.** 2009. Loanwords in the world's languages. In Haspelmath M, Tadmor U, eds; *Loanwords in the World's Languages*. Berlin and New York: de Gruyter. p. 55–75.
42. **Dunn M**, **Terrill A**, **Reesink G**, **Foley RA**, et al**.** 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science* **309**: 2072–5.
43. **Colonna V**, **Boattini A**, **Guardiano C**, **Dall'ara I**, et al**.** 2010. Long-range comparison between genes and languages based on syntactic distances. *Hum Hered* **70**: 245–54.
44. **Longobardi G**, **Guardiano C**, **Silvestri G**, **Boattini A**, et al**.** 2013. Toward a syntactic phylogeny of modern Indo-European languages. *J Hist Linguist* **3**: 122–52.
45. **Campbell L**, **Poser WJ.** 2008. *Language Classification: History and Method*. Cambridge: Cambridge University Press.
46. **Dybo A**, **Starostin G.** 2008. In defense of the comparative method, or the end of the Vovin controversy. In Smirnov IS, ed; *Aspekty Komparativistiki [Aspects of Comparativistics]*. Moscow: RGGU. p. 119–258.
47. **Swadesh M.** 1950. Salish internal relationships. *Int J Am Linguist* **16**: 157–67.
48. **Grant AP.** 2009. Loanwords in British English. In Haspelmath M, Tadmor U, eds; *Loanwords in the World's Languages*. Berlin and New York: de Gruyter. p. 360–82.
49. **Holm HJ.** 2011. "Swadesh lists" of Albanian revisited and consequences for its Position in the Indo-European Languages. *J Indo-Eur Stud* **39**: 43–99.
50. **Kluge F**, **Seebold H.** eds; 2002. *Etymologisches Wörterbuch der Deutschen Sprache [Etymological dictionary of the German language]*. Berlin and New York: de Gruyter.
51. **Barbançon F**, **Evans SN**, **Ringe D**, **Warnow T.** 2013. An experimental study comparing linguistic phylogenetic reconstruction methods. *Diachronica* **30**: 143–70.
52. **Dyen I**, **Kruskal JB**, **Black P.** 1992. An Indoeuropean classification. *Trans Am Philos Soc* **82**: 1–132.
53. **Greenhill SJ**, **Blust R**, **Gray RD.** 2008. The Austronesian basic vocabulary database: from bioinformatics to lexomics. *Evol Bioinforma* **4**: 271–83.
54. **Dunn M**, **Greenhill SJ**, **Levinson SC**, **Gray RD.** 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* **473**: 79–82.
55. **Zhang J.** 2003. Evolution by gene duplication: an update. *Trends Ecol Evol* **18**: 292–8.
56. **Treangen TJ**, **Rocha EP.** 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet* **7**: e1001284.
57. **Popa O**, **Dagan T.** 2011. Trends and barriers to lateral gene transfer in prokaryotes. *Curr Opin Microbiol* **14**: 615–23.
58. **Lukjancenko O**, **Wassenaar TM**, **Ussery DW.** 2010. Comparison of 61 sequenced *Escherichia coli* genomes. *Microb Ecol* **60**: 708–20.
59. **Bapteste E**, **O'Malley M**, **Beiko R**, **Ereshefsky M**, et al**.** 2009. Prokaryotic evolution and the tree of life are two different things. *Biol Direct* **4**: 34.
60. **Puigbo P**, **Wolf YI**, **Koonin EV.** 2010. The tree and net components of prokaryote evolution. *Genome Biol Evol* **2**: 745–56.
61. **Southworth FC.** 1964. Family-tree diagrams. *Language* **40**: 557–65.

62. **Holzer G.** 1996. *Das Erschließen Unbelegter Sprachen [Reconstructing unattested languages]*. Frankfurt am Main: Lang.
63. **Huson DH**, **Rupp R**, **Scornavacca C.** 2010. *Phylogenetic Networks*. Cambridge: Cambridge University Press.
64. **Morrison DA.** 2011. *An Introduction to Phylogenetic Networks*. Uppsala: RJR Productions.
65. **Bandelt HJ**, **Dress AW.** 1992. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Mol Phylogenet Evol* **1**: 242–52.
66. **Bryant D**, **Moulton V.** 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol* **21**: 255–65.
67. **Hamed MB**, **Wang F.** 2006. Stuck in the forest: trees, networks and Chinese dialects. *Diachronica* **23**: 29–60.
68. **Hamed MB.** 2005. Neighbour-nets portray the Chinese dialect continuum and the linguistic legacy of China's demic history. *Proc R Soc B* **272**: 1015–22.
69. **Heggarty P**, **Maguire W**, **McMahon A.** 2010. Splits or waves? Trees or webs? How divergence measures and network analysis can unravel language histories. *Philos Trans R Soc B* **365**: 3829–43.
70. **McMahon A**, **Heggarty P**, **McMahon R**, **Slaska N.** 2005. Swadesh sublists and the benefits of borrowing: an Andean case study. *Trans Philol Soc* **103**: 147–70.
71. **Bowern C.** 2010. Historical linguistics in Australia: trees, networks and their implications. *Philos Trans R Soc B* **365**: 3845–54.
72. **Saitou N**, **Nei M.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**: 406–25.
73. **Ronquist F.** 2004. Bayesian inference of character evolution. *Trends Ecol Evol* **19**: 475–81.
74. **Gray RD**, **Bryant D**, **Greenhill SJ.** 2010. On the shape and fabric of human history. *Philos Trans R Soc B* **365**: 3923–33.
75. **Wichmann S**, **Holman EW**, **Raman T**, **Walker RS.** 2011. Correlates of reticulation in linguistic phylogenies. *Lang Dyn Chang* **1**: 205–40.
76. **Koonin EV**, **Makarova KS**, **Aravind L.** 2001. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol* **55**: 709–42.
77. **Huson DH**, **Scornavacca C.** 2011. A survey of combinatorial methods for phylogenetic networks. *Genome Biol Evol* **3**: 23–35.
78. **Wang WS-Y**, **Minett JW.** 2005. Vertical and horizontal transmission in language evolution. *Trans Philol Soc* **103**: 121–46.
79. **Minett JW**, **Wang WS-Y.** 2003. On detecting borrowing. *Diachronica* **20**: 289–330.
80. **Nakhleh L**, **Ringe D**, **Warnow T.** 2005. Perfect phylogenetic networks: a new methodology for reconstructing the evolutionary history of natural languages. *Language* **81**: 382–420.
81. **Nichols J**, **Warnow T.** 2008. Tutorial on computational linguistic phylogeny. *Linguist Compass* **2**: 760–820.
82. **Dagan T**, **Artzy-Randrup Y**, **Martin W.** 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc Natl Acad Sci USA* **105**: 10039–44.
83. **Mirkin BG**, **Fenner TI**, **Galperin MY**, **Koonin EV.** 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol* **3**: 2.
84. **Dagan T**, **Martin W.** 2007. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci USA* **104**: 870–5.
85. **Cohen O**, **Pupko T.** 2011. Inference of gain and loss events from phyletic patterns using stochastic mapping and maximum parsimony – a simulation study. *Genome Biol Evol* **3**: 1265–75.
86. **Nelson-Sathi S**, **List J-M**, **Geisler H**, **Fangerau H**, et al**.** 2011. Networks uncover hidden lexical borrowing in Indo-European language evolution. *Proc R Soc B* **278**: 1794–803.
87. **List J-M**, **Moran S.** 2013. An open source toolkit for quantitative historical linguistics. *Proc ACL 2013 Syst Demonstr* 13–8.
88. **Lewis MP**, **Fennig CD.** 2013. *Ethnologue*. Dallas: SIL International.
89. **Harbert W.** 2007. *The Germanic Languages*. Cambridge: Cambridge University Press.

Using Phylogenetic Networks to Model Chinese Dialect History*

Johann-Mattis List
Forschungszentrum Deutscher Sprachatlas,
Philipps University Marburg, Marburg, Germany
mattis.list@uni-marburg.de

Shijulal Nelson-Sathi
Institute of Molecular Evolution,
Heinrich Heine University Düsseldorf, Düsseldorf, Germany
shijulalns@uni-duesseldorf.de

William Martin
Institute of Molecular Evolution,
Heinrich Heine University Düsseldorf, Düsseldorf, Germany
w.martin@uni-duesseldorf.de

Hans Geisler
Institute of Romance Languages and Literature,
Heinrich Heine University Düsseldorf, Düsseldorf, Germany
geisler@uni-duesseldorf.de

Abstract

The idea that language history is best visualized by a branching tree has been controversially discussed in the linguistic world and many alternative theories have been proposed. The reluctance of many scholars to accept the tree as the natural metaphor for language history was due to conflicting signals in linguistic data: many resemblances would simply not point to a unique tree. Despite these observations, the majority of automatic approaches applied to language data has been based on the tree model,

* First published in Søren Wichmann and Jeff Good (eds.). 2014. *Quantifying Language Dynamics: On the Cutting Edge of Areal and Phylogenetic Linguistics*, 125–154. Leiden: Brill.

while network approaches have rarely been applied. Due to the specific sociolinguistic situation in China, where very divergent varieties have been developing under the roof of a common culture and writing system, the history of the Chinese dialects is complex and intertwined. They are therefore a good test case for methods which no longer take the family tree as their primary model. Here we use a network approach to study the lexical history of 40 Chinese dialects. In contrast to previous approaches, our method is character-based and captures both vertical and horizontal aspects of language history. According to our results, the majority of characters in our data (about 54%) cannot be readily explained with the help of a given tree model. The borrowing events inferred by our method do not only reflect general uncertainties of Chinese dialect classification, they also reveal the strong influence of the standard language on Chinese dialect history.

Keywords

Chinese languages – Chinese linguistics – tree model – phylogenetic networks – lexical borrowing

1 Introduction

1.1 *Languages and Dialects*

What exactly is a language, and what is a dialect? One tends to say that the people from Shànghǎi, Běijīng, and Měixiàn all speak 'Chinese,' while people from Scandinavia speak 'Norwegian,' 'Swedish,' or 'Danish.' Looking at the phonetic transcriptions of the first sentence of Aesop's fable 'The Northwind and the Sun' in the three Chinese 'dialects' and the three Scandinavian 'languages' given in Table 1, the clear-cut distinction suggested by the different ways we name the varieties starts to become blurred. As the transcriptions show, the Chinese varieties differ from each other to a similar or even greater degree than the Scandinavian ones.

The reason for the fuzziness of the terms 'dialect' and 'language' can be found in the daily use of the terms in non-linguistic contexts. What is called a language and what a dialect does not necessarily depend on pure linguistic criteria, but often also on culture and politics (Barbour and Stevenson, 1998: 8). The problem of culture and politics, however, is that they have an impact on both languages and dialects. Although it certainly makes sense to state that Chinese dialects differ as much as the Scandinavian languages, it does not tell the whole truth about the sociolinguistic situation in China, where a large part

TABLE 1 *The first sentence of Aesop's fable 'The Northwind and the Sun' in different speech varieties. The words are semantically aligned, i.e. all translational equivalents are placed in the same column. Words shaded in gray are etymologically related*

Běijīng Chinese	iou^{21}	i^{55}	xuei^{35}	pei^{21}fəŋ^{55}	kən^{55}	tʰai^{51}iaŋ^{11}	ʈʂəŋ^{55}	tsai^{53}
Měixiàn Chinese	iu^{33}	it^{55}	pai^{33}a^{11}	pet^{33}fuŋ^{33}	tʰuŋ^{11}	ɲit^{11}tʰeu^{11}	hɔk^{33}	
Shànghǎu Chinese	ɦi^{22}		tʰã^{55} tsɪ^{21}	poʔ^{3}foŋ^{44}	taʔ^{5}	tʰa^{33}ɦiã^{44}	tsəŋ^{33}	hɔ^{44}

Běijīng Chinese (*cont.*)	naə̯^{51}	ʈʂəŋ^{55}luən^{51}	
Měixiàn Chinese	e^{53}	au^{55}	
Shànghǎu Chinese		ləʔ^{1}lə^{23}tsa^{53}	

Norwegian	nuːrɑʊinˈn̩	ɔ	suːlŋ̩					krɑŋlət	ɔm
Swedish	nuːɖanvɪndən	ɔ	suːlən		tɣɪstadə	ən gɔŋ			ɔm
Danish	noˀʌnvenˀn̩	ʌ	soːlˀn	kʰʌm		eŋ̊gɑŋ	i sd̥ʁiðˀ		ʌmˀ

of the population is bilingual, using a common language for writing and—if necessary—also for verbal communication. In order to describe such complex heterogeneous structures as modern languages, sociolinguists have proposed the *diasystem* model (Branner, 2006: 209). According to this model, a language is a complex aggregate of different linguistic systems coexisting and mutually influencing each other (Coseriu, 1973: 40). Usually, a diasystem is determined by a *Dachsprache* (*roof language*), a linguistic variety that serves as a standard for interdialectal communication (Goossens, 1973: 11).

In the case of the Chinese diasystem, the *Dachsprache* is the modern standard language (henceforth called Standard Chinese), which was originally derived from the dialect of Běijīng, but—being used as second language throughout China—has long started to live a life of its own. Its influence can be noticed in almost all dialects. Lexically, it often appears in terms of multiple words for a single concept, with one representing the word originally used in the dialect, and one being borrowed from Standard Chinese. In the example given in Table 1, for example, Shànghǎi [tʰa^{33}ɦiã^{44}] 'sun' has been borrowed from Standard Chinese 太阳 *tàiyáng* [tʰai^{51}iaŋ^{11}] 'sun.' This can be seen from the fact that there is another word for 'sun' in Shànghǎi: [ɳjɪʔ^{11}dʑ^{23}]. This word is much older than the former and is cognate with Měixiàn [ɲit^{11}tʰeu^{11}] 'sun.' Cases where dialects borrow from the *Dachsprache* are very frequent in almost all Chinese dialects, while cases of borrowing between neighboring dialects are probably even more frequent.

1.2 *Trees, Waves, and Networks*

Ever since August Schleicher first proposed the idea that the evolution of languages is best visualized by a branching tree ('dem Bilde eines sich verästelnden Baumes'; Schleicher, 1853: 787), this view has been controversially discussed in the linguistic world, leading to various opposing theories ranging from wave-like evolutionary scenarios (Schmidt, 1872) to early network proposals (Bonfante, 1931). Since most alternative approaches remained static, disregarding the time dimension in favor of the spatial dimension, the tree was never completely abandoned, and both the family tree (*Stammbaum*) and the wave theory (*Wellentheorie*) became standard models of language change that were used interchangeably, depending on the respective questions that scholars wanted to elaborate. Although, during the history of linguistics, the idea of combining both models into a single framework was often discussed (Schuchardt, 1900; Southworth, 1964), linguists failed to propose a formal model for phylogenetic networks that would have allowed both vertical and horizontal language relations to be captured. As historical linguistics took a quantitative turn at the beginning of the third millennium, many methods that had originally been designed to model and infer biological evolution were repeatedly applied to linguistic problems. While most of these approaches continued with the tree model, comparing languages with species (Gray and Atkinson, 2003; McMahon and McMahon, 2005; Atkinson and Gray, 2006), recent research has shown (Nelson-Sathi et al., 2011, List et al., 2014) that network approaches originally used to model microbial evolution (Dagan and Martin, 2007; Dagan et al., 2008) might be even more apt for modeling language history. Network approaches not only offer a formal way to model vertical and horizontal language relations, but also provide different methods for inferring these relations from linguistic data. So far, however, phylogenetic network approaches are still in their infancy, both with respect to the methods that have been proposed and with respect to their applications.

The Chinese dialects seem to be a good test case for these new approaches. Given their complex history, their 'close proximity to one another for two millennia and the pervasive influence of various quasi-standards and koinés on all Chinese dialects over a very long period' (Norman, 2003: 76), it is obvious that they are 'not entirely amenable to a Stammbaum formulation' (ibid.). Here we apply a network approach to model the history of 40 Chinese dialect varieties. In contrast to previous network analyses of Chinese dialects that were based on split distances and only measured the uncertainty of trees (Ben Hamed and Wang, 2006), our approach is character-based: it automatically infers hidden borrowings in the data and thus captures both the vertical and horizontal aspects of language history.

2 Materials

2.1 *Data*

The data that we used for our analysis is taken from the *Hànyǔ Fāngyán Yīnkù* (Hóu, 2004), a CD-ROM that offers different resources for Chinese dialects including phonological descriptions, phonetic transcriptions, and sound recordings for 40 different dialect varieties. From the CD-ROM we extracted a lexical subset, consisting of 180 glosses ('concepts') translated into the respective varieties. Chinese dialects often have multiple synonyms for one concept; therefore the resulting dataset comprises 10,201 words. Since the word lists were compiled for dialect studies where the selection of lexical items is usually based on phonetic criteria, only 48 of the 180 glosses (26%) belong to the basic vocabulary in the strict sense of Swadesh (1952 and 1955). The source material was obtained in a format not suitable for computational analyses, requiring the extraction procedure to be carried out semi-automatically, with additional manual cleaning by the researchers/present authors. All entries were double-checked by comparing the phonetic transcription for each word with its corresponding sound recording. The data was further enriched by looking up the geographic coordinates of the central cities where the varieties are spoken, translating the glosses into English, adapting the phonetic transcriptions to standard IPA, and applying a rough procedure for automatic cognate detection that is described in detail in the following section. Table 2 shows an excerpt of the data in its current format.

TABLE 2 *The basic format of the input data*[1]

ID	Variety	Concept	St. Chinese	IPA	Char.	Cogn. Set
1	Shànghǎi	'sun'	*tàiyáng* 太阳	$t^ha^{34-33}ɦiã^{13-44}$	太阳	2
2	Shànghǎi	'sun'	*tàiyáng* 太阳	$ȵjɪʔ^{1-11}dɤ^{13-23}$	日头	1
3	Sūzhou	'sun'	*tàiyáng* 太阳	$ȵiəʔ^{3}dʏ^{13-21}$	热头	3
4	Sūzhou	'sun'	*tàiyáng* 太阳	$t^hɑ^{513-55}ɦiã^{13-21}$	太阳	2
5	Hángzhōu	'sun'	*tàiyáng* 太阳	$t^hɛ^{445}ɦiɑŋ^{213-31}$	太阳	2
6	Wēnzhōu	'sun'	*tàiyáng* 太阳	t^ha^{42-22}ji	太阳	2

2.2 *Cognate Judgments*

Along with the recent quantitative turn in historical linguistics, one can also observe a shift from the interest in *proto-forms* to an interest in *cognates*. This likewise holds for our approach, which requires sets of cognate words as input data. Cognates are usually defined as words or morphemes that are derived from a common ancestor form via vertical inheritance (Trask, 2000: 62). Our input requirements are less strict, however: the method only requires that the words are etymologically related, or *homolog* in the biological sense, i.e. that they share a common ancestry, no matter whether this is due to vertical transfer or borrowing (Koonin, 2005: 311). In Chinese dialectology, it is common to specify not only the pronunciation of a given dialect word, but also give an assessment regarding its homology. Homology assessments are usually coded by providing the Chinese characters corresponding to a given word.[2] Since for most Chinese characters the Middle Chinese readings (spoken around the 6th century CE) can be reconstructed from old rime books, a character is somewhat similar to a proto-form. Thus, Táoyuán [ŋit^{22}tʰeu^{11}] and Hǎikǒu [zit^{3}hau^{31}] 'sun' are both written as 日头, and the proto-form would have been pronounced as *ɳit^{4}duw^{1} in Middle Chinese times (if the compound was already present during that time).[3] Note that the character assignments in Chinese dialectology are homologs in the strict sense, since no distinction is drawn between borrowing and vertical inheritance.

While the postulation of a proto-form for a given set of words is—ideally—a full statement regarding their phonetic and phylogenetic history, being a short-cut formulation for known, regular sound change processes, the postulation of cognate relations between words is much simpler, being merely a statement that there *is* a history relating them. It is usually emphasized that the nature of this history should only involve vertical transmission. The details of vertical transmission are usually ignored, and no further distinction between the

1 Note that the character assignment correctly claims that Sūzhou [ɳiəʔ^{3}dʏ$^{13-21}$] and Shànghǎi [ɳjɪʔ$^{1-11}$dɤ$^{13-23}$] are not cognate, with the initial syllable of the former going back to Middle Chinese *ɳet 'hot' and the initial syllable of the latter going back to Middle Chinese *ɳit 'sun'. The words are, however, closely related, since it is not impossible that the original form in Sūzhou was a reflex of Middle Chinese *ɳit 'sun', but was later reinterpreted as Middle Chinese *ɳet 'hot'. However, this does not influence our strict criterion for cognacy assignments.

2 The procedure for choosing the characters is not always clear-cut. See Kurpaska (2010: 118–120) for details.

3 Middle Chinese character readings follow an IPA adaptation of the system of Baxter (1992).

different types is drawn. Thus, in lexicostatistical databases, such as the *Tower of Babel Database* (http://starling.rinet.ru) or the *Indo-European Lexical Cognacy Database* (http://ielex.mpi.nl/), the Italian and French words for 'give,' *dare* and *donner* respectively, are usually placed in the same cognate set, although they go back to two different Latin words (*dare* 'give' and *dōnare* 'give as a present'). The reason for this cognate assignment is that the Latin forms themselves go back to a common Indo-European root, with *dare* being a reflex of Proto-Indo-European *deh$_3$- 'give' and *dōnare* being a reflex of its nominalized form *deh$_3$-no- 'what is given' (cf. Meiser, 1998). Trask (2000: 234 f.) proposes the term *oblique cognates* to address these specific cases of indirect cognate relations, but the term is rarely used in the literature, and direct and indirect cognacy are usually treated identically in practice.

Another problem of cognate assignment that is ignored in most quantitative approaches is the problem of *partial cognacy*. Is it justified to say that compound words such as Spanish *porque* and Russian *potomu čto* 'because' are cognate, since certain parts of them (*-que* and *čto*) can be traced back to Proto-Indo-European **kwi-* 'what'? And, if so, what is their relation when adding more words to the comparison, such as Danish *fordi* 'because,' which is partially cognate with the Spanish word (*for-* ≈ *por-*) but not the Russian? In most datasets, this problem is solved by assigning compound words to multiple cognate sets, one for each morpheme. Such an approach, however, can become problematic when dealing with languages where compounding is frequent. In Table 3, the words denoting 'moon' in seven Chinese dialects are contrasted in such a way that all cognate morphemes are aligned, with the characters in the first row representing the cognate set. As can be seen from this Table, the assignment of all morphemes to a specific cognate set yields as many cognate sets as there are dialects. Given that quantitative approaches to phylogenetic reconstruction usually assume the development of all cognate sets to be independent, an assignment of all cognate morphemes to a single cognate set would therefore not only drastically increase the amount of cognate sets, but would also be entirely unrealistic, since these cognate morphemes surely did not evolve independently from each other.

In order to cope with the problems of indirect and partial cognacy, we decided to apply a very strict procedure of cognate assignment, grouping only those terms into cognate sets that correspond to identical sequences of Chinese characters. Since the data contained 244 entries for which no corresponding Chinese character was identified (and therefore no cognate assignment could be made), we excluded these entries. The remaining 9,957 words were grouped into 3,061 cognate sets. The cognate sets were then converted into a binary presence-absence matrix, where the columns represented the taxa, and the

TABLE 3 *Problem of partial cognacy in the Chinese dialects. The table shows cognate morphemes of translations of the concept 'moon' in seven Chinese dialects. As can be seen from the table, no two words are completely cognate, although all words share at least one cognate morpheme.*

	Cognate Sets						
Dialect	月	亮	光	呢	奶	明	爷
Shànghǎi	ɦyɪʔ$^{1-11}$	liã$^{13-23}$					
Wēnzhōu	ɳʏy$^{213-21}$		kuɔ^{33}				
Xiàmén	geʔ$^{5-21}$						
Jiàn'ōu	ŋyɛ^{42}			ni^{44}	nai^{33}		
Tàiyuán	yəʔ$^{2-54}$					mi^{45}	
Píngyáo	yʌʔ^{53}					mi$^{13-53}$	iɛ$^{13-31}$
Zhèngzhōu	yɛ^{24}				nai^{53}nai$^{53-24}$		

rows corresponded to distinct presence-absence patterns for a given cognate set, with 1 indicating the presence of a reflex and 0 indicating its absence. Since our method requires that a given cognate set has reflexes in at least two taxa, we excluded 2,005 cognate sets that were reflected only in one taxon. Our presence-absence matrix was thus reduced to a total of 1,056 presence-absence patterns.

2.3 *Reference Trees*

Our method estimates the extent to which the evolution of a set of characters (cognate sets reflected in the presence-absence patterns) can be explained by an evolutionary scenario that allows for only the vertical inheritance of characters. This scenario has to be defined with the help of a *reference tree* that captures the history of the language varieties under investigation. Given the specific sociolinguistic situation in China, the classification of the Chinese dialects is extremely difficult, and the opinions of scholars differ to a great extent (see Karlgren, 1954; Lǐ, 2005; Norman, 2003; Wáng, 2009, and the overview in Kurpaska, 2010: 36–62). The most common grouping distinguishes seven major dialect groups, namely (1) Mandarin (Guānhuà), (2) Xiāng, (3) Gàn, (4) Wú, (5) Hakka (Kèjiā), (6) Cantonese (Yuè), and (7) Mǐn (Norman, 1988: 181). However, alternative approaches that subdivide these varieties further are also quite popular, and at least three additional groups, namely Jìn (otherwise assigned to Mandarin), Huī (otherwise assigned to either Wú or Mandarin), and Pínghùa (otherwise assigned to Cantonese), are often proposed and discussed in the lit-

TABLE 4 *The dialect groups in our sample*

Group	Chinese	Altern. Grouping	# Dialects
Mandarin (Guānhuà)	管话		17
Jìn	晋	Mandarin	3
Xiāng	湘		2
Gàn	赣		1
Huī	徽	Wú, Mandarin	2
Wú	吴		4
Hakka (Kèjiā)	客家		2
Cantonese (Yuè)	粵		2
Pínghuà	平话	Cantonese (Yuè)	1
Mǐn	闽		6

erature (Kurpaska, 2010: 64–73). The ten major dialect groups are summarized in Table 4, along with alternative classifications and the number of varieties in our sample that belong to each group.

Most classifications group the Chinese dialects by comparing their deviation from the phonological system of Middle Chinese. One of the most salient features is the series of voiced plosives (*b, *d, *g, etc.) in Middle Chinese (Kurpaska, 2010: 35). These plosives show varying reflexes in the Chinese dialects. Sometimes they are retained completely (> b, d, g), sometimes all of them are devoiced (> p, t, k), sometimes the devoicing is accompanied by aspiration (> p^h, t^h, k^h), and sometimes the reflexes are split into a voiceless unaspirated and a voiceless aspirated series (> p/p^h, t/t^h, k/k^h). As Lǐ (2005) demonstrates, these reflexes are sufficient to distinguish six of the seven standard dialect groups, with Gàn and Hakka being merged into a single group.[4] However, the problem of this criterion (and most other classification criteria) is that they are merely used to *distinguish* certain dialect groups, while they do not *explain* how they developed. Although most classifications proposed thus far are based on historical criteria, few of them explicitly try to account for the genealogical development of the Chinese dialects.

4 Lǐ (2005) distinguishes different contexts in which the split of voiced to voiceless unaspirated and voiceless aspirated plosives occurred in order to distinguish Mǐn, Cantonese, and Mandarin.

Different theories have been proposed regarding the history of the major dialect groups. Among the most popular is Karlgren's (1954: 212) theory that almost all of today's Chinese dialects (except from the Mǐn dialects) go back to a *koiné* that was very widespread during the 6th century. He further states that this language was identical to Middle Chinese, the language whose phonological characteristics are recorded in the rime books that were compiled during that time. Norman (1988 and 2003) proposes a different theory, according to which Hakka, Cantonese, and Mǐn can be traced back to a common ancestor which split from the remaining dialects before the Middle Chinese period.

Based on these two different theories, we created two reference trees, one reflecting Norman's *Southern Chinese hypothesis*, and one reflecting Karlgren's *Common Chinese hypothesis*. In order to increase the distance between the trees, and since we could not determine the exact subgrouping of all major dialect groups from the literature alone, we added further differences to the subgroupings. Thus, in the Southern Chinese tree we grouped Wú and Huī dialects together, while in the Common Chinese tree we placed Huī closest to the Mandarin-Jìn group. In a similar way, we merged Hakka and Gàn in the Common Chinese tree following a reasonably popular proposal (see Sagart, 2002: 129–132), while assigning them to separate groups in the Southern Chinese tree. We also classified the Jìn dialects as a Northern Mandarin group in the Southern Chinese tree, while classifying them as first outgroup of Mandarin in the Common Chinese tree. For the internal subgrouping of the major dialect groups in both hypotheses, we generally employed the groupings proposed in the *Language Atlas of China* (Wurm and Liú, 1987). In cases where these groupings were too shallow and additional information was available, this internal subgrouping was further modified. Here, the internal classification of the Mǐn dialects was changed according to the classification in Norman (1991), and the eight groups of Mandarin dialects were further subdivided following Norman (1988).[5] Both reference trees for the major groups are given in Fig. 1. In order to test for possible differences between these 'traditional' reference trees and reference trees calculated from automatic approaches, we reconstructed two additional reference trees automatically. We applied the UPGMA algorithm (Sokal and Michener, 1958) and the Neighbor-joining

5 We are well aware of the fact that neither of the two trees can really claim to represent the true history of the Chinese dialects. However, as long as there are no detailed proposals regarding the genealogical classification of the Chinese dialects, we think it is more fruitful to accept uncertainties and possible mistakes resulting from the given trees than to abstain from the analysis in general.

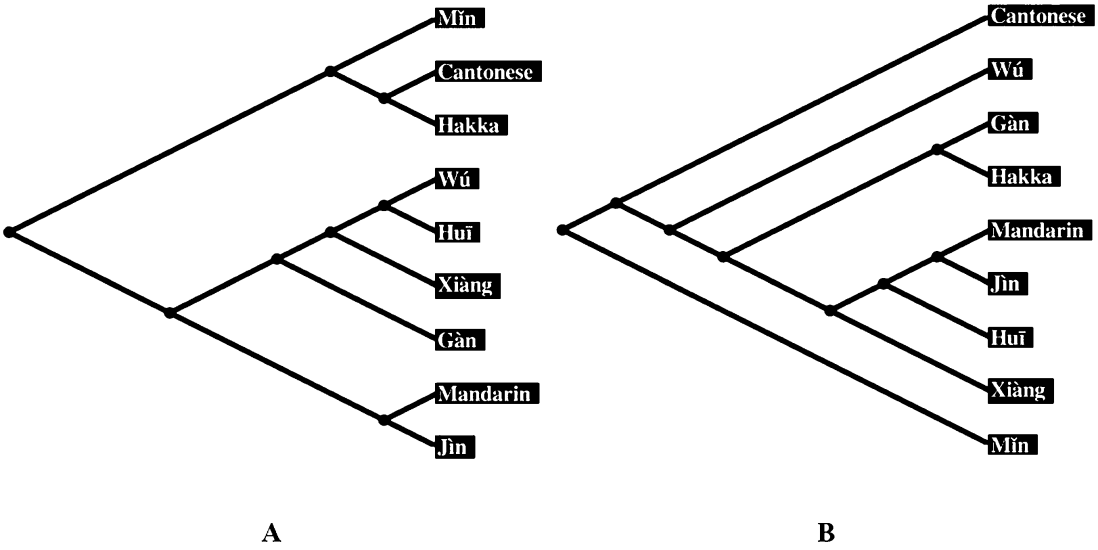


FIGURE 1 *Reference trees of the major groups for the Southern Chinese* (A) *and the Common Chinese* (B) *hypotheses. The reference trees are broadly based on the classifications of Norman* (*1988 and 2003*) *and Karlgren* (*1954*)*, respectively, with the topologies expanded and adapted to accommodate the present sample* (*see text*).

algorithm (Saitou and Nei, 1987) to distance matrices derived from shared cognate percentages between all dialect pairs. The complete reference trees for all four analyses are given in Supplementary Material 1.

3 Methods

Building on the *minimal lateral network* (MLN) approach by Nelson-Sathi et al. (2011), our methods are based on an improved framework for the reconstruction of rooted phylogenetic networks (List et al., 2014). In contrast to the original approach, we introduce a refined method for *gain-loss mapping*. This method offers more flexible models with varying numbers of gain and loss events, captures multifurcation in reference trees, and also handles a certain amount of parallel evolution. Furthermore, we present a new method that derives *spatial networks* from rooted phylogenetic networks by plotting the results of the MLN approach to geographic maps. The new method is implemented as part of LingPy, an open source Python library for automatic tasks in historical linguistics (List and Moran, 2013, Version 2.2).

3.1 *Gain-Loss Mapping*

As pointed out before, any model of language evolution must take into account vertical as well as horizontal relations—i.e., borrowing. Borrowing processes

can be incredibly complex. Nevertheless, they usually leave observable traces, so that the borrowed word is often phonetically quite similar to the donor word. Furthermore, since the process of borrowing itself is not tree-like, borrowings that are mistaken for cognates can show up in form of presence-absence patterns that cannot be readily explained by the branching patterns of a family tree alone. As an example, compare the most widespread words for 'mountain' in the Germanic languages (German *Berg*, Dutch, Swedish *berg*, Danish *bjerg*) with the English word *mountain*. Assuming that English is a Germanic language, we see an astonishing difference to supposedly related languages. However, there is a striking similarity with words meaning 'mountain' in Romance languages such as Italian *montagna*, Spanish *montaña*, Portuguese *montanha*, and French *montagne*. If we had further evidence regarding the history of the languages and their branching patterns, there are two possible scenarios which could account for this coincidence: (1) English *mountain* is truly cognate with the Romance words, and reflexes of the word came to be lost in all other Germanic languages, or (2) English *mountain* was borrowed from one of the Romance languages, thereby replacing Old English *beorg*, the regular English reflex of Proto-Germanic *bergan 'mountain.' Given the branching pattern of the Germanic languages, it is much more plausible to assume the latter scenario (and indeed, historical evidence shows that English 'mountain' was borrowed from Old French *montaigne*). Thus, if languages show patterns of shared cognates that are in conflict with a given family tree, these patterns may be taken as a heuristic device for the detection of hitherto unrecognized borrowings.

As the example of English *mountain* shows, it is possible to gain some basic insights into language history by simply investigating the dynamics of gain and loss events. In evolutionary biology, the analysis of gain-loss scenarios (also called *presence-absence patterns* or *phyletic patterns*) is a common heuristic to identify possible instances of lateral gene transfer, and different methods for analyzing such patterns have been proposed in the recent past (see the overview in Cohen et al., 2010).

The basic idea of all these approaches is to create *gain-loss scenarios* for a given set of characters. A gain-loss scenario explains how a particular phyletic pattern could have evolved along a given reference tree. For a given pattern, each node of the tree is assigned to one of two possible states indicating the presence (1) or the absence (0) of the character in the pattern. *Events* are changes in the states from ancestral nodes to their direct descendants. A *gain event* (also called *origin*) is defined as the change from state 0 to state 1, and a *loss event* is defined as the change from state 1 to state 0. If the most appropriate analysis of a given phyletic pattern supports multiple gains (origins) of a character, this is usually taken as evidence for possible events of

TABLE 5 *Phyletic patterns of the cognate sets for 'mountain'*

Language	Spanish	Portuguese	French	English	German	Swedish
'mountain'	*montaña*	*montanha*	*montagne*	*mountain*	*Berg*	*berg*
Pattern M	1	1	1	1	0	0
Pattern B	0	0	0	0	1	1

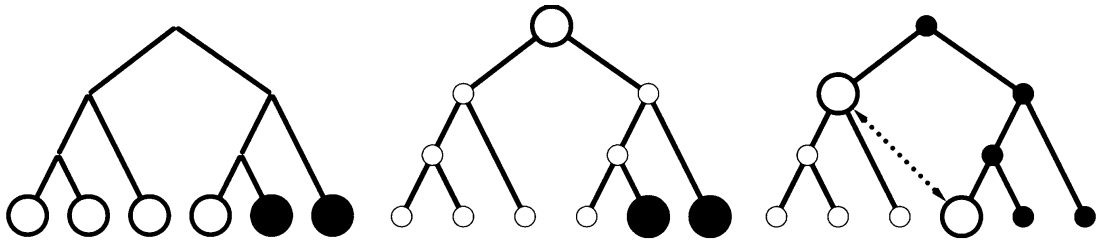


FIGURE 2 *Comparing alternative gain-loss scenarios. White nodes indicate the presence of a character, black nodes its absence. Large nodes indicate the respective event (gain or loss). In A, no scenario is inferred, B assumes one gain and two loss events, and C assumes two gain events and no loss event.*

lateral transfer (borrowing) that occurred during the evolution of the character. Table 5 illustrates how phyletic patterns are derived from the translation of 'mountain' into six Indo-European languages. For this group of languages, there are two different phyletic patterns, labeled M and B for convenience. Given the history of the six languages, Pattern B is unproblematic, supporting only a single origin hypothesis, with a loss of the character in English, and the gain of the character in the root. Pattern M (see Fig. 2A), however, can be mapped in two different ways: using a two-loss scenario as illustrated in Fig. 2B (scenario (1) above), or a two-gain scenario (scenario (2)), as illustrated in Fig. 2C. While the two-loss scenario infers that the character originated only once (in the root), the two-gain scenario infers two distinct origins for the character. Therefore, a lateral link between the two origins can be drawn, illustrated by the dotted line in Fig. 2C. This link is basically undirected, since it is not clear in which direction the borrowing event occurred. With this inference procedure, it is also not possible to determine when the link occurred, which explains why the link is drawn between the nodes in the tree where the characters originate.

Gain-loss scenarios can be inferred in different ways. Nelson-Sathi et al. (2011) follow Dagan and Martin (2007) in employing a *binary-branching top-*

down approach with different basic models, allowing for varying amounts of gains in a given phyletic pattern. The drawback of this approach is that the number of origins per phyletic pattern can only be an exponentiation of the base 2 (1, 2, 4, 8, 16, etc.), which results in a drastic restriction of the number of origins allowed by each model. A further drawback of this approach is that it can only be applied to bifurcating reference trees. This requirement is less problematic in biological applications since bifurcating reference trees are usually reconstructed automatically from the data. In linguistics, however, scholars are very cautious to propose detailed phylogenies, and multifurcating language trees (*soft polytomies* in the terms of Nunn, 2011: 22) are often used to reflect their uncertainty.

In order to overcome these shortcomings, we developed a *parsimony-based bottom-up approach* that allows for varying numbers of gains, depending on the phyletic pattern under investigation. In comparison with the top-down approach, our approach offers an increased number of models that can be tested on a given dataset. It also no longer restricts the maximal number of gain events that can be inferred by a given model, and—since the method is based on an exhaustive search of all possible scenarios—its application to multifurcating reference trees does not result in theoretical or practical problems.

Our approach is quite simple: given a phyletic pattern (a cognate set), there can be different gain-loss scenarios that could explain the evolution of the pattern. In order to find a consistent way of selecting the most parsimonious scenario, we test different *models* that assign different penalties for the scenarios, depending on the number of gain and loss events proposed by them. A model is defined as the ratio between penalties for gain and loss events. The model 2–1, for example, penalizes gain events with 2 and loss events with 1. The most parsimonious scenario for a given model is the one which minimizes the overall penalty. In order to compute all possible gain-loss scenarios, we use a bottom-up approach that starts from the leaves and climbs up to the root, thereby storing all different possibilities of character evolution. Basically, our approach is brute-force.

The search space can, however, be efficiently restricted. Firstly, when climbing up the reference tree in order to calculate the possible scenarios, we can exclude those which exceed the *maximum number of gain events* allowed on each path from the root of the tree to its leaves. If this number is set to 1 (as it is by default in our approach), this means that, on a given path, characters cannot be gained, lost, and gained again. This is a simplifying requirement, since it is possible that characters on a given lineage are lost and afterwards reintroduced as borrowings—an example being English 'flower,' which was borrowed from Old French *flour* which goes back in turn to Latin *flōre*(*m*). The Latin word is

cognate with English *blossom* and German *Blume* 'flower,' all being reflexes of Proto-Indo-European *b^hleh_3- 'blossom' (de Vaan, 2008: 227). A strict modeling of the complicated history of these words with help of gain-loss scenarios would require us to assume that the character was lost and gained again in English. However, given that these cases are very rare, allowing for them would not only bloat our search space, but also affect the results in a way that is difficult to control.

Secondly, having determined the scenarios that do not exceed our *maximum gain criterion*, we can filter them further by storing only those scenarios with minimal weight. Here, it is important to keep in mind that a scenario with minimal weight on a given subtree is not necessarily a scenario with minimal weight in general. Since, when climbing the reference tree, one cannot tell whether the character state of the temporary root node is an event (a change of the character state) or not, it is possible that a given scenario seems to be cheap at a certain point in the calculation but later turns out to be much more expensive. In order to prevent the model from missing good scenarios, we carry out a separate filtering of those scenarios in which the character in the temporary root node is present and those in which it is absent. Since unpredictable costs of subtree scenarios depend only on the state of the temporary root character, this guarantees that our approach always finds the most parsimonious scenario. It is possible that there is more than one scenario that minimizes the penalty. In such a case we first select the scenario with the minimal amount of gain events, and if there is still more than one scenario, we follow the proposal by Mirkin et al. (2003) and select the scenario in which the gain events are closest to the leaves of the reference tree.

As an example, compare the two-loss scenario in Fig. 2B with the two-gain scenario in Fig. 2C. For the two-loss scenario, the 2–1 model yields a total score of 4 ($1\times2+2\times1$), since there are two losses and one gain.[6] The two-gain scenario in Fig. 2C also yields a score of 4 ($2\times2+0\times1$). In this case, we choose the model which infers the minimal amount of gains, and the two-loss model is chosen as the most parsimonious one. Changing the model to 1–1 yields penalties of 3 ($1\times1+2\times1$) for the two-loss scenario and 2 ($2\times1+0\times1$) for the two-gain scenario. In this case, the two-gain scenario is the most parsimonious.

6 We follow Mirkin et al. (2003) in counting the presence of a character in the root as a normal gain event.

3.2 *Finding Optimal Gain-Loss Models*

Gain-loss mapping is useful for testing possible scenarios of character evolution. However, as long as there is no direct criterion that helps to choose the best of many solutions, the method hardly gives us any new insights. Here, we follow Nelson-Sathi et al. (2011) in using the distribution of *ancestral vocabulary sizes* as a criterion to determine the best model for a given dataset. The basic idea behind this criterion for model selection is that the number of words that ancestral languages use to express a given set of concepts should not differ greatly from the number of words used by the contemporary languages. When assuming that English *mountain* is not a borrowing but a retention (two-loss scenario), this would force us to trace the word back to Proto-Germanic. However, since the counterparts of 'mountain' in the rest of the Germanic languages also point to a common origin, this would necessitate the assumption that there were two words denoting the concept 'mountain' in Proto-Germanic. Although multiple synonyms for a given concept are not impossible, they are rather unlikely to occur frequently; and since our approach is applied to large datasets and not to single items, it seems reasonable to assume that a model explaining the given data adequately should be preferred to a model that yields much larger amounts of synonyms in the ancestral languages than are attested in the contemporary ones. In the case of *mountain*, this means that the 1–1 model should be preferred to the 2–1 model, since the latter favors the two-loss scenario and thus entails the assumption of more synonyms in the ancestral languages.

One could argue that the growing amounts of synonyms in ancestral languages can be explained by assuming the words had different meanings in those languages. English *mountain*, for example, could be derived from Proto-Indo-European *mon-ti 'protrusion, height,' which is the presumed ancestor of Latin *mōns* (de Vaan, 2008: 388). Such a scenario, however, is rather unlikely, since it presupposes that the same semantic shift from 'height' to 'mountain' occurred in the Romance languages and in English. While parallel semantic shift is not improbable *per se*, it is rather unlikely when involving the *same* source forms in *independent* branches of a language family. Furthermore, even if it was frequent, it would not disfavor vocabulary size distributions as a criterion for model selection. It would merely change what gain-loss mapping techniques can infer.

In order to compare how well a given model accounts for the vocabulary size criterion, we compute the number of characters present in the ancestral nodes of the reference tree by tracing all origins inferred by the model back to the respective nodes. We then use the Wilcoxon rank-sum test (see the description in Kruskal, 1957) to test the hypothesis that the ancestral and the contemporary

TABLE 6 *Patchy cognate sets for 'mountain.' In contrast to the cognate set in Table 5, pattern* M *is now split into two distinct patterns:* M_1 *and* M_2.

Language	**Spanish**	**Portuguese**	**French**	**English**	**German**	**Swedish**
'**mountain**'	*montaña*	*montanha*	*montagne*	*mountain*	*Berg*	*berg*
Pattern M_1	1	1	1	0	0	0
Pattern M_2	0	0	0	1	0	0
Pattern B	0	0	0	0	1	1

vocabulary distributions are likely to be drawn from the same sample. Since we cannot exclude the possibility that parallel evolution influences our results, we modified our method in such a way that it allows for a certain amount of parallel evolution. This can be done in a very straightforward way by using a scaling factor to decrease the ancestral vocabulary sizes before the Wilcoxon rank-sum test is applied. As a default, this scaling factor is set to 5%. Thus, we allow ancestral vocabulary size distributions to grow up to 5% larger than contemporary ones.

Having determined a model that explains the phyletic patterns of a given dataset in such a way that the distribution of ancestral and contemporary vocabulary sizes does not differ significantly, the results of the analysis can then be displayed by splitting all cognate sets for which more than one origin was inferred into secondary subsets, as illustrated in Table 6. These *patchy cognate sets* (PCS) can then be further analyzed in different ways. One could, for example, compare the correctness of the original cognate assignments by checking the sound correspondences between the distinct subsets for irregular patterns. In the case of English *mountain*, there is an irregular correspondence between the English [t] and the [t] in the Romance languages, where we would expect a [d] if it were a regular correspondence (compare English tooth [tu:θ] vs. French *dent* [dɑ̃] 'tooth').

3.3 *Minimal Lateral Network*

Another way to analyze the results further is to reconstruct a *minimal lateral network* (MLN) from the inferred gain-loss scenarios (Dagan et al., 2008; Nelson-Sathi et al., 2011). The MLN is a weighted network that displays patterns of vertical and lateral inheritance. The reference tree is used to represent patterns of vertical inheritance between the contemporary and the ancestral languages. Additional edges drawn between the nodes of the reference tree represent possible borrowing events. Borrowing events are assumed for all patterns for which

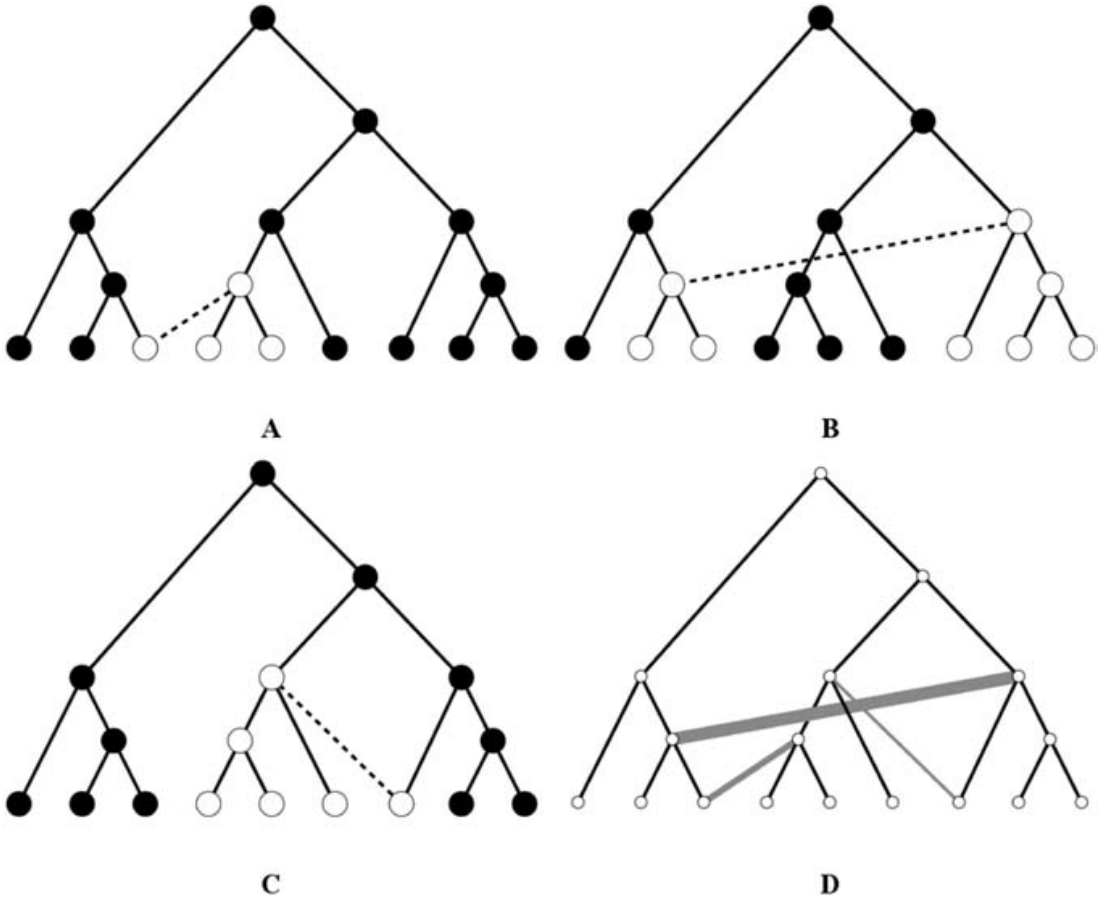


FIGURE 3 *Minimal lateral network reconstruction. If more than one origin is inferred for a given phyletic pattern, the nodes where the characters originate are connected by lateral edges (A–C). In the MLN (D), the edges inferred for all patterns are combined, with edge weights (visualized as differences in line width) reflecting the number of occurrences.*

more than one origin was inferred by a given gain-loss model, and links are drawn between the nodes in which the characters originate. The weights of these lateral edges reflect the number of patterns that support a given link. Figure 3 illustrates this procedure. In Figs 3A–C, three different links are drawn between nodes from which different characters originate more than once on the reference tree. If the number of patterns supporting these scenarios in a given dataset differs, with Fig. 3A occurring twice, Fig. 3B four times, and Fig. 3C once, we arrive at a weighted network for the whole dataset as shown in Fig. 3D.

Drawing lateral links between characters that originate from two different nodes is easy, since there is only one link that can be drawn to connect them. However, if a gain-loss scenario yields more than two separate origins for a given character, there are as many as $(n^2–n)/2$ possible edges which can be drawn to connect n nodes. While drawing all possible edges would surely cover all possibilities, it is obviously unrealistic: since borrowing is a directed

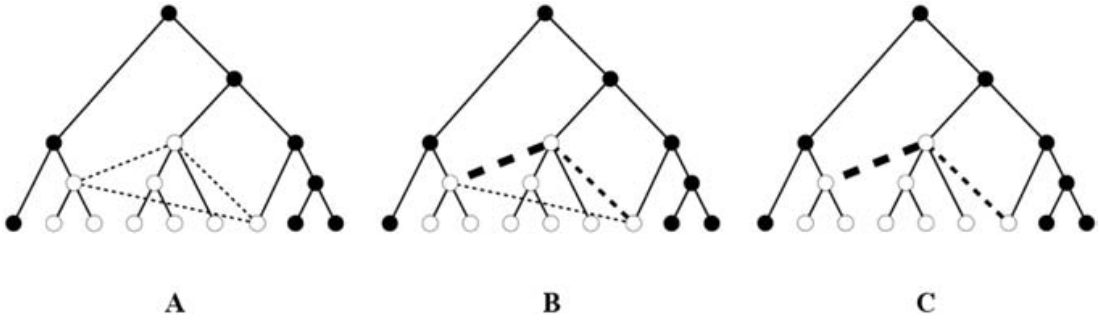


FIGURE 4 *Removing redundant lateral edges in the minimal lateral network. A shows the initial stage. B shows the intermediate stage after edge weights have been inferred for all lateral edges. C shows the resulting minimum spanning tree.*

process that involves a donor and a recipient language, such a scenario would indicate that all languages are both donors and recipients. In order to solve this problem, the complete graph representing all hypothetical connections has to be reduced to a graph consisting of n–1 edges that connects all nodes (a spanning tree). Given that, according to Cayley's (1889) formula, a complete graph of n edges has as many as n^{n-2} spanning trees, it is important to apply a consistent criterion to select one of these trees. The most straightforward way to do so is to select a *minimum spanning tree*, that is, a tree that minimizes the sum of the edge weights.[7] For gain-loss scenarios involving more than two origins, we determine the edge weights for all node pairs *ni* and *nj* by calculating the number of shared multiple origins of *ni* and *nj* in all phyletic patterns of the data. We then convert these weights to distances and use Kruskal's (1956) algorithm to calculate the minimum spanning tree between the nodes. This procedure is illustrated in Fig. 4. This is equivalent to assuming that potential donor lineages with a high frequency of occurrence in the sample have a higher probability of donating than low-frequency potential donor lineages.

3.4 *Minimal Spatial Network*

A minimal lateral network is useful to evaluate the degree to which the evolution of a set of characters follows the presumed branching pattern of a set of languages. However, since languages are not only spoken at a specific time, but also in a specific place, it seems useful to plot the inferred lateral connections onto a geographic map. This may be helpful both for evaluating the results of a given analysis and getting an impression of major diffusion areas. When

7 In our case it would be more appropriate to call it a 'maximum spanning tree,' since the edge weights in the MLN do not represent distances but similarities between nodes.

reconstructing a *minimal spatial network* (MSN) from a given MLN, only the leaves can be plotted because the ancestral nodes have few geographical constraints, so their inclusion in the graph would add too much cluttering information. Therefore the internal nodes of the MLN (the ancestral taxa) are removed and, as a result, internal edges (edges between contemporary and ancestral taxa, and edges between ancestral taxa) are lost. In order to retain information that is connected within them when constructing the spatial network, we project information from internal nodes onto leaves. As a selection criterion to link information from internal to external nodes, we use a simplified approach based on geographic distance. If an edge originally connects an internal node *ni* and an external node *ne*, we first determine all descendant nodes of *ni* on our geographic map. We then draw a convex hull around all descendant nodes of of *ni* and connect the descendant node of *ni* that is (a) on the hull and (b) geographically closest to our external node *ne*. For two internal nodes, we proceed in a similar way, the difference being that two convex hulls are drawn around the descendants of the two internal nodes, and the two geographically closest nodes which appear on the hulls are connected. The central idea behind this approach is that ancestral languages can be represented by the area covered by their descendants.

4 Results

4.1 *Gain-Loss Models for Southern and Common Chinese*

We applied our analysis to the Southern Chinese, the Common Chinese, and the two automatically reconstructed reference trees, using five different gain-loss models with varying penalties for gains and losses: 3–1, 5–2, 2–1, 3–2, and 1–1. We then compared the resulting distributions of ancestral and contemporary vocabulary sizes in order to determine which of the models would fit the data best. For all reference trees, there are two gain-loss models (5–2 and 2–1) in which the vocabulary size distributions do not differ significantly ($\alpha = 0.05$). In all cases, the 2–1 model is the one with the highest probability ($p = 0.73$ for Southern Chinese, $p = 0.76$ for Common Chinese, $p = 0.84$ for UPGMA, and $p = 0.55$ for Neighbor-joining).[8]

As far as the gain-loss models are concerned, the differences between the four trees do not seem to alter gain-loss mapping analyses greatly. Basically,

8 A comparison of the vocabulary size distributions inferred for all analyses is provided in Supplementary Material II.

TABLE 7 *Basic results of the analyses*

Comparandum	Southern Chinese	Common Chinese	Neighbor-joining	UPGMA
Best model	2–1	2–1	2–1	2–1
p-value	0.73	0.76	0.55	0.84
Patchy cognates	567 (54%)	557 (53%)	510 (48%)	585 (55%)
Average n. of origins	1.97	1.81	1.81	2.00
Maximal n. of origins	9	9	8	8

this also holds for some further general characteristics of the models, such as the average number of origins, the number of patchy cognate sets, and the maximum number of origins, all of which are displayed in Table 7.[9] The Neighbor-joining reference tree outperforms the other trees by yielding the lowest percentage of patchy cognate sets. However, since the Neighbor-joining tree itself is in conflict with traditional dialect classification, this merely shows that the Neighbor-joining method is good in maximizing the tree-like signal in the data. It does not mean that the results are necessarily more realistic. Comparing these results with those of List et al. (2014) for Indo-European languages, it is interesting to note that the percentage of patchy cognate sets is quite different (48–55% for the Chinese analyses, but 31% for Indo-European). Given the complex history of the Chinese dialects, this is not surprising but, rather, in agreement with our expectations.

4.2 *MLN and MSN*

Having determined a gain-loss model that brings ancestral and contemporary vocabulary size distributions closely together, we can use this scenario to reconstruct a minimal lateral network. Figure 5 shows the MLN reconstructed for the Southern Chinese reference tree. Interestingly, the heaviest edges occur inside the Mandarin and the Jìn dialects. Here, the Zhèngzhōu dialect plays a central role, having a remarkable number of connections not only to the ancestral node

9 Note that in Table 7 and throughout this paper, the term 'origins' refers to events that distribute a given cognate across dialects and geographical ranges. Thus, inferring 8 or 9 origins in Table 7 does not suggest 8 or 9 independent origins, it simply means that 8 or 9 events are inferred, under our minimizing premises, to underlie its current geographical and dialectic distribution.

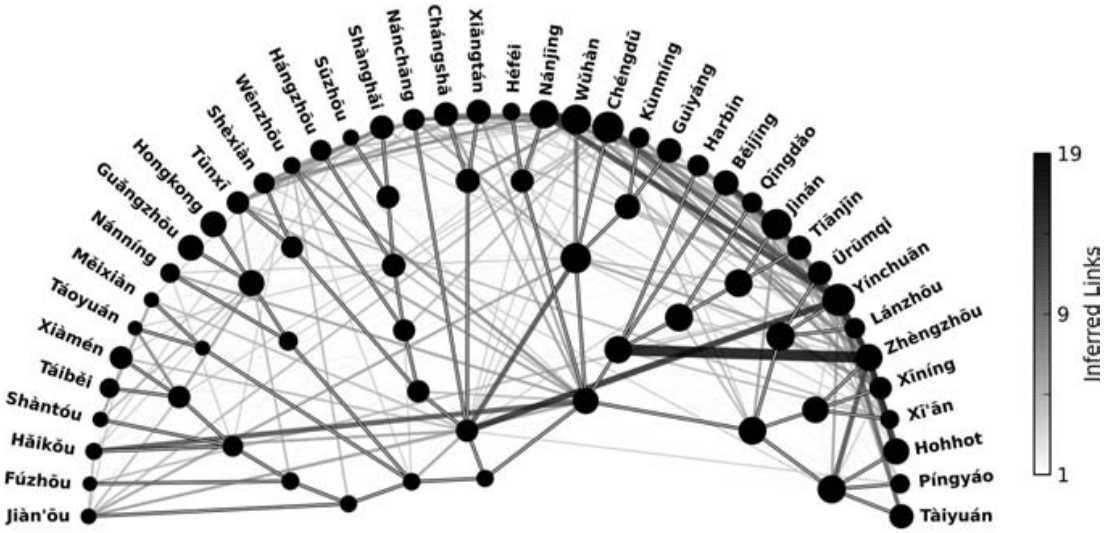


FIGURE 5 *The minimal lateral network of the Southern Chinese reference tree. The node size reflects the inferred number of cognate sets in each language variety. The links reflect the minimal number of lateral transfer events that is required to minimize the differences between the ancestral and the contemporary vocabulary size distribution.*

of the Northern Mandarin dialects (19 shared patchy cognate sets, PCSs), but also to Lányín Mandarin (11 PCSs with Yínchuān), and Jìn (11 links with Hohhot, 9 links with the ancestral node of Jìn). The fact that Zhèngzhōu is not grouped with the Zhōngyuán Mandarin dialects in both automatic analyses (see Supplementary Material I) further reflects the uncertain status of this dialect. Apart from the central role that Zhèngzhōu plays in the Southern Chinese MLN, there is a remarkable number of inferred connections between the Jìn dialects and the Northern and Northwestern Mandarin dialects. Both the role of Zhèngzhōu and the multitude of links between Jìn dialects and Northern and Northwestern Mandarin can also be reported for the Common Chinese analysis (see Supplementary Material III). The status of the Jìn dialects as a group separate from Mandarin is highly disputed in Chinese dialectology (Kurpaska, 2010: 74 f.). If their separation is justified, our method shows that they are highly influenced by neighboring varieties.

The heavy links between Northern and Northwestern Mandarin and Jìn dialects can be more easily recognized in the minimal spatial network shown in Fig. 6. Apart from the high and also quite unexpected diversity in the north, one can find interesting connections in the south-east, where the greatest number of generally recognized distinct dialect groups is found. Thus, Xiāngtán and Chángshā, the two Xiāng dialects in our sample, show their strongest connections to neighboring Mandarin dialects. That the Xiāng dialects have undergone a strong influence from Mandarin dialects has been noticed in the literature for a long time (Norman, 1988: 207 f.). Even more interesting is the strong

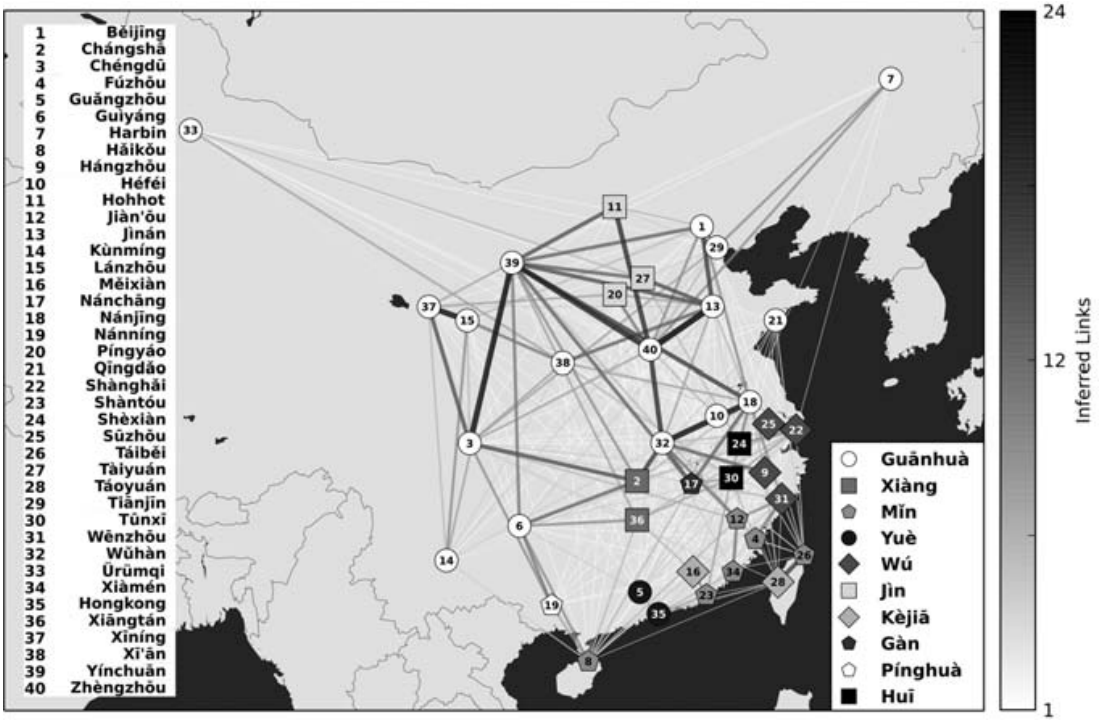


FIGURE 6 *The minimal spatial network of the Southern Chinese reference tree. The links reflect the external and the internal edges between all contemporary language varieties as inferred in the minimal lateral network.*

link between the Wú dialect Wēnzhōu and its neighboring Mǐn dialects.[10] This link is surprising, since in Chinese dialectology it is usually assumed that the border between the Mǐn and the Wú dialects is rather strict (ibid.: 189). However, a closer inspection of the words in Wēnzhōu that are patchily distributed shows that it is indeed very likely that they have been borrowed from the Mǐn dialects, since they are not found in the other Wú dialects, but are quite representative of the Mǐn varieties. Thus, among others, we find that the Wēnzhōu word for 'chopsticks' is [dzei^{22}] with the corresponding character 箸. This is a very archaic expression for 'chopsticks' that is almost exclusively reflected in the Mǐn dialect area. Most other dialects (including all other Wú dialects in our sample) have replaced it with cognate forms of Common Chinese *kuàizi* 筷子 (see Norman, 1988: 76 for details regarding the origin of *kuàizi*). Similar examples where Wēnzhōu has a form that is not reflected in the other Wú dialects, but common in the Mǐn dialects include:

10 In the MSN, the link is drawn between Wēnzhōu and Jiàn'ōu. This is, however, an artifact of the spatial representation. In the underlying MLN, the link is between Wēnzhōu and several ancestral nodes of the Mǐn dialects.

Wēnzhōu [dɤu^{31}] 头 'classifier (for cows and pigs)' (compare Shànghǎi [tsaʔ^{5}] 只), Wēnzhōu [tɕaŋ^{33}ko^{33}] 金瓜 'pumpkin' (compare Shànghǎi [ve^{22}ko^{44}] 南瓜), and Wēnzhōu [liɛ^{35}bu^{13}] 龙雹 'hail' (compare Shànghǎi [piŋ^{55}bɔ^{21}] 冰雹).[11]

Above we have seen that differences in the reference trees did not affect the gain-loss models. This was also observed in Nelson-Sathi et al.'s (2011) analysis of the Indo-European languages and is indicative of a high level of patchiness in the cognate distribution—for data with a comparatively large component of non-treelike structure, the influence of the reference tree becomes less crucial. What was also noted in the study of Nelson-Sathi et al. (2011), however, is that changes in the reference tree may have an impact on the concrete predictions of a given model, indicating in turn that there are detectable vertical components in the data. For our two reference trees in the present Chinese dataset, we can report similar findings. Although the agreement between the Southern and the Common Chinese analyses regarding the detection of patchy cognate sets is rather high, with 966 out of 1056 cognate sets (91%) being identically identified as either patchy or non-patchy cognate sets, many differences in the specific individual scenarios are still observable. Table 8 gives unweighted and weighted degrees for the five most connective nodes in the MLNs for Southern Chinese and Common Chinese.[12] Although four of the five most connected nodes appear in both analyses, they differ greatly regarding their unweighted and their weighted degrees. Since we do not know which of the two scenarios reflects the historical process more closely, we are currently limited to noting the differences. In future studies, it could be of interest to identify independent criteria by which to compare the probabilities of different weighted degrees for given (sets of) nodes, and to use these criteria to evaluate the attributes of different reference trees.

4.3 *Influence of Standard Chinese*

One point we have not addressed so far is the role of the *Dachsprache* in our data. Given that Standard Chinese derived from the dialect of Běijīng, it is surprising that this dialect only plays a minor role in the networks shown in Figs 5 and 6. Běijīng does not appear among the top five nodes with the highest

11 A full account of all the inferred patchy cognates for the Southern Chinese analysis is given in Supplementary Material IV.

12 The degree is the number of edges connecting to a given node in a graph. The weighted degree is calculated by summing up the weights for all edges of a given node (cf. Newman, 2004).

TABLE 8 *Comparing the nodes with the highest degrees for the Southern Chinese* (A) *and the Common Chinese analysis* (B)

Taxon	Degree		Taxon	Degree	
	Unweighted	Weighted		Unweighted	Weighted
Nánjīng	29	81	Shèxiàn	27	58
Zhèngzhōu	29	105	Chéngdū	26	69
Yínchuān	27	114	Yínchuān	26	93
Chéngdū	26	72	Jìnán	24	58
Jìnán	26	70	Nánjīng	24	70
A			B		

degrees (either unweighted or weighted), nor is it involved in any of the heaviest edges. The fact that Běijīng and Standard Chinese played a less pronounced role than expected might be due to a certain shortcoming in our method. Gain-loss mapping requires that borrowing events are still detectable due to patterns that cannot be explained by a reference tree. Borrowing, however, can become so frequent that patchy distributions are no longer detectable.[13] If a word is borrowed (or is actively introduced) by all taxa of a given branch so that the existence of its predecessors is masked, the gain-loss mapping approach assumes that these words are all inherited from a common ancestor language and so no patchy distributions are detected. If, however, the ancestral words have not died out and still exist in refugia that can be detected through more thorough geographical sampling, these effects should be detectable and, in principle, quantifiable.

Although the networks themselves do not give us a hint, the influence of Standard Chinese on Chinese dialect history can still be identified when comparing how many of the cognate sets in each dialect are actually patchy. In Table 9, the five dialects that show the largest frequencies of patchy cognate sets per number of words are listed. In this list, the Běijīng dialect as the closest representative of Standard Chinese occupies the first position, showing the highest ratio of patchy cognate sets per word in both the Southern Chinese

13 In genetics, there is the term 'selfish DNA' to describe genes that can rapidly increase their frequency through spread, because they are readily able to spread (*transposons*). There is also the concept of positive selection, which can lead to the very rapid spread and fixation of new alleles in a population.

TABLE 9 *Comparing the average number of patchy cognates per dialect in the Southern Chinese (A) and the Common Chinese analysis (B)*

Taxon	# Words	PCS	Ø	Taxon	# Words	PCS	Ø
Běijīng	236	95	0.40	Běijīng	236	99	0.42
Zhèngzhōu	278	108	0.39	Chéngdū	320	127	0.40
Tiānjīn	253	97	0.38	Zhèngzhōu	278	110	0.40
Jìnán	315	120	0.38	Tiānjīn	253	100	0.40
Chéngdū	320	121	0.38	Nánjīng	276	107	0.39
A				B			

and the Common Chinese analysis. This shows that Běijīng and Standard Chinese play a definite role in our network, although this role is currently not quantifiable in terms of degree and heavily weighted edges, but only in the patchy cognate sets themselves.

5 Discussion

In evolutionary biology and historical linguistics, the term *phylogenetic network* is often used in a very broad sense, referring to 'any graph used to represent evolutionary relationships (either abstractly or explicitly) between a set of taxa that labels some of its nodes (usually the leaves)' (Huson et al., 2010: 69). Given the fuzziness of this definition, Morrison (2011: 42) suggests drawing a further distinction between two types of phylogenetic networks: *data-display networks* and *evolutionary networks*. Data-display networks are merely a data summary, while evolutionary networks represent a direct phylogenetic hypothesis which 'should display evolutionary relationships between ancestors and descendants' (ibid.: 43). According to this definition, the popular *split networks* (Huson et al., 2010: 71f.), which were also applied to Chinese dialect data (Ben Hamed and Wang, 2006), are data-display networks; the networks we reconstructed with our method come close to evolutionary networks, since they display both patterns of vertical and lateral inheritance. Nevertheless, while our method appears to be pointing in the right direction with regard to uncovering vertically and horizontally shared components in phylogenetic analyses, it is clear that there are still many problems that need to be addressed in future studies.

Our method relies heavily on the accuracy of proposed assessments of etymological relatedness. If the data is incorrectly coded, the results will be off

the mark, but that is true of any analytic method, not just networks. The fact that differences regarding homology judgments can have a great impact on the results reported for gene distributions across genomes was shown in a study by Dagan and Martin (2007: 873), where varying sizes of gene families had a deeper impact on gain-loss models and estimated rates of lateral gene transfer than differences in reference trees. Our current approach to conducting cognate judgments is very strict. Even the slightest morphological variation that might result from regular processes of affixation will force us to separate words into different cognate sets. Although we think that the requirement of direct cognacy as opposed to partial or oblique cognacy is a necessary and reasonable requirement for our method, we recognize that the borders can overlap. Furthermore, it is highly likely that we missed many cases of valid, direct cognacy by conducting cognate judgments on the basis of the identity of the Chinese character sequences. This is a parameter that can be varied in future analyses.

The fact that our networks alone did not uncover the influence of Standard Chinese, and that its influence could only be shown when comparing the number of patchy cognate sets per number of words in a given variety, points to a general problem of the current method for network reconstruction. At the moment, our method simply connects those nodes on the reference tree for which a patchy cognate set has been inferred by a given gain-loss model. In this sense, our approach is greedy. The specific borrowing process, however, cannot be inferred with the method, since it neither points to a direction of the process, nor does it point to a concrete source, since in many cases the gain-loss model infers that characters originate on internal (ancestor) rather than external (contemporary) nodes. Although our method is an improvement over data-display networks, it is still an effort to translate its results into inferred historical processes.

Despite these drawbacks, we are confident that it is worthwhile to pursue this road further. Borrowing is an integral component of language history and the networks can accommodate this mechanism in a way that no bifurcating tree can. Our method clearly shows that the tree model also fails to explain the majority of the lexical data of the Chinese dialects in our sample. Not only does it confirm general uncertainties of Chinese dialect classification that have been long discussed, it also reveals the strong influence of the standard language on the diatopic varieties of Chinese, uncovering a small sketch of the complexity of Chinese dialect history.

Supplementary Material and Software

The Supplementary Material accompanying this study contains figures of all reference trees that were used for this study (Supplementary Material I), the vocabulary size distributions inferred for all analyses (Supplementary Material II), the MLN and MSN for the Common Chinese analysis (Supplementary Material III), and a full account of all patchy cognate sets inferred for the Southern Chinese analysis (Supplementary Material IV). The materials can be downloaded from:

http://www.molevol.de/resources/index.html?id=011list2014/

A Python script that replicates the analyses upon which this study was based along with the input data is available under:

https://gist.github.com/LinguList/7481097.

Acknowledgments

This research was supported by the *German Federal Ministry of Education and Research* (BMBF, research project 'Evolution and Classification in Biology, Linguistics, and the history of the Sciences' http://www.evoclass.de) and the ERC grants 240816 and F020515005.

References

Atkinson, Quentin D. and Russell D. Gray. 2006. How old is the Indo-European language family? Illumination or more moths to the flame? In Peter Forster and Colin Renfrew (eds.), *Phylogenetic Methods and the Prehistory of Languages*, 91–109. Cambridge: McDonald Institute for Archaeological Research.

Barbour, Stephen and Patrick Stevenson. 1998. *Variation im Deutschen. Soziolinguistische Perspektiven*. Berlin: de Gruyter.

Baxter, William H. 1992. *A Handbook of Old Chinese Phonology*. Berlin: Mouton de Gruyter.

Ben Hamed, Mahé and Feng Wang. 2006. Stuck in the forest: Trees, networks and Chinese dialects. *Diachronica* 23: 29–60(32).

Bonfante, Giuliano. 1931. I dialetti indoeuropei. *Annali del R. Istituto Orientale di Napoli* 4: 69–185.

Branner, David Prager. 2006. Some composite phonological systems in Chinese. In David Prager Branner (ed.), *The Chinese Rime Tables. Linguistic Philosophy and Historical-comparative Phonology*, 209–232. Amsterdam: Benjamins.

Cayley, Arthur. 1889. A theorem on trees. *Quarterly Journal of Pure and Applied Mathematics* 13: 376–378.

Cohen, Ofir, Haim Ashkenazy, Frida Belinky, Dorothée Huchon, and Tal Pupko. 2010. GLOOME: Gain loss mapping engine. *Bioinformatics* 26.22: 2914–2915.

Coseriu, Eugenio. 1973. *Probleme der strukturellen Semantik. Vorlesung gehalten im Wintersemester 1965/66 an der Universität Tübingen*. Tübingen: Narr.

Dagan, Tal, Yael Artzy-Randrup, and William Martin. 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proceedings of the National Academy of Sciences of the U.S.A.*. 105.29: 10039–10044.

Dagan, Tal and William Martin. 2007. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proceedings of the National Academy of Sciences of the U.S.A.*. 104.3: 870–875.

Goossens, Jan. 1973. *Niederdeutsch. Sprache und Literatur. Eine Einführung*. Neumünster: Karl Wachholtz.

Gray, Russell D. and Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426.6965: 435–439.

Hóu Jīng 侯精 (ed.). 2004. *Xiàndài Hànyǔ fāngyán yīnkù* 现代汉语方言音库 [Phonological database of Chinese dialects]. Shànghǎi 上海: Shànghǎi Jiàoyù 上海教育.

Huson, Daniel H., Regula Rupp, and Celine Scornavacca. 2010. *Phylogenetic Networks. Concepts, Algorithms, and Applications*. Cambridge: Cambridge University Press.

Karlgren, Bernhard. 1954. Compendium of phonetics in ancient and archaic Chinese. *Bulletin of the Museum of Far Eastern Antiquities* 26: 211–367.

Koonin, Eugene V. 2005. Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics* 39: 309–338.

Kruskal, Joseph B. 1956. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society* 7.1: 48–50.

Kruskal, William H. 1957. Historical notes on the Wilcoxon unpaired two-sample test. *Journal of the American Statistical Association* 52.279: 356–360.

Kurpaska, Maria. 2010. *Chinese Language(s). A Look through the Prism of The Great Dictionary of Modern Chinese Dialects*. Berlin and New York: de Gruyter.

Lǐ Xiǎofán 李小凡. 2005. *Hànyǔ fāngyán fēnqū fāngfǎ zài rènshi* 汉语方言分区方法再认识 [Reevaluating the classification of the Chinese dialects]. *Fāngyán* 方言 4: 356–363.

List, Johann-Mattis and Steven Moran. 2013. An open source toolkit for quantitative historical linguistics. In *51th Annual Meeting of the Association for Computational Linguistics (ACL 2013), Proceedings of the Conference System Demonstrations, Aug. 4–9, 2013, Sofia, Bulgaria*, 13–18. Stroudsburg, PA: Association for Computational Linguistics.

List, Johann-Mattis, Shijulal Nelson-Sathi, Hans Geisler, and William Martin. 2014.

Networks of lexical borrowing and lateral gene transfer in language and genome evolution. *BioEssays* 36.2: 141–150.

McMahon, April and Robert McMahon. 2005. *Language Classification by Numbers*. Oxford: Oxford University Press.

Meiser, Gerhard. 1998. *Historische Laut- und Formenlehre der lateinischen Sprache*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Mirkin, Boris G., Trevor I. Fenner, Michael Y. Galperin, and Eugene V. Koonin. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evolutionary Biology* 3: 2 (doi: 10.1186/1471-2148-3-2).

Morrison, David A. 2011. *An Introduction to Phylogenetic Networks*. Uppsala: RJR Productions.

Nelson-Sathi, Shijulal, Johann-Mattis List, Hans Geisler, Heiner Fangerau, Russell D. Gray, William Martin, and Tal Dagan. 2011. Networks uncover hidden lexical borrowing in Indo-European language evolution. *Proceedings of the Royal Society B. Biological Sciences* 278.1713: 1794–1803.

Newman, Mark E.J. 2004. Analysis of weighted networks. *Physical Review E* 70.5: 056131.

Norman, Jerry. 1988. *Chinese*. Cambridge: Cambridge University Press.

Norman, Jerry. 1991. The Mǐn dialects in historical perspective. In William S.-Y. Wang (ed.), *Languages and Dialects of China* (Journal of Chinese Linguistics, Monograph Series Number 3), 325–360. Berkeley: Project on Linguistic Analysis.

Norman, Jerry. 2003. The Chinese dialects: Phonology. In Graham Thurgood and Randy LaPolla (eds.), *The Sino-Tibetan Languages*, 72–83. London: Routledge.

Nunn, Charles L. 2011. *The Comparative Approach in Evolutionary Anthropology and Biology*. Chicago: University of Chicago Press.

Sagart, Laurent. 2002. Gan, Hakka and the formation of Chinese dialects. In Dah-an Ho (ed.), *Dialect Variations in Chinese. Papers from the Third International Conference on Sinology, Linguistics Section*, 129–153. Taipei: Academia Sinica.

Saitou, Naruya and Masatoshi Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4.4: 406–425.

Schleicher, August. 1853. Die ersten Spaltungen des indogermanischen Urvolkes. *Allgemeine Monatsschrift für Wissenschaft und Literatur*: 786–787.

Schmidt, Johannes. 1872. *Die Verwandtschaftsverhältnisse der indogermanischen Sprachen*. Weimar: Hermann Böhlau.

Schuchardt, Hugo. 1900. *Über die Klassifikation der romanischen Mundarten. Probe-Vorlesung gehalten zu Leipzig am 30. April 1870*. Graz: Styria. Downloadable at http://schuchardt.uni-graz.at/werk/pdf/309 (accessed June 4, 2014).

Sokal, Robert R. and Michener, Charles D. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin* 28: 1409–1438.

Southworth, Franklin C. 1964. Family-tree diagrams. *Language* 40.4: 557–565.

Swadesh, Morris. 1952. Lexico-statistic dating of prehistoric ethnic contacts. With special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society* 96.4: 452–463.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21.2: 121–137.

Trask, Robert L. 2000. *The Dictionary of Historical and Comparative Linguistics*. Edinburgh: Edinburgh University Press.

de Vaan, Michiel. 2008. *Etymological Dictionary of Latin and the Other Italic Languages*. Leiden and Boston: Brill.

Wáng Hóngjūn 王洪君. 2009. *Jiāngù yǎnbiàn, tuīpíng hé céncì de Hànyǔ fāngyán lìshǐ guānxì móxíng* 兼顾演变、 推平和层次的汉语方言历史关系模型 [A historical relation model of Chinese dialects with multiple perspectives of evolution, level and stratum]. *Fāngyán* 方言 3: 204–218.

Wurm, Stephen A. and Liú Yǒngquán 刘涌泉 (eds.). 1987. *Zhōngguó yǔyán dìtújí* 中国语言地图集 [Language atlas of China]. Hongkong: Longman Group.

2.2 Ancestral State Reconstruction

The two previous studies on phylogenetic networks both make use of a rather simple and straightforward two-step procedure. First, an algorithm is used to plot individual word histories in a given reference three. Second, individual word histories are systematically compared and analyzed by inferring major lateral connections and visualizing the results in form of phylogenetic networks or networks plotted in geographic space. The first step in the procedure, which was called *character mapping* in the two previous studies, is also known as *ancestral state reconstruction* in evolutionary biology, and beyond doubt the more important of the two steps.

The two studies by List et al. (2014a) and List et al. (2014b) make exclusive use of *weighted parsimony techniques* for binary-state characters in reconstructing the ancestral states. Given that parsimony is considered problematic among scholars in evolutionary biology and computational historical linguistics, it is therefore important to improve the techniques for ancestral state reconstruction further. First attempts in this direction are presented in the following two studies.

The first study, titled "Beyond cognacy: historical relations between words and their implication for phylogenetic reconstruction" (List 2016), takes the notion of *cognacy* in historical linguistics as a starting point to explain how a more realistic modeling of lexical evolution might be achieved. By comparing *cognacy* with the notion of *homology* in evolutionary biology, a more fine-grained model of lexical evolution is developed that allows to distinguish more processes than merely gain and loss of words. In order to handle more complex processes, involving specifically derivation and compounding, in computational analyses for ancestral state reconstruction, the study proposes the use of multi-state models for character evolution which allow for evolutionary processes with a strong directional tendency by employing asymmetric step matrices in a parsimony framework of ancestral state reconstruction. The results show that these improved models have a much higher success rate in reconstructing the lexical evolution scenarios of a gold standard test set of Chinese dialects where ancestral states are known from historical records.

Given the well-known disadvantages of parsimony-based techniques for ancestral state reconstruction, the second study, titled "Using ancestral state reconstruction methods for onomasiological reconstruction in multilingual word lists" (Jäger and List 2018), systematically compares recent approaches for ancestral state reconstruction as they are common in evolutionary biology, testing binary and multi-state models of character evolution in weighted parsimony frameworks, the weighted parsimony techniques underlying the *minimal lateral network* approach discussed before (List et al. 2014a, List et al. 2014b), as well as ancestral state reconstruction based on *maximum likelihood*. The results show that maximum likelihood approaches clearly outperform all other approaches. However, a detailed qualitative comparison of the computational results with the known scenarios of lexical evolution in the Indo-European and the Sinitic sample of the data also showed that the test data themselves, which was based on judgments not provided by the authors of the study of Jäger and List (2018), had several errors.

Beyond cognacy: historical relations between words and their implication for phylogenetic reconstruction

Johann-Mattis List*

Centre des Recherches Linguistiques sur l'Asie Orientale, École des Hautes Études en Sciences Sociales, 2 rue de Lille, Paris 75007, France, and Team Adaptation, Integration, Reticulation, Evolution, Université Pierre et Marie Curie, 9 quai St Bernard, Paris 75005, France

*Corresponding author: mattis.list@lingpy.org

Abstract

This article investigates the terminology and the processes underlying the fundamental historical relations between words in linguistics (*cognacy*) and genes in biology (*homology*). The comparison between linguistics and biology shows that there are major inconsistencies in the analogies drawn between the two research fields and the models applied in phylogenetic reconstruction in linguistics. Cognacy between words is treated as a binary relation which is either present or not. Words, however, can exhibit different degrees of cognacy which go beyond the distinction between orthologous and paralogous genes in biology. The complex nature of cognacy has strong implications for the models used for phylogenetic reconstruction. Instead of modeling lexical evolution as a process of cognate gain and cognate loss, we need to go beyond the cognate relation and develop models which take the *degrees of cognacy* into account. This opts for the use of evolutionary models which handle multistate characters and allow to define potentially asymmetrical transition tendencies among the character states instead of time-reversible binary state models in phylogenetic approaches. The benefit of multistate models with asymmetric transition tendencies is demonstrated by testing how well different models of lexical change perform in semantic reconstruction on a lexicostatistical dataset of 23 Chinese dialects in a parsimony framework. The results show that the improved models largely outperform the popular gain–loss models. This suggests that improved models of lexical change may have strong consequences for phylogenetic approaches in linguistics.

1. Introduction

Evolutionary biology and historical linguistics both deal with the evolution of objects. Evolutionary biology investigates the evolution of species, morphological characters, and genes, and historical linguistics investigates the evolution of language varieties, grammatical features, and words. In both disciplines, *historical relations* are an important way to describe the consequences of evolutionary processes. Historical relations are defined for evolving objects which share a common history. The most general historical relation is the relation of *common descent*. This relation can hold both for lineages and for their characteristics. If the relation concerns the latter, biologists call it *homology*. In linguistics, this relation is often compared with the relation of *cognacy*. In contrast to historical relations, we can define various

nonhistorical relations between evolving objects. We can compare species for phenotypic similarity and language varieties for typological similarity. We can compare species for the similarity of their habitat, and language varieties for their geographic closeness. Although these similarities can give us hints regarding deeper historical relations, they are neither a necessary nor a sufficient condition for them.

Evolutionary biology has a rich terminological framework describing fundamental historical relations between genes and morphological characters. Discussions regarding the epistemological and ontological aspects of these relations are frequent and fruitful (Jensen 2001; Koonin 2001; Petsko 2001; Sonnhammer and Koonin 2002; Morrison 2015). In historical linguistics, terminological questions regarding historical relations have occasionally been raised in the past (Katičić 1966; Arapov and Xerc 1974), and recent discussions about the cognacy of grammatical features in historical syntax have emerged (Campbell and Harris 2002; Barðdal and Eythórsson 2012; Walkden 2013). In quantitative applications, however, the fundamental historical relations between words, morphemes, or grammatical features are usually assumed to be self-evident, not deserving specific attention. As a result, our traditional terminology dealing with relatedness, inheritance, and descent is often used imprecisely, frequently leading to confusion in quantitative applications. Computational approaches in historical linguistics are often based on software originally designed for bioinformatics. Scholars justify the use of bioinformatics software in linguistics by drawing analogies between historical relations in the two disciplines. Unfortunately, these analogies often ignore the peculiarities of biological evolution and language history. Instead, they offer a simplified mapping between terms in both disciplines and disregard the underlying processes.

In the following, I will try to illustrate the problems in phylogenetic reconstruction in more detail. I will try to show that the models which are currently used to infer phylogenies from linguistic data suffer from a loss of valid information arising from the superficial analogy between homology and cognacy and a simplification of the processes underlying lexical change. Since terminological misunderstandings are the core of the problem, I will first carry out a brief comparison of biological and linguistic terminology on historical relations, pointing to similarities and differences in the two fields (Section 2). By discussing the complexities of lexical change, I will point to further pitfalls that should be avoided when modeling lexical change with biological software (Section 3). I will then propose improvements to the models currently used in computational historical linguistics (Section 4), and illustrate for a small lexical dataset of Chinese dialects how complex historical relations between words can be modeled in computational approaches to phylogenetic reconstruction (Section 5).

2. Terminology for historical relations in biology and linguistics

Scholars have often compared biological and linguistic terminology (Gray 2005; Croft 2008; Pagel 2009; Geisler and List 2013). The analogies that have been made are, however, not necessarily very precise. This becomes especially evident in the analogies drawn between the terms which are used to describe historical relations between evolving objects in both fields. The most popular analogy in this context is that between *homology* in biology and *cognacy* in linguistics (Pagel 2009). In the following, I will carry out a detailed comparison between the terminology used in both fields, thereby showing that the analogy between homology and cognacy is essentially misleading.

2.1 Homology

Homology is a fundamental concept in evolutionary biology, designating a 'relationship of common descent between any entities, without further specification of the evolutionary scenario' (Koonin 2005: 311). The term was first defined by Richard Owen (1804–92), who distinguished 'homologues', as 'the same organ in different animals under every variety of form and function' (Owen 1843: 379), from 'analogues' as an 'organ in one animal which has the same function as another part or organ in a different animal' (Owen 1843: 374). Homology is a very general historical relation between evolving objects. It does not specify the process from which the relation originated. Geneticists distinguish three subtypes of homology based on processes underlying the homology of genes in molecular evolution: *orthology*, *paralogy*, and *xenology*. Orthology refers to 'genes related via speciation' (Koonin 2005: 311), paralogy refers to 'genes related via duplication' (Koonin 2005: 311), and xenology refers to genes 'whose history, since their common ancestor, involves an interspecies (horizontal) transfer of the genetic material for at least one of those characters' (Fitch 2000: 229).

In a paper from 1970, Fitch suggested to distinguish two kinds of homology in molecular evolution: homology as the 'result of speciation so that the history of the gene reflects the history of the species' should be called 'orthology', and homology as the 'result of gene duplication so that both copies have descended side by

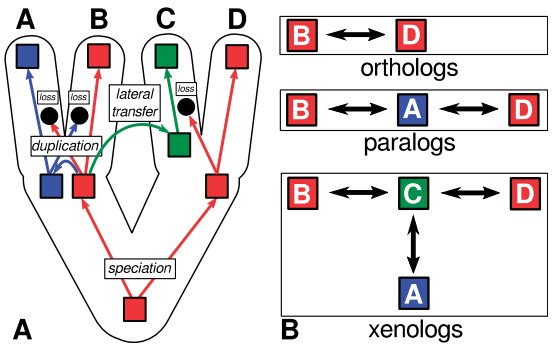


Figure 1. Subtypes of homology in molecular biology. Three processes, *speciation*, *duplication*, and *lateral transfer* underly the three basic types of homology in molecular evolution. The processes are illustrated in (A), the resulting relations are illustrated in (B).

side during the history of an organism' should be called 'paralogy' (Fitch 1970: 113). First evidence that genome evolution does not only involve the mutation of individual genes but also the duplication of genes as a whole was reported in the 1930s (Zhang 2003; Taylor and Raes 2004).

In 1983, Gray suggested to use the term *xenology* as a third subtype of homology in order to distinguish those cases in which genes are homologous, but neither orthologous nor paralogous, since 'cells and organisms have acquired foreign genes in the past' (Gray and Fitch 1983: 64). It is now a well-established fact that prokaryotes (bacteria) may acquire genetic material from 'their neighborhood or [...] environment and incorporate it into their genomes' (Nelson-Sathi et al. 2013: 166). Lateral gene transfer processes were first detected and described in the 1950s (Freeman 1951). Only 30 years later, however, scholars began to emphasize the importance of lateral gene transfer for microbial evolution (Syvanen 1985). Figure 1 contrasts the three basic processes of speciation, duplication, and lateral transfer with the resulting historical relations in evolutionary biology.

2.2 Cognacy

In historical linguistics, the only relation which is explicitly defined is *cognacy* (also called *cognation*). Cognacy usually refers to words related via 'descent from a common ancestor' (Trask 2000: 63) and it is strictly distinguished from descent involving lateral transfer (*borrowing*). The term cognacy itself, however, covers both direct and indirect descent. Hence, German *Zahn* 'tooth' is cognate with English *tooth*, as is German *Kopf* 'head' with English *cup*, and German *Getränk* 'drink' with English *drink*, although the historical processes that shaped the present appearance of these three word pairs are quite different: apart from the sound shape, *Zahn* and *tooth* have regularly developed from Proto-Germanic **tan P* (Kroonen 2013: 509f); *Kopf* and *cup* both go back to Proto-Germanic **kuppa-* 'vessel' (Pfeifer 1993; Kluge and Seebold 2002),[1] but the meaning of the German word has changed greatly; *Getränk* and *drink* go ultimately back to Proto-Germanic **drinkan* 'to drink' (Kroonen 2013: 100f), but the German noun was built as a collective (with prefix *Ge-*) from the nominalized form of the verb (Pfeifer 1993), while the English noun was directly built from the verb. The nominalized form, Proto-Germanic **dranka-* is still reflected in German *Trank* 'potion'. Thus, of the three examples of cognate words, only the first would qualify as having evolved by direct inheritance. Starostin (2013: 140) suggests to distinguish 'etymological cognacy' from 'lexicostatistical cognacy', the former denoting words whose 'forms go back to the same protoform', and the latter denoting words whose 'meanings go back to the same meaning in the proto-language as well'. Trask (2000: 234) suggests the term *oblique cognacy* to label cases in which 'two or more words in related languages [...] continue alternant forms of a single root in the ancestral language', but this term is rarely used and most of the time linguists simply use the term *cognacy* without further specifying what they actually mean.

2.3 Beyond homology and cognacy

In an earlier paper (List 2014: 38–46) I abstracted from the processes underlying the historical relations between genes to contrast the biological and the linguistic terminology. In this comparison, I took *common descent* as the most basic relation, with homology as a direct counterpart. The term 'common descent' may be a bit misleading, but what I had in mind by then were all forms of *historical relations*, including those resulting from lateral transfer. Common descent was further subdivided into *direct common descent* (corresponding to orthology), *indirect common descent* (corresponding to paralogy), and *common descent involving lateral transfer* (corresponding to xenology). I then contrasted the abstract relations and the biological terminology with the terminology currently found in linguistics, thereby pointing to missing slots in the linguistic terminology, for which new terms are proposed. Table 1 illustrates this comparison by contrasting the abstract basic

1 Most likely the word is an early borrowing from Latin which happened before the split of English and German (see Pfeifer 1993).

Table 1. Comparing biological and linguistic terminology for historical relatedness (with modifications taken from List 2014). Terms in red are suggested to make up for missing terminology in historical linguistics

Historical relations		Terminology					
		Biology		Linguistics			
Common descent	Direct	Homology	Orthology	Etymological relation	Cognacy	Direct cognacy
	Indirect		Paralogy			Indirect cognacy
	Involving lateral transfer		Xenology		Indirect etymological relation	

relations with the terminology in biology and linguistics. Relations for which proper terms are missing in linguistics and for which I proposed new terms are colored in red (List 2014: 44). As one can easily see from the table, historical linguistics does not offer direct counterparts for the abstract relations underlying *homology*, *orthology*, and *xenology* in evolutionary biology. Cognacy in historical linguistics is often deemed to be identical with homology in evolutionary biology (Gray 2005; Pagel 2009), but if we follow the comparison, this is only true if one ignores common descent involving lateral transfer, since borrowings are explicitly excluded from the classical definition of cognacy in historical linguistics (Trask 2000: 63).

As we can see from the table, linguistics lacks a proper term for a historical relation between words regardless of whether they are inherited or borrowed (homology in biology, etymological relation according to Table 1). There is also no term denoting the relation between words of which one has been borrowed during its history (xenology in biology). This does not mean, of course, that the relations do not occur in the linguistic domain. Lateral transfer, the process underlying the relation of xenology in molecular biology is also common in language history.[2] In contrast to a relation between two words which involves lateral transfer, the term *borrowing* refers to distinct processes involving a donor and a recipient. As an example for such a relation, consider the words German *kurz* 'short' and English *short* (List 2014: 40). These words are not cognate. German *kurz* is

2 We should, of course, be careful with analogies, and it is clear that the specific processes of lexical borrowing are completely different from the processes of lateral gene transfer in biology. On an abstract level, however, the analogy between lateral gene transfer and lexical borrowing holds, in so far as both processes involve the direct transfer of material between evolving objects.

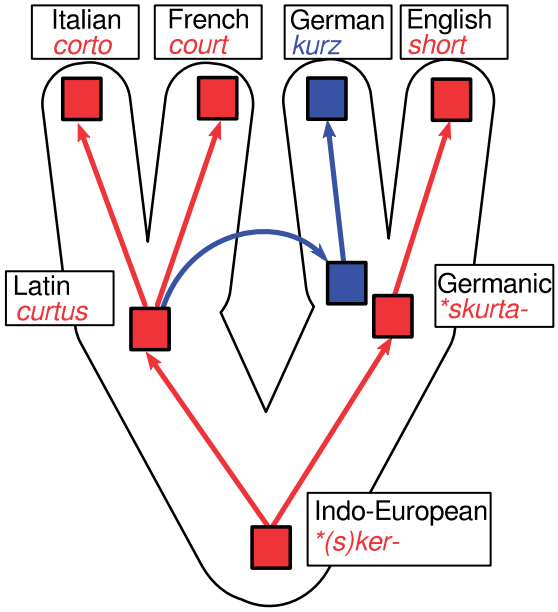


Figure 2. Complex historical relations between reflexes of Proto-Indo-European **(s)ker-* 'cut off'.

a borrowing from Latin *curtus* 'mutilated' (Pfeifer 1993), but English *short* probably goes back to Proto-Indo-European **(s)sker-* 'cut off' (Rix et al. 2001), and so does Latin *curtus* (Vaan, 2008). The specific history behind these relations is illustrated in Fig. 2. Since German *kurz* was borrowed early from Latin, we cannot say that *kurz* has been borrowed from French *court*, but we also cannot say that both words are cognate. Yet since both words share a common history, it would be likewise wrong to label them as unrelated, in lack of a proper terminology.

3. Modeling lexical change

In the previous section, I have introduced the basic terminology which biologists and linguists use to denote specific relations between evolving objects. I have then presented an earlier approach of mine (List 2014), where I used the distinctions made in the biological domain in order to introduce new terms for specific historical relations between words. On the first look, the approach seems justified, and the proposed analogies between biological and linguistic relations seem to be fruitful. When looking into the details, however, it becomes clear that important questions are left unanswered. While it is obvious that cognacy in linguistics is not the same as homology in biology, it is less clear how we should understand the idea of *direct* and *indirect* cognacy.

What exactly is meant to be indirect here? Is it the fact that words differ in meaning, thus being akin to words which are *root-cognate* but not *lexicostatistically cognate*, following the distinction of Starostin (2013: 140), or should we instead concentrate on morphological differences, thus following the notion of *oblique cognacy* proposed by Trask (2000: 234)? And how does the idea of 'indirect descent' relate to paralogy and the process of gene duplication in biology? In the following, I will try to show that we need to go beyond my earlier proposal in order to develop a satisfying model of lexical change that can be used for phylogenetic reconstruction.

3.1 Degrees of cognacy

Morrison (2015: 50) points to the relative character of homology in evolutionary biology in emphasizing that evolving objects can exhibit homology at different levels, which may even be independent of each other:

> *The classic example is the comparison of bird wings and bat wings. These are homologous as forelimbs (structures), which are general throughout the tetrapods, but they are not homologous as wings (functions), because they represent independent modifications of those forelimbs in the ancestors of birds and bats. (Morrison 2015: 50)*

We can find similar situations in linguistics: if we consider words for 'to give' in the four Romance languages Portuguese, Spanish, Provencal, and French, we can state that both Portuguese *dar* and Spanish *dar* are homologous, as are Provencal *douna* and French *donner*. The former go back to the Latin word *dare* 'to give', the latter go back to the Latin word *dōnāre* 'to gift (give as a present)'. In times when Latin was spoken, both *dare* and *dōnāre* were clearly separated words denoting clearly separated concepts and being used in clearly separated contexts. The verb *dōnāre* itself was derived from Latin *dōnum* 'present, gift'. Similar to English where nouns can be easily used as verbs, Latin allowed for specific morphological processes to turn nouns into verbs. What the ancient Romans were not aware of is that Latin *dōnum* 'gift' and Latin *dare* 'to give' themselve go back to a common word form. This was no longer evident in Latin, but it was in Proto-Indo-European, the ancestor of the Latin language. Thus, Latin *dare* goes back to Proto-Indo-European **deh*$_3$- 'to give', and Latin *dōnum* goes back to Proto-Indo-European * *deh*$_3$ - *no*- 'that what is given (the gift)' (Meiser 1998). The word form **deh*$_3$ - *no*- is a regular derivation from **deh*$_3$-, so on the Indo-European level, both forms are *homologous*, since one is derived from the other. This means in turn, that Latin *dare* and *dōnum* are also homologues, since

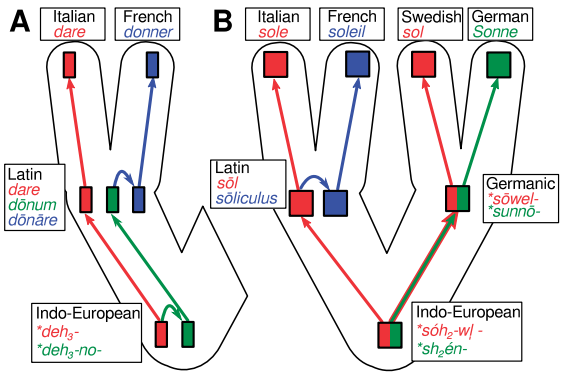


Figure 3. Degrees of cognacy in Indo-European language history: the development of words meaning 'to give' from Proto-Indo-European via Latin to Italian and French (A), and the development of words meaning 'sun' in from Proto-Indo-European to Italian, French, Swedish, and German (B).

they are the residual forms of the two homologous words in Proto-Indo-European. And since Latin *dōnāre* is a regular derivation of *dōnum*, it means, again, that Latin *dare* and *dōnāre* are also homologous, as are the words in the four descendant languages, Portuguese *dar*, Spanish *dar*, Provencal *douna*, and French *donner*. Depending on the time depth we apply, we will arrive at different homology decisions. The history of the words is depicted in Fig. 3A.

An even more complex example are words like Italian *sole*, French *soleil*, Swedish *sol*, and German *Sonne*, all meaning 'sun'. Indo-European scholars assume that the Proto-Indo-European word for sun had a complex, stem-alternating paradigm with two different base forms, one for nominative and accusative case **séh*$_2$*u̯el*-, and one for the oblique cases, **sh*$_2$*én*- (Wodtko et al. 2008: 606). Proto-Germanic inherited this paradigm completely (**sōel*- versus **sunnōn*, Kroonen 2013: 463f), but it was simplified via the process known as *analogy* in historical linguistics, and the nominative stem was taken as the base form in Latin *sōl* (Meyer-Lübeke 1911: §8059). In Swedish and German, the complex base form was also simplified, but in different directions, with the Swedish form taking the nominative stem as the basis of analogy, and the German form taking the oblique stem. While Italian *sole* is the regular reflex of Latin *sōl*, French *soleil* goes directly back to Latin *sōliculus* 'small sun', a Latin diminutive of *sol* (Meyer-Lübeke 1911: §8067). From this perspective, Italian *sole* is more closely related to Swedish *sol* than to French *soleil*, although French and Italian are, of course, much closer genetically related than are Swedish and Italian. The history of the reflexes of the Indo-European word for 'sun' is depicted in Fig. 3B.

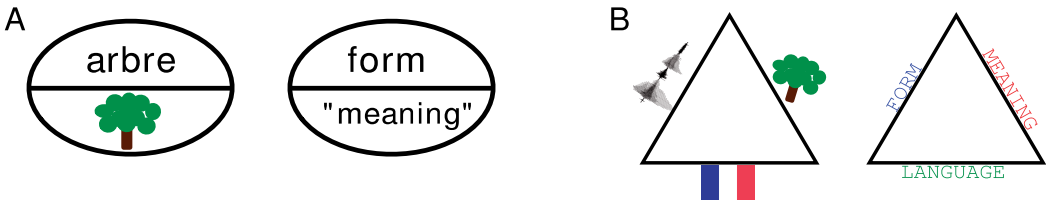


Figure 4. The different dimensions of the linguistic sign: (A) Shows the classical model after Saussure (1916). (B) Shows an extended sign model in which the language, the system in which a sign is used was added as a third component.

3.2 Dimensions of lexical change

In a very simple language model, the lexicon of a language can be seen as a bag of words. A word is further defined by two aspects: its *form* and its *meaning*. Thus, the French word *arbre* can be defined by its written form *arbre* or its phonetic form [ɑrbrə], and its meaning 'tree'. This is reflected in the famous sign model of Ferdinand de Saussure (1857–1913, Saussure 1916), which I have reproduced in Fig. 4A. In order to emphasize the importance of the two aspects, linguists often say that form and meaning of a word are like two sides of the same coin, but we should not forget that a word is only a word if it belongs to a certain language. From the perspective of the German or the English language, for example, the sound chain [ɑrbrə] is just meaningless. So instead of two major aspects of a word, we may better talk of three major aspects: *form*, *meaning*, and *language* (Ternes 1987: 22f; List 2014: 15–18). As a result, our bilateral sign model becomes a trilateral one, as illustrated in Fig. 4B.

Gévaudan (2007) distinguishes three dimensions of lexical change: The *morphological dimension*, the *semantic dimension*, and the *stratic dimension*. The morphological dimension points to changes in the form of words which are not due to regular sound change. As an example, consider German *Getränk* 'drink' and its ancestral form Old High German *tranc* 'drink'. While the meaning of the word is the same, the German word *Getränk* is a collective derivation of the Old High German source form (Pfeifer 1993). The derivation process involved prefix *Ge-*, and the modification of the main vowel. The semantic dimension is illustrated by changes like the one from Proto-Germanic **kuppa-* 'vessel' to German 'Kopf'. The stratic dimension refers to changes which involve lexical material *outside* the *historical continuum* of a given language (Gévaudan 2007: 141f). In the terminology of Gévaudan (2007: 141f), *stratum* refers to languages as historical continua, and should not be confused with the way the term is used in sociolinguistics, where it refers to language varieties used in certain layers of a linguistic society (Coseriu 1973; Oesterreicher 2001), but rather in opposition to the term *adstratum* in historical and areal linguistics (Gévaudan 2007: 141). Usually, changes along the stratic dimension belong to the class of *borrowing processes*. (Gévaudan 2007: 141–63) argues, however, that processes like onomatopoeia, antonomasy, and folk etymology can also be characterized as processes which involve the stratic dimension of lexical change, since they are based on material which does not stem from the historical continuum of a given language. An example for a simple type of stratic change is English *mountain* which was borrowed from Old French *montaigne* 'mountain'. An example for a more complex type of stratic change is German *Maus* 'mouse (for a computer)' which was not directly transferred from English but rather received a broadened semantic function under the influence of the English word (compare Weinreich (1974: 47–62) and Gévaudan (2007: 143–51) for more details on different types of lexical interference).

Note that these three dimensions of lexical change correspond directly to the three major aspects constituting the linguistic sign: the morphological dimension changes the *form* of a word, the semantic dimension its *meaning*, and the stratic dimension its *language*. Thus, the three dimensions of lexical change, as proposed by Gevaudan find their direct reflection in the major dimensions along which words can vary.

3.3 27 Shades of cognacy

When looking at the different historical relations from the perspective of the three dimensions of lexical change, it becomes clear that the new terms I proposed earlier (List 2014) do not necessarily solve our problem of reflecting the different aspects of lexical change and lexical variation adequately. Although it seems justified to point to the difference between cognacy in linguistics and homology in biology, it proposes a problematic analogy between paralogy and indirect cognacy without further specifying how indirect cognacy should be defined in the end. When investigating the different uses of the

Table 2. 27 shades of cognacy: the table shows exemplarily how cognacy can be modeled according to the three dimensions of lexical change, highlighting potential analogies in biology.

		Stratic	Morphological	Semantic
Relation	*Biol. Term*	*continuity*		
traditional notion of cognacy	-	+	+/-	+/-
cognacy à la Swadesh (1952, 1955)	-	+	+/-	+
direct cognate relation (List 2014)	orthology	+	+	+
oblique cognate relation (Trask 2000)		+	-	+/-
etymological relation (List 2014)	homology	+/-	+/-	+/-
oblique etymological relation (List 2014)	xenology	-	+/-	+/-
...	...	...	...	...

term 'cognacy', for example, it becomes obvious that the differences result from *controlling* for one or more of the three dimensions of lexical change proposed by Gévaudan (2007).[3] The notion of cognacy of a classical Indo-Europeanist, for example, controls the stratic dimension by requiring *stratic continuity* (no borrowing), but at the same time it is indifferent regarding the other two dimensions. This is what Starostin (2013: 140) called 'etymological cognacy'. Cognacy à la Swadesh (especially Swadesh 1952,1955), as we know it from lexicostatistics (Swadesh 1952, 1955) and its modern derivations (Gray and Atkinson 2003), is indifferent regarding *morphological continuity*, but controls the semantic and the stratic dimensions by only considering words that have the same meaning and have not been borrowed. This is what Starostin, (2013: 140) called 'lexicostatistical cognacy'.

'Traditional cognacy' and 'cognacy à la Swadesh', however, are but two ways to control for the three dimensions of lexical variation, and one can easily think of more perspectives on historical relations between words, including the terminology that is used in evolutionary biology. In Table 2, I have attempted to illustrate in which way the different terms, including the biological terms of homology, orthology, and xenology, cover processes by controlling each for one or more of the three dimensions of lexical change (with + indicating that continuity is required, − indicating that change is required, and +/− indicating indifference). Note that paralogy was not included in the comparison, since the process of gene duplication is a very specific event that probably has no fruitful analogy in historical linguistics. Contrasting the different dimensions of lexical change with the terminology used to refer to different relations between words shows the arbitrariness of the traditional linguistic terminology. Why do we only cover two out of $3 * 3 * 3 = 27$ different possible types? Why do we only control by requiring continuity, not change? It also shows the fundamental difference between change processes in linguistics and biology.

4. Models of lexical change in phylogenetic reconstruction

In the previous sections, I have tried to show that not only the terminology that we use to denote historical relations between evolving entities in linguistics and biology shows some important differences, but also that the processes underlying lexical change in language history are very particular, involving three major dimensions of lexical variation which themselves can be further subdivided into a multitude of minor process types.[4] In the following, I will try to illustrate how our models can be modified in order to account for more complex historical relations between words.

4.1 Gain loss models and morphological variation

The majority of automatic methods for phylogenetic reconstruction in historical linguistics employ lexical data to infer language phylogenies. When employing these

3 Note that, in this context, 'controlling' for a dimension means to consider only those historically related words in which *no variation* along that very dimension occurred *during their history since separation*. If we compare French *soleil* 'sun' with Italian *sole* 'sun', for example, we would need to state that the French word changed its meaning from *small sun* to *sun*, and although both forms are identical regarding their synchronic meaning, their history involves variation along the semantic dimension (see Starostin 2013 for more examples on cases of *unilateral independent semantic development*). In practice, when linguists prepare lexicostatistical databases, however, controlling for meaning is usually reduced to checking for identity along a given dimension. It is clear that this can be problematic. In the absence of counterevidence the majority of linguists would probably assume that meaning identity in cognate word forms is good evidence that no semantic change happened since the separation of the forms, but it is obvious that semantic identity is only a necessary for semantic continuity since separation.

4 Already a brief overview of some classical work on the complexities of semantic change (Wilkins 1996), morphological change (Koch 1996), and stratic change (Weinreich 1974) shows that the three-dimensional model of lexical change only touches the tip of the huge iceberg of lexical change.

Table 3. Lexicostatistical scheme of data-encoding and the creation of presence-absence matrices. The table shows how lexicostatistical word lists are produced, how cognates are assigned to words by using numerical identifiers, and how the data are then converted into binary presence absence matrices for the purpose of phylogenetic comparison. Note that the proto-form which is given for each cognate set in the table below is not necessarily included in lexicostatistical datasets, but it, nevertheless, is implicitly assumed.

Basic Concept	German	ID	English	ID	Italian	ID	French	ID
HAND	Hand	1	hand	1	mano	2	main	2
BLOOD	Blut	3	blood	3	sangue	4	sang	4
HEAD	Kopf	5	head	6	testa	7	tête	7
...	...	...	...	...	...	...	...	...

ID	Proto-Form	Basic Concept	German	English	Italian	French
1	PGM *xanda-	HAND	1	1	0	0
2	LAT *mānus*	HAND	0	0	1	1
3	PGM *bloda-	BLOOD	1	1	0	0
4	LAT *sanguis*	BLOOD	0	0	1	1
5	PGM *kuppa-	HEAD	1	0	0	0
6	PGM *xawbda-	HEAD	0	1	0	0
7	LAT *tēsta*	HEAD	0	0	1	1
...	...	...	...	...	...	...

methods, it is important to specify a model of lexical change that the algorithms can use to infer the trees or the networks that fit the data best. Most datasets employ a lexicostatistical scheme of data-coding (Dyen et al. 1992; Ringe et al. 2002; Greenhill et al. 2008; Bouckaert et al. 2012; Greenhill 2015). This means, that they are based on concept lists of 100 and more items which are translated into the languages under investigation. By comparing all translations in each concept slot with each other, linguists then annotate which words are cognate. The notion of cognacy that is underlying these databases is usually the notion of 'cognacy à la Swadesh' in Table 2, that is, annotators try to filter out borrowings, consider only semantically identical items, and do not necessarily regard morphological variation.

The methods which are then used to analyze the data, be they based on probabilistic approaches (Felsenstein 1981; Huelsenbeck et al. 2001), or parsimony (Fitch 1971; Sankoff 1975), are almost exclusively based on gain–loss models of lexical change (Pagel 2009). They reduce the change of phylogenetic characters to processes of gain and loss and essentially assume that during evolution a language can either gain a new word or loose an existing one. In these models, each phylogenetic character has only two states, presence, or absence, and presence–absence matrices of cognate sets are fed to the algorithms in order to infer language phylogenies. Presence–absence matrices are retrieved from the original data by breaking up the semantic slots into sets of cognate words, and listing for each language whether it has a word belonging to the respective cognate set or not (Atkinson and Gray 2006). This way of data preparation and encoding is further illustrated in Table 3.

The binary coding practice has strong consequences, since it is vulnerable to historical word relations with variation along the semantic and the morphological dimension. First, the general procedure by which lexicostatistical data is binary encoded and concepts are split into several independent characters creates dependencies which cannot be observed by the algorithms. It deprives the analysis of the essential criterion for gain and loss, since presence and absence are defined with respect to meaning identity. Gain and loss need to be essentially interpreted as gain and loss with respect to a certain concept slot, not with respect to the entire language. The loss of a word means that the word is no longer used to express a certain meaning, and the gain of a word implies that a new word is used to express a certain meaning. Yet since meaning is discarded by the binarisation procedure (see Table 3), the models are given no clue to handle instances of parallel semantic shift. A more realistic gain–loss analysis should include a larger sample of words and annotate cognates regardless of differences in meaning (Michael et al. 2015).

Second, the lexicostatistical coding practice is vulnerable with respect to morphological change, since morphological variation is deliberately ignored when assigning words to cognate sets. This was not the case in the early days of lexicostatistics. Hattori (1961), for example, distinguished clearly between true 'orthologues' and morphologically derived words. Recalling the example of Italian *dare* and French *donner* given in Fig. 3, it is clear that we can annotate the words quite differently, depending not only on the "shade" of cognacy we choose, but also on the desired depth of analysis. In current practice, words like *dare* and *donner* are usually assigned to the same cognate set, and their morphological differences are ignored.[5] When annotating the words, however, we should ask ourselves which kind of annotation would be the best for the underlying model that we use. From this perspective, we would do best in coding Italian *dare* and French *donner* as being not cognate, since by the time that *donner* replaced earlier *dare* in the ancestor of French, the word *dare* was lost with respect to the meaning 'to give', and the word *donner* was gained.

5 Compare the coding in the Indo-European Lexical Cognacy Database at http://ielex.mpi.nl/cognate/405/, version accessed on 2016-04-08 available at WebCite: http://www.webcitation.org/6dGAxAG9r.

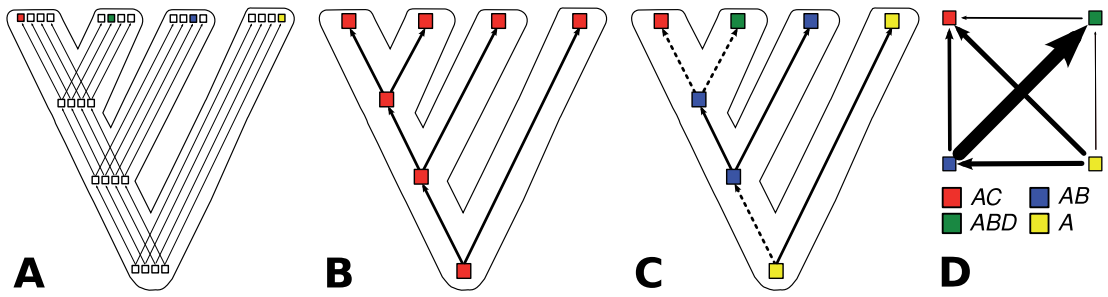


Figure 5. From gain–loss models to weighted directed character-state transitions: (A) Shows a strict approach in which four partially related compound words (as show at the bottom of D) are modeled as four different characters. (B) Shows the consequences of a lumping approach when partially cognate words are treated as fully cognate in binary presence–absence models. (C) Shows weighted directed character–state transitions, based on known transition tendencies displayed at the top of (D), with arrows indicating directions and edge width indicating the relative strength of transition tendencies.

Table 4. Complex etymological structure in word compounds. The table shows partial etymological relations of words for 'moon' in four Chinese dialects. Dialect data Hóu (2004), Middle Chinese (MC) readings follow Baxter (1992) with modifications.

Variety	Form	Character	Etymological structure			
			MC *ŋiot 月	MC *kwaŋ 光	MC *bjut 佛	MC *ljaŋɦ 亮
Fúzhōu 福州	ŋuoʔ^{5}	月	ŋ u o ʔ ^{5}			
Měixiàn 梅縣	ŋiat^{5} kuoŋ^{44}	月光	ŋ i a t ^{5}	k u o ŋ ^{44}		
Wēnzhōu 溫州	ɲy^{21} kuɔ^{35} vai^{13}	月光佛	ɲ - y - ^{21}	k u ɔ - ^{35}	v a i ^{13}	
Běijīng 北京	yɛ^{51} liɑŋ^{1}	月亮	– y ɛ - ^{51}			l i ɑ ŋ ^{1}

The problem of morphological variation in lexicostatistical datasets becomes even more evident when looking at more specific processes of morphological change like *compounding*. While compounding is less characteristic for the Indo-European language familiy (at least as far as the stable parts of the lexicon are concerned), it plays an important role in the Sino-Tibetan language family (Matisoff 2000: 341f; Chung et al. 2014; List 2015: 56–58). In the Chinese dialects, for example, the majority of words is only indirectly related, as illustrated in Table 4 where the words for 'moon' in four Chinese dialects share the same base morpheme, but differ regarding the further parts of their compounds. When investigating these patterns, we can immediately infer processes of lexical change that link these patterns. Fúzhōu [ŋuoʔ^{5}] 月, for example, reflects the oldest stage in which Chinese was still predominantly monosyllabic. Měixiàn [ŋiat^{5} kuoŋ^{44}] 月光 reflects a younger stage in which bisyllabic structures were gaining ground, and Wēnzhōu [ɲjy^{21}kuɔ^{35}vai^{13}] 月光佛 reflects an even later stage, since it builds on the form in Měixiàn, adding a suffix that marks nominalization (compare Wēnzhōu [ɲji^{21}dʏu^{35}vai^{13}] 日頭佛 'sun').[6] In the 'classical' lexicostatistical view of cognacy and the 'classical' models of word gain and word loss, these processes are all ignored, although they may bear important phylogenetic information. One would either label all four words as cognate, since they share the same base morpheme (Satterthwaite-Phillips 2011: 95–103), or label them all as not being cognate, since their parts to not match completely (Ben Hamed and Wang 2006; Gates 2012: 51). If we want to model the evolution of the four words for 'moon' in the four dialects realistically, neither of the two encoding practices will be of use. In both cases, all phylogenetic signal will be lost and the analysis cannot tell us how the words really developed (see Fig. 5A and B).

4.2 From binary to multistate models

In principle, phylogenetic methods can handle semantic and morphological variation sufficiently. All we need to

6 Note that in this case, as in general when dealing with lexical change in a classical lexicostatistical framework, sound change is ignored as a factor of change, since regular sound change involves the sound system and not individual phonetic material (Gévaudan, 2007: 14).

do is to switch from binary gain–loss models to multistate models. In a binary state model each character can only be present or absent in a given language, like the cognate set 1 in Table 3, for example, which is present in German and English but absent in Italian and Spanish. In a multistate model, a character cannot only be present or absent, but it can also *vary* among languages and occur in different shapes. Instead of labeling French *donner* and Italian *dar* either as exclusively cognate or as exlusively noncognate, we could assign both words to the same character but assign them different states. In this way, we could handle both variation along the semantic and the morphological dimension of lexical change. If we can further determine how likely it is for the character to switch from one particular state to another, we can force our algorithms to prefer certain transitions and to ignore others. In the case of the Chinese words for 'moon' in Table 4, for example, we already saw that Měixiàn [ŋiat^{5} kuoŋ^{44}] 月光 is particularly close to Wēnzhōu [ɲjy^{21}kuɔ^{35}vai^{13}] 月光佛, since the latter was only extended by one suffix. When comparing the Wēnzhōu form with the form [ŋuoʔ^{5}] 月 in Fúzhōu, we can further easily say that the transition from the Fúzhōu form to the Měixiàn form should be easier to accomplish than the direct transition to the Wēnzhōu form. If we further know that the process we are dealing with has strong *unidirectional tendencies*, as it is the case for many processes of sound change and grammaticalization (Haspelmath 2004), but also in inflectional morphology (Wurzel 1985), and potentially even in analogy (Jacques 2016), we can model this by using *irreversible models* in our analyses (Huelsenbeck et al. 2002; Bohl and Lancaster 2003).

In a parsimony framework of phylogenetic reconstruction (Fitch 1971; Sankoff 1975), the difficulty of switching between the different states of a character is handled by defining specific *weights* for character state transitions. If we further know that the process we are dealing with has strong unidirectional tendencies, we can model this by assigning *asymmetric weights* for the transition preferences between the states of a character. The differences between gain–loss models and multistate models allowing for asymmetric transition preferences in a parsimony framework are exemplified in Fig. 5, but multistates and asymmetric transition tendencies can essentially also be handled in probabilistic frameworks.

5. Using improved models to study Chinese dialect history

In order to illustrate the benefits of improved models for lexical change, I have prepared a small experiment on Chinese dialect history. In this experiment, I test how well different models of lexical change with varying degrees of complexity perform on the task of *semantic reconstruction*. In classical historical linguistics, semantic reconstruction seeks to infer the original meaning of a set of cognate words (Fox 1995: 115–6). The experiment I designed follows lexicostatistical approaches in which semantic reconstruction seeks to identify the word form which was used to express a certain concept in an ancestral language (Kassian et al. 2015: 304–6). In this context, semantic reconstruction can be treated as a specific type of *ancestral state reconstruction* (Pagel 1999) applied to lexicostatistical data. The starting point is a lexicostatistical wordlist, consisting of a list of concepts which are translated into a set of language varieties. Concepts comprise phylogenetic characters, and the counterparts of the concepts in the respective language varieties reflect different states of the characters. Semantic reconstruction starts from a *reference phylogeny* (a phylogenetic tree) and tries to infer which character state was present at the root. Chinese is attested through its contemporary dialects, whose diversity is at least comparable to that of the Romance languages (Wang 1997), but also in ancient texts predating the diversification of the modern dialect varieties by several hundred years.[7] Therefore, in the majority of cases, there is independent evidence regarding the words which were originally used to express a given concept. For this reason, Chinese is an ideal candidate to test the performance of different models of lexical change.

7 There is some disagreement among Chinese linguists regarding the exact dating of the ancestor of all Chinese dialects. Some scholars assume that the modern dialects developed from a *koine* spoken in the early Táng 唐 dynasty (618–907 AD) around 600 AD (Karlgren 1954; Pulleyblank 1984). Other scholars propose an earlier diversification. Assuming that the very conservative Mǐn 閩 dialect group had much earlier split off from the rest of Chinese (Norman and Coblin 1995; Handel 2010), they place their common ancestor in the late Hàn 漢 dynasty (206 BC–220 AD) some time around 200 AD. Nevertheless, with ancient Chinese texts dating back to 1000 BC and earlier, with rich collections of classical texts being available from the sixth century BC onwards, Ancient Chinese is clearly ancestral to all Chinese dialects, as is also reflected in its sound system (Baxter and Sagart 2014).

Table 5. The concepts selected for the study

1. ash / 灰	2. back / 背	3. belly / 腹	4. bird / 鸟	5. bone / 骨	6. claw / 爪
7. cloud / 云	8. day / 天	9. dog / 犬	10. ear / 耳	11. earth / 地	12. eat / 食
13. egg / 卵	14. eye / 目	15. fire / 火	16. flesh / 肉	17. flower / 花	18. fog / 雾
19. fruit / 果	20. guts / 肠	21. hand / 手	22. heart / 心	23. horn / 角	24. ice / 冰
25. knee / 膝	26. lake / 湖	27. leaf / 叶	28. leg / 脚	29. liver / 肝	30. louse / 虱
31. man / 男	32. moon / 月	33. mouth / 口	34. name / 名	35. neck / 颈	36. night / 夜
37. nose / 鼻	38. path / 路	39. person / 人	40. river / 江	41. rope / 索, 绳	42. sand / 沙
43. seed / 种	44. skin / 皮	45. sky / 天	46. smoke / 烟	47. snake / 蛇	48. star / 星
49. stone / 石	50. sun / 日	51. tail / 尾	52. tongue / 舌	53. tooth / 牙	54. water / 水
55. wing / 翼	56. woman / 女	57. worm / 虫			

5.1 Materials

The data for the experiment were originally compiled for the study of Ben Hamed and Wang (2006). It comprises 200 concepts translated into 23 Chinese dialect varieties. The concept list is largely identical with the list of 200 items proposed by Swadesh (1952).[8] In the data, partial cognate relations are annotated by listing the 'etymological character' for each morpheme of a word (*běnzì* 本字, Branner 2000: 35). This information is regarded as problematic by some Chinese dialectologists (Branner 2000), since it is not necessarily clear how consistently the morphemes in dialect words are identified. Datasets like the one by Ben Hamed and Wang (2006) are, nevertheless, a useful starting point for experiments on morphological processes in lexical evolution, especially since other collections which list information on partial cognacy in such great detail are not available. For most of the cases, however, we can assume that the assignments are correct. In an earlier study (List 2015), I used the data by Ben Hamed and Wang (2006) and converted it into a machine-readable text format, which I used for this experiment. All data were thoroughly checked and refined, since the partial cognate assignments were not the primary target of my earlier study and therefore only inconsistently converted into text format.

Ben Hamed and Wang (2006) also give the ancestral forms for the concepts in Old Chinese. Since Old Chinese is supposed to be the ancestor of all dialect varieties in the sample, the data can be used as a 'gold standard' to test the accuracy of ancestral state reconstruction methods. Since processes of lexical evolution are quite different for nouns and verbs, with compounding and partial cognacy occurring almost exclusively on nouns, only nouns were considered for this study. Of the 85 concepts denoting nouns in the sample, 28 were excluded. Either the reflexes were all different from the Old Chinese forms and it would be impossible to reconstruct them, or the reflexes were all identical with the Old Chinese form, and reconstruction would be no challenge at all. The 57 forms considered for the experiment are listed in Table 5 along with the supposed ancestral forms in Old Chinese.

Ancestral state reconstruction requires a reference phylogeny as input. Here I build on an earlier approach (List 2015) where I compared reference phylogenies for three independent hypotheses on Chinese dialect history, namely Laurent Sagart's *Arbre des Dialectes Chinois* (Sagart 2011), the *Hànyǔ Fāngyán Shùxíngtú* 漢語方言樹形圖 ('Tree chart of Chinese dialects') by Yóu Rǔjié 游汝傑 (Yóu 1992: 91–106), and Jerry Norman's *Southern Chinese Hypothesis* (Norman 1988: 210–4). These reference phylogenies differ regarding the subgrouping of the seven major dialect groups of Chinese and are based on competing criteria for subgrouping (see List 2015: 36f for details).

5.2 Methods

The experiment employs a *parsimony framework* for character transitions (Nunn 2011: 59–63). Parsimony was used for reasons of simplicity and data sparseness. Parsimony applications can be easily implemented from scratch, while there are no available ready-to-use implementations of probabilistic approaches which handle asymmetric transitions between multiple character states. Given the sparseness of the data available for testing, it is also not clear whether probabilistic applications would converge. Four different models of varying complexity were defined for the experiment:

a. BINARY: Character states which are not completely identical in their compound structure are split into sets of binary characters following the classical procedure described in Atkinson and Gray (2006).

8 The list is included into the Concepticon resource (http://concepticon.clld.org, see List et al. 2016).

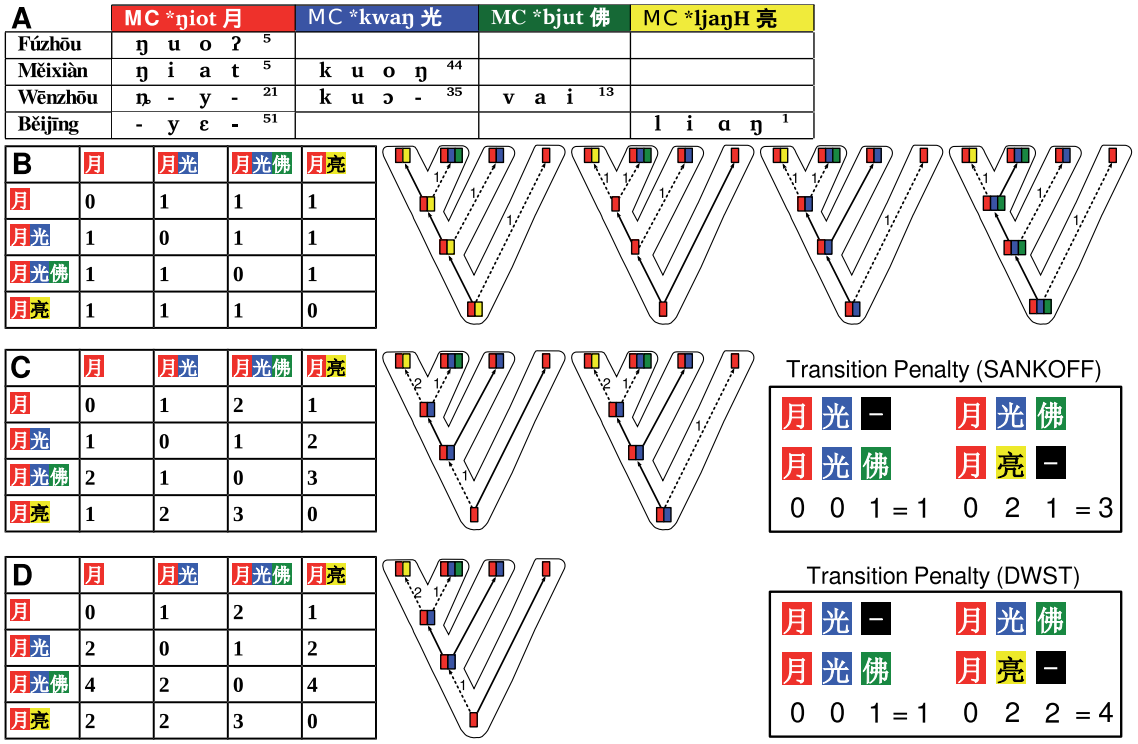


Figure 6. Comparing multistate models for lexical change. The figure shows how the evolution of the four words for 'moon' is inferred within a parsimony framework. On top, the etymological structure of the words is displayed, and unique colors are assigned to refer to the morpheme structure in the remainder of the figure (A). On the left, the penalties for character transitions (step matrices) are shown for the FITCH (B), the SANKOFF (C), and the DWST model (D). For SANKOFF and DWST, example calculations for transition penalties are displayed on the right (see also the main text). For each model, all trees with optimal weight are displayed. Dashed edges in the trees indicate a transition involving a change. Numbers on dashed lines denote the weight, as derived from the corresponding matrix of transition penalties on the left.

Character transitions are modeled as a gain–loss process.

b. FITCH (multistate): Lexical evolution is modeled as a process of character transitions with equal weights, following the classical model by Fitch (1971).

c. SANKOFF (multistate, weighted): Lexical evolution is modeled as a process of character transitions with unequal weights, following the classical model by Sankoff (1975).

d. DWST ('directed weighted state-transitions', multistate, weighted, directed): Lexical evolution is modeled as a process of character transitions with unequal weights and in dependence of the direction of the transition.

The BINARY and the FITCH model are straightforward in their implementation. The BINARY model only handles gains and losses with losses being favored over gains. The parsimony weight for gain events was set to 2, and the penalty for loss events was set to 1, since these penalties yielded the most plausible scenarios in earlier experiments on the data (List 2015). The FITCH model gives equal weights to transitions between all states. In the case of SANKOFF and DWST, transitions are weighted differently depending on the character states. Since we lack exhaustive linguistic accounts on processes of compounding in the Chinese dialects, a very simple approach for the computation of the weights was employed. In a first step, the morpheme representation of two words, which is given in Chinese character readings, with identical characters representing cognate morphemes, was aligned using the Needleman–Wunsch algorithm (Needleman and Wunsch 1970). In a second step, it was counted in how many positions the aligned sequences differ. This distance, commonly known as the Hamming distance (Hamming 1950), was further refined by counting substitutions (those instances where two different morphemes are aligned) twice, and insertions and deletions (those instances where a morpheme was aligned with a gap symbol or vice versa) only once.

Table 6. Comparing the results for the four analyses and the three reference trees. The first number in the *hits* and the *fails* column indicates the proportion, the second number indicates the absolute values. As mentioned in the text, hits and fails are computed by comparing for all proposed forms reconstructed back to the root whether they are identical with the forms in the gold standard. If they are, this counts as a *hit*, if not, this counts as a *fail*. If more than one form are proposed for a given concept, results are averaged.

Model	*Arbre*		*Shùxíngtú*		*Southern Chinese*	
	Hits	Fails	Hits	Fails	Hits	Fails
BINARY	0.55 / 31.04	0.45 / 24.96	0.52 / 29.04	0.48 / 26.96	0.52 / 28.95	0.48 / 27.05
FITCH	0.63 / 35.51	0.37 / 20.49	0.51 / 28.31	0.49 / 27.69	0.47 / 26.40	0.53 / 29.60
SANKOFF	0.76 / 42.83	0.24 / 13.17	0.67 / 37.50	0.33 / 18.50	0.62 / 34.50	0.38 / 21.50
DWST	0.82 / 45.70	0.18 / 10.30	0.82 / 46.00	0.18 / 10.00	0.79 / 44.50	0.21 / 11.50

In contrast to the SANKOFF model, the computation of weights for the DWST model only reduces the weights for insertions (a gap aligned with a morpheme), but not for deletions. This transition schema accounts for the tendency of *disyllabification* in the history of Chinese, during which most of the monosyllabic words in the Chinese dialects were replaced by bisyllabic compounds. Figure 6 gives examples for the differences in the transition penalties of the multistate-models (FITCH, SANKOFF, and DWST) and the calculation of the transition penalties for the SANKOFF and the DWST model. It is beyond doubt that the models could be further refined, and potentially also trained. For the purpose of the experiment, however, it is advisable to keep the models as abstract as possible. This guarantees that we do not overly fit the models to the data, and it also makes it easier to determine the major factors that determine differences in their performances.

The models and the code to optimize the parsimony score were implemented in Python. The code requires the LingPy software package for quantitative tasks in historical linguistics (List and Moran 2013) to calculate the alignments between the characters states and the transition probabilites. The source code along with the data, the results, and further instructions on how to replicate all analyses presented in this article are provided as supplementary data.

5.3 Results

With four different models and three different reference phylogenies, 12 different tests needed to be carried out. In order to evaluate the quality of semantic reconstruction, a simple approach was used. In this approach, one counts the amount of *hits* and *fails*. For each concept, all ancestral forms proposed by a given test were considered and compared with the known forms in the 'gold standard'. If only one form was proposed, this form can either be a hit or a fail, that is, it can either be identical with the form in the gold standard, or not. If multiple forms are proposed by an algorithm, the score is divided among hits and fails, following the proportion of correctly and incorrectly proposed ancestral forms. If, for example, two forms are proposed of which only one is correct, this would be scored as a 50% hit and a 50% fail. The results were evaluated separately for each meaning slot and then averaged across all 57 concepts in the sample.

Table 6 shows the detailed results for all 12 different analyses, including the overall parsimony scores obtained. The DWST model performs best in all respects, regardless of the reference phylogeny. The SANKOFF model outperforms the remaining two models, but only when applied to the *Arbre* reference phylogeny, it comes close to the high scores of the DWST model. Whether the BINARY or the FITCH model performs better is hard to say, given that the differences are minimal on average, and both models seem to rely heavily on the reference phylogeny. What is remarkable is that the DWST model does not only show the highest scores, but also a high resistency regarding the underlying reference phylogeny. According to the analysis by List (2015), the *Arbre* gives a more realistic picture of Chinese dialect history. This is reflected by the highly improved scores of all models (except from DWST) for the *Arbre* phylogeny as opposed to *Shùxíngtú* and *Southern Chinese*.

Parsimony approaches may yield multiple solutions which are all optimal with respect to the transition penalties defined in a model. Depending on the character and tree topology, the amount of optimal scenarios may vary greatly. In the FITCH analyses, for example, the number of possible scenarios for all characters ranges from 1 (for 'ash') to 4 797 (for 'night'). As expected, the

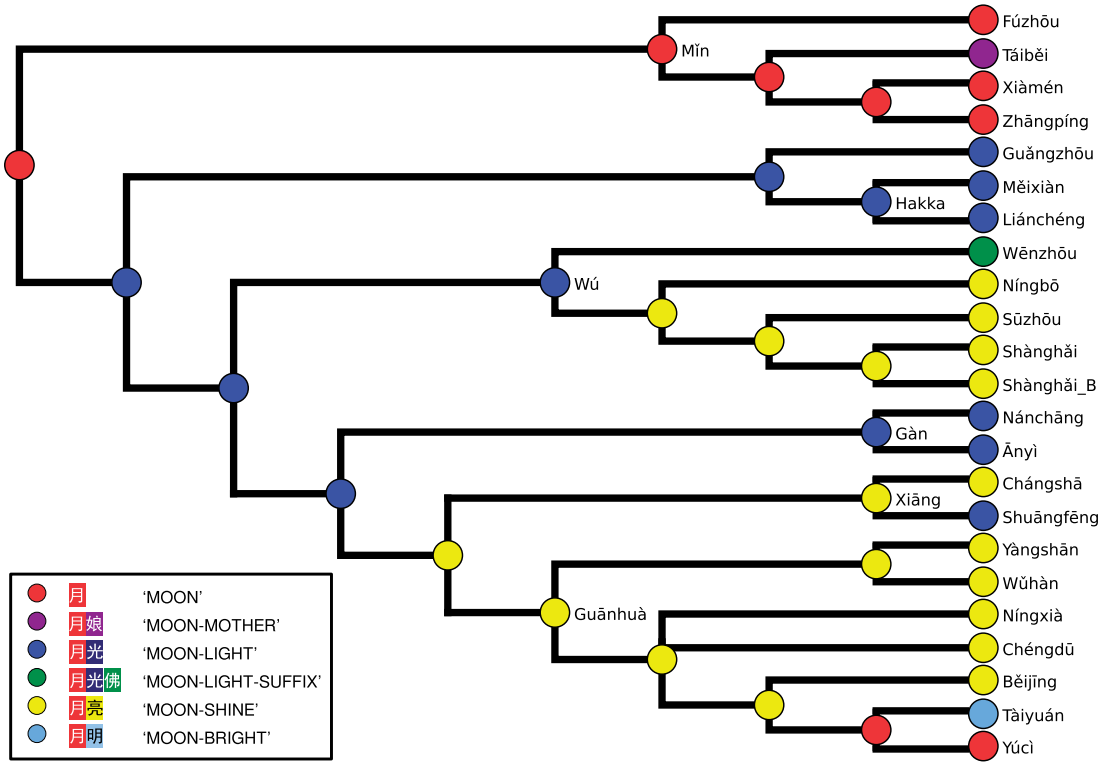


Figure 7. One of four optimal scenarios for the development of words for 'moon' along the *Arbre* reference phylogeny.

Table 7. Comparing the proposed proto-forms and the number of optimal scenarios based on the *Arbre* reference phylogeny for three exemplary concepts. Forms with an asterisk represent 'hits', that is, forms which are identical with the gold standard.

Models	'belly'		'ear'		'moon'	
	Forms	Scen.	Forms	Scen.	Forms	Scen.
FITCH	肚, 腹老, 腹肚, 腹*	39	耳*, 耳朵, 耳仔, 耳菇, 耳公	34	月*, 月光	48
SANKOFF	肚*, 腹肚, 腹	5	耳*	3	月*, 月光	8
DWST	腹*, 肚	2	耳*	1	月*	4

number of possible scenarios decreases when increasing the complexity of the models. This is shown in Table 7 where the proposed proto-forms and the number of possible scenarios for the analysis of three concepts using the three multistate models are displayed. The table shows clearly that complex models reduce the uncertainty with respect to alternative scenarios.

Figure 7 shows one of four possible scenarios for the development of reflexes of 'moon' inferred by the DWST model for the *Arbre* reference phylogeny. The scenario proposes a pattern in which the word form *yuè* 月 'moon' was replaced by the compound *yuèguāng* 月光 'moon-light' in all dialects except from the Mǐn subgroup. While this may well reflect a realistic scenario, we also find homoplastic (reoccurring) transitions, especially from *yuèguāng* 月光 to *yuèliàng* 月亮 'moon-shine' in the Wú subgroup. Homoplasy may point to lateral transfer events (List et al. 2014, Dagan and Martin 2007), but our knowledge regarding lexical evolution during the history of the Chinese dialects is still very limited. It is extremely difficult to tell with certainty whether the common reflexes of *yuèliàng* 月亮 in the Běijīng-Xiāng and the Wú subgroup reflect independent parallel developments or areal influence.

6. Conclusion

In this article, I have pointed to problems in the models used for phylogenetic reconstruction in linguistics resulting from a superficial treatment of historical relations between words. *Cognacy* is not a binary relation which is either present or not. Instead, we can distinguish different subtypes of cognacy, just as biologists can identify specific types of homology between genes. In an earlier paper, I proposed to compare the biological subtypes of homology (orthology, paralogy, xenology) directly with potential subtypes of historical word relations in linguistics (List 2014), but by concentrating on the major dimensions of lexical change proposed by Gévaudan (2007), namely morphological, semantic, and stratic change, I have shown that we can even go beyond the biological terminology and set up fine-grained schemas for historical relations in linguistics.

Which notion of cognacy we use for phylogenetic reconstruction crucially depends on the data we have at hand and the algorithms we intend to employ. I have shown that the inconsistencies in the treatment of historical relations between words have a direct impact on the way cognates are coded and data are analyzed in phylogenetic approaches. This was illustrated in detail for historical relations involving morphological change, especially compounding. If compounding is frequent and characteristic for a given language family, phylogenetic approaches which model lexical change merely as a process of cognate gain and cognate loss are inadequate and unrealistic. In order to take the different *degrees of cognacy* into account, I proposed to employ multistate instead of binary state models, and to further allow for potentially asymmetric transition tendencies among character states. The benefits of these models were demonstrated in a small experiment on semantic reconstruction applied to a lexicostatistical dataset of 23 Chinese dialect varieties. The results of this experiment strongly suggest that multistate models with asymmetric transition tendencies are superior to binary state models. What I have presented is, however, but a small step toward improved models of lexical change. More experiments including more language families need to be carried out. Instead of ancestral state reconstruction, we need to test the potential of multistate models for phylogenetic reconstruction in general. Probabilistic models, be they based on Maximum Likelihood (Felsenstein 1981) or Bayesian inference (Huelsenbeck et al. 2001), may prove really useful in this regard. In parsimony, we need to provide exact models for the transition between characters, and we always run the danger of overfitting our step matrices on a given dataset. Probabilistic models can help to estimate transition probabilities and could thus even provide new insights which go beyond cognacy and help us to detect major trends in lexical evolution, including morphological, semantic, and stratic change. In order to allow for these improved models of lexical change, however, we need to rethink the way we handle cognacy in our databases and start being more explicit in our annotations.

Supplementary data

The most recent release of the accompanying software application can be found at https://zenodo.org/badge/latestdoi/5137/digling/beyond-cognacy-paper. An interactive application showing all inferred evolutionary scenarios for the Arbre phylogeny by Sagart (2011) is available at http://digling.github.io/beyond-cognacy-paper/.

Acknowledgements

I thank three anonymous reviewers for challenging critics and helpful advice, and Gerhard Jäger for helpful comments made on an earlier version of this article. I am very grateful to Hans Geisler, who originally pointed me to many of the examples on lexical change which were treated in this article, to David Morrison, for important comments on the article and numerous fruitful discussions on questions of homology and cognacy, and to Laurent Sagart, who shared his ideas and his profound knowledge of Chinese dialect classification with me.

Funding

This research was supported by the DFG research fellowship grant 'Vertical and lateral aspects of Chinese dialect history' (Grant No. 261553824), which is gratefully acknowledged.

References

Arapov, M. V. and Xerc, M. M. (1974) *Matematičeskie metody v istoričeskoj lingvistike* [Mathematical Methods in Historical Linguistics]. Moscow: Nauka.

Atkinson, Q. D. and Gray, R. D. (2006) 'How Old is the Indo-European Language Family? Illumination or More Moths to the Flame?', in P. Forster and C. Renfrew (eds.) *Phylogenetic Methods and the Prehistory of Languages*, pp. 91–109. Cambridge and Oxford and Oakville: McDonald Institute for Archaeological Research.

Barðdal, J. and Eythórsson, T. (2012) 'Reconstructing Syntax: Construction Grammar and the Comparative Method', in H. Boas and I. A. Sag (eds.) *Sign-based Construction Grammar*, pp. 257–308. Stanford: CSLI Publications

Baxter, W. H. (1992) *A Handbook of Old Chinese Phonology*. Berlin: de Gruyter.

—— and Sagart, L. (2014) *Old Chinese. A New Reconstruction*. Oxford: Oxford University Press.

Ben Hamed, M. and Wang, F. (2006) 'Stuck in the Forest: Trees, Networks and Chinese Dialects', *Diachronica*, 23: 29–60.

Bohl, E. and Lancaster, P. (2003) 'Irreversible Markov Processes for Phylogenetic Models', *Numerical Linear Algebra with Applications*, 10: 577–93.

Bouckaert, R., et al. (2012) 'Mapping the Origins and Expansion of the Indo-European Language Family', *Science*, 337: 957–60.

Branner, D. P. (2000) *Problems in Comparative Chinese Dialectology. The Classification of Miin and Hakka*. Berlin and New York: Mouton de Gruyter.

Campbell, L. and Harris, A. C. (2002) 'Syntactic Reconstruction and Demythologizing 'Myths and the Prehistory of Grammars', *Journal of Linguistics* 38: 599–618.

Chung, K. S., Hill, N. W. and Sun, J. T.-S. (2014) 'Sino-Tibetan', in R. Lieber and P. Štekauer (eds.) *The Oxford Handbook of Derivational Morphology*, pp. 619–50. Oxford: Oxford University Press.

Coseriu, E. (1973) *Probleme der strukturellen Semantik* [Problems of Structural Semantics]. Tübingen: Narr.

Croft, W. (2008) 'Evolutionary Linguistics', *Annual Review of Anthropology*, 37: 219–34.

Dagan, T. and Martin, W. (2007) 'Ancestral Genome Sizes Specify the Minimum Rate of Lateral Gene Transfer During Prokaryote Evolution', *Proceedings of the National Academy of Sciences*, 104: 870–75.

de Saussure, F. (1916) *Cours de linguistique générale* [Course in General Linguistics]. Lausanne: Payot.

Dyen, I., Kruskal, J. B. and Black, P. (1992) 'An Indoeuropean Classification', *Transactions of the American Philosophical Society*, 82: iii–132.

Felsenstein, J. (1981) 'Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach', *Journal of Molecular Evolution*, 17: 368–76.

Fitch, W. M. (1970) 'Distinguishing Homologous from Analogous Proteins', *Systematic Zoology*, 19: 99–113.

——. (1971) 'Toward 'Defining the Course of Evolution: Minimum Change for a Specific Tree Topology', *Systematic Biology*, 20: 406–16.

——. (2000) 'Homology. A Personal View on Some of the Problems', *Trends in Genetics*, 16: 227–31.

Fox, A. (1995) *Linguistic Reconstruction*. Oxford: Oxford University Press.

Freeman, V. J. (1951) 'Studies on the Virulence of Bacteriophage-infected Strains of Corynebacterium Diphtheriae', *Journal of Bacteriology*, 61: 675–88.

Gates, J. P. (2012) 'Situ in Situ. Towards a Dialectology of Jiāróng (rGyalrong)', PhD thesis, Trinity Western University.

Geisler, H. and List, J.-M. (2013) 'Do Languages Grow on Trees? The Tree Metaphor in the History of Linguistics', in H. Fangerau, H. Geisler, T. Halling and W. Martin (eds.) *Classification and Evolution in Biology, Linguistics and the History of Science. Concepts – Methods – Visualization*, pp. 111–24. Stuttgart: Franz Steiner Verlag.

Gévaudan, P. (2007) *Typologie des lexikalischen wandels* [Typology of Lexical Change]. Tübingen: Stauffenburg.

Gray, G. S. and Fitch, W. M. (1983) 'Evolution of Antibiotic Resistance Genes', *Molecular Biology and Evolution*, 1: 57–66.

Gray, R. D. (2005) 'Evolution: Pushing the Time Barrier in the Quest for Language Roots', *Science*, 309: 2007–8.

Gray, R. D. and Atkinson, Q. D. (2003) 'Language-tree Divergence Times Support the Anatolian Theory of Indo-European Origin', *Nature*, 426: 435–9.

Greenhill, S. J. (2015) 'TransNewGuinea.org: An Online Database of New Guinea Languages', *PLoS One*, 10: e0141563.

Greenhill, S. J., Blust, R. and Gray, R. D. (2008) 'The Austronesian Basic Vocabulary Database: From Bioinformatics to Lexomics', *Evol. Bioinformatics*, 4: 271–83.

Hamming, R. W. (1950) 'Error Detection and Error Detection Codes', *Bell System Technical Journal*, 29: 147–60.

Handel, Z. (2010) 'Old Chinese and Min', *Chûgoku Gogaku* 中國語學 [Bulletin of the Chinese Language Society of Japan], 257: 34–68.

Haspelmath, M. (2004) 'On Directionality in Language Change with Particular Reference to Grammaticalization', in O. Fischer, M. Norde and H. Perridon (eds.) *Up and Down the Cline – The Nature of Grammaticalization*, pp. 17–44. Amsterdam and Philadelphia: John Benjamins.

Hattori, S. (1961) 'A Glottochronological Study on Three Okinawan Dialects', *International Journal of American Linguistics*, 27: 52–62.

Hóu, J. (2004) *Xiàndài Hànyǔ fāngyán yīnkù* 現代漢語方言音庫 [Phonological Database of Chinese Dialects]. Shànghǎi: Shànghǎi Jiàoyù.

Huelsenbeck, J. P., Bollback, J. P. and Levine, A. M. (2002) 'Inferring the Root of a Phylogenetic Tree', *Systems Biology*, 51: 32–43.

——, et al. (2001) 'Bayesian Inference of Phylogeny and its Impact on Evolutionary Biology', *Science*, 294: 2310–14.

Jacques, G. (2016) 'On the Directionality of Analogy in a Dhegiha Paradigm', *International Journal of American Linguistics*, 82: 239–48.

Jensen, R. A. (2001) 'Orthologs and Paralogs – We Need to get it Right', *Genome Biology*, 2: 1002.1–1002.3.

Karlgren, B. (1954) 'Compendium of Phonetics in Ancient and Archaic Chinese', *Bulletin of the Museum of Far Eastern Antiquities*, 26: 211–367.

Kassian, A., Zhivlov, M. and Starostin, G. S. (2015) 'Proto-Indo-European-Uralic Comparison from the Probabilistic Point of View', *The Journal of Indo-European Studies*, 43: 301–47.

Katičić, R. (1966) 'Modellbegriffe in der vergleichenden Sprachwissenschaft *[The Conception of Models in Historical Linguistics]*', *Kratylos*, 11: 49–67.

Kluge, F. and Seebold, E. (2002) *Etymologisches Wörterbuch der deutschen Sprache* [Etymological Dictionary of German]. 24 edn. Berlin: de Gruyter.

Koch, H. (1996) Reconstruction in Morphology, in M. Durie, (ed.) *The Comparative Method Reviewed. Regularity and Irregularity in Language Change*, pp. 218–63. New York: Oxford University Press.

Koonin, E. V. (2001) 'An Apology for Orthologs – or Brave New Memes', *Genome Biology*, 2: 1005.1–1005.2

Koonin, E. V. (2005) 'Orthologs, Paralogs, and Evolutionary Genomics', *Annual Review of Genetics*, 39: 309–38.

Kroonen, G. (2013) *Etymological dictionary of Proto-Germanic*. Leiden and Boston: Brill.

List, J-M. (2014) *Sequence Comparison in Historical Linguistics*. Düsseldorf: Düsseldorf University Press.

List, J.-M.. (2015) 'Network Perspectives on Chinese Dialect History', *Bulletin of Chinese Linguistics*, 8: 42–67.

——, Cysouw, M. and Forkel, R. (2016) *Concepticon: A Resource for the Linking of Concept Lists*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://concepticon.clld.org>.

—— and Moran, S. (2013) 'An Open Source Toolkit for Quantitative Historical Linguistics', in *Proceedings of the ACL 2013 System Demonstrations*, pp. 13–18. Stroudsberg: Association of Computational Linguistics

—— et al. (2014) 'Networks of Lexical Borrowing and Lateral Gene Transfer in Language and Genome Evolution', *Bioessays*, 36: 141–50.

Matisoff, J. A. (2000) 'On the Uselessness of Glottochronology for the Subgrouping of Tibeto-Burman', in C. Renfrew, A. McMahon and L. Trask (eds.), *Time Depth in Historical Linguistics*, pp. 333–71. Cambridge: McDonald Institute for Archaeological Research.

Meiser, G. (1998) *Historische Laut- und Formenlehre der lateinischen Sprache* [Historical Studies of the Sounds and the Forms of Latin]. Darmstadt: Wissenschaftliche Buchgesellschaft.

Meyer-Lübke, W. (1911) *Romanisches etymologisches Wörterbuch* [Etymological Dictionary of Romance]. Heidelberg: Winter.

Michael, L., et al. (2015) 'A Bayesian Phylogenetic Classification of Tupí-Guaraní', *LIAMES*, 15: 193–221.

Morrison, D. (2015) 'Molecular Homology and Multiple-Sequence Alignment: An Analysis of Concepts and Practice', *Australian Systematic Botany*, 28: 46–62.

Needleman, S. B. and Wunsch, C. D. (1970) 'A Gene Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins', *Journal of Molecular Biology*, 48: 443–53.

Nelson-Sathi, S., et al. (2013) 'Reconstructing the Lateral Component of Language History and Genome Evolution using Network Approaches', in H. Fangerau, H. Geisler, T. Halling and W. Martin (eds.), *Classification and Evolution in Biology, Linguistics and the History of Science. Concepts – Methods – Visualization*, pp. 163–80. Stuttgart: Franz Steiner Verlag.

Norman, J. (1988) *Chinese*. Cambridge: Cambridge University Press.

—— and Coblin, W. S. (1995) 'A New Approach to Chinese Historical Linguistics', *Journal of the American Oriental Society*, 115: 576–84.

Nunn, C. L. (2011) *The Comparative Approach in Evolutionary Anthropology and Biology*. Chicago and London: University of Chicago Press.

Oesterreicher, W. (2001) 'Historizität, Sprachvariation, Sprachverschiedenheit, Sprachwandel *[*Historicity, Language Variation, Language Diversity, Language Change*]*', in M. Haspelmath (ed.) *Language Typology and Language Universals*, pp. 1554–95. Berlin and New York: Walter de Gruyter.

Owen, R. (1843) *Lectures on Comparative Anatomy*. London: Longman, Brown, Green, and Longmans.

Pagel, M. (2009) 'Human Language as a Culturally Transmitted Replicator', *Nature Reviews Genetics*, 10: 405–15.

Pagel, M. D. (1999) 'Inferring the Historical Patterns of Biological Evolution', *Nature*, 401: 877–84.

Petsko, G. A. (2001) 'Homologuephobia', *Genome Biology*, 2: 1002.1–1002.2.

Pfeifer, W. (1993) *Etymologisches Wörterbuch des Deutschen* [Etymological Dictionary of German], 2 edn. Berlin: Akademie. <http://www.dwds.de/>.

Pulleyblank, E. (1984) *Middle Chinese: A Study in Historical Phonology*. Vancouver: UBC Press.

Ringe, D., Warnow, T. and Taylor, A. (2002) 'Indo-European and Computational Cladistics', *Transactions of the Philological Society*, 100: 59–129.

Rix, H. et al. (2001) *Lexikon der Indogermanischen Verben* [Lexicon of Indo-European Verbs], Wiesbaden: Reichert.

Sagart, L. (2011) 'Classifying Chinese Dialects / Sinitic Languages on Shared Innovations', Paper, presented at the Séminaire Sino-Tibétain du CRLAO. <https://www.academia.edu/19534510/Chinese_dialects_classified_on_shared_innovations>.

Sankoff, D. (1975) 'Minimal Mutation Trees of Sequences', *SIAM Journal on Applied Mathematics*, 28: 35–42.

Satterthwaite-Phillips, D. (2011) 'Phylogenetic Inference of the Tibeto-Burman Languages or on the Usefuseful of Lexicostatistics (and "megalo"-comparison) for the Subgrouping of Tibeto-Burman', PhD thesis, Stanford University, Stanford.

Sonnhammer, E. L. and Koonin, E. V. (2002) 'Orthology, Paralogy and Proposed Classification for Paralog Subtypes', *Trends in Genetics*, 18: 619–20.

Starostin, G. S. (2013) 'Lexicostatistics as a Basis for Language Classification', in H. Fangerau, H. Geisler, T. Halling and W. Martin (eds.), *Classification and Evolution in Biology, Linguistics and the History of Science. Concepts – Methods – Visualization*, pp. 125–46. Stuttgart: Franz Steiner Verlag.

Swadesh, M. (1952) 'Lexico-statistic Dating of Prehistoric Ethnic Contacts', *Proceedings of the American Philosophical Society*, 96: 452–63.

——. (1955) 'Towards Greater Accuracy in Lexicostatistic Dating', *International Journal of American Linguistics*, 21: 121–37.

Syvanen, M. (1985) 'Cross-species Gene Transfer. Implications for a New Theory of Evolution', *Journal of Theoretical Biology*, 112: 333–43.

Taylor, J. S. and Raes, J. (2004) 'Duplication and Divergence: The Evolution of New Genes and Old Ideas', *Annual Review of Genetics*, 38.

Ternes, E. (1987) *Einführung in die Phonologie* [Introduction to Phonology]. Darmstadt: Wissenschaftliche Buchgesellschaft.

Trask, R. L. (2000) *The Dictionary of Historical and Comparative Linguistics*. Edinburgh: Edinburgh University Press.

Vaan, M., (ed.) (2008) *Etymological Dictionary of Latin and the Other Italic Languages*. Leiden and Boston: Brill.

Walkden, G. (2013) 'The Correspondence Problem in Syntactic Reconstruction', *Diachronica*, 30: 95–122.

Wang, W. S-Y. (1997) 'Languages or Dialects?', *The CUHK Journal of Humanities*, 1: 54–62.

Weinreich, U. (1974) *Languages in Contact*, 8th edn. The Hague and Paris: Mouton.

Wilkins, D. P. (1996) 'Natural Tendencies of Semantic Change and the Search for Cognates', in M. Durie (ed.), *The Comparative Method Reviewed. Regularity and Irregularity in Language Change*, pp. 264–304. New York: Oxford University Press.

Wodtko, D., Irslinger, B. and Schneider, C. (2008) *Nomina im Indogermanischen Lexikon* [Nouns in the Indo-European Lexicon]. Heidelberg: Winter.

Wurzel, W. U. (1985) 'Morphologische Natürlichkeit und morphologischer Wandel. Zur Vorhersagbarkeit von Sprachveränderungen [Morphologigal Naturalness and Morphological Change. On the Predictability of Language Change]', in J. Fisiak (ed.) *Papers from the 6th International Conference on Historical Linguistics*, pp. 587–99. Amsterdam: John Benjamins.

Yóu, R. (1992) *Hànyǔ fāngyánxué dǎolùn* 漢語方言學導論 [Chinese Dialectology]. Shànghǎi: Shànghǎi Jiàoyù

Zhang, J. (2003) 'Evolution by Gene Duplication: an Update', *Trends in Ecology and Evolution*, 18.

Using ancestral state reconstruction methods for onomasiological reconstruction in multilingual word lists*

Gerhard Jäger
Eberhard-Karls Universität Tübingen
gerhard.jaeger@uni-tuebingen.de

Johann-Mattis List
Max-Planck-Institute for the Science of Human History, Jena
list@shh.mpg.de

Abstract

Current efforts in computational historical linguistics are predominantly concerned with phylogenetic inference. Methods for ancestral state reconstruction have only been applied sporadically. In contrast to phylogenetic algorithms, automatic reconstruction methods presuppose phylogenetic information in order to explain what has evolved when and where. Here we report a pilot study exploring how well automatic methods for ancestral state reconstruction perform in the task of onomasiological reconstruction in multilingual word lists, where algorithms are used to infer how the words evolved along a given phylogeny, and reconstruct which cognate classes were used to express a given meaning in the ancestral languages. Comparing three different methods, Maximum Parsimony, Minimal Lateral Networks, and Maximum Likelihood on three different test sets (Indo-European, Austronesian, Chinese) using binary and multi-state coding of the data as well as single and sampled phylogenies, we find that Maximum Likelihood largely outperforms the other methods. At the same time, however, the general performance was disappointingly low, ranging between 0.66 (Chinese) and 0.79 (Austronesian) for the F-Scores. A closer linguistic evaluation of the

* This research was supported by the ERC Advanced Grant 324246 EVOLAEMP (GJ), the DFG-KFG 2237 *Words, Bones, Genes, Tools* (GJ), the DFG research fellowship grant 261553824 (JML) and the ERC Starting Grant 715618 CALC (JML). We thank our anonymous reviewers for helpful comments on earlier versions of this article, as well as all the colleagues who made their data and code publicly available.

reconstructions proposed by the best method and the reconstructions given in the gold standards revealed that the majority of the cases where the algorithms failed can be attributed to problems of independent semantic shift (homoplasy), to morphological processes in lexical change, and to wrong reconstructions in the independently created test sets that we employed.

Keywords

ancestral state reconstruction – lexical reconstruction – computational historical linguistics – phylogenetic methods

1 Introduction

Phylogenetic reconstruction methods are crucial for recent quantitative approaches in historical linguistics. While many scholars remain skeptical regarding the potential of methods for automatic sequence comparison, phylogenetic reconstruction, be it of networks using the popular SplitsTree software (Huson, 1998), or family trees, using distance- (Sokal and Michener, 1958; Saitou and Nei, 1987) or character-based approaches (Edwards and Cavalli-Sforza, 1964; Fitch, 1971; Ronquist et al., 2012; Bouckaert et al., 2014), has entered the mainstream of historical linguistics. This is reflected in a multitude of publications and applications on different language families, from Ainu (Lee and Hasegawa, 2013) and Australian (Bowern and Atkinson, 2012) to Semitic (Kitchen et al., 2009) and Chinese (Ben Hamed and Wang, 2006). There is also a growing interest in the implications of phylogenetic analyses for historical linguistics, as can be seen from the heated debate about the dating of Indo-European (Gray and Atkinson, 2003; Atkinson and Gray, 2006; Bouckaert et al., 2014; Chang et al., 2015), and the recent attempts to search for deep genetic signals in the languages of the world (Pagel et al., 2013; Jäger, 2015).

Given the boom of quantitative approaches in the search for language trees and networks, it is surprising that methods which infer the ancestral states of linguistic characters have been rarely applied and tested so far. While methods for phylogenetic reconstruction infer how related languages evolved into their current shape, methods for *ancestral state reconstruction* (ASR) use a given phylogeny to infer the previous appearance of the languages. This is illustrated in Fig. 1 for the reconstruction of lexical conceptualization patterns (more on this specific kind of ancestral state reconstruction below). What is modeled as ancestral state in this context is open to the researcher's interest, ranging from

the original pronunciation of words (Bouchard-Côté et al., 2013), the direction of sound change processes (Hruschka et al., 2015), the original expression of concepts (List, 2016), or even linguistic and cultural aspects beyond the lexicon, such as ancestral color systems (Haynie and Bowern, 2016), numeral systems (Zhou and Bowern, 2015) or cultural patterns, e.g., matrilocality (Jordan et al., 2009). While methods for ancestral state reconstruction are commonly used in evolutionary biology, their application is still in its infancy in historical linguistics. This is in strong contrast to classical historical linguistics, where the quest for proto-forms and proto-meanings is often given more importance than the search for family trees and sub-groupings. In the following, we will report results of a pilot study on ancestral state reconstruction applied to lexicostatistical word list data. Our goal is to infer which words were used to *express* a given concept in the ancestral languages.

This task is not to be confused with *semantic reconstruction*, where linguists try to infer the original meaning of a given word. Our approach, in contrast, reflects the onomasiological perspective on the linguistic sign, as we try to infer the original *word* that expressed a given meaning. Since no commonly accepted name exists for this approach, we chose the term "onomasiological reconstruction."[1] Classical semantic reconstruction in historical linguistics starts from a set of cognate words and tries to identify the original meaning of the ancestral word form (Wilkins, 1996). For this purpose, scholars try to take known directional tendencies into account. These tendencies are usually based on the author's intuition, despite recent attempts to formalize and quantify the evidence (Urban, 2011). Following the classical distinction between *semasiology* and *onomasiology* in semantics, the former dealing with 'the meaning of individual linguistic expressions' (Bussmann, 1996: 1050), and the latter dealing with the question of how certain concepts are expressed (ibid.: 834), semantic reconstruction is a *semasiological approach* to lexical change, as scholars start from the *meaning* of several lexemes in order to identify the meaning of the proto-form and its later development.

Instead of investigating lexical change from the semasiological perspective, one could also ask which of several possible word forms was used to denote a certain meaning in a given proto-language. This task is to some degree similar to proper semantic reconstruction, as it deals with the question of which meaning was attached to a given linguistic form. The approach, however, is

1 We chose this term for lack of alternatives, not because we particularly like it, and we are aware that it may sound confusing for readers less familiar with discussions on semantic change and lexical replacement, but we try to explain this in more detail below.

onomasiological, as we start from the concept and search for the "name" that was attached to it. *Onomasiological semantic reconstruction*, the reconstruction of former *expressions*, has been largely ignored in classical semantic reconstruction.[2] This is unfortunate, since the onomasiological perspective may offer interesting insights into lexical change. Given that we are dealing with two perspectives on the same phenomenon, the onomasiological viewpoint may increase the evidence for semantic reconstruction.

This is partially reflected in the "topological principle in semantic [i.e. onomasiological, GJ and JML] reconstruction" proposed by Kassian et al. (2015). This principle uses phylogenies to support claims about the reconstruction of ancestral expressions in historical linguistics, trying to choose the 'most economic scenario' (ibid.: 305) involving the least amount of semantic shifts. By adhering to the onomasiological perspective and modifying our basic data, we can model the problem of onomasiological reconstruction as an ancestral state reconstruction task, thereby providing a more formal treatment of the topological principle. In this task, we (1) start from a multilingual word lists in which a set of concepts has been translated into a set of languages (a classical "Swadesh list" or lexicostatistic word list; Swadesh, 1955), (2) determine a plausible phylogeny for the languages under investigation, and (3) use ancestral state reconstruction methods to determine which word forms were most likely used to *express* the concepts in the ancestral languages in the tree. This approach yields an analysis as the one shown in Fig. 1.

Although we think that such an analysis has many advantages over the manual application of the topological principle in onomasiological reconstruction employed by Kassian et al. (2015), we should make very clear at this point that our reformulation of the problem as an ancestral state reconstruction task also bears certain shortcomings. First, since ancestral state reconstruction models character by character independently from each other, our approach relies on identical meanings only and cannot handle semantic fields with fine-grained meaning distinctions. This is a clear disadvantage compared to qualitative analyses, but given that models always simplify reality, and that neither algorithms nor datasets for testing and training are available for the extended task, we think it is justified to test how close the available ancestral state reconstruction methods come to human judgments. Second, our phylogenetic approach to onomasiological reconstruction does not answer any questions regarding semantic change, as we can only state which words are likely to have been

2 Notable exceptions include work by S. Starostin and colleagues, compare, for example, Starostin (2016).

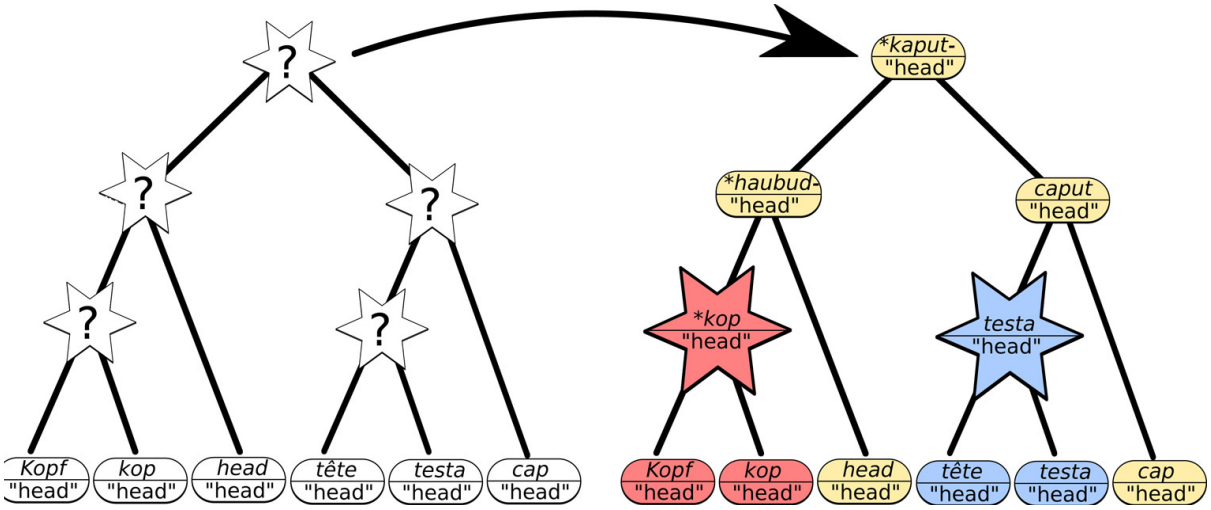


FIGURE 1 *Ancestral state reconstruction: The graphic illustrates the key idea of ancestral state reconstruction. Given six words in genetically related languages, we inquire how these words evolved into their current shape. Having inferred a phylogeny of the languages as shown on the left of the figure, ancestral state reconstruction methods use this phylogeny to find the best way to explain how the six words have evolved along the tree, thereby proposing ancestral* states *of all words under investigation. The advantage of this procedure is that we can immediately identify not only the original nature of the characters we investigate, but also the changes they were subject to. Ancestral state reconstruction may thus yield important insights into historical processes, including sound change and lexical replacement.*

used to express certain concepts in ancestral languages. This results clearly from the data and our phylogenetic approach, as mentioned before, and it is an obvious shortcoming of our approach. However, since the phylogenetic onomasiological reconstruction provides us with concrete hypotheses regarding the meaning of a given word on a given node in the tree, we can take these findings as a starting point to further investigate how words changed their meaning afterwards. By providing a formal and data-driven way to apply the topological principle, we can certainly contribute to the broader tasks of semantic and onomasiological reconstruction in historical linguistics. As a third point, we should not forget that our method suffers from the typical shortcomings of all data-driven disciplines, namely the shortcomings resulting from erroneous data assembly, especially erroneous cognate judgments, such as undetected borrowings (Holm, 2007) and inaccurate translations of the basic concepts (Geisler and List, 2010) which are investigated in all approaches based on lexicostatistical data. The risk that errors in the data have an influence on the inferences made by the methods is obvious and clear. In order to make sure that we evaluate the full potential of phylogenetic methods for ancestral state reconstruction, we therefore provide an exhaustive error analysis not only for the inferences made in our tests, but also for the data we used for testing.

In the following, we illustrate how ancestral state reconstruction methods can be used to approximate onomasiological reconstruction in multilingual word lists. We test the methods on three publicly available datasets from three different language families and compare the results against experts' assessments.

2 Materials and methods

2.1 *Materials*

2.1.1 Gold standard

In order to test available methods for ancestral state reconstruction, we assembled lexical cognacy data from three publicly available sources, offering data on three different language families of varying size:

1. Indo-European languages, as reflected in the *Indo-European lexical cognacy database* (IELex; Dunn, 2012, accessed on September 5, 2016),
2. Austronesian languages, as reflected in the *Austronesian Basic Vocabulary Database* (ABVD; Greenhill et al., 2008, accessed on December 2, 2015), and
3. Chinese dialect varieties, as reflected in the *Basic Words of Chinese Dialects* (BCD; Wang, 2004, provided in List, 2016).

All datasets are originally classical word lists as used in standard approaches to phylogenetic reconstruction: They contain a certain number of concepts which are translated into the target languages and then annotated for cognacy. In order to be applicable as a test set for our analysis, the datasets further need to list proto-forms of the supposed ancestral language of all languages in the sample. All data we used for our studies is available from the supplementary material.

The BCD database was used by Ben Hamed and Wang (2006) and is no longer accessible via its original URL, but it has been included in List (2015) and later revised in List (2016). It comprises data on 200 basic concepts (a modified form of the concept list by Swadesh, 1952) translated into 23 Chinese dialect varieties. Additionally, Wang (2004) lists 230 translations in Old Chinese for 197 of the 200 concepts. Since Old Chinese is the supposed ancestor of all Chinese dialects, this data qualifies as a gold standard for our experiment on ancestral state reconstruction. We should, however, bear in mind that the relationship between Old Chinese, as a variety spoken some time between 800 and 200 BC, and the most recent common ancestor of all Chinese dialects, spoken between

200 and 400 CE, is a remote one. We will discuss this problem in more detail in our linguistic evaluation of the results in section 4. Given that many languages contain multiple synonyms for the same concept, the data, including Old Chinese, comprises 5,437 words, which can be clustered into 1,576 classes of cognate words; 980 of these are "singletons," that is, they comprise classes containing only one single element. Due to the large time span between Old Chinese and the most recent common ancestor of all Chinese dialects, not all Old Chinese forms are technically reconstructible from the data, as they reflect words that have been lost in all dialects. As a result, we were left with 144 reconstructible concepts for which at least one dialect retains an ancestral form attested in Old Chinese.

For the IELex data,[3] we used all languages and dialects except those marked as "Legacy" and two creole languages (*Sranan* and *French Creole Dominica*, as lexical change arguably underlies different patterns under creolization than it does in normal language change). This left us with 134 languages and dialects, including 31 ancient languages (*Ancient Greek, Avestan, Classical Armenian, Gaulish, Gothic, Hittite, Latin, Luvian, Lycian, Middle Breton, Middle Cornish, Mycenaean Greek, Old Persian, Old Prussian, Old Church Slavonic, Old Gutnish, Old Norse, Old Swedish, Old High German, Old English, Old Irish, Old Welsh, Old Cornish, Old Breton, Oscan, Palaic, Pali, Tocharian A, Tocharian B, Umbrian, Vedic Sanskrit*). The data contain translations of 208 concepts into those languages and dialects (often including several synonymous expressions for the same concept from the same language). Most entries are assigned a *cognate class label*. We only used entries containing an unambiguous class label, which left us with 26,524 entries from 4,352 cognate classes. IELex also contains 167 reconstructed entries (for 135 concepts) for Proto-Indo-European. These reconstructions were used as gold standard to evaluate the automatically inferred reconstructions.

ABVD contains data from a total of 697 Austronesian languages and dialects. We selected a subset of 349 languages (all taken from the 400-language sample used in Gray et al., 2009), each having a different ISO code which is also covered in the Glottolog database (Hammarström et al., 2015). ABVD covers 210 concepts, with a total of 44,983 entries from 7,727 cognate classes for our 349-language sample. It also contains 170 reconstructions for Proto-Austronesian (each denoting a different concept) including cognate-class assignments. An overview of the data used is given in Table 1.

3 IELex is currently being thoroughly revised as part of the *Cognates in the Basic Lexicon* (COBL) project, but since this data has not yet been publicly released, we were forced to use the IELex data which we retrieved from ielex.mpi.nl.

TABLE 1 *Datasets used for ancestral state reconstruction. "Reconstructible" states in the column showing the number of concepts refer to the amount of concepts in which the proto-form is reflected in at least one of the descendant languages. "Singletons" refer to cognate sets with only one reflex, which are not informative for the purpose of certain methods of ancestral state reconstruction, like the MLN approach, and therefore excluded from the analysis.*

Dataset	Languages	Concepts	Cognate classes	Singletons	Words
IELex	134	207 (135 reconstructible)	4,352	1,434 singletons	26,524
ABVD	349	210 (170 reconstructible)	7,727	2,671 singletons	44,983
BCD	24	200 (144 reconstructible)	1,576	980 singletons	5,437

2.2 *Methods*

2.2.1 Reference phylogenies

All ASR methods in our test (except the baseline) rely on phylogenetic information when inferring ancestral states, albeit to a different degree. Some methods operate on a single tree topology only, while other methods also use branch lengths information or require a sample of trees to take phylogenetic uncertainty into account. To infer those trees, we arranged the cognacy information for each data set into a presence-absence matrix. Such a data structure is a table with languages as rows and cognate classes occurring within the data set as columns. A cell for language *l* and cognate class *cc* for concept *c* has entry

- 1 if *cc* occurs among the expressions for *c* in *l*,
- 0 if the data contain expressions for *c* in *l*, but none of them belongs to *cc*, and
- undefined if *l* does not contain any expressions for *c*.

Bayesian phylogenetic inference was performed on these matrices. For each data set, tree search was constrained by *prior* information derived from the findings of traditional historical linguistics. More specifically, we used the following prior information:

- **IELex.** We used 14 topological constraints (see Fig. 2), age constraints for the 31 ancient languages, and age constraints for 11 of the 14 topological constraints. The age constraints for *Middle Breton, Middle Cornish, Mycenaean Greek, Old Breton, Old Cornish, Old Welsh*, and *Palaic* are based on information from Multitree (The LINGUIST List, 2014, accessed on October 14, 2016).

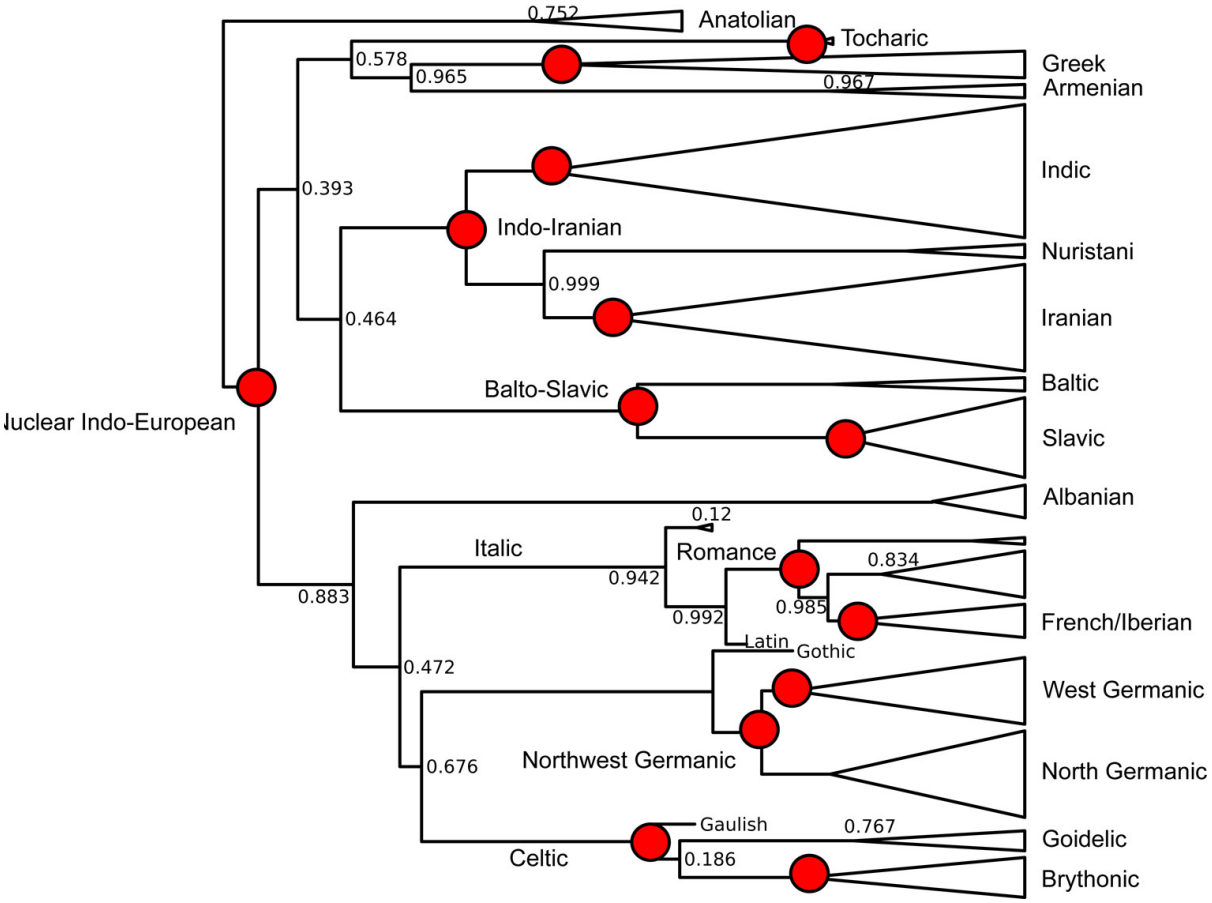


FIGURE 2 *Maximum Clade Credibility tree for IELex (schematic). Topological constraints are indicated by red circles. Numbers at intermediate nodes indicate posterior probabilities (only shown if* < 1*).*

The age constraint for *Pali* is based on information from Encyclopaedia Britannica (2010, accessed on October 14, 2016). The constraints for *Old Gutnish* are taken from Wessen (1968) and those for *Old Swedish* and *Old High German* from Campbell and King (2013). All other age constraints are derived from the Supplementary Information of Bouckaert et al. (2012).

– **ABVD**. We only considered trees consistent with the Glottolog expert classification (Hammarström et al., 2015). This amounts to 213 topological constraints.
– **BDC**. We only considered trees consistent with the expert classification from Sagart (2011). This amounts to 20 topological constraints.

Analyses were carried out using the MrBayes software (Ronquist et al., 2012). Likelihoods were computed using ascertainment bias correction for all-absent characters and assuming Gamma-distributed rates (with 4 Gamma categories). Regarding the tree prior, we assumed a relaxed molecular clock model (more

specifically, the *Independent Gamma Rates* model (cf. Lepage et al., 2007), with an exponential distribution with rate 200 as prior distribution for the variance of rate variation). Furthermore we assumed a birth-death model (Yang and Rannala, 1997) and random sampling of taxa with a sampling probability of 0. 2. For all other parameters of the prior distribution, the defaults offered by the software were used.[4]

For each dataset, a *maximum clade credibility tree* was identified as the **reference tree** (using the software *TreeAnnotator*, retrieved on September 13, 2016; part of the software suite Beast, cf. Bouckaert et al., 2014). Additionally, 100 trees were sampled from the posterior distribution for each dataset and used as **tree sample** for ASR.

2.2.2 Ancestral state reconstruction

For our study, we tested three different established **algorithms**, namely (1) Maximum Parsimony (MP) reconstruction using the Sankoff algorithm (Sankoff, 1975), (2) the minimal lateral network (MLN) approach (Dagan et al., 2008) as a variant of Maximum Parsimony in which parsimony weights are selected with the help of the *vocabulary size criterion* (List et al., 2014b, 2014c), and (3) Maximum Likelihood (ML) reconstruction as implemented in the software *BayesTraits* (Pagel and Meade, 2014). These algorithms are described in detail below.

We tested two different ways to arrange cognacy information as *character matrices*:

- **Multistate characters**. Each concept is treated as a character. The value of a character for a given language is the cognate class label of that language's expression for the corresponding concept. If the data contain several non-cognate synonymous expressions, the language is treated as polymorphic for that character. If the data do not contain an expression for a given concept and a given language, the corresponding character value is undefined.
- **Binary characters**. Each cognate class label that occurs among the documented languages of a dataset is a character. Possible values are 1 (a language contains an expression from that cognate class), 0 (a language does not contain an exponent of that cognate class, but other expressions for the corre-

4 These defaults are: uniform distribution over equilibrium state frequencies; standard exponential distribution as prior for the shape parameter α of the Gamma distribution modeling rate variation; standard exponential distribution as prior over the tree age, measured in expected number of mutations per character.

sponding concept are documented) or undefined (the data do not contain an expression for the concept from the language in question).

All three algorithms rely on a reference phylogeny to infer ancestral states. To test the impact of **phylogenetic uncertainty**, we performed ASR both on the *reference tree* and on the *tree sample* for all three algorithms. The procedures are now presented for each algorithm in turn.

Maximum Parsimony (*MP*). A *complete scenario* for a character is a phylogenetic tree where all nodes are labeled with some character value. For illustration, three scenarios are shown in Fig. 3. The *parsimony score* of a scenario is the number of mutations, i.e., of branches where the mother node and the daughter node carry different labels. Now suppose only the labels at the leaves of the tree are given. The parsimony score of such a *partial scenario* is the minimal parsimony score of any complete scenario consistent with the given leaf labels. In the example in Fig. 3, this value would be 2. The ASR for the root of the tree would be the root label of the complete scenario giving rise to this minimal parsimony score. If several complete scenarios with different root labels give rise to the same minimal score, all their root labels are possible ASRs. This logic can be generalized to *weighted parsimony*. In this framework, each mutation from a state at the mother node to the state at the daughter node of a tree has a certain *penalty*, and these penalties may differ for different types of mutations. The overall parsimony score of a complete scenario is the sum of all penalties for all mutations in this scenario.[5]

5 There is a variant of MP called *Dollo parsimony* (Le Quesne, 1974; Farris, 1977) which is *prima facie* well-suited for modeling cognate class evolution. Dollo parsimony rests on the assumption that complex characters evolve only once, while they may be lost multiple times. If "1" represents presence and "0" absence of such a complex character, the weight of a mutation $1 \rightarrow 0$ should be infinitesimally small in comparison to the weight of $0 \rightarrow 1$. Performing ASR under this assumption amounts to projecting each character back to the latest common ancestor of all its documented occurrences. While this seems initially plausible since each cognate class can, by definition, emerge only once, recent empirical studies have uncovered that multiple mutations $0 \rightarrow 1$ can easily occur with cognate-class characters. A typical scenario is parallel semantic shifts. Chang et al. (2015), among others, point out that descendent words of Proto-Indo-European **pod-* 'foot' independently shifted their meaning to 'leg' both in Modern Greek and in Modern Indic and Iranian languages. So the Modern Greek *πόδι* and the Marathi *pāy*, both meaning 'leg,' are cognate according to IELex, but the latest common ancestor language of Greek and Marathi (Nuclear Proto-Indo-European or a close descendant of it) probably used a non-cognate word to express 'leg.' Other scenarios leading to the parallel emergence of cognate classes are loans and *incomplete lineage sorting*; see the discus-

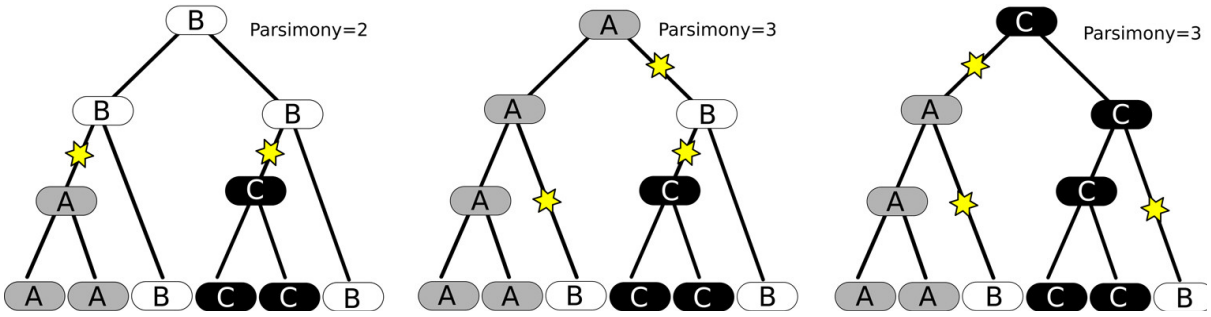


FIGURE 3 *Complete character scenarios. Mutations are indicated by yellow stars.*

The *Sankoff algorithm* is an efficient method to compute the parsimony score and the root ASR for a partial scenario. It works as follows. Let *states* be the ordered set of possible states of the character in question, and let n be the cardinality of this set. For each pair of states i, j, $w(i,j)$ is the penalty for a mutation from $states_i$ to $states_j$.

- **Initialization.** Each leaf l of the tree is initialized with a vector $\mathrm{wp}(l)$ of length n, with $\mathrm{wp}(l)_i = 0$ if l's label is $states_i$, and ∞ else. (If l is polymorphic, all labels occuring at l have the score 0.)
- **Recursion.** Loop through the non-leaf nodes of the tree bottom-up, i.e., visit all daughter nodes before you visit the mother node. Each non-terminal node *mother* with the set *daughters* as daughter nodes is annotated with a vector $\mathrm{wp}(mother)$ according to the rule

$$\mathrm{wp}(mother)_i = \sum_{d \in daughters} \min_{1 \le j \le n} (w(i,j) + \mathrm{wp}(d)_j) \qquad (1)$$

- **Termination.** The parsimony score is $\min_{1 \le i \le n} \mathrm{wp}(root)_i$ and the root ASR is $\arg\min_{1 \le i \le n} \mathrm{wp}(root)_i$.

If MP-ASR is performed on a sample of trees, the Sankoff algorithm is applied to each tree in the sample, and the vectors at the roots are summed up. The root ASR is then the state with the minimal total score. For our experiments, we used the following **weight matrices**:

- For multistate characters, we used uniform weights, i.e., $w(i,i) = 0$ and $w(i,j) = 1$ iff $i \neq j$.

sion in Section 4. Bouckaert et al. (2012) test a probabilistic version of the Dollo approach and conclude that a time-reversible model provides a better fit of cognate-class character data.

– For binary presence-absence characters, we assumed that the penalty of a gain is twice as high as the penalty for a loss: $w(i, i) = 0$, $w(1, 0) = 1$, and $w(0, 1) = 2$.[6]

For a given tree and a given character, the Sankoff algorithm produces a parsimony score for each character state. If the cognacy data are organized as multi-state characters, each state is a cognate class. The *reconstructed states* are those achieving the minimal value among these scores. If a tree sample, rather than a single tree, is considered, the parsimony scores are averaged over the results for all trees in the sample. The reconstructed states are those achieving the minimal average score. If the cognacy data are organized as presence-absence characters, we consider the parsimony scores of state "1" for all cognate classes expressing a certain concept. The reconstructed cognate classes are those achieving the minimal score for state "1." If a tree sample is considered, scores are averaged over trees.

Minimal Lateral Networks (*MLN*). The MLN approach was originally developed for the detection of lateral gene transfer events in evolutionary biology (Dagan et al., 2008). In this form, it was also applied to linguistic data (Nelson-Sathi et al., 2011), and later substantially modified (List et al., 2014b, 2014c). While the original approach was based on very simple gain-loss-mapping techniques, the improved version uses weighted parsimony on presence-absence data of cognate set distributions. In each analysis, several parameters (ratio of weights for gains and losses) are tested, and the best method is then selected, using the criterion of *vocabulary size distributions*, which essentially states that the amount of synonyms per concept in the descendant languages should not differ much from the amount of synonyms reconstructed for ancestral languages. Thus, of several competing scenarios for the development of characters along the reference phylogeny, the scenario that comes closest to the distribution of words in the descendant languages is selected. This is illustrated in Fig. 4. Note that this criterion may make sense intuitively, if one considers that a language with excessive synonymy would make it more difficult for the speakers to communicate. Empirically, however, no accounts on average synonym frequencies

6 The ratio between gains and losses follows from the experience with the MLN approach, which is presented in more detail below and which essentially tests different gain-loss scenarios for their suitability to explain a given dataset. In all published studies in which the MLN approach was tested (List et al., 2014b, 2014c; List, 2015), the best gain-loss ratio reported was 2:1.

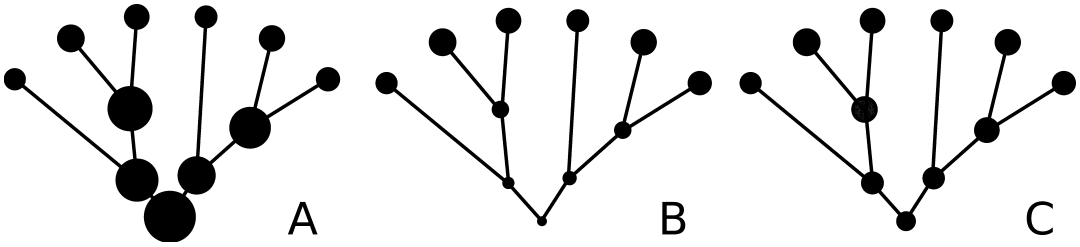


FIGURE 4 *Vocabulary size distributions as a criterion for parameter selection in the* MLN *approach.* A *shows an analysis which proposes far too many words in the ancestral languages,* B *proposes far to few words, and* C *reflects an optimal scenario.*

across languages are available, and as a result, this assumption remains to be proven in future studies.

While the improved versions were primarily used to infer borrowing events in linguistic datasets, List (2015) showed that the MLN approach can also be used for the purpose of ancestral state reconstruction, given that it is based on a variant of weighted parsimony. Describing the method in all its detail would go beyond the scope of this paper. For this reason, we refer the reader to the original publications introducing and explaining the algorithm, as well as the actual source code published along with the LingPy software package (List and Forkel, 2016). To contrast MLN with the variant of Sankoff parsimony we used, it is, however, important to note that the MLN method does not handle *singletons* in the data, that is, words which are not cognate with any other words.[7] It should also be kept in mind that the MLN method in its currently available implementation only allows for the use of *binary characters states*: multi-state characters are not supported and can therefore not be included in our test.

Maximum Likelihood (*ML*). While the Maximum Parsimony principle is conceptually simple and appealing, it has several shortcomings. As it only uses topological information and disregards branch lengths, it equally penalizes mutations on short and on long branches. However, mutations on long branches are intuitively more likely than those on short branches if we assume that branch length corresponds to historical time. Also, MP entirely disregards the possibility of multiple mutations on a single branch. It would go beyond the scope of this article to fully spell out the ML method in detail; the interested reader is

7 The technical question of parsimony implementations is here whether one should penalize the origin of a character in the root or not. The parsimony employed by MLN penalizes all origins. As a result, words that are not cognate with any other word can never be reconstructed to a node higher in the tree. For a discussion of the advantages and disadvantages of this treatment, see Mirkin et al. (2003).

referred to the standard literature on phylogenetic inference (such as Ewans and Grant, 2005, Section 15.7) for details. In the following we will confine ourselves to presenting the basic ideas.

The fundamental assumption underlying ML is that character evolution is a *Markov process*. This means that mutations are non-deterministic, stochastic events, and their probability of occurrence only depends on the current state of the language. For simplicity's sake, let us consider only the case where there are two possible character states, 1 (for presence of a trait) and 0 (absence). Then there is a probability p_{01} that a language gains the trait within one unit of time, and p_{10} that it loses it.

The probability that a language switches from state i to state j within a time interval t is then given by the *transition probability* $P(t)_{ij}$:[8]

$$\alpha = \frac{p_{01}}{p_{01} + p_{10}} \tag{2}$$

$$\beta = \frac{p_{10}}{p_{01} + p_{10}} \tag{3}$$

$$\lambda = -\log(1 - p_{01} - p_{10}) \tag{4}$$

$$P(t) = \begin{pmatrix} \beta + \alpha \cdot (-\lambda t) & \alpha - \alpha \cdot (-\lambda t) \\ \beta - \beta \cdot (-\lambda t) & \alpha + \beta \cdot (-\lambda t) \end{pmatrix} \tag{5}$$

α and β are the *equilibrium probabilities* of states 1 and 0 respectively, and λ is the *mutation rate*. If t is large in comparison to the minimal time step (such as the time span of a single generation), we can consider t to be a continuous variable and the entire process a *continuous time Markov process*. This is illustrated in Fig. 5 for $\alpha = 0.2$, $\beta = 0.8$, and $\lambda = 1$.

If a language is in state 0 at time 0, its probability to be in state 1 after time t is indicated by the solid line. This probability continuously increases and converges to α. This is the gross probability to start in state 0 and end in state 1; it includes the possibility of multiple mutations, as long as the number of mutations is odd. The dotted line shows the probability of ending up in state 1 after time t when a language starts in state 1. This quantity is initally close to 100%, but it also converges towards α over time. In other words, the absence of mutations (or a sequence of mutations that re-established the initial state) is predicted to be unlikely over long periods of time. In a complete scenario, i.e., a phylogenetic tree with labeled non-terminal nodes, the likelihood of a branch

8 We assume that the rows and columns of $P(t)$ are indexed with 0, 1.

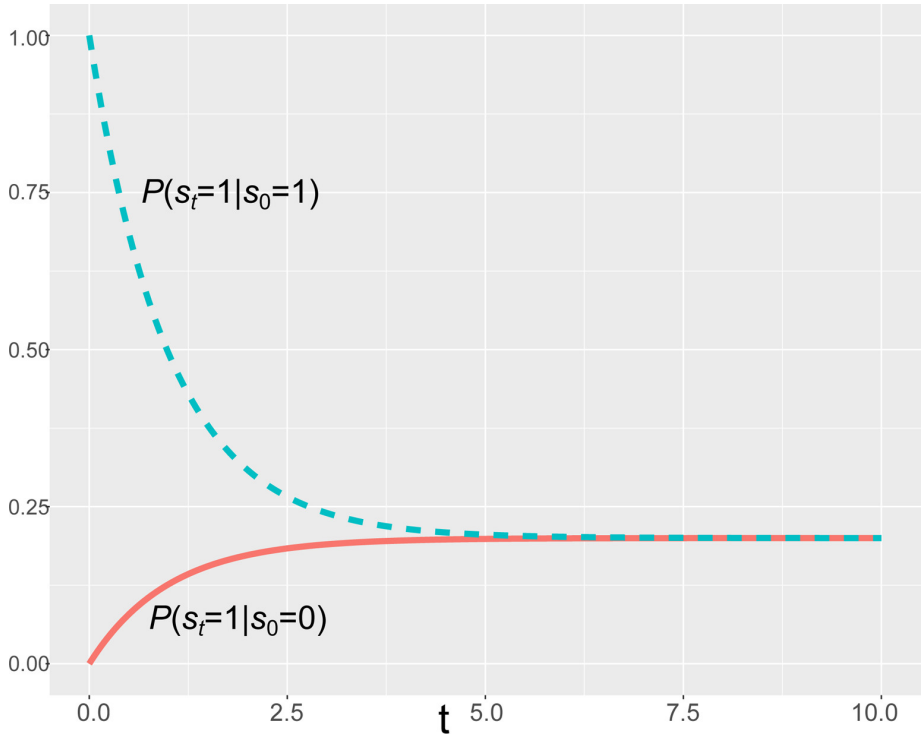


FIGURE 5 *Gain and loss probabilities under a continuous-time Markov process*

is the probability of ending in the state of the daughter node if one starts in the state of the mother node after a time interval given by the branch length.

The overall likelihood of a complete scenario is the product of all branch likelihoods, multiplied with the equilibrium probability of its root state. The likelihood of a partial scenario, where only the states of the leaves are known, is the sum of the likelihoods of all complete scenarios consistent with it. It can efficiently be computed in a way akin to the Sankoff algorithm. ($\mathscr{L}(x)$ is the likelihood vector of node x, and π_i is the equilibrium probability of state i.)

- **Initialization**. Each leaf l of the tree is initialized with a vector $\mathscr{L}(l)$ of length n, with $\mathscr{L}(l)_i = 1$ if l's label is $states_i$, and 0 else. (If l is polymorphic, all labels occuring at t have the same likelihood, and these likelihoods sum up to 1.)
- **Recursion.** Loop through the non-leaf nodes of the tree bottom-up, i.e., visit all daughter nodes before you visit the mother node. Each non-terminal node *mother* with the set *daughters* as daughter nodes is annotated with a vector $\mathscr{L}(mother)$ according to the rule

$$\mathscr{L}(mother)_i = \prod_{d \in daughters} \sum_{1 \le j \le n} (P(t)_{i,j} \mathscr{L}(d)_j), \tag{6}$$

  where t is the length of the branch connecting d to its mother node.

– **Termination**. The likelihood of the scenario is $\sum_{1 \leq i \leq n} \mathscr{L}(root)_i$. The ASR likelihood of state i is proportional to $\pi_i \mathscr{L}(root)_i$.[9]

The likelihood of the scenario calculated this way is the sum of the likelihoods of all scenarios compatible with the information at the leaves. The overall likelihood of a tree for a character matrix is the product of the likelihoods for the individual characters. (This captures the simplifying assumption that characters are mutually stochastically independent.)

As the model parameters (λ and the equilibrium probabilities) are not known *a priori*, they are estimated from the data. This is done by choosing values that maximize the overall likelihood of the tree for the given character matrix, within certain constraints. In our experiments we used the following constraints:

– For multistate characters, we assumed a uniform equilibrium distribution for all characters, and identical rates for all character transitions.
– For binary characters, we assumed equilibrium probabilities to be identical for all characters. Those equilibrium probabilities were estimated from the data as the empirical frequencies. We assumed *gamma-distributed rates*, i.e., rates were allowed to vary to a certain degree between characters.

Once the model parameters are fixed, the algorithm produces a probability distribution over possible states for each character. The *reconstructed states* are identified in a similar way as for Sankoff parsimony. First these probabilities are averaged over all trees if more than one tree is considered. For multistate characters, the state(s) achieving the highest probability are selected. For binary presence-absence characters, those cognate classes for a given concept are selected that achieve the highest average probability for state 1.

2.3 *Evaluation*

For all three datasets considered, the gold standard contains cognate class assignments for a common ancestor language. For the Chinese data, these are documented data for Old Chinese. For the other two datasets, these are reconstructed forms of the supposed latest common ancestor (LCA), Proto-Indo-European and Proto-Austronesian respectively. The Old Chinese variety

9 Note that this approach can only be used to compute the *marginal likelihood* of states at the *root of the tree*. To perform ASR at interior nodes or joint ASR at several nodes simultaneously, a more complex approach is needed. These issues go beyond the scope of this article.

is not identical with the latest common ancestor of all Chinese dialects, but predates it by several hundred years. Due to the rather stable character of the written languages as opposed to the vernaculars throughout the history of Chinese, it is difficult to assess with certainty which exact words were used to denote certain basic concepts, and Old Chinese as reflected in classical sources is a compromise solution as it allows us to consider written evidence rather than reconstructed forms (see Section 4 for a more detailed discussion).

For the evaluation, we only consider those concepts for which (a) the LCA data identify a cognate class and (b) this cognate class is also present in one or more of the descendant languages considered in the experiment. The gold standard defines a set of cognate classes that were present in the LCA language. Let us call this set *LCA*. Each ASR algorithm considered defines a set of cognate classes that are reconstructed for the LCA. We denote this set as *ASR*. In the following we will deploy evaluation metrics established in machine learning to assess how well these two sets coincide:

$$precision \doteq \frac{|LCA \cap ASR|}{|ASR|} \tag{7}$$

$$recall \doteq \frac{|LCA \cap ASR|}{|LCA|} \tag{8}$$

$$F\text{-}score \doteq 2 \times \frac{precision \times recall}{precision + recall} \tag{9}$$

The *precision* expresses the proportion of correct reconstructions among all reconstructions. The *recall* gives the proportion of ancestral cognate classes that are correctly reconstructed. The *F-score* is the harmonic mean between precision and recall.

Results for the various ASR algorithms are compared against a *frequency baseline*. According to the baseline, a cognate class *cc* for a given concept *c* is reconstructed if and only if *cc* occurs at least as frequently among the languages considered (excluding the LCA language) as any other cognate class for *c*. This baseline comes very close to the current practice in classical historical linguistics, as presented in Starostin (2016), although it is clear that trained linguists practicing onomasiological reconstruction may take many additional factors into account. For IELex, we also considered a second baseline, dubbed the *sub-family baseline*. A cognate class *cc* is deemed reconstructed if and only if it occurs in at least two different sub-families, where sub-families are *Alba-*

nian, Anatolian, Armenian, Balto-Slavic, Celtic, Germanic, Greek, Indo-Iranian, Italic, and *Tocharian*.

3 Results

The individual results for all datasets and algorithm variants are given in Tables 2, 3 and 4. Note that MLN does not offer a multi-state variant, so for MLN, only results for binary states are reported. The effects of the various design choices—coding characters as multi-state or binary; using a single reference tree or a sample of trees—as well as the differences between the three ASR algorithms considered here are summarized in Fig. 6. The bars represent the average difference in F-score to the frequency baseline, averaged over all instances of the corresponding category across datasets.

It is evident that there are major differences in the performance of the three algorithms considered. While the F-score for MLN-ASR remains, on average, below the baseline, Sankoff-ASR and ML-ASR clearly outperform the baseline. Furthermore, ML-ASR clearly outperforms Sankoff-ASR. Given that both MLN-ASR and Sankoff-ASR deal with Maximum Parsimony, the rather poor performance of the MLN approach shows that the basic vocabulary size criterion may not be the best criterion for penalty selection in parsimony approaches. It may also be related to further individual choices introduced in the MLN algorithm or our version of Sankoff parsimony. Given that the MLN approach was not primarily created for the purpose of ancestral state reconstruction, our findings do not necessarily invalidate the approach per se, yet they show that it might be worthwhile to further improve on its application to ancestral state reconstruction.

The impact of the other choices is less pronounced. Binary character coding provides slightly better results on average than multistate character coding, but the effect is minor. Likewise, capturing information about phylogenetic uncertainty by using a sample of trees leads, on average, to a slight increase in F-scores, but this effect is rather small as well.

To understand why ML is superior to the two parsimony-based algorithms tested here, it is important to consider the conceptual differences between parsimony-based and likelihood-based ASR. Parsimony-based approaches operate on the tree topology only, disregarding branch lengths. Furthermore, the numerical parameters being used, i.e. the mutation penalties, are fixed by the researcher based on intuition and heuristics. ML, in contrast, uses branch length information, and it is based on an explicit probabilistic model of character evolution.

TABLE 2 *Evaluation results for Chinese*

Algorithm	Characters	Tree	Precision	Recall	F-score
frequency baseline	multi	–	0.599	0.590	0.594
MLN	bin	single	0.568	0.729	0.638
MLN	bin	sample	0.568	0.729	0.638
Sankoff	multi	single	0.484	0.743	0.586
Sankoff	multi	sample	0.510	0.722	0.598
Sankoff	bin	single	0.596	0.688	0.639
Sankoff	bin	sample	0.651	0.660	0.655
ML	multi	single	0.669	0.660	**0.664**
ML	multi	sample	0.669	0.660	**0.664**
ML	bin	single	0.634	0.625	0.629
ML	bin	sample	0.641	0.632	0.636

TABLE 3 *Evaluation results for IELex*

Algorithm	Characters	Tree	Precision	Recall	F-score
frequency baseline	multi	–	0.607	0.497	0.547
sub-family baseline	bin	–	0.402	0.885	0.553
MLN	bin	single	0.781	0.303	0.437
MLN	bin	sample	0.781	0.303	0.437
Sankoff	multi	single	0.367	0.739	0.491
Sankoff	multi	sample	0.566	0.594	0.580
Sankoff	bin	single	0.542	0.630	0.583
Sankoff	bin	sample	0.597	0.503	0.546
ML	multi	single	0.741	0.606	0.667
ML	multi	sample	0.763	0.624	0.687
ML	bin	single	0.778	0.636	0.700
ML	bin	sample	0.785	0.642	**0.707**

This point is illustrated in Fig. 7, which schematically displays ASR for the concept *eat* for the Chinese dialect data. The left panel visualizes Sankoff ASR and the right panel shows Maximum-Likelihood ASR. The guide tree identifies two sub-clades, shown as the upper and lower daughter of the root node. The dialects in the upper part of the tree represent the large group of North-

TABLE 4 *Evaluation results for ABVD*

Algorithm	Characters	Tree	Precision	Recall	F-score
frequency baseline	multi	–	0.618	0.618	0.618
MLN	bin	single	0.843	0.412	0.553
MLN	bin	sample	0.882	0.394	0.545
Sankoff	multi	single	0.688	0.849	0.760
Sankoff	multi	sample	0.726	0.816	0.768
Sankoff	bin	single	0.723	0.771	0.746
Sankoff	bin	sample	0.757	0.749	0.753
ML	multi	single	0.788	0.788	**0.788**
ML	multi	sample	0.788	0.788	**0.788**
ML	bin	single	0.776	0.776	0.776
ML	bin	sample	0.771	0.771	0.771

ern and Central dialects, including the dialect of Beijing, which comes close to standard Mandarin Chinese. The dialects in the lower part of the tree represent the diverse Southern group, including the archaic Mǐn 闽 dialects spoken at the South-Eastern coast as well as Hakka and Yuè 粵 (also referred to as Cantonese), the prevalent variety spoken in Hong Kong. All Southern dialects use the same cognate class (*eat.Shi.1327*, Mandarin Chinese *shí* 食, nowadays only reflected in compounds) and all Northern and Central dialects use a different cognate class (*eat.Chi.243*, Mandarin Chinese *chī* 吃, regular word for 'eat' in most Northern varieties). Not surprisingly, both algorithms reconstruct *eat.Shi.1327* for the ancestor of the Southern dialects and *eat.Chi.243* for the ancestor of the Northern and Central dialects. Sankoff ASR only uses the tree topology to reconstruct the root state. Since the situation is entirely symmetric regarding the two daughters of the root, the two cognate classes are tied with exactly the same parsimony score at the root. Maximum-Likelihood ASR, on the other hand, takes branch lengths into account. Since the latest common ancestor of the Southern dialects is closer to the root than the latest common ancestor of the Northern and Central dialects, the likelihood of a mutation along the lower branch descending from the root is smaller than along the upper branch. Therefore the lower branch has more weight when assigning probabilities to the root state. Consequently, *eat.Shi.1327* comes out as the most likely state at the root—which is in accordance with the gold standard. Our findings indicate that the more fine-grained, parameter-rich Maximum-Likelihood approach is generally superior to the simpler parsimony-based approaches.

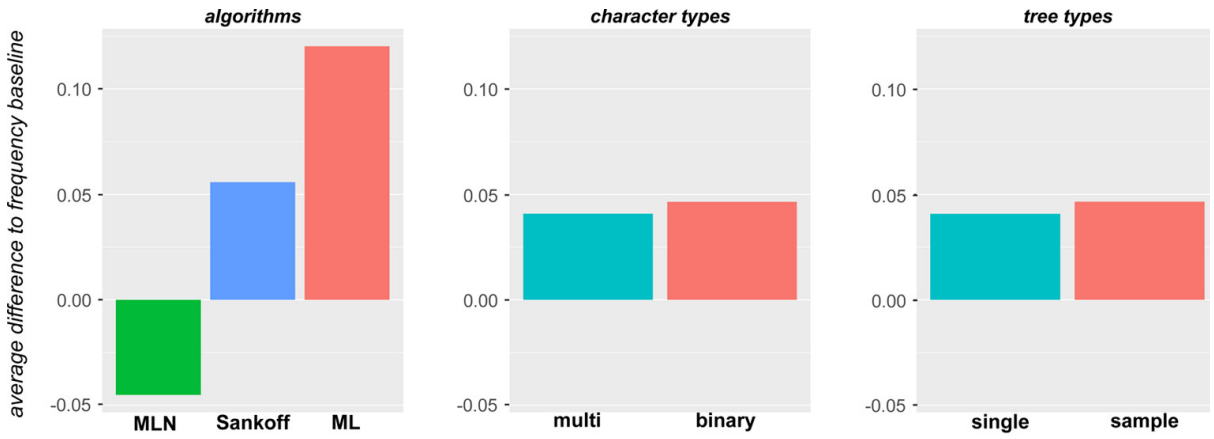


FIGURE 6 *Average differences in F-score to frequency baseline*

The parameters of the Maximum-Likelihood model, as well as the branch lengths, are estimated from the data. Our findings underscore the advantages of an empirical, stochastic and data-driven approach to quantitative historical linguistics as compared to more heuristic methods with few parameters.

4 Linguistic evaluation of the results

The evaluation of the results against a gold standard can help us to understand the general performance of a given algorithm. Only a careful linguistic evaluation, however, helps us to understand the specific difficulties and obstacles that the algorithms have to face when being used to analyze linguistic data. We therefore carried out detailed linguistic evaluations of the results proposed for IELex and BCD: we compared the results of the best methods for each of the datasets (Binary ML Sample for IELex, and Multi ML for BCD) with the respective gold standards, searching for potential reasons for the differences between automatic method and gold standard. The results are provided in Appendix B. In each of the two evaluations, we compared those forms which were reconstructed back to the root in the gold standard but missed by the algorithm, and those forms proposed by the algorithm but not by the gold standard. By consulting additional literature and databases, we could first determine whether the error was due to the algorithm or due to a problem in the gold standard. In a next step, we tried to identify the most common sources of errors, which we assigned to different error classes. Due to the differences in the histories and the time depths, the error classes we identified differ slightly, and while a rather common error in IELex consisted in erroneous cognate judgments in the gold standard,[10] we find many problematic meanings that are rarely expressed

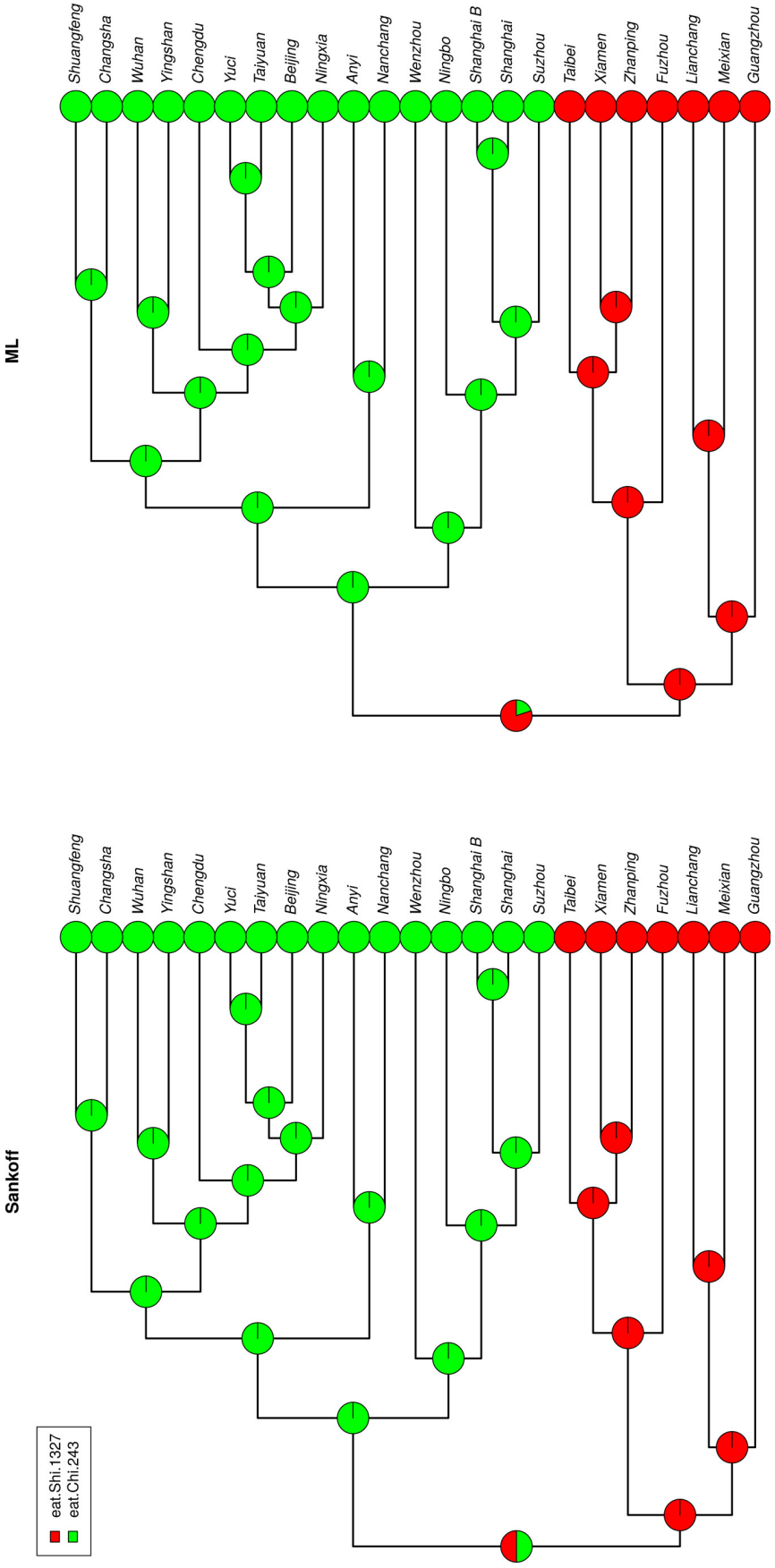


FIGURE 7 *Maximum-Likelihood* ASR *and Sankoff Parsimony* ASR *for the concept* eat *for Chinese dialect data*

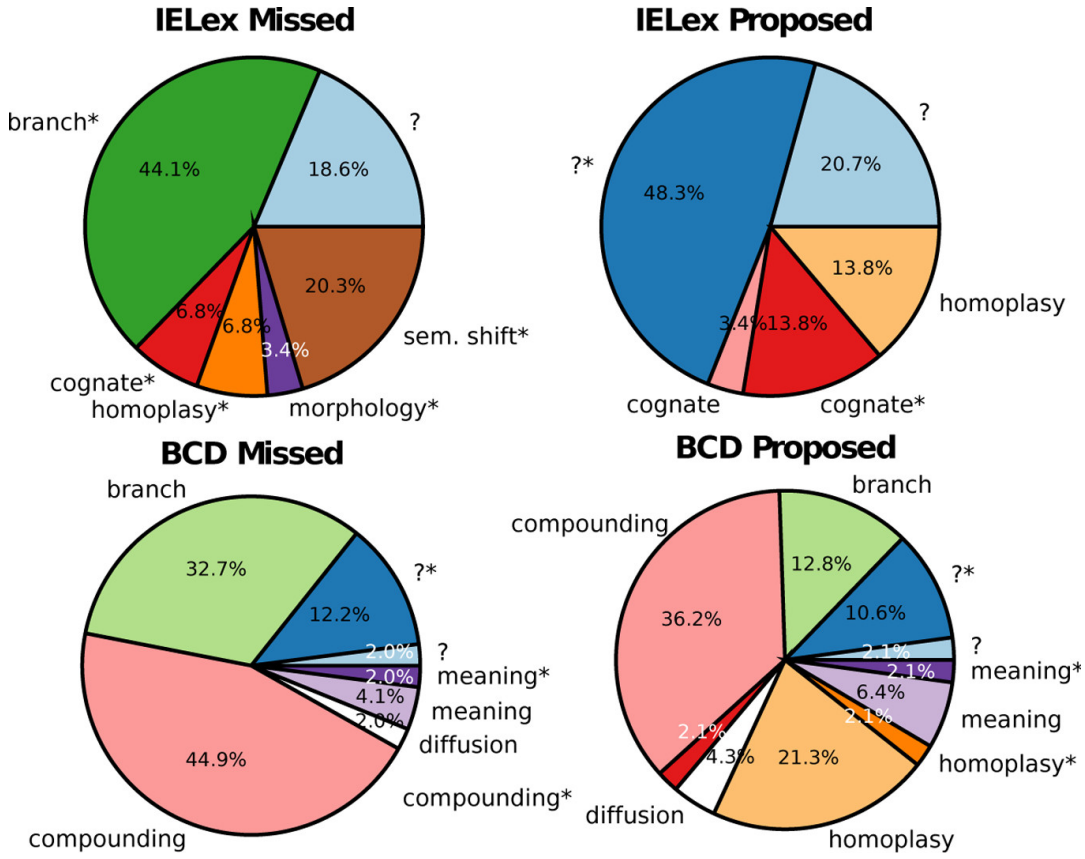


FIGURE 8 *Detailed error analysis of the algorithmic performance on IELex and BCD. If a certain error class is followed by an asterisk, this means that we attribute the error to the gold standard rather than to the algorithm. For a detailed discussion of the different error classes mentioned in this context, please see the detailed analysis in the supplementary material.*

overtly in Chinese dialects in BCD.[11] Apart from errors which were hard to classify and thus not assigned to any error class, problems resulting from the misinterpretation of branch-specific cognate sets as well as problems resulting from parallel semantic shift (homoplasy) were among the most frequent problems in both datasets.

Figure 8 gives detailed charts of the error analyses for missed and erroneously proposed items in the two datasets. The data is listed in such a way that mismatches between gold standard and algorithms can be distinguished. When inspecting the findings for IELex, we can thus see that the majority of the 59 cognates missed by the algorithm can be attributed to cognate sets that are only reflected in one branch in the Indo-European languages and therefore do

11 Examples include meanings for 'if,' 'because,' etc., which may be expressed but may as well be omitted in normal speech, see Appendix B2 for details.

not qualify as good candidates to be reconstructed back to the proto-language. As an example, consider the form *pneu̯- (cognate class breathe : P), which is listed as onomasiological reconstruction for the concept 'to breathe' in the gold standard. As it only occurs in Ancient Greek and has no reflexes in any other language family, this root is highly problematic, as is also confirmed by the *Lexicon of Indo-European Verbs*, where the root is flagged as questionable (Rix et al., 2001: 489). Second, the error statistics for Indo-European contain cognate sets whose onomasiological reconstruction is not confirmed by plausible semantic reconstructions in the gold standard. As an example for this error class, consider the form *dhōĝh-e/os- (cognate class day : B) proposed for the meaning slot 'day.' While Kroonen (2013: 86 f.) confirms the reconstruction of the root, as it occurs in Proto-Germanic and Indo-Iranian, the meaning 'day' is by no means clear, as the PIE root *di̯eu̯- 'heavenly deity, day' is a more broadly reflected candidate for the 'day' in PIE (Meier-Brügger, 2002: 187 f.).

Of the 29 cognates missed, the majority cannot be readily classified, as these comprise cases where a reconstruction back to the proto-language *in* the given meaning slot seems to be highly plausible. Thus, the form *kr̥-m-i- (cognate class worm : A) is not listed in the gold standard, but proposed by the Binary ML approach. The root is reflected in both Indo-Iranian and in Slavic (Derksen, 2008: 93 f.) and generally considered a valid Indo-European root with the meaning 'worm, insect' (Mallory and Adams, 2006: 149). Given that 'worm' and 'insect' are frequently expressed by one polysemous concept in the languages of the world (see the CLICS database of cross-linguistic polysemies, List et al., 2014a), we see no reason why the form is not listed in the gold standard. Second in frequency of the items proposed by the algorithm are cases of clear homoplasy that were interpreted as inheritance by the ML approach. As an example, consider the form *serp- (cognate class snake : E), which the algorithm proposes as a candidate for the meaning 'snake.' While the cognate set contains the Latin word *serpens*, as well as reflexes in Indo-Iranian and Albanian, it may seem like a good candidate. According to Vaan (2008: 558), however, the verbal root originally meant 'to crawl,' which would motivate the parallel denotation in Latin and Albanian. Instead of assuming that the noun already denoted 'snake' in PIE times, it is therefore much more likely that we are dealing with independent semantic shift.

Turning to our linguistic evaluation of the results on the Chinese data, we also find branch-specific words as one of the major reasons for the 49 forms which were proposed in the gold standard but not recognized by the best algorithm (Multi ML). However, here we cannot attribute these to questionable decisions in the gold standard, but rather to the fact that many Old Chinese words are often reflected only in some of the varieties in the sample. As an

example for a challenging case, consider the form 口 *kǒu* 'mouth' (cognate class `mouth-Kou-222`, #31). The regular word for 'mouth' in most dialects today is 嘴 *zuǐ*, but the Mǐn dialects, the most archaic group and the first to branch off the Sinitic family, have 喙 *huì* as an innovation, which originally meant 'beak, snout.' While *kǒu* survives in many dialects and also in Mandarin Chinese in restricted usage (compare 住口 *zhùkǒu* 'close' + 'mouth' = 'shut up') or as part of compounds (口水 *kǒushuǐ* 'mouth' + 'water' = 'saliva'), it is only in the Yuè dialect Guǎngzhōu that it appears with the original meaning in the BCD. Whether *kǒu*, however, is a true retention in Guǎngzhōu is quite difficult to say, and comparing the data in the BCD with the more recent dataset by Liú et al. (2007), we can see that *zuǐ*, in the latter, is given for Guǎngzhōu instead of *kǒu*. The differences in the data are difficult to explain, and we see two possible ways to account for them: (1) If *kǒu* was the regular term for 'mouth' in Guǎngzhōu in the data by Wang (2004), and if this term is not attested in any other dialect, we are dealing with a *retention* in the Yuè dialects, and with a later diffusion of the term *zuǐ* across many other dialect areas apart from the Mǐn dialects, which all shifted the meaning of *huì*. (2) If *kǒu* is just a variant in Guǎngzhōu as it is in Mandarin Chinese, we are dealing with a methodological problem of *basic word translation* and should assume that *kǒu* is completely lost in its original meaning. In both cases, however, the history of 'mouth' is a typical case of *inherited variation* in language history. Multiple terms with similar reference potential were already present in the last common ancestor of the Chinese dialects. They were later individually resolved, yielding patterns that remind of *incomplete lineage sorting* in evolutionary biology (see List et al., 2016 for a closer discussion of this analogy).

The problem of inherited variation becomes even more evident when we consider the largest class of errors in both the items missed and the items proposed by the algorithm: the class of errors due to *compounding*. Compounding is a very productive morphological process in the Chinese dialects, heavily favored by the shift from a predominantly monosyllabic to a bisyllabic word structure in the history of Chinese (see Sampson, 2015 and replies to the article in the same volume for a more thorough discussion on potential reasons for this development). This development led to a drastic increase of bisyllabic words, which is reflected in almost all dialects, affecting all parts of the lexicon. Thus, while the regular words for 'sun' and 'moon' in Ancient Chinese texts were 日 *rì* and 月 *yuè*, the majority of dialects nowadays uses 日頭 *rìtóu* (lit. 'sun-head') and 月光 *yuèguāng* (lit. 'moon-shine'). These words have developed further in some dialect areas and yield a complex picture of patterns of lexical expression that are extremely difficult to resolve historically. Given that we find the words even in the most archaic dialects, but *not* in ancient texts

of the late Hàn time and later (around 200 and 300 CE), the time when the supposed LCA of the majority of the Chinese dialects was spoken, it is quite difficult to explain the data in a straightforward way. We could either propose that the LCA of Chinese dialects already had created or was in the stage of creating these ancient compound words, and that written evidence was too conservative to reflect it; or we could propose that the words were created later and then diffused across the Chinese dialects. Both explanations seem plausible, as we know that spoken and written language often differed quite drastically in the history of Chinese. Comparing modern Chinese dialect data, as provided by Liú et al. (2007), with dialect surveys of the late 1950s, as given in Běijīng Dàxué (1964), we can observe how quickly Mandarin Chinese words have been diffusing recently: while we find only *rìtóu*[12] as a form for 'sun' in Guǎngzhōu, Liú et al. only list the Mandarin form 太陽 *tàiyáng*, and Hóu (2004), presenting data collected in the 1990s, lists both variants. We can see from these examples that the complex interaction between morphological processes like compounding and intimate language contact confronts us with challenging problems and may explain why the automatic methods perform worst on Chinese, despite the shallow time depths of the language family.

5 Conclusion

What can we learn from these experiments? One important point is surely the striking superiority of Maximum Likelihood, outperforming both parsimony approaches. Maximum Likelihood is not only more flexible, as parameters are estimated from the data, but in some sense, it is also more realistic, as we have seen in the reconstruction of the scenario for 'eat' (see Fig. 7) in the Chinese dataset, where the branch lengths, which contribute to the results of ML analyses, allow the algorithm to find the right answer. Another important point is the weakness of all automatic approaches and what we can learn from the detailed linguistic evaluation. Here, we can see that further research is needed to address those aspects of lexical change which are poorly handled by the algorithms. These issues include first and foremost the problem of independent semantic shift, but also the effects of morphological change, especially in the Chinese data. List (2016) uses weighted parsimony with polarized (directional) transition penalties for multi-state characters for ancestral state recon-

12 In the Yuè dialects, this form has been reinterpreted as 'hot-head' 熱頭 *rètóu* instead of 'sun-head.'

struction of Chinese nouns and reports an increased performance compared to unweighted parsimony. However, since morphological change and lexical replacement are clearly two distinct processes, we think it is more promising to work on the development of stochastic models, which are capable of handling two or more distinct processes and may estimate transition tendencies from the data. Another major problem that needs to be addressed in future approaches is the impact of language contact on lexical change processes, as well as the possibility of language-internal variation, which may obscur tree-like divergence even if the data evolved in a perfectly tree-like manner. These instances of *incomplete lineage sorting* (List et al., 2016) became quite evident in our qualitative analysis of the Chinese and Indo-European data. Given their pervasiveness, it is likely that they also have a major impact on classical phylogenetic studies, which only try to infer phylogenies from the data. As a last point, we should mention the need for increasing the quality of our test data in historical linguistics. Given the multiple questionable reconstructions we found in the test sets during our qualitative evaluation, we think it might be fruitful, both in classical and computational historical linguistics, to intensify the efforts towards semantic and onomasiological reconstruction.

Supplementary materials

All data used for this study, along with the code that we used and the results we produced, are available at https://dx.doi.org/10.5281/zenodo.1173120.

The appendices contain a list of all age constraints for Indo-European that were used in our phylogenetic reconstruction study (Appendix A) as well as a detailed, qualitative analysis of all differences between the automatic and the gold standard assessments in IElex (Appendix B1) and BCD (Appendix B2). They are available as supplementary materials and can be accessed through the following link: http://doi.org/10.6084/m9.figshare.6580382.v1.

References

Atkinson, Quentin D. and Russell D. Gray. 2006. How old is the Indo-European language family? Illumination or more moths to the flame? In Peter Forster and Colin Renfrew (eds.) *Phylogenetic Methods and the Prehistory of Languages*, 91–109. Cambridge/Oxford/Oakville: McDonald Institute for Archaeological Research.

Ben Hamed, Mahe and Feng Wang. 2006. Stuck in the forest: Trees, networks and Chinese dialects. *Diachronica* 23: 29–60.

Bouchard-Côté, Alexandre, David Hall, Thomas L. Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences of the U.S.A.* 110(11): 4224–4229.

Bouckaert, Remco, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A. Suchard, Andrew Rambaut, and Alexei J. Drummond. 2014. BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Computational Biology* 10(4): e1003,537. 10.1371/journal.pcbi.1003537. Accessible at http://beast2.org (accessed February 4, 2018).

Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* 337(6097): 957–960.

Bowern, Claire and Quentin D. Atkinson. 2012. Computational phylogenetics of the internal structure of Pama-Nyungan. *Language* 88: 817–845.

Bussmann, Hadumod (ed.). 1996. *Routledge Dictionary of Language and Linguistics*. London/New York: Routledge.

Běijīng Dàxué 北京大学. 1964. *Hànyǔ fāngyán cíhuì* 汉语方言词汇 [*Chinese dialect vocabularies*]. Běijīng 北京: Wénzì Gǎigé.

Campbell, George L. and Gareth King. 2013. *Compendium of the World's Languages*, vol. 1. London/New York: Routledge.

Chang, Will, Chundra Cathcart, David Hall, and Andrew Garret. 2015. Ancestry-constrained phylogenetic analysis support the Indo-European steppe hypothesis. *Language* 91(1): 194–244.

Dagan, Tal, Yael Artzy-Randrup, and William Martin. 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proceedings of the National Academy of Sciences of the U.S.A.* 105(29): 10,039–10,044.

Derksen, Rick. 2008. *Etymological Dictionary of the Slavic Inherited Lexicon*. Leiden/Boston: Brill.

Dunn, Michael. 2012. Indo-European lexical cognacy database (IELex). Accessible at http://ielex.mpi.nl/ (accessed September 5, 2016).

Edwards, Anthony W.F. and Luigi Luca Cavalli-Sforza. 1964. Reconstruction of evolutionary trees. In Vernon H. Heywood and John McNeill (eds.) *Phenetic and Phylogenetic Classification*, 67–76. London: Systematics Association Publisher.

Encyclopaedia Britannica, Inc. 2010. *Encyclopaedia Britannica*. Edinburgh: Encyclopaedia Britannica, Inc. https://www.britannica.com.

Ewans, Warren and Gregory Grant. 2005. *Statistical Methods in Bioinformatics: An Introduction*. New York: Springer.

Farris, James S. 1977. Phylogenetic analysis under Dollo's law. *Systematic Biology* 26(1): 77–88.

Fitch, Walter M. 1971. Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology* 20(4): 406–416.

Geisler, Hans and Johann-Mattis List. 2010. Beautiful trees on unstable ground. Notes on the data problem in lexicostatistics. In Heinrich Hettrich (ed.) *Die Ausbreitung des Indogermanischen. Thesen aus Sprachwissenschaft, Archäologie und Genetik* [*The spread of Indo-European. Theses from linguistics, archaeology, and genetics*]. Wiesbaden: Reichert. Downloadable at https://hal.archives-ouvertes.fr/hal-01298493/document (accessed February 4, 2018). Document has been submitted in 2010 and is still waiting for publication.

Gray, Russell D. and Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426(6965): 435–439.

Gray, Russell D., Alexei J. Drummond, and S.J. Greenhill. 2009. Language phylogenies reveal expansion pulses and pauses in pacific settlement. *Science* 323(5913): 479–483.

Greenhill, Simon J., Robert Blust, and Russell D. Gray. 2008. The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics. *Evolutionary Bioinformatics* 4: 271–283. Accessible at http://language.psy.auckland.ac.nz/austronesian/ (accessed February 4, 2018).

Hammarström, Harald, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2015. Glottolog. Accessible at http://glottolog.org (accessed February 4, 2018).

Haynie, Hanna J. and Claire Bowern. 2016. Phylogenetic approach to the evolution of color term systems. *Proceedings of the National Academy of Sciences of the U.S.A.* 113(48): 13,666–13,671.

Holm, Hans J. 2007. The new arboretum of Indo-European "trees". *Journal of Quantitative Linguistics* 14(2–3): 167–214.

Hóu, Jīngyī (ed.). 2004. *Xiàndài Hànyǔ fāngyán yīnkù* 现代汉语方言音库 [*Phonological database of Chinese dialects*]. Shànghǎi 上海: Shànghǎi Jiàoyù 上海教育.

Hruschka, Daniel J., Simon Branford, Eric D. Smith, Jon Wilkins, Andrew Meade, Mark Pagel, and Tanmoy Bhattacharya. 2015. Detecting regular sound changes in linguistics as events of concerted evolution. *Current Biology* 25(1): 1–9.

Huson, Daniel H. 1998. Splitstree: Analyzing and visualizing evolutionary data. *Bioinformatics* 14(1): 68–73.

Jordan, Fiona M., Russell D. Gray, Simon J. Greenhill, and Ruth Mace. 2009. Matrilocal residence is ancestral in Austronesian societies. *Proceedings of the Royal Society B* 276: 1957–1964.

Jäger, Gerhard. 2015. Support for linguistic macrofamilies from weighted alignment. *Proceedings of the National Academy of Sciences of the U.S.A.* 112(41): 12,752–12,757.

Kassian, Alexei, Mikhail Zhivlov, and George S. Starostin. 2015. Proto-Indo-European-Uralic comparison from the probabilistic point of view. *The Journal of Indo-European Studies* 43(3–4): 301–347.

Kitchen, Andrew, Christopher Ehret, Shiferaw Assefa, and Connie J. Mulligan. 2009. Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. *Proceedings of the Royal Society B* 276(1668): 2703–2710.

Kroonen, Guus. 2013. *Etymological Dictionary of Proto-Germanic*. Leiden/Boston: Brill.

Le Quesne, Walter J. 1974. The uniquely evolved character concept and its cladistic application. *Systematic Biology* 23(4): 513–517.

Lee, Sean and Toshikazu Hasegawa. 2013. Evolution of the Ainu language in space and time. *PLoS ONE* 8(4): e62,243.

Lepage, Thomas, David Bryant, Hervé Philippe, and Nicolas Lartillot. 2007. A general comparison of relaxed molecular clock models. *Molecular Biology and Evolution* 24(12): 2669–2680.

List, Johann-Mattis. 2015. Network perspectives on Chinese dialect history. *Bulletin of Chinese Linguistics* 8: 42–67.

List, Johann-Mattis. 2016. Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction. *Journal of Language Evolution* 1(2): 119–136. 10.1093/jole/lzw006.

List, Johann-Mattis and Robert Forkel. 2016. LingPy. A Python library for historical linguistics. https://zenodo.org/badge/latestdoi/5137/lingpy/lingpy. Accessible at http://lingpy.org (accessed February 4, 2018).

List, Johann-Mattis, T. Mayer, A. Terhalle, and M. Urban. 2014a. Clics: Database of Cross-Linguistic Colexifications. Accessible at http://clics.lingpy.org (accessed February 4, 2018).

List, Johann-Mattis, Shijulal Nelson-Sathi, Hans Geisler, and William Martin. 2014b. Networks of lexical borrowing and lateral gene transfer in language and genome evolution. *Bioessays* 36(2): 141–150.

List, Johann-Mattis, Shijulal Nelson-Sathi, William Martin, and Hans Geisler. 2014c. Using phylogenetic networks to model Chinese dialect history. *Language Dynamics and Change* 4(2): 222–252.

List, Johann-Mattis, Jananan Sylvestre Pathmanathan, Philippe Lopez, and Eric Bapteste. 2016. Unity and disunity in evolutionary sciences: Process-based analogies open common research avenues for biology and linguistics. *Biology Direct* 11(39): 1–17.

Liú Lìlǐ 刘俐李, Wáng Hóngzhōng 王洪钟, and Bǎi Yíng 柏莹. 2007. *Xiàndài Hànyǔ fāngyán héxīncí, tèzhēng cíjí* 现代汉语方言核心词·特征词集 [*Collection of basic vocabulary words and characteristic dialect words in modern Chinese dialects*]. Nánjīng 南京: Fènghuáng 凤凰.

Mallory, James P. and Douglas Q. Adams. 2006. *The Oxford Introduction to Proto-Indo-European and the Proto-Indo-European World*. Oxford: Oxford University Press.

Meier-Brügger, Michael. 2002. *Indogermanische Sprachwissenschaft* [*Indo-European linguistics*]. Berlin/New York: de Gruyter, 8th ed.

Mirkin, Boris G., Trevor I. Fenner, Michael Y. Galperin, and Eugene V. Koonin. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evolutionary Biology* 3: 2.

Nelson-Sathi, Shijulal, Johann-Mattis List, Hans Geisler, Heiner Fangerau, Russell D. Gray, William Martin, and Tal Dagan. 2011. Networks uncover hidden lexical borrowing in Indo-European language evolution. *Proceedings of the Royal Society of London B: Biological Sciences* 278(1713): 1794–1803.

Pagel, Mark, Quentin D. Atkinson, Andreea S. Calude, and Andrew Meade. 2013. Ultraconserved words point to deep language ancestry across Eurasia. *Proceedings of the National Academy of Sciences of the U.S.A.* 110(21): 8471–8476.

Pagel, Mark and Andrew Meade. 2014. BayesTraits 2.0. Software distributed by the authors. Downloadable at http://www.evolution.rdg.ac.uk/BayesTraitsV2.html (accessed September 5, 2016)

Rix, Helmut, Martin Kümmel, Thomas Zehnder, Reiner Lipp, and Brigitte Schirmer. 2001. *LIV. Lexikon der Indogermanischen Verben. Die Wurzeln und ihre Primärstammbildungen* [*Lexicon of Indo-European Verbs. The roots and their primary stems*]. Wiesbaden: Reichert.

Ronquist, Fredrik, Maxim Teslenko, Paul van der Mark, Daniel L. Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc A. Suchard, and John P. Huelsenbeck. 2012. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* 61(3): 539–542.

Sagart, Laurent. 2011. Classifying chinese dialects/sinitic languages on shared innovations. Paper presented at the Séminaire Sino-Tibétain du CRLAO (2011-03-28). Downloadable at https://www.academia.edu/19534510 (accessed February 4, 2018).

Saitou, Naruya and Masatoshi Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4(4): 406–425.

Sampson, Geoffrey. 2015. A Chinese phonological enigma. *Journal of Chinese Linguistics* 43(2): 679–691.

Sankoff, David. 1975. Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics* 28(1): 35–42.

Sokal, Robert R. and Charles D. Michener. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin* 28: 1409–1438.

Starostin, George S. 2016. From wordlists to proto-wordlists: Reconstruction as 'optimal selection'. *Faits de langues* 47(1): 177–200. 10.3726/432492177.

Swadesh, Morris. 1952. Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* 96(4): 452–463.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21(2): 121–137.

The LINGUIST List. 2014. Multitree: A digital library of language relationships. Accessible at http://multitree.org (accessed February 4, 2018).

Urban, Matthias. 2011. Asymmetries in overt marking and directionality in semantic change. *Journal of Historical Linguistics* 1(1): 3–47.

Vaan, Michiel. 2008. *Etymological Dictionary of Latin and the Other Italic Languages*. Leiden and Boston: Brill.

Wang, Feng. 2004. BCD: Basic-words of Chinese dialects. Formerly available at http://chinese.pku.edu.cn/wangf/wangf.htm.

Wessen, Elias. 1968. *Die nordischen Sprachen* [*The Nordic languages*]. Berlin: de Gruyter.

Wilkins, David P. 1996. Natural tendencies of semantic change and the search for cognates. In Mark Durie and Malcom Ross (eds.) *The Comparative Method Reviewed. Regularity and Irregularity in Language Change*, 264–304. New York: Oxford University Press.

Yang, Ziheng and Bruce Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov Chain Monte Carlo method. *Molecular Biology and Evolution* 14(7): 717–724.

Zhou, Kevin and Claire Bowern. 2015. Quantifying uncertainty in the phylogenetics of Australian numeral systems. *Proceedings of the Royal Society B* 282(1815): 20151,278.

3 Data Formats and Annotation Frameworks

3.1 Cross-Linguistic Data Formats

In order to make it possible to integrate quantitative and qualitative methods, it is of great importance that data are available in computer- and human-readable form at the same time. While it is often straightforward to make computer-readable data human-readable, the opposite is much harder to achieve, specifically when data have been originally only collected for the purpose of being accessible to humans alone. A classical example for this *lack of formalization* can be found in etymological dictionaries in historical linguistics, which present complex etymological relations between words in form of scientific prose. Although etymological dictionaries often *have* a certain degree of formality and scholars aim at presenting the data in a formal way, the majority of etymological dictionaries is produced with help of word editors alone, and no additional software to store the underlying data in a machine-readable database system are being made. As a result, it may at times be possible to reverse-engineer the intended relations in a given etymological dictionary, and to store them in a database, but in many cases, this turns out to be impossible, also because scholars barely check to which degree they always conform to the style rules they intend to use.

In order to arrive at a level of formalization of data in historical linguistics that allows us to parse them by a computer while at the same time to inspect them ourselves, we need to first establish rigorous standards that allow for a general cross-linguistic comparability of linguistic data. In order to make sure that data are comparable across datasets and resources, general properties of linguistic data, such as common language names, common identifiers for concepts used in the elicitation of wordlists, or common symbols to represent speech sounds, need to be established in form of *reference catalogs*, such as *Glottolog* (Hammarström et al. 2020) as a reference catalog for languages, or *Concepticon* (List et al. 2020) as a reference for concepts (List et al. 2016a).

In the following two studies, first attempts to data standardization are being discussed. The first study, titled "Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics" (Forkel et al. 2018) presents the *Cross-Linguistic Data Formats* initiative (CLDF), which proposes general formats for typical linguistic data types, such as wordlists, structural datasets, dictionaries, or parallel texts. A major strategy of the CLDF formats is to propagate the use of reference catalogs when preparing new datasets. That means that instead of providing only a table with language names, scholars are encouraged to also add the Glottocodes for each language, in order to make sure that confusion, resulting from idiosyncratic or ambiguous language names can be avoided. Similarly, the CLDF initiative recommends to add Concepticon Concept Set identifiers when dealing with wordlist data in addition to elicitation glosses, in order to avoid confusion with respect to the concepts that were elicited in a collection of wordlists.

To further enhance the CLDF data formats, the second study, titled "A cross-linguistic database of phonetic transcription systems" (Anderson et al. 2018), proposes a new reference catalog, called Cross-Linguistic Transcription Systems (CLTS). This reference catalog provides identifiers for speech sounds

and links them to grapheme representations across different transcription systems, such as the International Phonetic Alphabet (IPA 1999) or the North-American Phonetic Alphabet (Pullum and Ladusaw 1996), and additionally also to various transcription datasets, such as Phoible (Moran et al. 2014) or LAPSyD (Maddieson et al. 2013). The data collection is accompanied by a software package that can be used to explore the data, or to convert between the different transcription systems.

Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics

Robert Forkel[1], Johann-Mattis List[1], Simon J. Greenhill[1,2], Christoph Rzymski[1], Sebastian Bank[1], Michael Cysouw[3], Harald Hammarström[1,4], Martin Haspelmath[1,5], Gereon A. Kaiping[6] & Russell D. Gray[1,7]

The amount of available digital data for the languages of the world is constantly increasing. Unfortunately, most of the digital data are provided in a large variety of formats and therefore not amenable for comparison and re-use. The Cross-Linguistic Data Formats initiative proposes new standards for two basic types of data in historical and typological language comparison (word lists, structural datasets) and a framework to incorporate more data types (e.g. parallel texts, and dictionaries). The new specification for cross-linguistic data formats comes along with a software package for validation and manipulation, a basic ontology which links to more general frameworks, and usage examples of best practices.

[1]Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Jena, Germany. [2]ARC Centre of Excellence for the Dynamics of Language, Australian National University, Canberra, Australia. [3]Research Center Deutscher Sprachatlas, Philipps University Marburg, Marburg, Germany. [4]Department of Linguistics and Philology, Uppsala University, Uppsala, Sweden. [5]Department of English Studies, Leipzig University, Leipzig, Germany. [6]Centre for Linguistics, Leiden University, Leiden, Germany. [7]School of Psychology, University of Auckland, Auckland, New Zealand. Correspondence and requests for materials should be addressed to R.F. (email: forkel@shh.mpg.de) or J.-M.L. (email: list@shh.mpg.de)

Introduction

The last two decades have witnessed a dramatic increase in language data, not only in form of monolingual resources[1] for the world's biggest languages, but also in form of *cross-linguistic datasets* which try to cover as many of the world's languages as possible. Creating datasets in linguistics is currently *en vogue*, and apart from traditional ways of linguistic data collection in form of etymological dictionaries, user dictionaries, and grammatical surveys, data are now being published in form of *online databases* (the most complete list of such databases is curated at http://languagegoldmine.com/) and *online appendices or supplements to published papers*, addressing topics as diverse as cross-linguistic lexical associations (cf. http://clics.lingpy.org and http://clics.clld.org), etymologically annotated word lists for large language families like Austronesian (cf. https://abvd.shh.mpg.de[2] and http://www.trussel2.com/acd/) and Indo-European (cf. http://ielex.mpi.nl), inventories of speech sounds (cf. http://phoible.org), or grammatical features compared across a large sample of the world's languages (cf. http://wals.info). Along with the increase in the amount of data, there is also an increased interest in linguistic questions, with scholars from both linguistic and non-linguistic disciplines (e.g. archaeology, anthropology, biology, economics, and psychology) now trying to use linguistic data to answer a wide variety of questions of interest to their disciplines. For example, large-scale cross-linguistic studies have recently been conducted to test how robustly languages are transmitted[3] and which forces drive change[4,5]. Cross-linguistic data have proven useful to detect semantic structures which are universal across human populations[6], and how semantic systems like color terminology have evolved[7,8]. Another group of studies have analysed cross-linguistic data using quantitative phylogenetic methods to investigate when particular language families started to diverge[9–12]. Cross-linguistic studies have even explored proposed non-linguistic factors shaping languages from climate[13,14], to population size[15–17], to genes[18,19], and how these factors may or may not shape human social behavior at a society level[20]. (All URLS mentioned in this paragraph were accessed July 26, 2018).

Despite this gold rush in the creation of linguistic databases and their application reflected in a large number of scholarly publications and an increased interest in the media, linguistic data are still far away from being "FAIR" in the sense of Wilkinson *et al.*[21]: Findable, Accessible, Interoperable, and Reusable. It is still very difficult to *find* particular datasets, since linguistic journals often do not have a policy on supplementary data and may lack resources for hosting data on their servers. It is also often difficult to *access* data, and many papers which are based on original data are still being published without the data[1] and having to request the data from the authors is sometimes a more serious obstacle than it should be[22,23]. Due to idiosyncratic formats, linguistic datasets also often lack *interoperability* and are therefore not *reusable*.

Despite the large diversity of human languages, often linguistic data can be represented by very simple data types which are easy to store and manipulate. Word lists and grammatical surveys, for example, can usually be represented by triples of *language*, *feature*, and *value*. The simplicity, however, is deceptive, as there are too many degrees of freedom which render most of the data that have been produced hard to compare. Due to the apparently simple structure, scholars rarely bother with proper serialization, assuming that their data will be easy to re-use. Although there are recent and long-standing standardization efforts, like the establishment of the *International Phonetic Alphabet* (IPA) as a unified alphabet for phonetic transcription[24], which goes back to the end of the 19th century[25], or the more recent publication of reference catalogues for languages[26] and word meanings[27], linguists often forgo these standards when compiling their datasets and use less strictly specified documentation traditions.

While certain standards, such as the usage of unified transcription systems, are generally agreed upon but often not applied (or mis-applied) in practice, other types of linguistic data come along with a multitude of different standards which make data interoperability extremely difficult (see Fig. 1 for examples on different practices of *cognate coding in wordlists* in historical linguistics).

At the same time, funding agencies such as the *German Academic Research Council* emphasize that 'the use of open or openly documented formats [to enable] free public access to data deriving from research should be the norm'[28], mirroring the European Research Council's guidelines for *Open Access to Research Data* in the *Horizon 2020* programme[29]. Since the importance of cross-linguistic data is constantly increasing, it is time to re-evaluate and improve the state of standardization of linguistic data[30].

While we have to ask ourselves whether adding another standard might worsen the situation[31], it is also clear that the current problems of "data-FAIR-ness" in comparative and typological linguistics persist and that standardization is the only way to tackle them. What may set our attempt apart from previous trials is a focus on data re-use scenarios as motivating use cases.

Previously, the focus of standardization attempts was often on comprehensiveness (cf. the GOLD ontology http://linguistics-ontology.org/, accessed July 27, 2018) which led to problems with adoption. Our proposal is more modest, targeting mainly the specific case of tool-based re-use (i.e. analysis, visualization, publication, etc.) of linguistic data. While this may seem overly specific, it is central to the scientific method and reproducible research[32]. This approach may also be particularly successful, because it puts the burden of early adoption on a sample of the linguistics community which may be best equipped to deal with it: the computationalists. The line between computational and non-computational linguists is diffuse enough for the former to act as catalysts for adoption, in particular because tools which

a ***One Value per Cell***

Many datasets that have been published in the past place multiple values in the same cell of their data. This is most frequently the case with elicitation meanings for which multiple translations could be found. Since scholars are rarely explicit about the separators or the techniques by which they handle these problems, many different ways to address multiple translations per meaning have been used in the past, ranging from additional columns up to secondary characters indicating multiple values in a cell (commas, slashes, pipes), and datasets may even mix the different techniques. To avoid these problems, CLDF specifies to use long tables throughout all applications.

NEITHER:

Meaning	English	German	Dutch
bark	bark	Rinde, Borke	bast

NOR:

Meaning	English	German	Dutch
bark	bark	Rinde	bast
bark	*	Borke	---

BUT:

ID	Meaning	Language	Form
1	*bark*	English	bark
2	*bark*	German	Rinde
3	*bark*	German	Borke
4	*bark*	Dutch	bast

b ***Anticipate the Need of Multiple Tables***

When a certain complexity of analysis is reached, multiple tables become inevitable in linguistic datasets. Unfortunately, the need of multiple tables may not be readily anticipated, and datasets do not transparently state how to link across tables.Formats for cognate coding show great variation in this regard, ranging from multiple sheets in spreadsheet software that were manually created up to customized formats in which additional information is encoded in form of markup, such as colored cells or text in italic or bold font. All these attempts are very error prone and lead to data-loss, especially if only certain parts of the data are shared. To avoid these problems, CLDF specifies to turn to multiple tables whenever this is needed, but to make it explicit in the metadata, how tables should be linked.

NEITHER:

--SHEET-B

Meaning	English	German	Dutch
bark	A	B, A	C

NOR:

--SHEET-A

Meaning	English	German	Dutch
bark	bark	Rinde	Borke	bast

BUT:

--TABLE-A

ID	Meaning	Language	Form
1	*bark*	English	bark
2	*bark*	German	Rinde
3	*bark*	German	Borke
4	*bark*	Dutch	bast

--TABLE-B

ID	Cognacy
1	bark-A
2	bark-B
3	bark-A
4	bark-c

Figure 1. **Basic rules of data coding, taking cognate coding in wordlists as an example**. (**a**) Illustrates why long tables[53] should be favored throughout all applications. (**b**) Underlines the importance of anticipating multiple tables along with metadata indicating how they should be linked[44].

can be built on standardized cross-linguistic data include web applications to make data publicly accessible to speaker communities and the general public (cf. http://clld.org, accessed July 27, 2018).

Results

To address the above-mentioned obstacles of sharing and re-use of cross-linguistic datasets, the *Cross-Linguistic Data Formats* initiative (CLDF) offers modular specifications for common data types in language typology and historical linguistics, which are based on a shared data model and a formal ontology.

Data Model

The data model underlying the CLDF specification is simple, yet expressive enough to cover a range of data types commonly collected in language typology and historical linguistics. The core concepts of this model have been derived from the data model which was originally developed for the *Cross-Linguistic Linked Data project* (cf. http://clld.org, accessed July 27, 2018), which aimed at developing and curating interoperable data publication structures using linked data principles as the integration mechanism for distributed resources. The CLLD project resulted in a large number of online datasets which provide linguists with a uniform "look-and-feel" despite their diverse content (see Table 1).

The main entities in this model are: (a) *Languages* - or more generally *languoids* (cf. http://glottolog.org, accessed July 27, 2018), which represent the objects under investigation; (b) *Parameters*, the comparative concepts[33], which can be measured and compared across languages; and (c) *Values*, the "measurements" for each pairing of a language with a parameter. In addition, each triple should have at least one (d) *Source*, as cross-linguistic data are typically aggregated from primary sources which themselves are the result of language documentation based on linguistic fieldwork. This reflects the observation of Good and Cysouw[34] that cross-linguistic data deal with *doculects*, i.e. languages as they are documented in a specific primary source - rather than languages as they are spoken directly by the speakers.

In this model, each *Value* is related to one *Parameter* and one *Language* and can be based on multiple *Sources*. The many-to-many relation between *Value* and *Source* is realized via *References* which can carry

Name	URL	Description
World Atlas of Language Structures	wals.info	Grammatical survey of more than 2000 languages world-wide.
Comparative Siouan Dictionary	csd.clld.org	Etymological dictionary of Siouan languages.
Phoible	phoible.org	Cross-linguistic survey of sound inventories for more than 2000 languages world-wide.
Glottolog	glottolog.org	Reference catalogue of language names, geographic locations, and affiliations.
Concepticon	concepticon.clld.org	Reference catalogue of word meanings and concepts used in cross-linguistic surveys and psycholinguistic studies.

Table 1. Examples of popular databases produced within the CLLD framework.

an additional *Context* attribute, which is typically represented by page numbers when dealing with printed sources.

The CLDF Specification

CLDF is a package format, describing various types of cross-linguistic data; in other words, a CLDF dataset is made up by a set of data files (i.e. files holding tabular data, or tables) and a descriptive file, wrapping this set and defining relations between tables. Each linguistic data type is modeled via a CLDF *module*, with additional, orthogonal aspects of the data modeled as CLDF *components*. "Orthogonal" here refers to aspects of the data which recur across different data types, e.g. references to sources, or glossed examples. This approach mirrors the way Dublin Core metadata terms (a common way of describing metadata, cf. http://dublincore.org, accessed July 27, 2018) are packaged into meaningful sets using *Application Profiles* (cf. http://dublincore.org/documents/2009/05/18/profile-guidelines/, accessed July 27, 2018): a well known technique to support custom, modular - yet interoperable - metadata specifications devised by the Dublin Core Metadata Initiative. CLDF modules are profiles of cross-linguistic data types, consisting of CLDF components and terms from the CLDF ontology.

CLDF Ontology. The CLDF specification recognizes certain objects and properties with well-known semantics in comparative linguistics. These are listed in a "vocabulary" or "ontology" (cf. https://www.w3.org/standards/semanticweb/ontology for a description of vocabularies in the context of the Semantic Web) - the CLDF Ontology - thereby making them available for reference by URI - the key mechanism of the Semantic Web (that is, the "Web of Data", cf. https://www.w3.org/standards/semanticweb/data). Wherever possible, this ontology builds on existing ontologies like the *General Ontology for Linguistic Description* (cf. http://linguistics-ontology.org/, accessed July 27, 2018). In particular, the CLDF Ontology makes it easy to link entities in a CLDF dataset to a reference catalogue by providing corresponding reference properties.

Basic Modules in CLDF. Currently, CLDF defines two modules which handle the most basic types of data which are frequently being used, collected, and shared in historical linguistics and typology (cf. http://clld.org/datasets.html). The *Wordlist* module handles lexical data which are usually based on a *concept list* that has been translated into a certain number of different languages, wich are often further analysed by adding information on cognate judgments or by further aligning the cognate words[35]. The *StructureDataset* module handles grammatical features in a very broad sense, which are usually collected to compare languages typologically.

Two more modules are in an early stage of standardisation: The *ParallelText* module can be used to encode texts which were translated into different languages and are split into functional units (like similar sentences or paragraphs) to render them comparable. The *Dictionary* module makes it possible to encode the lexicon of individual languages.

While these modules are usable in this stage as well, they also serve as examples of the extensibility of the standard: CLDF is intended as iterative, evolving standard, providing a short feedback loop between standardization, implementation and non-standard extensions - thus allowing new data types to be integrated easily.

Each of the modules defines additional components which define relations among the values across languages, inside a language, or value-internally.

Components. CLDF modules can include *components*. *Components* are pre-defined tables or custom, that is non-standardized, tables. While *components* can have different interpretations, depending on the *module* they are combined with, in the *Wordlist* module they are typically interpreted as concepts and in the *StructureDataset* module they most often interpreted as categorical variables.

Package Format of CLDF. CLDF is built on the World Wide Web Consortium (W3C) recommendations *Model for Tabular Data and Metadata on the Web* (cf. https://www.w3.org/TR/tabular-data-model/, accessed July 27, 2018) and *Metadata Vocabulary for Tabular Data* (cf. https://www.w3.org/TR/tabular-metadata/, accessed July 27, 2018, henceforth referred to as CSVW for "CSV on the Web"), which provide a

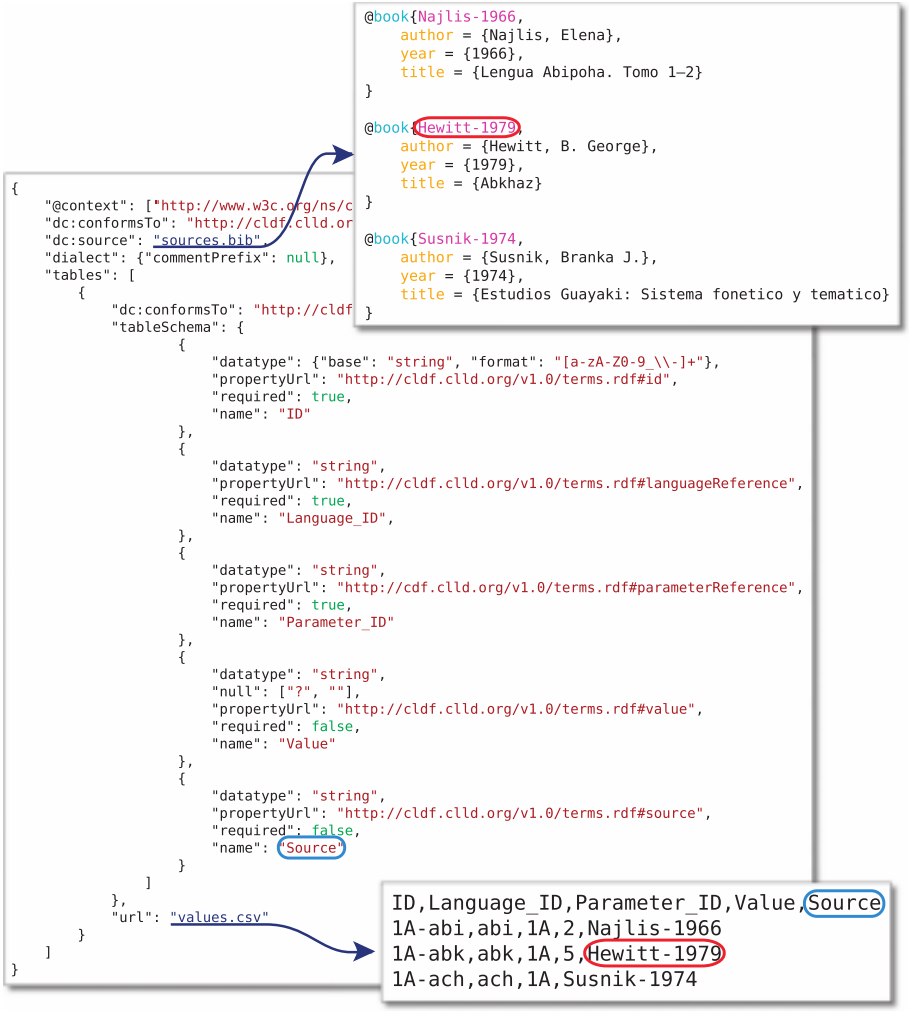


Figure 2. **Using CSVW metadata to describe the files making up a CLDF dataset.**

package format allowing us to tie together multiple files containing tabular data (see Fig. 2). Thus, each CLDF dataset is described by a JSON (Javascript Object Notation, see http://json.org/) metadata file according to CSVW tabular metadata specification.

This means that there are standard ways of including metadata: *Common properties* on *table* or *table group* descriptions can be used to add (a) bibliographic metadata using terms from the Dublin Core namespace (cf. http://purl.org/dc/terms/), (b) provenance information using terms from the PROV namespace (cf. https://www.w3.org/ns/prov), (c) catalogue information using terms from Data Catalog Vocabulary (cf. http://www.w3.org/ns/dcat#). Thus, by providing a way to specify such metadata in a machine-readable way, CLDF complements the efforts of the RDA Linguistics Interest Group (cf. http://site.uit.no/linguisticsdatacitation/austinprinciples, accessed July 27, 2018).

Extensibility of CLDF. The CLDF specification is designed for extensibility. A CLDF dataset can comprise any number of additional tables (by simply adding corresponding table definitions in the metadata file), or by adding additional columns to specified tables. Thus, we expect to see further standardization by converging usage, much like Flickr machine tags evolved (cf. https://www.flickr.com/groups/api/discuss/72157594497877875, accessed July 27, 2018). A dataset may, for example, specify scales for its parameters to guide appropriate visualization. If more and more users employ this new specification, it will become a candidate for standardization within the CLDF specification.

As an example for future enhancement, CLDF could build on extensive metadata schemes like the COREQ standards for qualitative social science research[36] to allow for an explicit annotation of basic attributes related to language informants when handling original fieldwork data (such as age, gender, multilingualism, etc.). In a similar way, existing semantic web ontologies could be further integrated into the CLDF specification, provided adapters of CLDF find them useful and important.

This extension mechanism (and backwards compatible, frequent releases) allows us to start out small and focused on a handful of use cases and data types for which there is already tool support.

Reference Catalogues

Creating a lean format like CLDF has been made easier by using reference catalogues to specify entities like languages or concepts. This, in turn, is made possible by employing the linking mechanism built into the W3C model and by leveraging JSON-LD, a JSON serialization of the RDF model underlying the Linked Data principles (cf. https://www.w3.org/TR/json-ld/, accessed July 26, 2018).

Linking to the corresponding properties in the CLDF Ontology allow for unambiguous references to standard catalogues like Glottolog and ISO 639-3[26] for languoids and Concepticon for lexical concepts. While Glottolog is now well-established among linguists concentrating on cross-linguistic language comparison, Concepticon is a rather young attempt to standardize the reference to lexical concepts as they can be encountered in numerous questionnaires that scholars use in fieldwork and comparative studies. Similar to Glottolog, Concepticon offers unique identifiers for currently 3144 lexical concepts, along with definitions and additional metadata. The lexical concepts defined by Concepticon, however, are not meant to reflect concepts that are expressed by the words in any specific language, but instead link to various resources (so-called *concept-lists*) in which these concepts were elicited. Similar to language names, which show many different variants in the linguistic literature, the glosses which scholars use to elicit a certain concept in cross-linguistic studies may also drastically vary. Linking these elicitation glosses to the Concepticon thus allows for a rapid aggregation of highly diverse datasets. As an example, consider the recently published new version of the CLICS database (cf. http://clics.clld.org), providing information on recurring polysemies for more than 1500 concepts, in which currently 15 different datasets have been aggregated with help of Glottolog and Concepticon. We are currently working on additional reference catalogues for phonetic transcriptions (*Cross-Linguistic Transcription Systems*, cf. https://github.com/cldf/clts, accessed July 27, 2018) and grammatical features (working title *Grammaticon*,[37]) and hope to make them available to CLDF data descriptions by providing corresponding reference properties in future versions of the CLDF Ontology.

However, while including reference properties for certain catalogues facilitates data aggregation and re-use, the CLDF specification does not require the use of any or all reference catalogues. Instead, users should decide what is most applicable to the dataset itself.

Interacting with CLDF Datasets

The main goal of CLDF is connecting cross-linguistic data and tools. The constituent file formats of CLDF - CSV, JSON and BibTeX -- enjoy ample support for reading and writing on many platforms and in many computing environments. Thus, reading and writing CLDF dataset should be easily achieved in any environment. A sufficiently standardized data format like CLDF means that general data editing tools (e.g. https://visidata.org/) can be used for working with CLDF data (see https://csvconf.com for more information about CSV in science, accessed July 26, 2018). A standardized format allows the community to move from ad-hoc tools programmed by a proficient minority for their particular use case, towards more and better applications, making their functionality available also to researchers without programming skills.

A few such tools already exist. LingPy (cf. http://lingpy.org, accessed July 27, 2018), a suite of open source Python modules, provides state-of-the-art algorithms and visualizations for quantitative historical linguistics; BEASTLing[38], a Python package, translates human-readable descriptions of phylogenetic inference into the complex driver files for the popular BEAST software; EDICTOR[39], a graphical JavaScript application, allows scholars to edit etymological dictionary data in a machine- and human-readable way. While the development on these examples began before the CLDF standard, all three of them were originally using CSV dialects for easy data exchange and are now in the process of adding support for CLDF data, thus showing the value of interoperability.

Further, CLDF is standardised such that scripts can easily become shareable and reusable tools for other researchers, rather than one-use scripts. To collect and publish such tools, we initiated a GitHub repository called the CLDF Cookbook (cf. https://github.com/cldf/cookbook). Currently, the cookbook contains recipes for visualization of CLDF datasets, for reading and writing data in CLDF-format from within the LingPy library, and for accessing CLDF data from R.

A Python API: pycldf

In many research disciplines the Python programming language has become the de-facto standard for data manipulation (often including analyses[40],). Thus, providing tools for programmatic access to CLDF data from Python programs increases the usefulness of a format specification like CLDF. We implemented a Python package pycldf (cf. https://github.com/cldf/pycldf, accessed July 27, 2018), serving as reference implementation of the CLDF standard, and in particular supporting reading, writing and validating CLDF datasets (cf. https://github.com/cldf/pycldf/tree/master/examples, accessed July 26, 2018).

By making use of the table descriptions in a CLDF metadata file, pycldf can do a lot more. For example, based on the datatype descriptors and foreign key relations specified in table schemas, pycldf can provide a generic conversion of a CLDF dataset into an SQLite database; thereby allowing analysis of

Abbr.	Requirement	Note
P	PEP 20	"Simple things should be simple, complex things should be possible" (cf. https://www.python.org/dev/peps/pep-0020/, accessed July 27, 2018): Datasets can be one simple CSV file, encoding language-parameter-value-triples.
R	Referencing	If entities and parameters can be linked to reference catalogues such as Glottolog or Concepticon, this should be preferred to duplicating information.
A	Aggregability	Data should be simple to concatenate, merge, and aggregate in order to guarantee their reusability.
H	Human- and machine-readability	Data should be both editable *by hand* and amenable to reading and writing by software (preferable software which typical linguists can be expected to use).
T	Text	Data should be encoded as UTF-8 text files or in formats that provide full support for UTF-8.
I	Identifiers	Identifiers should be resolvable HTTP-URLs, where possible, if not, this should be documented in the metadata.
C	Compatibility	Compatibility with existing tools, standards, and practices should always be kept in mind and never easily given up.
E	Explicitness	One row should only store one data point, and each cell should only have one type of data, unless specified in the metadata.

Table 2. Practical demands regarding cross-linguistic data formats.

CLDF datasets using SQL - one of the work horses of data science. Another example for the usefulness of programmatic access to CLDF data is validation. Having a Python library available for CLDF means validation can be built into LibreOffice's spreadsheet application or easily run via continuous integration services like Travis on datasets hosted in public repositories (see, for example, https://github.com/lexibank/birchallchapacuran, accessed July 26, 2018).

Discussion

At the beginning of the CLDF initiative we developed a list of practitioner requirements for cross-linguistic data, based on the experiences of linguists who have worked and are regularly working with cross-linguistic datasets. These practical principles are summarized in Table 2[41], and when comparing them with our first version of CLDF, it can be seen that CLDF still conforms to all of them. Furthermore, when comparing our initial requirements with the criteria for file formats and standards put forward in guidelines for research data management such as the ones proposed by the WissGrid project[42], one can also see that the perspectives are largely compatible, thus corroborating our hope that while being sufficiently specific to be of use for linguists, CLDF will also be generic enough to blend in with current best practices for research data management across disciplines.

Following a similar line of reasoning as Gorgolewski *et al.*[43] lay out in their proposal of a unified data structure for brain imaging data, and building on recommendations from the "Good Practices of Scientific Computing" by Wilson *et al.*,[44] we decided to base CLDF on well-known and well-supported serialization formats, namely CSV and JSON, with their specific shortcomings being outbalanced by building on CSVW, including its concept of CSV dialects, which allows us to support more variation in tabular data files and help with adaptation of the format. CSVW and its support for foreign keys between tables also allows us to seamlessly implement the recommendation to "anticipate the need to use multiple tables, and use a unique identifier for every record"[43].

Since CSVW is specified as a JSON-LD dialect (i.e. grounded in the Resource Description Framework RDF, cf. https://www.w3.org/TR/rdf11-primer/, accessed July 27, 2018), it can be combined with an RDF *Vocabulary* or *Ontology* to provide (a) the syntax of a relational serialization format via CSVW, as well as (b) the semantics of the entities in the data model via the ontology. Thus, the CLDF Ontology provides answers to the two questions of "Which things do exist?" and "Which things are based on others?", which are considered crucial to assess the identification needs for data collections[42].

Being able to build on Linked Data technologies to attach custom semantics to CSV data is the main advantage for us of CSVW over the similar *Data Package* Standard (cf. https://frictionlessdata.io/specs/data-package/), with its pure JSON package descriptions. It should also be noted that the overlap between these two data packaging specifications is so big and the specifications so similar, that the authors of the *Data Package* standard "imagine increasing crossover in tool and specification support"[45].

When adopting CSVW as the basis of the specification, it may seem counter-intuitive to model source information via BibTeX - rather than as just another CSV table, linked to with foreign keys. But given that (a) Glottolog - the most extensive bibliography of language descriptions - disseminates BibTeX and (b) the many-to-many relation between values and sources would have required an additional association table, (c) BibTeX is a standard format readable and usable by most citation software programs, BibTeX seemed to be the right choice when maximizing maintainability of datasets.

Another design decision taken with CLDF was to not specify a single-file format. Instead of forcing users to provide their data in database formats, like SQLite (cf. https://sqlite.org/appfileformat.html, accessed July 27, 2018), or in pure text formats with extensible markup, like the NEXUS format in biology[46], we opted for specifying a multi-file format - and deliberately chose to not define any packaging. Instead, we regard packaging of usually rather small sets of small text files as a problem for which multiple solutions with particular use cases have already been proposed (e.g. *zip* for compression, *bagit* for archiving, etc., cf. https://tools.ietf.org/html/draft-kunze-bagit-14, accessed July 27, 2018). We do not

even have to specify a particular directory layout for the multiple files forming a CLDF dataset, because the description file references data files using URIs, thereby turning CLDF into a multi-file format almost as flexible as HTML. While this decision goes against the idea of "self-describing data" - underlying formats like XML - it works well with databases with established curation workflows, because it provides an inobtrusive way to enhance the existing dataset: For example the "traditional" WALS Online tab-separated format (e.g. http://wals.info/feature/1A.tab) can be turned into a CLDF dataset (by anyone) by providing a separate description file, just referencing the tab-separated file as data file.

Since CLDF has been developed in close collaboration with researchers working on different ends of data-driven research in historical linguistics and language typology, CLDF is already being used by large linguistic projects (cf. http://clics.clld.org/ and http://www.model-ling.eu/lexirumah/, both accessed July 27, 2018) and as the data format for publishing supporting information[11,47]. CLDF is the native format for the forthcoming global language databases *Grambank*, *Lexibank* and *Parabank* (cf. http://glottobank.org/) being developed by a consortium of research centers and universities. Further, CLDF is by now already supported by a larger number of software packages and applications, ranging from libraries for automatic sequence comparison in historical linguistics (LingPy), via packages for phylogenetic analyses (BEASTLing[38]), up to interfaces for data inspection and curation (EDICTOR[39]).

Since the CLDF initiative was born out of the Cross-Linguistic Linked Data (CLLD) project, it is readily integrated into the CLLD framework and will allow users to publish their data without efforts on the web, making their data *findable* by exposing data and metadata to the major search engines, and increasing thus their interoperability. An important part of enabling data re-use is making data discoverable. In today's digital environment this means largely being "present" on the web. Basing CLDF on the recommendations of W3C's *Tabular Data on the Web* working group is a partial answer to this requirement.

Making it simple to publish CLDF datasets as CLLD applications goes a step further, because CLLD applications improve the visibility of datasets by exposing data and metadata to the major search engines, but also to field-specific aggregators such as OLAC, the *Open Language Archives Community*. More specifically, since CLLD applications implement the data provider part of the OAI-PMH protocol (cf. http://www.openarchives.org/OAI/openarchivesprotocol.html, accessed July 27, 2018) a CLDF dataset served by a CLLD application will be discoverable from OLAC and other portals.

It is important to note that CLDF is not limited to linguistic data alone. By embracing reference catalogues like Glottolog which provide geographical coordinates and are themselves referenced in large-scale surveys of cultural data, such as D-PLACE[48], CLDF may drastically facilitate the testing of questions regarding the interaction between linguistic, cultural, and environmental factors in linguistic and cultural evolution.

Methods

Efforts to standardize cross-linguistic data, in particular typological datasets and with the aim of comparability across datasets, have been undertaken since at least 2001, when Dimitriadis presented his *Typological Database System*[49] (cf. http://languagelink.let.uu.nl/tds/index.html, accessed July 27, 2018). One initial step was to introduce general database principles to database design in linguistic typology[50].

Rather than standardizing data formats, the CLLD project largely tried to standardize the software stack for cross-linguistic databases. Still, the core data model which could be extracted from these database software implementations served as one of the inspirations when standard data formats were discussed at the workshop *Language Comparison with Linguistic Databases*, held 2014 at the Max Planck Institute for Psycholinguistics in Nijmegen.

The followup workshop *Language Comparison with Linguistic Databases 2* - held in 2015 at the Max Planck Institute for Evolutionary Anthropology in Leipzig - saw concrete proposals towards what now is CLDF[41]; and later this year, the workshop *Capturing Phylogenetic Algorithms for Linguistics* - held at the Lorentz Center in Leiden - brought together people interested in analysis of cross-linguistic data, thus providing a test bed for the proposals.

Apart from these larger meetings where scholars discussed ideas of standardization, the CLDF-initiative profited from the numerous Glottobank meetings organized by the Department of Linguistic and Cultural Evolution at the Max Planck Institute for the Science of Human History (Jena), in which big-picture ideas of standards for linguistic data were discussed in more concrete terms by smaller teams which came forward to work on specific aspects of the specification, including reference catalogues like Concepticon, the handling of etymological data, and linking to external projects like D-PLACE.

These events formed a group representing the main institutions in the small field of large-scale comparison of cross-linguistic data, which contributed to the CLDF specification.

When a Linguistics Data Interest Group was endorsed by Research Data Alliance (RDA) in 2017, echoing RDA's call to 'develop and apply common standards across institutions and domains to ensure greater interoperability across datasets' in Linguistics, this coincided nicely with the progress of CLDF 1.0.

Code Availability

The CLDF specification is curated using a GitHub repository (cf. https://github.com/cldf/cldf). Released versions are published and archived via Zenodo under the Apache 2.0 license. The current version of the specification is CLDF 1.0.1[51].

The pycldf package is maintained in a GitHub repository (cf. https://github.com/cldf/cldf). Released versions are available from the Python Package Index (cf. https://pypi.python.org/pypi/pycldf) and archived with Zenodo[52] under the Apache 2.0 license.

References

1. Gawne, L., Kelly, B. F., Berez-Kroeker, A. L. & Heston, T. Putting practice into words: the state of data and methods transparency in grammatical descriptions. *Lang. Documentation Conserv* **11,** 157–189 (2017).
2. Greenhill, S. J., Blust, R. & Gray, R. D. The Austronesian basic vocabulary database: from bioinformatics to lexomics. *Evol. Bioinform* **4,** 271–283 (2008).
3. Blasi, D. E., Michaelis, S. M. & Haspelmath, M. Grammars are robustly transmitted even during the emergence of creole languages. *Nature Human Behaviour* **1,** 723–729 (2017).
4. Newberry, M. G., Ahern, C. A., Clark, R. & Plotkin, J. B. Detecting evolutionary forces in language change. *Nature* **551,** 223–226 (2017).
5. Greenhill, S. J. *et al.* Evolutionary dynamics of language systems. *P. Natl. Acad. Sci. USA* **114,** E8822–E8829 (2017).
6. Youn, H. *et al.* On the universal structure of human lexical semantics. *P. Natl. Acad. Sci. USA* **113,** 1766–1771 (2016).
7. Haynie, H. J. & Bowern, C. Phylogenetic approach to the evolution of color term systems. *P. Natl. Acad. Sci. USA* **113,** 13666–13671 (2016).
8. Gibson, E. *et al.* Color naming across languages reflects color use. *P. Natl. Acad. Sci. USA* **114,** 10785–10790 (2017).
9. Bouckaert, R. *et al.* Mapping the origins and expansion of the Indo-European language family. *Science* **337,** 957–960 (2012).
10. Chang, W., Cathcart, C., Hall, D. & Garret, A. Ancestry-constrained phylogenetic analysis support the Indo-European steppe hypothesis. *Language* **91,** 194–244 (2015).
11. Kolipakam, V. *et al.* A Bayesian phylogenetic study of the Dravidian language family. *Roy. Soc. Open Sci* **5,** 171504 (2018).
12. Grollemund, R. *et al.* Bantu expansion shows habitat alters the route and pace of human dispersals. *P. Natl. Acad. Sci. USA* **112,** 13296–13301 (2015).
13. Everett, C., Blasi, D. E. & Roberts, S. G. Climate, vocal folds, and tonal languages: connecting the physiological and geographic dots. *P. Natl. Acad. Sci. USA* **112,** 1322–1327 (2015).
14. Maddieson, I. & Coupé, C. Human spoken language diversity and the acoustic adaptation hypothesis. *J. Acoust. Soc. Am.* **138,** 1838 (2015).
15. Lupyan, G. & Dale, R. Language structure is partly determined by social structure. *PloS One 5* e8559 (2010).
16. Bromham, L., Hua, X., Fitzpatrick, T. G. & Greenhill, S. J. Rate of language evolution is affected by population size. *P. Natl. Acad. Sci. USA* **112,** 2097–2102 (2015).
17. Greenhill, S. J., Hua, X., Welsh, C. F., Schneemann, H. & Bromham, L. Population size and the rate of language evolution: a test across Indo-European, Austronesian, and Bantu languages. *Front. Psychol* **9,** 576 (2018).
18. Dediu, D. & Ladd, D. R. Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, aspm and microcephalin. *P. Natl. Acad. Sci. USA* **104,** 10944–10949 (2007).
19. DeMille, M. M. C. *et al.* Worldwide distribution of the DCDC2 READ1 regulatory element and its relationship with phoneme variation across languages. *P. Natl. Acad. Sci. USA* **115,** 4951–4956 (2018).
20. Roberts, S. G., Winters, J. & Chen, K. Future tense and economic decisions: controlling for cultural evolution. *PLoS One* **10,** e0132145 (2015).
21. Wilkinson, M. D. *et al.* The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3,** 160018 (2016).
22. Tamburelli, M. & Brasca, L. Revisiting the classification of Gallo-Italic: a dialectometric approach. *Digit. Scholarsh. Hum* **33,** 442–455 (2018).
23. Saxena, A., Borin, L. In *Approaches To Measuring Linguistic Differences* eds Borin, L. & Saxena, A. *Carving Tibeto-Kanauri by its joints: using basic vocabulary lists for genetic grouping of languages*. (De Gruyter Mouton, 2013).
24. IPA, International Phonetic Association. *Handbook Of The International Phonetic Association*. (Cambridge Univ. Press, 1999).
25. Kalusky, W. *Die Transkription Der Sprachlaute Des Internationalen Phonetischen Alphabets: Vorschläge Zu Einer Revision Der Systematischen Darstellung Der IPA-Tabelle*. (LINCOM Europa, 2017).
26. Lewis M. P. & Fennig C. D. eds *Ethnologue*. 17th edn, (SIL International, 2013).
27. List, J.-M., Cysouw, M. & Forkel, R. In *Proceedings Of The Tenth International Conference on Language Resources and Evaluation Concepticon:* a resource for the linking of concept lists. (European Language Resources Association, 2016).
28. Deutsche Forschungsgemeinschaft. *Guidelines On The Handling Of Research Data In Biodiversity Research* https://is.gd/Oofm6W (2015).
29. European Commission. Directorate-General for Research & Innovation. *H2020 Programme: Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020* https://is.gd/BUkJLJ (2017).
30. Berez-Kroeker, A. L. *et al.* Reproducible research in linguistics: a position statement on data citation and attribution in our field. *Linguistics* **56,** 1–18 (2018).
31. xkcd. *Standards* http://xkcd.com/927/ (2011).
32. Stodden, V., Seiler, J. & Ma, Z. An empirical analysis of journal policy effectiveness for computational reproducibility. *P. Natl. Acad. Sci. USA* **115,** 2584–2589 (2018).
33. Haspelmath, M. Comparative concepts and descriptive categories. *Language* **86,** 663–687 (2010).
34. Good, J. & Cysouw, M. Languoid, doculect, glossonym: formalizing the notion of 'language'. *Lang. Documentation Conserv* **7,** 331–359 (2013).
35. List, J.-M., Walworth, M., Greenhill, S. J., Tresoldi, T. & Forkel, R. Sequence comparison in computational historical linguistics. *J. Language Evolution* **3** (2018).
36. Tong, A., Sainsbury, P. & Craig, J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *Int. J. Qual. Health C* **19,** 349–357 (2007).
37. Haspelmath, M. & Forkel, R. Toward a standard list of grammatical comparative concepts: The Grammaticon https://is.gd/WGF36N (2017).
38. Maurits, L., Forkel, R., Kaiping, G. A. & Atkinson, Q. D. Beastling: a software tool for linguistic phylogenetics using BEAST 2. *PLoS One* **12,** e0180908 (2017).
39. List, J.-M. In *Proceedings Of The 15th Conference Of The European Chapter Of The Association for Computational Linguistics. System Demonstrations* A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets. (Association for Computational Linguistics, 2017).
40. Millman, K. J. & Aivazis, M. Python for scientists and engineers. *Comput. Sci. Eng.* **13,** 9–12 (2011).
41. Hammarström, H. A Proposal for Data Interface Formats for Cross-Linguistic Data https://github.com/clld/lanclid2/raw/master/presentations/hammarstrom.pdf (2015).
42. Ludwig, J. & Enke, H. Leitfaden zum forschungsdatenmanagement. *Ergebnisse aus dem WissGrid-Projekt* **15** (2013).

43. Gorgolewski, K. J. *et al.* The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci. Data* **3,** 160044 (2016).
44. Wilson, G. *et al.* Good enough practices in scientific computing. *PLOS. Comput. Biol.* **13,** e1005510 (2017).
45. Fowler, D., Barratt, J. & Walsh, P. Frictionless data: making research data quality visible. *Int. J. Digit. Curation* **12** (2017).
46. Maddison, D. R., Swofford, D. L. & Maddison, W. P. Nexus: an extensible file format for systematic information. *Syst. Biol* **46,** 590–621 (1997).
47. Hill, N. & List, J.-M. Challenges of annotation and analysis in computer-assisted language comparison: a case study on burmish languages. *Yearbook of the Poznań, Linguistic Meeting* **3,** 47–76 (2017).
48. Kirby, K. R. *et al.* D-PLACE: a global database of cultural, linguistic and environmental diversity. *PLoS One* **11,** e0158391 (2016).
49. Dimitriadis, A., Windhouwer, M., Saulwick, A., Goedemans, R., Bíró, T. In. The Use of Databases in Cross-Linguistic Studies (eds Everaert M., Musgrave, S. & Dimitriadis, A.) How to integrate databases without starting a typology war: the typological database system. (De Gruyter Mouton, 2009).
50. Dimitriadis, A., Musgrave, S. In *The Use of Databases in Cross-Linguistic Studies* (eds Everaert, M., Musgrave, S. Dimitriadis, A.) Designing linguistic databases: A primer for linguists. (De Gruyter Mouton, 2009).
51. Forkel, R., List, J.-M., Cysouw, M., Rzymski, C. & Greenhill, S. J. Source code for: CLDF 1.0.1. *Zenodo* https://doi.org/10.5281/zenodo.1252097 (2018).
52. Forkel, R., Bank, S., Greenhill, S. J., Rzymski, C. & Kaiping, G. Source code for: pycldf 1.5.0. *Zenodo* https://doi.org/10.5281/zenodo.1324189 (2018).
53. Wickham, H. Tidy data. *J. Stat. Softw.* **59,** 1–23 (2014).

Acknowledgements

This research would not have been possible without the generous support by many institutes and funding agencies. As part of the CLLD project (cf. http://clld.org) and the Glottobank project (cf. http://glottobank.org), the work was supported by the Max Planck Society, the Max Planck Institute for the Science of Human History, and the Royal Society of New Zealand (Marsden Fund grant 13-UOA-121, RF). JML was funded by the DFG research fellowship grant 261553824 *Vertical and lateral aspects of Chinese dialect history* (2015-2016) and the ERC Starting Grant 715618 *Computer-Assisted Language Comparison* (cf. http://calc.digling.org). SJG was supported by the Australian Research Council's Discovery Projects funding scheme (project number DE 120101954) and the ARC Center of Excellence for the Dynamics of Language grant (CE140100041). MH was supported by the ERC Advanced Grant 670985 *Grammatical Universals*. GAK was funded by NWO Vici project 277-70-012 *Reconstructing the past through languages of the present: the Lesser Sunda Islands*.

Author Contributions

Author R.D.G., R.F., M.C., H.H., M.H., and J.M.L. initiated the CLDF initiative. By making CLDF a key initiative for data handling at the Department of Linguistic and Cultural Evolution (MPI-SHH, Jena), R. D.G. provided financial, administrative, and conceptual support for C.L.D.F. R.F., S.J.G., and J.M.L. conceptualized the specification. R.F. conceptualized and designed the implementation. C.R., R.F., M.C., S.J.G., H.H., M.H., J.M.L., and G.K. contributed to the specification. R.F. and S.B. wrote the code for the pycldf package. R.F., J.M.L., and S.G. wrote the first draft. C.R., R.F., S.G., G.K., and J.M.L. expanded the first draft. All authors revised the second draft and agree with the final version.

Additional Information

Competing interests: The authors declare no competing interests.

How to cite this article: Forkel, R. *et al.* Cross-Linguistic Data Formats, advancing data sharing and reuse in comparative linguistics. *Sci. Data*. 5:180205 doi: 10.1038/sdata.2018.205 (2018).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

A cross-linguistic database of phonetic transcription systems

Cormac Anderson[1], Tiago Tresoldi[1], Thiago Chacon[2], Anne-Maria Fehn[134], Mary Walworth[1], Robert Forkel[1] and Johann-Mattis List[1]*
[1]Max Planck Institute for the Science of Human History, Jena; [2]Universidade de Brasilia; [3]Goethe University, Frankfurt; [4]CIBIO/InBIO: Research Center in Biodiversity and Genetic Resources, Vairão, Portugal
** list@shh.mpg.de*

Abstract
Contrary to what non-practitioners might expect, the systems of phonetic notation used by linguists are highly idiosyncratic. Not only do various linguistic subfields disagree on the specific symbols they use to denote the speech sounds of languages, but also in large databases of sound inventories considerable variation can be found. Inspired by recent efforts to link cross-linguistic data with help of reference catalogues (Glottolog, Concepticon) across different resources, we present initial efforts to link different phonetic notation systems to a catalogue of speech sounds. This is achieved with the help of a database accompanied by a software framework that uses a limited but easily extendable set of non-binary feature values to allow for quick and convenient registration of different transcription systems, while at the same time linking to additional datasets with restricted inventories. Linking different transcription systems enables us to conveniently translate between different phonetic transcription systems, while linking sounds to databases allows users quick access to various kinds of metadata, including feature values, statistics on phoneme inventories, and information on prosody and sound classes. In order to prove the feasibility of this enterprise, we supplement an initial version of our cross-linguistic database of phonetic transcription systems (CLTS), which currently registers five transcription systems and links to fifteen datasets, as well as a web application, which permits users to conveniently test the power of the automatic translation across transcription systems.
Keywords: phonetic transcription; phoneme inventory databases; cross-linguistically linked data; reference catalog; dataset.

1. Introduction

Phonetic transcription has a long tradition in historical linguistics. Efforts to design a unified transcription system capable of representing and distinguishing all

the sounds of the languages of the world go back to the late 19th century. Early endeavours included Bell's Visible Speech (1867) and the Romic transcription system of Henry Sweet (1877). In 1886, Paul Passy (1859–1940) founded the Fonètik Tîtcerz' Asóciécon (Phonetic Teachers' Association), which later became the *International Phonetic Association* (see Kalusky 2017: 7f). In contrast to writing systems targeted at encoding the speech of a single language variety in a visual medium, phonetic transcription aims at representing different kinds of speech in a unified system, which ideally would enable those trained in the system to reproduce foreign speech directly.

Apart from the primary role which phonetic transcription plays in teaching foreign languages, it is also indispensable for the purposes of language comparison, both typological and historical. In this sense, the symbols that scholars use to transcribe speech sounds, that is, the *graphemes*, which we understand as sequences of one or more glyphs, serve as *comparative concepts*, in the sense of Haspelmath (2010). While the usefulness of phonetic transcription may be evident to typologists interested in the diversity of speech sounds (although see critiques of this approach to phonological typology, i.a. Simpson 1999), the role of unified transcription systems like the *International Phonetic Alphabet* (IPA) is often regarded as less important in historical linguistics, where scholars often follow the algebraic tradition of Saussure (1916, already implicit in Saussure 1878). This emphasises the *systematic aspect* of historical language comparison, in which the distinctiveness of sound units within a system is more important than how they compare in substance across a sample of genetically related languages. If we leave the language-specific level of historical language comparison, however, and investigate general patterns of sound change in the languages of the world, it is obvious that this can only be done with help of comparable transcription systems serving as comparative concepts.

Here, we believe that use can be made of cross-linguistic reference catalogues, such as Glottolog (http://glottolog.org, Hammarström et al. 2017), a reference catalogue for language varieties, and Concepticon (http://concepticon.clld.org, List et al. 2016), a reference catalogue for lexical glosses taken from various questionnaires. Both projects serve as standards by linking metadata to the objects they define. In the case of Glottolog, geo-coordinates and reference grammars are linked to language varieties (*languoids* in the terminology of Glottolog), in the case of Concepticon, lexical glosses taken from questionnaires are linked to *concept sets*, and both *languoids* and *concept sets* are represented by unique identifiers to which scholars can link when creating new cross-linguistic resources. We think that it is time that linguists strive to provide simi-

lar resources for speech sounds, in order to increase the comparability of phonetic transcription data in historical linguistics and language typology.

2. Phonetic transcription and transcription data

When dealing with phonetic transcriptions, it is useful to distinguish *transcription systems* from *transcription data*. The former describe a set of symbols and rules for symbol combinations which can be used to represent speech in the medium of writing, while the latter result from the application of a given transcription system and aim to display linguistic diversity in terms of sound inventories or lexical datasets. While transcription systems are *generative* in that they can be used to encode sounds by combining the basic material, transcription data are *static* and fixed in size (at least for a given version published at a certain point in time). Transcription data have become increasingly important, with recent efforts to provide cross-linguistic accounts of sound inventories (Moran et al. 2014; Maddieson et al. 2013), but we can say that every dictionary or word list that aims at representing the pronunciation of a language can be considered as transcription data in a broad sense.

In the following, we give a brief overview of various transcription traditions that have commonly been used to document the languages of the world, and then introduce some notable representatives of cross-linguistic transcription data. Based on this review, we then illustrate how we try to reference the different practices to render phonetic transcriptions comparable across transcription systems and transcription datasets.

2.1. Phonetic transcription systems

When talking about transcription systems, we are less concerned with actual orthographies, which are designed to establish a writing tradition for a given language, but more with scientific descriptions of languages as we find them in *grammars*, *word lists*, and *dictionaries* and which are created for the purpose of language documentation. Despite the long-standing efforts of the International Phonetic Association to establish a standard reference for phonetic transcription, only a small proportion of current linguistic research actually follows IPA guidelines consistently.

2.1.1. The International Phonetic Alphabet

The *International Phonetic Alphabet* (IPA 1999, IPA 2015), devised by the *International Phonetic Association*, is the most common system of phonetic notation. As an alphabetic system, it is primarily based on the Latin alphabet, following conventions that were oriented towards 20th century mechanical typesetting practices; it consists of *letters* (indicating "basic" sounds), *diacritics* (adding details to basic sounds), and *suprasegmental markers* (representing features such as *stress*, *duration*, or *tone*). The IPA's goal is to serve as a system capable of transcribing all languages and speech realisations, eventually extended with additional systems related to speech in a broader sense, such as singing, acting, or speech pathologies. The IPA has been revised multiples times, with the last major update in 1993 and the last minor changes published in 2005.

2.1.2. Transcription systems in the Americas

In the Americas, although IPA has become more prevalent of late, there is only a minimum level of standardisation in the writing systems used for the transcription of local languages. While in North America most of the transcription systems of the twentieth century generally comprised different versions of what is generally known as the *North American Phonetic Alphabet* (NAPA, Pullum and Laduslaw 1996[1986]), in South America the picture is murkier. Although Americanist linguists have occasionally tried to harmonise the transcription systems in use (Herzog et al. 1934), we find a plethora of local traditions that have been greatly influenced by varying objectives, ranging from the goal of developing practical orthographies (often with an intended closeness to official national language orthographies), via the desire to represent phonemic generalisations in transcriptions, up to practical concerns of text production with typewriting machines (Smalley1964).[1] As a result, it is extremely difficult to identify a common Americanist tradition of phonetic transcription.

[1] Other kinds of adaptations involved modification of standard symbols such as the use of "stroke" in some letters representing stops in order to create a grapheme for a fricative sound lacking in the Latin based typography (e.g., ‹ꝑ› for voiceless bilabial fricative [ɸ], ‹ꝺ› for dental voiced fricative).

2.1.3. Transcription systems in African linguistics

Attempts to standardise the transcription of previously unwritten African languages with Latin-based writing systems date back to the middle of the 19th century (Lepsius 1854). In 1928, a group of linguists led by Diedrich Westermann (1875–1956) developed what came later to be known as the *African Alphabet*, an early attempt to enable both practical writing and scientific documentation of African languages with a minimal number of diacritic characters (International Institute 1930). In subsequent years, the system gained popularity among linguists and eventually served as the basis for the *African Reference Alphabet* (ARA, UNESCO 1978; Mann and Dalby 1987). Despite their relative success, most transcription systems and practical orthographies in use today are *mixed systems*, which inherit different parts from the IPA and the ARA, as well as alphabets of former colonial languages, alongside idiosyncratic elements. Although some areas developed regional conventions, languages with similar phoneme inventories may still be transcribed with widely diverging systems.[2]

2.1.4 Transcription systems in the Pacific

Among Oceanic languages, transcription conventions are extremely varied and are frequently based on regional orthographic conventions or the preferences of the respective linguists. In West Oceania, there is an increasing use of IPA in recent linguistic descriptions, however most existing descriptions are highly inconsistent, particularly when it comes to features that are typologically rare.[3] While Polynesian languages arguably maintain more straightforward phonological systems than their westerly cousins, they have been described with equal ambiguity. The various transcriptions include *outdated conventions*, *regional orthographic conventions*, and *individual linguists' inventions*. These have result-

[2] For instance, while most "Khoisan" (cf. Güldemann 2014) and Bantu languages of Southern Africa follow the African Reference Alphabet in transcribing clicks with Latin letters, linguistic treatments tend to use the IPA (following suggestions by Köhler et al. 1988). For example, the palatal click is indicated by ‹tc› in the first case and by ‹ǂ› in the second.

[3] For example, the *linguo-labial stop* of some Vanuatu languages has been described using an apostrophe following the labial ‹p'› (Lynch 2016), by using a subscript seagull diacritic under the labial ‹p̼› (Dodd 2014), and by using a subscript turned-bridge diacritic under the labial ‹p̪› (Crowley 2006a); the *doubly articulated labio-velar stop* in Vurës (Banks Islands) has been described as ‹p͡ʷ› (Malau 2016), whereas in the Avava language of Malekula, it has been transcribed with a tilde over the labial ‹p̃› (Crowley 2006b).

ed in highly ambiguous representations that easily lead to incorrect interpretations of the data, especially when being used by comparative linguists who are not familiar with the traditions.[4]

2.1.5 Transcription systems in South-East Asian languages

South-East Asian languages have a number of features that lend themselves to idiosyncratic phonetic transcription. A prominent example is tone, for which most scholars tend to prefer superscript or subscript numbers (e.g., ‹^{35}›) instead of the iconic IPA tone letters (‹˥˩›) originally designed by Chao (1930). Since scholars also use superscript numbers to indicate phonological tone (ignoring actual tone values) tone assignment can be easily confused. In addition to the transcription of tone, many language varieties have some peculiar sounds, which are not easy to be rendered in IPA and are therefore often transcribed with specific symbols common only in SEA linguistics.[5] Although especially younger field workers tend to transcribe their data consistently in IPA, we find many datasets and textbooks employing older versions of the IPA.[6]

2.1.6 Summary of transcription systems

Designing and applying phonetic transcription systems is not an easy enterprise, especially in cases where the goal is to provide a global standard. When com-

[4] Examples include, among others: (1) characters associated with a given sound being used to represent an entirely different sound (‹h› used for the *glottal stop*, Tregear 1899; ‹y› used for [ð], Salisbury 2002); (2) one character being used to represent various sound qualities (‹g› used for the *velar nasal* in Tregear 1899, and the *voiced uvular stop* in Charpentier and François 2015); (3) diacritics on vowels ambiguously used to indicate duration (Stimson and Marshall 1964) or *glottal stops* (Kieviet 2017).

[5] Among these are the symbols ‹ɿ› and ‹ʅ›, which are commonly used to denote vowels pronounced with friction. They could be transcribed as syllabic sibilant fricatives [z̩] and [ʐ̩], respectively, but given the problems of readability with these symbols, as well as the relative frequency of these sounds across Chinese dialects and in other Sino-Tibetan languages, scholars continue to use the symbols ‹ɿ› and ‹ʅ›.

[6] The most prominent difference is the usage of ‹’› as an aspiration marker [ʰ], which can be found in many sources (Beijing Daxue 1964), reflecting an older IPA standard which is also still in use in Americanist transcription systems and occasionally still taught in recent textbooks on Chinese linguistics (see, for example, Huáng and Liào 2002). Contrast this with the frequent use of the same symbol to represent ejectives in other traditions.

paring the particular problems of transcription systems and transcription practice in different parts of the world, one can identify many similar obstacles that linguists face when trying to preserve speech in writing. The most prominent ones include (a) the influence of the orthography of the dominant language (in many parts of the world the colonial language of the oppressors), (b) a tendency to favour tradition over innovation (which results in many practices that were once considered standard now having been abandoned), (c) specific challenges in transcribing local language varieties with the material provided by the standard, (d) systemic (phonological) considerations which would entice linguists to favor symbols which reflect the phonology of the language varieties under question more properly, and (e) technical considerations (as transcription systems devised up until the mid-20th century were forced to consider the limitations of mechanical typesetting).[7] While these technical considerations should have now become largely obsolete with the introduction of the Unicode standard, this is not always the case. Judging from practical experience it is obvious that Unicode has made many things a lot easier, but since the majority of linguists are less acquainted with questions of computation and coding, the problem of typesetting is still an important factor in linguistic transcription practice.

2.2. Transcription data

In addition to transcription systems as they are used by scholars and teachers, a number of datasets offer transcription data. Usually these datasets represent typological surveys of phoneme inventories (Maddieson et al. 1984; Maddieson et al. 2013; Moran et al. 2014; Ruhlen 2008). Originally they are taken from grammatical descriptions of the languages of the world and also tend to contain an introduction into the typical sound systems of the languages under investigation. Another type of frequently available transcription data (in the sense of fixed sets of sounds which are provided in the form of transcriptions) are feature descriptions of individual collections of speech sounds which can range from single-language descriptions (Chomsky and Halle 1968), up to large collections directed towards cross-linguistic, computer-assisted applications (Mortensen 2017).

[7] This includes the IPA itself, which has many glyphs that are rotated versions of letters, e.g. IPA (1912). Further, restrictions in the early days of computing led to limited by encoding schemes such as ASCII (which led to the development of ASCII representations of IPA, such as X-SAMPA).

In a broader sense, all data collections that provide *metadata* for a given set of sounds can be qualified as transcription data. When applying this extended definition of transcription data, we can think of many further examples, including diachronic datasets of sound change (Kümmel 2008, Index Diachronica), interactive illustrations of speech sounds (Multimedia IPA chart, Wikipedia), or lexical datasets that offer phonetic transcriptions (List and Prokić 2014).

2.3. Comparability of transcription systems and data

When dealing with transcription systems and transcription data, linguists face several problems. Some of these are problems of a practical nature, which we explore further below, while others are of a theoretical nature, and touch upon long-standing issues in phonology and phonetics, and the relationship between the two. Among these theoretical problems, are those of *commensurability*, of *context*, and of *resolution*.

In spite of frequent attempts to compare phonemic inventories in phonological typology (Dryer and Haspelmath 2011; Maddieson 1984) these efforts are beset by serious difficulties. The classical structuralist treatment of the phoneme considers it to be a *relational entity* (Trubetzkoy 1939), the value of which is dependent on its place with respect to other phonemes within a system. In this understanding, the phonemes of one language are not commensurate to those of another language: it is only as a member of a system that a phoneme finds its value. This critique is taken up by Simpson (1999) who argues that the allophone replaces the phoneme in large databases, thereby reducing "the phonemic system of a language to a small, arbitrary selection of its phonetics". Although this problem cannot really be resolved, we note that different phonological databases have attempted to address it in different ways. In LAPSyD (Maddieson et al. 2013), the symbols chosen for the phonemes are often frequently occurring ones, abstracting away from too much phonetic detail. In PHOIBLE (Moran et al. 2014), on the other hand, phonemes are often transcribed with great phonetic detail, with numerous diacritics. While at first glance the latter approach might appear preferable, as it gives more information, it runs into serious difficulties, given Simpson's critique above.

The crux of this problem is that the realisation of a given phoneme depends considerably on *context*. For example, the German stops typically transcribed /b/, /d/, and /g/ are pronounced *voiceless* when in final position, whereas between vowels they are pronounced with voice. In European Spanish, while the

voiced stops /b/, /d/ and /g/ occur with the phonetic values [b], [d], and [g] in initial position, elsewhere they are more often pronounced as fricatives [β], [ð], and [ɣ]. It is not clear, in such cases, which set of symbols should be used, and even if a principled decision could be made (e.g. based on frequency, Bybee 2001), a great loss of information is involved in choosing one symbol over the other – it is equally misleading to characterise Spanish as a language without voiced stops or as a language without voiced fricatives. Such difficulties are not only of relevance in phonological typology, but can have serious repercussions in historical linguistics as well. To take an example, linguists typically transcribe two series of stops in Scottish Gaelic – aspirated /pʰ/, /tʰ/, and /kʰ/ and unaspirated /p/, /t/, and /k/. In Modern Irish, on the other hand, the convention is to transcribe rather voiceless /p/, /t/, and /k/ and voiced /b/, /d/, and /g/. In reality, however, the voiceless stops of Irish are also aspirated, and the voiced ones are only passively voiced, i.e. it is an "aspirating" language in the parlance of laryngeal typology (Honeybone 2005). The difference between these two very closely related languages lies solely in the fact that in Irish there is perhaps a greater tendency to passively voice the second series. To a naïve historical linguist, however (or indeed, to an even more naïve algorithm), this minor difference would seem a highly significant one, and would require the postulation of entirely spurious sound changes ("deaspiration" and "voicing" of the two Irish series, for example) to account for the difference.

This last example leads to a further difficulty: the level of *resolution* of the different transcription datasets available varies widely. Sapir (1930) gives an extremely detailed account of the phonological system of Southern Paiute, very rich in phonetic detail. However, in our only description of the closely related language Chemehuevi (Press 1980) there is a comparative paucity of discussion of phonetic particulars. This is not to criticise her grammar (indeed one could make exactly the opposite statement about the quality of the syntactic description in her grammar and Sapir's),[8] but rather to recognise that these two sets of transcription data have a very different level of resolution. Obviously, there are great difficulties inherent in comparing datasets of differing levels of resolution: *absence of evidence* (e.g. in some phonetic particular of Chemehuevi) does not equate to *evidence of absence*. Our degree of knowledge about the phonetics

[8] One might suggest that one of the reasons for which Press did not go into great detail on the phonetics of this language was because Sapir had already provided an extremely in-depth account of a very closely-related idiom, and thus comparatively less was known about the syntax than the phonetics of this language cluster.

and phonology of the languages of the world varies greatly, from practically nothing to voluminous descriptions detailing small sociolectal, dialectal, and idiolectal divergences.

One might ask then, given these difficulties we recognise, what the value of this enterprise is. We believe that notwithstanding these theoretical difficulties, some practical progress can still be made. Given that transcription systems are rarely standardised in a rigid manner, and allow for a certain amount of freedom of choice, scholars have come up with many ad-hoc solutions, which are reflected in specific traditions that have developed in different sub-fields of comparative linguistics. As we have seen in Section 2.1, in different linguistic traditions there are various particularities in the representation of sounds in a written medium. Scholars are usually aware of these differences in their field of expertise, but when it comes to global accounts of phonetic and phonological diversity, the particularities of the different traditions may easily introduce errors into our analyses. A great number of the practical difficulties encountered in comparative studies arise not from the broader theoretical problems outlined above, but from exactly these idiosyncrasies of tradition or personal taste. In some cases, different linguists represent sounds that are fundamentally the same in different ways (see, for instance, the examples from Pacific languages in Section 2.1.4). Convenience also plays a role here: as it is inconvenient to write a superscript ‹^{h}› to mark aspiration of a stop, scholars often just use the normal ‹h› instead, assuming that their colleagues will understand, when reading the introduction to their field work notes or grammars.[9] An ‹h› following a stop, however, does not necessarily point to aspiration in all linguistic traditions. In Australian linguistics, for example, it often denotes a laminal stop (Dench 2002).

Further problems that scholars who work in a qualitative framework may not even realise arise from the nature of Unicode, which offers different encodings for characters that look the same (Moran and Cysouw 2017: 54). While scholars working qualitatively will have no problems to see that ‹ə› (Unicode 0259, *Latin Small Letter Schwa*) and ‹ɘ› (Unicode 01DD, *Latin Small Letter Turned E*) are identical, the two symbols are different for a computer, as they are represented internally by different code points. As a result, an automatic aggregation of data will treat these symbols as different sounds when comparing languages automatically, or when aggregating information on the sound inventories of the languages in the world.

[9] We recognise however, that in some cases it may be more principled to write e.g. /ph/ rather than /p^{h}/. An example is Khmer, where there is good evidence that these aspirated stops are actually clusters, as the /p/ and the /h/ can be separated by an infix (Jakob 1963).

Judging from the above-mentioned examples, we can identify three major problems which make it hard for us to compare phonetic transcriptions cross-linguistically: (a) errors introduced due to the wrong application of the Unicode standard; (b) general incomparability due to the use of different transcription systems; and (c) ambiguities introduced by scholars due to individual transcription preferences. In order to render our transcription systems and datasets cross-linguistically comparable, both for humans and for machines, it therefore seems indispensable to work on a system that normalises transcriptions across different transcription systems and transcription data by linking existing transcription systems and datasets to a unified standard. Such a system should ideally (a) ease the *process of writing phonetic transcriptions* (e.g. by providing tools that automatically check and normalise transcriptions while scholars are creating them), (b) ease the *comparison of existing transcriptions* (e.g. by providing an internal reference point for a given speech sound which links to different grapheme representations in different transcription systems and datasets), and (c) provide a *standard* against which scholars can test existing data. While such an approach cannot solve the theoretical issues of comparability discussed above, it can nonetheless be of considerable practical benefit.

3. The Framework of Cross-Linguistic Transcription Systems

In the spirit of *reference catalogues* for cross-linguistically linked data (Glottolog and Concepticon, see Section 1), we have established a preliminary version of a reference catalogue for phonetic transcription systems and datasets, called *Cross-Linguistic Transcription Systems* (CLTS). The goal of the CLTS framework is to systematically increase the comparability of linguistic transcriptions by linking graphemes generated by transcription systems and graphemes documented in transcription datasets to unique feature bundles drawn from a simple but efficient feature system. With due respect to all obstacles which the documentation of speech through transcription may face in theory and practice, the CLTS system can be seen as a first step towards identifying graphemes across transcription systems and transcription datasets with unique speech sounds. In this sense, CLTS also aids the *translation* between transcription systems and datasets, and can further serve as a *standard* for transcription in practice. Figure 1 illustrates this integrative role of CLTS.

In the following, we will briefly introduce the basic techniques by which we try to render linguistic transcription data comparable. Apart from the data itself

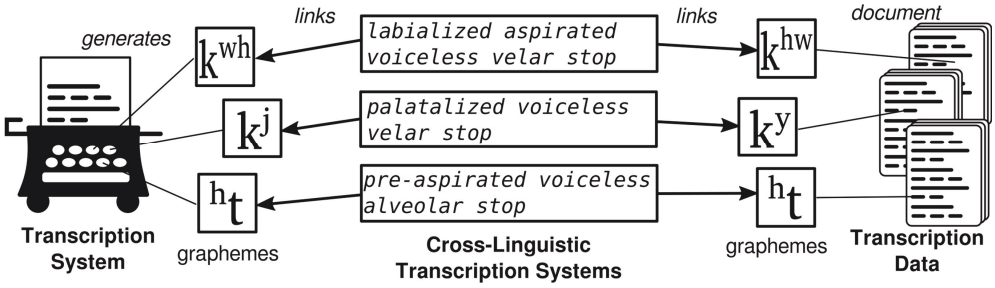


Figure 1. Basic idea behind the CLTS reference catalogue.

(discussed in Section 3.1), which we assemble and annotate in our reference catalogue, we also introduce a couple of different techniques which help to check the consistency of our annotations and ease the creation of new data to which we can link (Section 3.2).

3.1. Materials

3.1.1 Sound classes in CLTS

In order to link graphemes in transcription systems and transcription datasets to feature bundles, it is useful to distinguish rudimentary *classes* of sounds.[10] We distinguish three basic sound classes (*consonants*, *vowels*, and *tones*), a specific class of *markers* (to indicate syllable or morpheme breaks or word boundaries) and two derived classes (*consonant clusters* and *diphthongs*). As of the moment, we do not allow for triphthongs and clusters of more than two consonants (although they could be added at a later stage), in order to keep the system manageable. Clicks are represented as a specific type of consonant that has *click* or *nasal-click* as its *manner*. The representation of tones as a sound class of itself is necessitated by the fact that many phonetic descriptions of tone languages (especially in South-East-Asian languages) represent tone separately. It is further justified by phonological theory, given that tones in many languages may change independently, often correlated with factors that cannot be tied to a seg-

[10] We know that the distinction between basic sound types like *vowels* and *consonants* is often disputed in discussions on phonology and phonetics. For the purpose of linking speech sounds across datasets, however, it is useful to maintain the distinction for practical reasons, as both transcription systems and transcription datasets often maintain these distinctions.

mental context. In addition, we allow tones to be represented with diacritics on vowels (e.g., ‹á› in IPA would be described as an *unrounded open front vowel with high tone*), but we do not encourage scholars to represent their data in this form, as it has many disadvantages when it comes historical language comparison in practice and does not account well for the largely suprasegmental nature of tones.

Complex sound classes in CLTS are not explicitly defined, but instead *automatically derived* by identifying the basic graphemes of which they consist. Diphthongs are thus defined by two vowels, and the grapheme ‹oe›, for example, is treated as a diphthong consisting of a *rounded close-mid back* and an *unrounded close-mid front* vowel. In a similar way, we allow complex consonant clusters to be defined in order to transcribe, for example, doubly articulated consonants or clicks containing a pulmonic release (see Table 1 for examples).[11]

Table 1. Examples for the basic classes of sounds represented in CLTS.

Class	Grapheme	Features
consonant	k^{wh}	labialised aspirated velar stop
vowel	u̻	creaky rounded close back
cluster	kp	from voiceless velar stop to voiceless bilabial stop
diphthong	au̯	from unrounded open front to non-syllabic rounded close back
tone	²¹⁴	contour from-mid-low via-low to-mid-high
marker	+	marker for morpheme boundaries

3.1.2. Features bundles as comparative concepts

In order to ensure that we can compare sounds across different transcription systems and datasets, a feature system that can be used to model sounds as feature bundles, serving as comparative concepts in the sense of Haspelmath (2010) is

[11] For clusters involving clicks, we follow Traill (1993), Güldemann (2001), and Nakagawa (2006), who identify two segments for these sounds, a lingual influx (consonant-onset), and a pulmonic efflux (consonant-offset). For example, [ǀχ] is analyzed as a cluster consisting of a dental click [ǀ] as C-onset, and a uvular fricative [χ] as C-offset.

indispensable. We therefore propose specific feature systems for each of our three sound classes (consonant, vowel, tone), which allow us to identify a large number of different sounds across transcription systems and transcription datasets. The features themselves can be roughly divided into *obligatory features* (like *manner*, *place*, and *phonation* in consonants, and *roundedness*, *height*, and *centrality* in vowels), and *optional features* (usually binary, i.e., present or absent, such as *duration*, *nasalisation*, *aspiration*). Our current feature system contains 25 consonant features,[12] 21 vowel features,[13] and 4 tonal features[14] (Appendix A gives a table with all features and their possible values).

Our choice of features derives from the graphemic representation of sounds in the system of the IPA. It is practically oriented and does not claim to represent any deeper truth about distinctive features in phonology. Instead we focus on being able to align the features as easily as possible with a given graphemic representation of a particular sound in a transcription system. As a result, some features may appear awkward and even naïve from a phonological perspective. For example, instead of distinguishing *ejectives* from plain consonants by manner only (contrasting "ejective stops" and "plain stops"), we code *ejectivity* as an additional feature with a binary value (present or absent). In a similar way, we do not distinguish between different *kinds* of phonation (*voiced*, *breathy-voiced*, *creaky-voiced*, etc.) but code separately for *breathiness*, *creakiness*, and *phonation* (*voiced* or *voiceless*). The advantage of this coding practice is that we can easily infer sounds that we have not yet listed in our database based on the combination of base graphemes and diacritics. In addition, we can also avoid discussions in those cases where linguists often disagree. If we explicitly treated the diacritic ‹ʱ› in the IPA transcription system as indicating breathiness and implying voiced phonation, we would have a problem in distinguishing the admittedly rare instances where scholars explicitly transcribe voiceless stops with breathy release using a voiceless stop in combination with the diacritic for breathy voice (‹pʱ›, ‹tʱ›, ‹kʱ›, etc.) in order to indicate a voiceless initial with

[12] The features are: articulation, aspiration, breathiness, creakiness, duration, ejection, glottalisation, labialisation, laminality, laterality, *manner, nasalisation, palatalisation, pharyngealisation, *phonation, *place, preceding, raising, relative articulation, release, sibilancy, stress, syllabicity, velarisation, and voicing (features with an asterisk are obligatory).

[13] These are: articulation, breathiness, *centrality, creakiness, duration, frication, glottalisation, *height, nasalisation, pharyngealisation, raising, relative articulation, rhotacisation, *roundedness, rounding, stress, syllabicity, tone, tongue root, velarisation, voicing (features with an asterisk are obligatory).

[14] Tonal features are: contour, end, middle, and start (all obligatory).

(breathy) voiced aspiration (Starostin 2017). We could of course argue that these pronunciations are impossible physiologically and impose a system that automatically normalises these graphemes by either treating them as breathy-voiced stops or by treating them as plain-aspirated stops. We prefer, however, to leave the system as inclusive as possible for the time being, following the general principle that it is easier to reduce a given system at a later point for a specific purpose (while preserving the more complex version) than to impose restrictions too early. Given the flexibility of our system (which is presented in more detail in Section 3.2), it would be straightforward to create a strict feature representation that normalises those segments articulatory phoneticians consider impossible. However, if we erroneously reduce the data now, based on assumptions about phonetics that may well be disputed among experts, we run the risk of making regrettable decisions that are difficult to reverse. For this reason, we describe the grapheme ‹pʱ› as a *breathy voiceless bilabial stop consonant*, knowing well that scholars might object to the existence of this sound.

3.1.3. Transcription systems

A transcription system is understood as a *generative entity* in CLTS, being capable of creating sounds that were not produced explicitly before (although the ultimate productivity of a transcription system is, of course, limited). Transcription systems are defined by providing graphemes for the basic sound classes (*consonants*, *vowels*, *tones*), which are explicitly defined and linked to our feature system. Additionally, *diacritics* can be defined and may precede or follow the base graphemes, adding one additional feature per symbol to the base grapheme, depending on their position and the sound class of the base grapheme. In the IPA system, for example, the diacritic ‹ʰ› can only be attached to consonants, but it will evoke different feature values when preceding ‹ʰt› (*pre-aspirated voiceless alveolar stop consonant*) or following ‹tʰ› (*aspirated voiceless alveolar stop consonant*) the base grapheme ‹t›.

Transcription systems can furthermore specify *aliases*, both for base graphemes and for diacritics. The IPA, for example, allows one to indicate *breathiness* by two diacritics, the ‹dʱ› which we mentioned above, and the ‹◌̤›, which is placed under the base grapheme. In the CLTS framework, both glyphs can be parsed, and both ‹dʱ› and ‹d̤› would be interpreted as a *breathy voiced alveolar stop*, but ‹dʱ› would be treated as the regular grapheme representation and ‹d̤› as

its alias.[15] Other important examples of aliases are affricates such as the *voiceless alveolar affricate*, which can be rendered as either a single symbol ‹ʦ› (Unicode 02A6) or two symbols ‹ts› (Unicode points 0074 and 0073, the preferred version in CLTS).[16] In these and many other cases, the CLTS framework correctly recognises the sounds denoted by the graphemes, while at the same time proposing a default representation of ambiguous graphemes in a given transcription system.

CLTS currently offers five different transcription systems, namely a *broad* version of the IPA (called BIPA), a preliminary version of the transcription system underlying the *Global Lexicostatistical Database* (GLD, http://starling.rinet.ru/new100/main.htm, Starostin and Krylov 2011), the transcription system employed by the *Automatic Similarity Judgment Project* (ASJPCODE, http://asjp.clld.org, Wichmann et al. 2016), an initial version of the *North American Phonetic Alphabet* (NAPA, Pullum and Ladusaw 1996), and an initial version of the Uralic Phonetic Alphabet (UPA, Setälä 1901). Most of our initial efforts went into the creation of the B(road)IPA system. This choice is justified, as most transcription datasets also follow the supposed IPA standards to a large degree. In the future, however, we hope that we can further expand the data by expanding both the generative power and the accuracy of the remaining transcription systems, and by adding new transcription systems.

3.1.4. Transcription data

CLTS currently links 15 different transcription datasets, summarised in Table 2. The datasets were selected for different reasons. We tried to assemble as many of the cross-linguistic sound inventory datasets as possible (Nikolaev 2015; Maddieson et al. 2013; Mielke 2008; Moran et al. 2014; Ruhlen 2008), since apart from the comparison of Phoible with Ruhlen's database by Dediu and Moisik (2016), these existing datasets have not yet been thoroughly compared. Linking them to CLTS should thus immediately illustrate the usefulness of our

[15] The decision of what we define as an alias and what we define as the regular symbol is mostly based on practical considerations regarding visibility. Since the glyph ‹◌̤› will be difficult if not impossible to spot when placed under certain consonants, we decided to define ‹ʱ› as the base diacritic to indicate breathiness for consonants, but kept ‹◌̤› for vowels.

[16] We know well that no single decision will ever satisfy all users, but given the flexibility of the system, users can always easily define their sub-standard while at the same time maintaining comparability via our feature system.

Table 2. Basic coverage statistics for transcription datasets linked by the CLTS framework.

ID	Name	Source	Graph.	CLTS	Cov.
APiCS	Atlas of Pidgin and Creole Language Structures Online	Michaelis et al. 2013	177	177	100
BDPA	Benchmark database of phonetic alignments	List and Prokić 2014	1466	1329	91
BJDX	Chinese Dialect Vocabularies	Beijing Daxue 1964	124	124	100
Chomsky	Sound Pattern of English	Chomsky and Halle 1968	45	45	100
Diachronica	Index Diachronica	Anonymous 2014, D. 2017	652	552	85
Eurasian	Database of Eurasian Phonological Inventories	Nikolaev 2015	1562	1366	87
LAPSyD	Lyon-Albuquerque Phonological Systems Database	Maddieson et al. 2013	795	712	90
Multimedia	Multimedia IPA Charts	Department of Linguistics 2017	138	134	97
Nidaba	Lexicon Analysis and Comparison	Eden 2018	1936	1872	97
PanPhon	PanPhon Project	Mortensen 2017	6334	6220	98
PBase	PBase Project	Mielke 2008	1068	859	80
Phoible	Phonetics Information Base and Lexicon	Moran et al. 2014	1843	1589	86
PoWoCo	Potential of Word Comparison	List et al. 2017	378	370	98
Ruhlen	Global Linguistic Database	Ruhlen 2008	701	437	62
Wiki	Wikipedia IPA Descriptions	Wikipedia contributors 2017	184	168	91

framework (see Section 4.3 for details). Furthermore, given the large number of sound segments which one can find in these datasets (most of them representing

a supposedly strict version of IPA), they provide a useful way to test how well our framework recognises sounds written in IPA which were not explicitly defined. Additional datasets were chosen to illustrate links to feature systems (Chomsky and Halle 1968), for illustrative purposes (Department of Linguistics 2017; Wikipedia contributors 2018), or to test our system by providing either large collections of graphemes (Eden 2018; Mortensen 2017; List and Prokić 2014; List et al. 2017), or for reasons of general interest and curiosity (Michaelis et al. 2013; Anonymous 2014).

Table 3. Small excerpt of Unicode confusables normalised in CLTS.

Source	Code	Target	Code	Sound Name
λ	03BB	ʎ	028E	palatalised alveolar lateral approximant consonant
ə	01DD	ə	0259	unrounded mid central vowel
ʔ	0242	ʔ	0294	voiceless glottal stop consonant
ε	03B5	ɛ	025B	unrounded open-mid front

3.2. Methods

3.2.1. Parsing and generating sounds

CLTS employs a sophisticated algorithm for the parsing and generation of graphemes for a given transcription system. The parsing algorithm employs a three-step procedure, consisting of (A) normalisation, (B) direct lookup, and (C) generation of graphemes.

In (A), all sounds are generally normalised, following Unicode's NFD normalisation in which diacritics and base graphemes are maximally dissolved (Moran and Cysouw 2017: 16). In addition, the algorithm uses system-specific normalisation tables of homoglyphs, which can be easily confused. The normalisation applies to single glyphs only and employs a simple lookup table in which source and target glyph are defined. In this way, one can easily prevent users from using the wrong character to represent, for example, the schwa-sound [ə], since the data is normalised beforehand. Table 3 gives a small list of examples for base graphemes normalised in CLTS.

In (B), the algorithm searches for direct matches of the grapheme with the base graphemes provided along with the transcription system. If a grapheme can be matched directly, the algorithm checks whether it is flagged as an alias and provides the corrected grapheme.

If the grapheme could not be resolved in (A), the algorithm tries to generate it in (C), by first using a regular expression to identify whether the unknown grapheme contains a known base grapheme. If this is the case, the algorithm searches to the left and the right of the base grapheme for known diacritics, looks up the features from the table of diacritic features, and then combines the features of the base grapheme with the new features supplied by the diacritics to a generated sound. The algorithm returns an unknown sound if either no base grapheme can be identified or if one of the diacritics cannot be interpreted correctly.[17]

The algorithm can be used in a reverse fashion by supplying a feature bundle from which the algorithm will then try to infer the underlying grapheme in a given transcription system. Here again, we can distinguish between sounds that were already defined as base graphemes of the transcription system, and sounds that are generated by identifying a base sound and then converting the remaining features to diacritic symbols. Since the order of features serving as diacritics is defined directly, the algorithm explicitly normalises phonetic transcriptions in those cases in which features are supplied in the wrong order. For example, if a transcription system provides the *labialised aspirated voiceless velar stop consonant* as ‹k^{hw}› (as, for example, APiCS), the algorithm will normalise the order of diacritics to ‹k^{wh}› and flag the grapheme as an alias.

3.2.2. Python API and online database

CLTS comes with a Python API which can be used from the command line or within Python scripts and offers a convenient way to test the framework both on large datasets and on an ad-hoc basis. It also comes along with a brief tutorial introducing the main aspects of the code as well as a “cookbook” containing a series of coding recipes to address specific tasks. The data is further presented

[17] The generation procedure is strictly *accumulative*, and no features of the base grapheme can be changed post-hoc. This explains most peculiarities of our feature system and reflects a deliberate design choice. Given the large number of speech sounds that we could identify in the different transcription datasets, we had to make sure to keep the complexity of the algorithm on a level that can still be easily understood.

online at https://clts.clld.org in the form of a database in the well-known *Cross-Linguistically Linked Data* framework (http://clld.org, Haspelmath and Forkel 2015), which provides interested users with the common look and feel of popular CLLD datasets such as Glottolog or WALS. There is also a web application, available at http://calc.digling.org/clts/, that allows users to quickly check if their data conforms to the standards defined in our database. More information on the Python API can be found in Appendices B. The full source code is available online at https://zenodo.org/record/1623511.

4. Examples

4.1. Normalisation and parsing of sounds

In order to illustrate how the parsing algorithm underlying CLTS works, let us consider the grapheme ‹^{w}t^{s}:^{h}› as a fictitious example which we want to parse with the B(road)IPA system of CLTS. In a first step, the algorithm normalises the grapheme, thereby replacing the normal colon ‹:› by its correct IPA equivalent ‹ː›. The colon is often confused with the correct IPA counterpart, and often we find both the colon and the correct glyph in the same dataset (e.g., in APiCS). The remaining sequence ‹^{w}t^{s}ː^{h}› is now tested for direct matches with the table of pre-defined base graphemes of BIPA. Since the algorithm does not find the sequence, it will apply a regular expression to check against potential base grapheme candidates and select the longest grapheme. In our case, this is the sequence ‹t^{s}› which itself is flagged as an alias whose correct version is ‹ts›. In terms of features, this sound is defined as a *voiceless alveolar sibilant affricate consonant*. Two subsequences are remaining, the ‹^{w}› to the left, and ‹ː^{h}› to the right. The first can be directly mapped to the feature value *pre-labialised*, the second subsequence maps to *long* and *aspirated*, respectively. The algorithm now assembles all features to a feature bundle and sorts them according to the pre-defined order of features when writing a grapheme. The resulting sound is now described as a *pre-labialised aspirated long voiceless alveolar sibilant affricate consonant* and the grapheme representation in BIPA is given as ‹^{w}ts^{h}ː›. The sound will be labeled as both *normalised* and *aliased*, accounting for the correction of the homoglyph ‹:›, the alias ‹t^{s}›, and the order of the original grapheme.

Table 4: Parsing examples for the CLTS algorithm.

Input	Norm.	Alias	Base	BIPA	Name
a:	: → :	–	–	a:	long unrounded open front vowel
t:s	: → :	t:s → ts:	–	ts:	long voiceless alveolar sibilant affricate consonant
k^{hw}	–	–	k	k^{wh}	labialised aspirated voiceless velar stop consonant
t^{hy}	^{y} → ^{j}		t	t^{jh}	palatalised aspirated voiceless alveolar stop consonant
t:^{sh}	–	–	t	?	unknown sound (‹^{s}› is not defined as a diacritic)

Table 4 gives more illustrations of the algorithm by showing the different stages of normalisation, alias lookup, identification of the base grapheme, and generation of the target sound. The last sound in the table cannot be parsed with the current transcription system, since the diacritic ‹^{s}› in the grapheme ‹t:^{sh}› is not defined as a valid diacritic (as its interpretation would be ambiguous, since in many transcription systems it is only used in combination with alveolars and dentals to indicate an affricate).

4.2. Looking at transcription datasets through CLTS

Table 2 above provides some general statistics regarding the number of graphemes which we find in the original transcription data, the number of items we could link to CLTS, and the number of unique sounds which we identify. The general statistics reveal a rather disappointing situation: instead of providing largely similar collections of graphemes for the speech sounds collected in the different transcription datasets, we find that only a small proportion effectively overlaps, blowing the number of supposedly unique sounds up to as many as 8754. While this might point to errors in our system, we are confident that it instead displays the general nature of linguistic transcription data, given that the 17403 graphemes of all transcription datasets themselves amount to 12384 unique graphemes *without* CLTS. We further checked the majority of the graphemes manually, finding that it is not the failure of the framework to merge sounds for which spelling variants exist, but rather the fact that many datasets list large numbers of sounds one might judge to be unlikely to be produced in any language and which are of low frequency in their respective datasets. These might well reflect idiosyncrasies of interpretation rather than real variation.

A further factor contributing to the large number of sounds in CLTS are transcription datasets like Nidaba and PanPhon which were at least in part automatically created in order to allow one to recognise and provide features for sounds which were not yet accounted for in the data. Since the CLTS framework has a strong generative component, linking these datasets to our framework is useful for two reasons. First, it allows us to generate a large number of potential sounds that might have already been used in some datasets we have not yet included and will help scholars in linking their data to CLTS. Second, it offers a test for the generative strength of our system. Since CLTS so far creates many more potential sounds, which can be uniquely identified, this is an important proof of concept that our system is already capable of integrating many different transcription datasets in an almost completely automated manner.

What we can also learn from linking transcription data to CLTS are obvious errors in the original datasets. Many datasets, for example, provide different graphemes for what CLTS assigns to the same sound. Examples are ‹ts› vs. ‹tˢ› for the voiceless alveolar sibilant affricate consonant in the Eurasian dataset, since ‹tˢ› only occurs one time in the data, and is assigned to Danish, where it reflects phonological convention rather than real pronunciation. Many datasets also confuse the order of diacritics, thus listing ‹kʰʷ› and ‹kʷʰ› as two separate sounds (Phoible, LAPSyD, Diachronica). Other datasets distinguish ‹t͡ʂ› from ‹ʈʂ› (Eurasian, PoWoCo, PBase), of which the latter is defined as alias in the B(road)IPA of CLTS and thus described as *voiceless retroflex sibilant affricate consonant*. Since CLTS normalises the order of diacritics, and provides a large alias system for the BIPA transcription system, these errors can be easily detected and help to improve future versions of the respective datasets.

5. Outlook

Given the theoretical difficulties inherent in phonetic transcription (elaborated in Section 2.3), readers may ask themselves whether linguistics really needs a reference catalogue such as the one we present here. Apart from the immediate benefit of increasing the comparability of large transcription datasets, which we have illustrated above, we see many interesting use-cases for our framework. Given the various methods for normalisation that CLTS offers, the framework can help scholars working with transcriptions to improve their data considerably. This does not only apply to the large phoneme inventory datasets, which

can directly profit from the problems which were identified when linking them to CLTS, but also to the increasing numbers of digitally available lexical datasets resulting from retro-digitisation of older sources or recent field work. With a growing interest in computer-assisted applications in historical linguistics and lexical typology, especially in automated methods for the identification of cognate words (List et al. 2017; Jäger et al. 2017), there is also an increased need for high-quality transcriptions that can be easily parsed by algorithms. With its inbuilt feature system and the feature systems supplied as metadata with the transcription datasets, providing coverage for a large number of sounds, advanced methods for cognate detection and linguistic reconstruction can be easily designed and tested. Last but not least, CLTS also has an educational component, since it rigorously exposes variation across transcription datasets, bringing the need for consistency and adherence to standards to our attention.

References

Anonymous. 2014. *Index Diachronica*. <https://chridd.nfshost.com/diachronica/>

Bell, A. 1867. *Visible speech: The science of universal alphabetics: Or, self-interpreting physiological letters, for the writing of all languages in one alphabet. Illustrated by tables, diagrams, and examples*. London: Simpkin, Marshall.

Běijīng Dàxué 北京大学. 1964. *Hànyǔ fāngyán cíhuì* [Chinese dialect vocabularies]. Běijīng: Wénzì Gǎigé 文字改革.

Bybee, J. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.

Chao, Y. 2006. A system of 'tone letters'. In: Wu, Z.-J. and X.-N. Zhao (eds.), *Linguistic essays by Yuenren Chao*. Běijīng: Shāngwù. 98–102.

Charpentier, J.-M. and A. François. 2015. *Linguistic atlas of French Polynesia / Atlas linguistique de la Polynésie française*. Berlin, Boston: De Gruyter Mouton.

Chomsky, N. and M. Halle. 1968. *The sound pattern of English*. New York: Harper and Row.

Crowley, T. 2006. *The Avava Language of Central Malakula (Vanuatu)*. The Australian National University: Pacific Linguistics, Research School of Pacific and Asian Studies.

Crowley, T. 2006. *Nese: A Diminishing Speech Variety of Northwest Malakula (Vanuatu)*. The Australian National University: Pacific Linguistics, Research School of Pacific and Asian Studies.

Dediu, D. and S. Moisik. 2016. Defining and counting phonological classes in cross-linguistic segment databases. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation*. 1955–1962.

Dench, A. 2002. Descent and diffusion: The complexity of the Pilbara situation. In: Aikhenvald, A. and R. Dixon (eds.), *Areal diffusion and genetic inheritance: Problems in comparative linguistics*. Oxford: Oxford University Press. 105–133.
Dodd, R. 2014. V'ënen Taut: Grammatical topics in the Big Nambas Language of Malekula. (PhD dissertation, University of Waikato.)
Dolgopolsky, A. 1964. Gipoteza drevnejšego rodstva jazykovych semej Severnoj Evrazii s verojatnostej točky zrenija [A probabilistic hypothesis concering the oldest relationships among the language families of Northern Eurasia]. *Voprosy Jazykoznanija* 2. 53–63.
Dryer, M. and M. Haspelmath. 2011. *The World Atlas of Language Structures online*. Munich: Max Planck Digital Library.
Eden, E. 2018. Measuring phonological distance between languages. (PhD dissertation, University College London.)
Güldemann, T. 2001. Phonological regularities of consonant systems across Khoesan lineages. *University of Leipzig Papers on Africa* 16. 1–50.
Güldemann, T. 2014. 'Khoisan' linguistic classification today. In: Güldemann, T. and A.-M. Fehn (eds.), *Beyond 'Khoisan'. Historical relations in the Kalahari Basin*. Amsterdam and Philadelphia: John Benjamins. 1–40.
Hammarström, H., R. Forkel, and M. Haspelmath. 2017. Glottolog. Version 3.0. Leipzig: Max Planck Institute for Evolutionary Anthropology.
Haspelmath, M. 2010. Comparative concepts and descriptive categories. *Language* 86(3). 663–687.
Haspelmath, M. and R. Forkel. 2015. CLLD – Cross-Linguistic Linked Data. Max Planck Institute for Evolutionary Anthropology: Leipzig.
Herzog, G., S. Newman, E. Sapir, M. Swadesh, M. Swadesh, and C. Voegelin. 1934. Some orthographic recommendations. *American Anthropologist* 36(4). 629–631.
Honeybone P. 2005. Diachronic evidence in segmental phonology: The case of laryngeal specifications. In: van Oostendorp, M. and J. van de Weijer (eds.), *The internal organisation of phonological segments*. Mouton de Gruyter: Berlin and New York. 319–354.
Hóu Jīngyī 侯精一 (ed.). 2004. *Xiàndài Hànyǔ fāngyán yīnkù* 现代汉语方言音库 [Phonological database of Chinese dialects]. Shànghǎi 上海: Shànghǎi Jiàoyù 上海教育.
Huáng, B. and X. Liào. 2002. *Xiàndài Hànyǔ* 现代汉语 [Modern Chinese]. Běijīng: Gāoděng Jiàoyù.
International Institute of African Languages and Cultures. 1930. *Practical orthography of African languages*. (Revised edition.) Oxford: Oxford University Press.
International Phonetic Association. 1912. *The Principles of the International Phonetic Association*. Bourg-la-Reine and London: Paul Passy and Daniel Jones.
International Phonetic Association. 1999. *Handbook of the International Phonetic Association*. Cambridge: Cambridge University Press.
International Phonetic Association. 2015. *The International Phonetic Alphabet*. (Revised to 2015.)
Department of Linguistics. 2017. *Multimedia IPA chart*. Victoria: University of Victoria.

Jacob, J.M. 1963. Prefixation and infixation in old Mon, old Khmer, and modern Khmer. *Linguistic comparison in Southeast Asia and the Pacific*. 62–70.
Jäger, G., J.-M. List and P. Sofroniev. 2017. Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. (Long papers.) 1204–1215.
Kalusky, W. 2017. *Die Transkription der Sprachlaute des Internationalen Phonetischen Alphabets: Vorschläge zu einer Revision der systematischen Darstellung der IPA-Tabelle*. München: LINCOM Europa.
Kieviet, P. 2017. *A Grammar of Rapa Nui*. Berlin: Language Science Press.
Köhler, O., P. Ladefoged, J. Snyman, A. Traill and R. Vossen. 1988. The symbols for clicks. *Journal of the International Phonetic Association* 18(2). 140–142.
Kümmel, M. 2008. *Konsonantenwandel* [Consonant change]. Reichert: Wiesbaden.
Lepsius, C. 1854. *Das allgemeine linguistische Alphabet: Grundsätze der Übertragung fremder Schriftsysteme und bisher noch ungeschriebener Sprachen in europäische Buchstaben*. Wilhelm Hertz: Berlin.
List, J.-M. 2014. *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.
List, J.-M. and J. Prokić. 2014. A benchmark database of phonetic alignments in historical linguistics and dialectology. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. 288–294.
List, J.-M., M. Cysouw, and R. Forkel. 2016. Concepticon. A resource for the linking of concept lists. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. 2393–2400.
List, J.-M., S. Greenhill, and R. Gray. 2017. The potential of automatic word comparison for historical linguistics. *PLOS ONE* 12(1). 1–18.
Lynch, J. 2016. Malakula internal subgrouping: Phonological evidence. *Oceanic Linguistics* 55(2). 399–431.
Maddieson, I. 1984. *Patterns of sounds*. Cambridge: Cambridge University Press.
Maddieson, I., S. Flavier, E. Marsico, C. Coupé and F. Pellegrino. 2013. LAPSyD: Lyon-Albuquerque Phonological Systems Database. In: *Proceedings of Interspeech*.
Malau, C. 2016. *A grammar of Vurës, Vanuatu*. Berlin: Walter de Gruyter:
Mann, M. and D. Dalby. 1987. *A thesaurus of African languages: A classified and annotated inventory of the spoken languages of Africa with an appendix on their written representation*. London: Zell Publishers.
Michaelis, S., P. Maurer, M. Haspelmath and M. Huber. 2013. *The Atlas of Pidign and Creole language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.
Mielke, J. 2008. *The emergence of distinctive features*. Oxford: Oxford University Press.
Moran, S. and M. Cysouw. 2017. *The Unicode cookbook for linguists. Managing writing systems using Orthography Profiles*. Zürich: Zenodo.
Moran, S., D. McCloy and R. Wright (eds.). 2014. *PHOIBLE Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.
Mortensen, D. 2017. *PanPhon. Python API for accessing phonological features of IPA Segments*. Pittsburgh: Carnegie Mellon School of Computer Science.

Nakagawa, H. 2006. Aspects of the phonetic and phonological structure of the Gui language. (PhD dissertation, University of the Witwatersrand, Johannesburg.)
Nikolaev, D., A. Nikulin and A. Kukhto. 2015. *The database of Eurasian phonological inventories*. Moscow: RGGU. <http://eurasianphonology.info>
Press, M. L. 1980. *Chemehuevi: A grammar and lexicon*. Berkeley: University of California Press.
Pullum, G. and W. Ladusaw. 1996. *Phonetic symbol guide*. Chicago: University of Chicago Press.
Ruhlen, M. 2008. *A global linguistic database*. Moscow: RGGU.
Salisbury, M.C. 2002. A grammar of Pukapukan. (PhD dissertation, The University of Auckland.)
Sapir, E. 1930. *Southern Paiute, a Shoshonean language*. Boston: Academic Press.
Saussure, F. de. 1878. *Mémoire sur le système primitif des voyelles dans les langues indo-européennes*. Leipzig: Teubner.
Saussure, F. de. 1916. *Cours de linguistique générale*. Lausanne: Payot.
Setälä, E. 1901. Über transskription der finnisch-ugrischen sprachen. *Finnisch-ugrische Forschungen* 1. 15–52.
Simpson, A. 1999. Fundamental problems in comparative phonetics and phonology: does UPSID help to solve them. In: *Proceedings of the 14th international congress of phonetic sciences*.
Starostin, G. and P. Krylov (eds.). 2011. *The global lexicostatistical database. Compiling, clarifying, connecting basic vocabulary around the world: From free-form to tree-form*. <http://starling.rinet.ru/new100/main.htm>
Starostin, G. (ed.) 2017. *Annotated Swadesh wordlists for the Hmong group* (Hmong-Mien family).
Stimson, J. F. and D.S. Marshall. 1964. *A dictionary of some Tuamotuan dialects of the Polynesian language*. Leiden: M. Nijhoff.
Sweet, H. 1877. *A handbook of phonetics, including a popular exposition of the principles of spelling reform*. Oxford: Clarendon Press.
Tadadjeu, M. and E. Sadembouo. 1979. *Alphabet Générale des langues Camerounaises*. Yaoundé: Departement des Langues Africaines et Linguistique, Université de Yaoundé.
Traill A. 1993. The feature geometry of clicks. In: van Staden, P.M.S. (ed.), *Linguistica: Festschrift E. B. van Wyk: 'n huldeblyk*. Pretoria: van Schaik. 134–140.
Tregear, E. 1899. *Dictionary of Mangareva: Or Gambier Islands*. Wellington: J. Mackay.
Trubetzkoy, N. 1939. *Grundzüge der Phonologie* [Foundations of phonology]. Prague: Cercle Linguistique de Copenhague.
UNESCO. 1978. African languages. In: *Proceedings of the meeting of experts on the transcription and harmonization of African languages*.
Wichmann, S., E. Holman and C. Brown. 2016. *The ASJP database*. Jena: Max Planck Institute for the Science of Human History.
Wikipedia contributors. 2018. International Phonetic Alphabet. Wikipedia, The Free Encyclopedia. <https://en.wikipedia.org/w/index.php?title=International_Phonetic_Alphabet&oldid=822828531>. Accessed 29 Jan 2018.

Acknowledgments

JML and TT were funded by the the ERC Starting Grant 715618 "Computer-Assisted Language Comparison" (http://calc.digling.org). We thank Gereon Kaiping for providing early support in testing and discussing the *pyclts* software package. We thank Adrian Simpson, Martin Haspelmath, Ludger Paschen, and Paul Heggarty for helpful comments on earlier versions of this draft, and we thank Simon J. Greenhill and Christoph Rzymski for providing support with the software.

Software and data

Software and data accompanying this paper have been hosted with Zenodo and can be found at https://doi.org/10.5281/zenodo.1617697. The source code for the Python API is curated on GitHub at https://github.com/cldf/clts/. The data can be further inspected and conveniently browsed at https://clts.clld.org.

Appendix

Current feature system underlying the CLTS framework.

Sound type	Feature	Value
vowel	relative_articulation	centralized
vowel	relative_articulation	mid-centralized
vowel	relative_articulation	advanced
vowel	relative_articulation	retracted
vowel	centrality	back
vowel	centrality	central
vowel	centrality	front
vowel	centrality	near-back
vowel	centrality	near-front

Sound type	Feature	Value
vowel	creakiness	creaky
vowel	rounding	less-rounded
vowel	rounding	more-rounded
vowel	stress	primary-stress
vowel	stress	secondary-stress
vowel	pharyngealization	pharyngealized
vowel	rhotacization	rhotacized
vowel	voicing	devoiced
vowel	nasalization	nasalized
vowel	syllabicity	non-syllabic
vowel	raising	lowered
vowel	raising	raised
vowel	height	close
vowel	height	close-mid
vowel	height	mid
vowel	height	near-close
vowel	height	near-open
vowel	height	open
vowel	height	open-mid
vowel	frication	with-frication
vowel	roundedness	rounded
vowel	roundedness	unrounded
vowel	duration	long
vowel	duration	mid-long
vowel	duration	ultra-long
vowel	duration	ultra-short

Sound type	Feature	Value
vowel	velarization	velarized
vowel	tongue_root	advanced-tongue-root
vowel	tongue_root	retracted-tongue-root
vowel	tone	with_downstep
vowel	tone	with_extra-high_tone
vowel	tone	with_extra-low_tone
vowel	tone	with_falling_tone
vowel	tone	with_global_fall
vowel	tone	with_global_rise
vowel	tone	with_high_tone
vowel	tone	with_low_tone
vowel	tone	with_mid_tone
vowel	tone	with_rising_tone
vowel	tone	with_upstep
vowel	articulation	strong
vowel	breathiness	breathy
vowel	glottalization	glottalized
consonant	aspiration	aspirated
consonant	sibilancy	sibilant
consonant	creakiness	creaky
consonant	release	unreleased
consonant	release	with-lateral-release
consonant	release	with-mid-central-vowel-release
consonant	release	with-nasal-release
consonant	ejection	ejective
consonant	place	alveolar

Sound type	Feature	Value
consonant	place	alveolo-palatal
consonant	place	bilabial
consonant	place	dental
consonant	place	epiglottal
consonant	place	glottal
consonant	place	labial
consonant	place	linguolabial
consonant	place	labio-palatal
consonant	place	labio-velar
consonant	place	labio-dental
consonant	place	palatal
consonant	place	palatal-velar
consonant	place	pharyngeal
consonant	place	post-alveolar
consonant	place	retroflex
consonant	place	uvular
consonant	place	velar
consonant	pharyngealization	pharyngealized
consonant	voicing	devoiced
consonant	voicing	revoiced
consonant	nasalization	nasalized
consonant	preceding	pre-aspirated
consonant	preceding	pre-glottalized
consonant	preceding	pre-labialized
consonant	preceding	pre-nasalized
consonant	preceding	pre-palatalized

Sound type	Feature	Value
consonant	labialization	labialized
consonant	syllabicity	syllabic
consonant	palatalization	labio-palatalized
consonant	palatalization	palatalized
consonant	phonation	voiced
consonant	phonation	voiceless
consonant	duration	long
consonant	duration	mid-long
consonant	stress	primary-stress
consonant	stress	primary-stress
consonant	stress	primary-stress
consonant	stress	primary-stress
consonant	stress	secondary-stress
consonant	laterality	lateral
consonant	velarization	velarized
consonant	manner	affricate
consonant	manner	approximant
consonant	manner	click
consonant	manner	fricative
consonant	manner	implosive
consonant	manner	nasal
consonant	manner	nasal-click
consonant	manner	stop
consonant	manner	tap
consonant	manner	trill
consonant	laminality	apical

Sound type	Feature	Value
consonant	laminality	laminal
consonant	articulation	strong
consonant	breathiness	breathy
consonant	glottalization	glottalized
consonant	raising	lowered
consonant	raising	raised
consonant	relative_articulation	centralized
consonant	relative_articulation	mid-centralized
consonant	relative_articulation	advanced
consonant	relative_articulation	retracted
tone	middle	via-high
tone	middle	via-low
tone	middle	via-mid
tone	middle	via-mid-high
tone	middle	via-mid-low
tone	start	from-high
tone	start	from-low
tone	start	from-mid
tone	start	from-mid-high
tone	start	from-mid-low
tone	start	neutral
tone	contour	contour
tone	contour	falling
tone	contour	flat
tone	contour	rising
tone	contour	short

Sound type	Feature	Value
tone	end	to-high
tone	end	to-low
tone	end	to-mid
tone	end	to-mid-high
tone	end	to-mid-low

Corresponding author:
Johann-Mattis List
Department of Linguistic and Cultural Evolution
Max Planck Institute for the Science of Human History
Kahlaische Straße 10
07745 Jena
Germany
list@shh.mpg.de

3.2 Annotation in Historical Linguistics

In addition to data standardization, data annotation is of crucial importance for the framework of computer-assisted language comparison. The major idea is that annotation is carried out with help of software tools that make it easier for linguists to manually process the data by correcting errors introduced by automated approaches, or by annotating data from scratch. Apart from making the manual annotation work of linguists a lot easier, annotation tools play another important role in computer-assisted language comparison, in so far as they make it possible to check the input from linguists upon submission, which allows to avoid various kinds of errors, resulting from wrong spellings or other inconsistencies.

The following two studies present a new framework for the annotation of etymological data in historical linguistics. It is based on web-based applications that make it possible to use the tools across all major platforms and employs general ideas for the annotation of historical linguistic data that have been developed in previous studies (List 2016). The first study, titled "A Web-Based Interactive Tool for Creating, Inspecting, Editing, and Publishing Etymological Datasets" (List 2017), presents the EDICTOR tool (`https://digling.org/edictor`), a web-based tool for the annotation of etymological datasets. The second study, titled "Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages" (Hill and List 2017), concentrates on some major issues related to the annotation of *partial cognates*, as they often result from compounding processes in South-East Asian languages. In addition to explaining how these relations can be annotated with help of the web-based EDICTOR application, it also introduces some novel ways of analysing the data, once annotation has been sufficiently carried out.

A Web-Based Interactive Tool for Creating, Inspecting, Editing, and Publishing Etymological Datasets

Johann-Mattis List
Max Planck Institute for the Science of Human History
Kahlaische Straße 10
07745 Jena
list@shh.mpg.de

Abstract

The paper presents the Etymological DICtionary ediTOR (EDICTOR), a free, interactive, web-based tool designed to aid historical linguists in creating, editing, analysing, and publishing etymological datasets. The EDICTOR offers interactive solutions for important tasks in historical linguistics, including facilitated input and segmentation of phonetic transcriptions, quantitative and qualitative analyses of phonetic and morphological data, enhanced interfaces for cognate class assignment and multiple word alignment, and automated evaluation of regular sound correspondences. As a web-based tool written in JavaScript, the EDICTOR can be used in standard web browsers across all major platforms.

1 Introduction

The amount of large digitally available datasets for various language families is constantly increasing. In order to analyse these data, linguists turn more and more to automatic approaches. Phylogenetic methods from biology are now regularly used to create evolutionary trees of language families (Gray and Atkinson, 2003). Methods for the comparison of biological sequences have been adapted and allow to automatically search for cognate words in multilingual word lists (List, 2014) and to automatically align them (List, 2014). Complex workflows are used to search for deep genealogical signals between established language families (Jäger, 2015).

In contrast to the large arsenal of software for automatic analyses, the number of tools helping to *manually* prepare, edit, and correct lexical datasets in historical linguistics is extremely rare. This is surprising, since automatic approaches still lag behind expert analyses (List et al., 2017). Tools for data preparation and evaluation would allow experts to directly interact with computational approaches by manually checking and correcting their automatically produced results. Furthermore, since the majority of phylogenetic approaches makes use of manually submitted expert judgments (Gray and Atkinson, 2003), it seems indispensable to have tools which ease these tasks.

2 The EDICTOR Tool

The Etymological DICtionary ediTOR (EDICTOR) is a free, interactive, web-based tool that was specifically designed to serve as an interface between quantitative and qualitative tasks in historical linguistics. Inspired by powerful features of STARLING (Starostin, 2000) and RefLex (Segerer and Flavier, 2015), expanded by innovative new features, and based on a very simple data model that allows for a direct integration with quantitative software packages like LingPy (List and Forkel, 2016), the EDICTOR is a lightweight but powerful toolkit for computer-assisted applications in historical linguistics.

2.1 File Formats and Data Structure

The EDICTOR was designed as a lightweight file-based tool that takes a text file as input, allowing to modify and save it. The input format is a plain tab-separated value (TSV) file, with a *header* indicating the value of the columns. This format is essentially identical with the format used in LingPy. Although the EDICTOR accepts all regular TSV files as input, its primary target are *multi-lingual word lists*, that is, datasets in which a given number of *concepts* has been translated into a certain range of *target languages*.

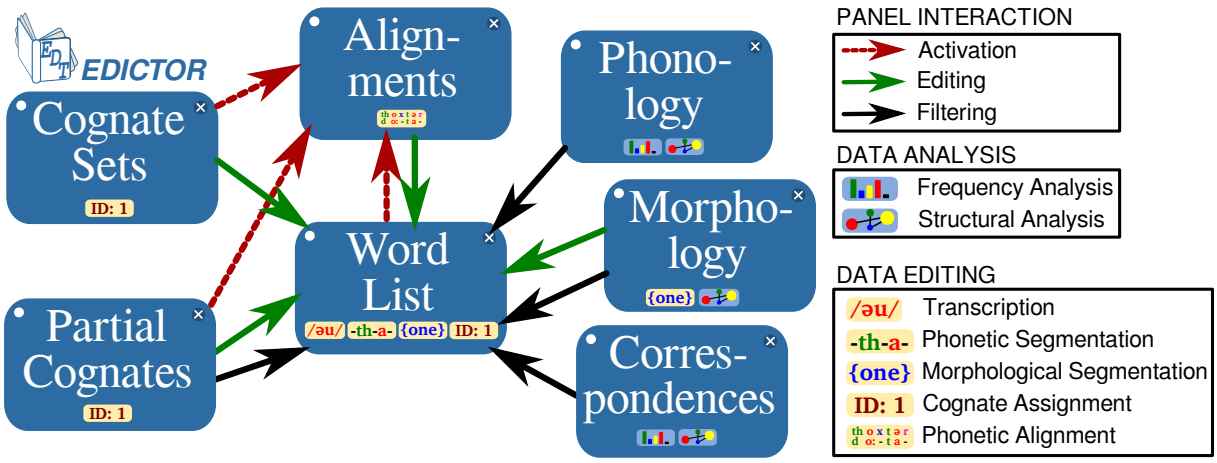


Figure 1: Basic panel structure of the EDICTOR.

ID	DOCULECT	CONCEPT	TRANSCRIPTION	...
1	German	Woldemort	valdəmar	...
2	English	Woldemort	wɔldəmɔrt	...
3	Chinese	Woldemort	fu^{51}ti^{51}mɔ^{35}	...
4	Russian	Woldemort	vladimir	...
...	...	...	...	...
10	German	Harry	haralt	...
11	English	Harry	hæri	...
12	Russian	Harry	gali	...
...	...	...	...	...

Figure 2: Basic file format in the EDICTOR

2.2 User Interface

The EDICTOR is divided into different *panels* which allow to edit or analyse the data in different ways. The core module is the *Word List panel* which displays the data in its original form and can be edited and analysed as one knows it from spreadsheet applications. For more complex tasks of data editing and analysis, such as cognate assignment or phonological analysis, additional panels are provided. Specific modes of interaction between the different panels allow for a flexible interaction between different tasks. Using drag-and-drop, users can arrange the panels individually or hide them completely. Figure 1 illustrates how the major panels of the EDICTOR interact with each other.

2.3 Technical Aspects

The EDICTOR application is written in plain JavaScript and was tested in Google Chrome, Firefox, and Safari across different operating systems (Windows, MacOS, Linux). For the purpose of offline usage, users can download the source code. For direct online usage, the tool can be accessed via its project website.

3 Data Editing in the EDICTOR

3.1 Editing Word List Data

Editing data in the Word List panel of the EDICTOR is straightforward by inserting values in text-fields which appear when clicking on a given field or when browsing the data using the arrow keys of the keyboard. Additional keyboard shortcuts allow for quick browsing. For specific data types, automatic operations are available which facilitate the input or test what the user inserts. Transcription, for example supports SAMPA-input. The segmentation of phonetic entries into meaningful sound units is also carried out automatically. Sound segments are highlighted with specific background colors based on their underlying sound class and sounds which are not recognized as valid IPA symbols are highlighted in warning colors (see the illustration in Figure 3). The users can decide themselves in which fields they wish to receive automatic support, and even Chinese input using an automatic Pīnyīn converter is provided.

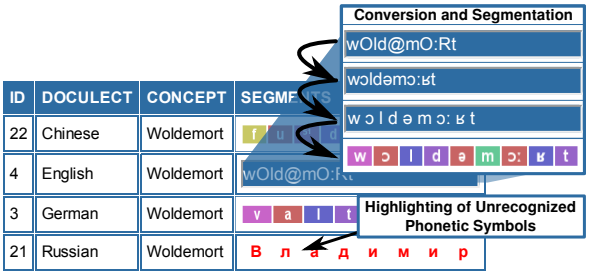


Figure 3: Editing word lists in the EDICTOR

3.2 Cognate Assessment

Defining which words in multilingual word lists are cognate is still a notoriously difficult task for machines (List, 2014). Given that the majority of datasets are based on manually edited cognate judgments, it is important to have tools which facilitate this task while at the same time controlling for typical errors. The EDICTOR offers two ways to edit cognate information, the first assuming complete cognacy of the words in their entirety, and the second allowing to assign only specific parts of words to the same cognate set. In order to carry out partial cognate assignment, the data needs to be morphologically segmented in a first stage, for example with help of the Morphology panel of the EDICTOR (see Section 4.2). For both tasks, simple and intuitive interfaces are offered which allow to browse through the data and to assign words to the same cognate set.

Figure 4: Aligning words in the EDICTOR

3.3 Phonetic Alignment

Since historical-comparative linguistics is essentially based on sequence comparison (List, 2014), alignment analyses, in which words are arranged in a matrix in such a way that corresponding sounds are placed in the same column, are underlying all cognate sets. Unfortunately they are rarely made explicit in classical etymological dictionaries. In order to increase explicitness, the EDICTOR offers an alignment panel. The alignment panel is essentially realized as a pop-up window showing the sounds of all sequences which belong to the same cognate set. Users can edit the alignments by moving sound segments with the mouse. Columns of the alignment which contain unalignable parts (like suffixes or prefixes) can be explicitly marked as such. In addition to manual alignments, the EDICTOR offers a simple alignment algorithm which can be used to pre-analyse the alignments. Figure 4 shows an example for the alignment of four fictive cognates in the EDICTOR.

4 Data Analysis in the EDICTOR

4.1 Analysing Phonetic Data

Errors are inevitable in large datasets, and this holds also and especially for phonetic transcriptions. Many errors, however, can be easily spotted by applying simple sanity checks to the data. A straightforward way to check the consistency of the phonetic transcriptions in a given dataset is provided in the Phonology panel of the EDICTOR. Here all sound segments which occur in the segmented transcriptions of one language are counted and automatically compared with an internal set of IPA segments. Counting the frequency of segments is very helpful to spot simple typing errors, since segments which occur only one time in the whole data are very likely to be errors. The internal segment inventory adds a structural perspective: If segments are found in the internal inventory, additional phonetic information (manner, place, etc.) is shown, if segments are missing, this is highlighted. The results can be viewed in tabular form and in form of a classical IPA chart.

4.2 Analysing Morphological Data

The majority of words in all languages consist of more than one morpheme. If historically related words differ regarding their morpheme structure, this poses great problems for automatic approaches to sequence comparison, since the algorithms usually compare words in their entirety. German *Großvater* ‘grandfather’, for example, is composed of two different morphemes, *groß* ‘large’ and *Vater* ‘father’. In order to analyse multi-morphemic words historically, it is important to carry out a morphological annotation analysis. In order to ease this task, the Morphology panel of the EDICTOR offers a variety of straightforward operations by which morpheme structure can be annotated and analysed at the same time. The core idea behind all operations is a search for similar words or morphemes in the same language. These *colexifications* are then listed and displayed in form of a bipartite *word family network* in which words are linked to morphemes, as illustrated in Figure 5. The morphology analysis in the EDICTOR is no miracle cure for morpheme detection, and morpheme boundaries need still to be annotated by the user. However, the dynamically produced word family networks as well as the explicit listing of words sharing the same subsequence of sounds greatly facilitates this task.

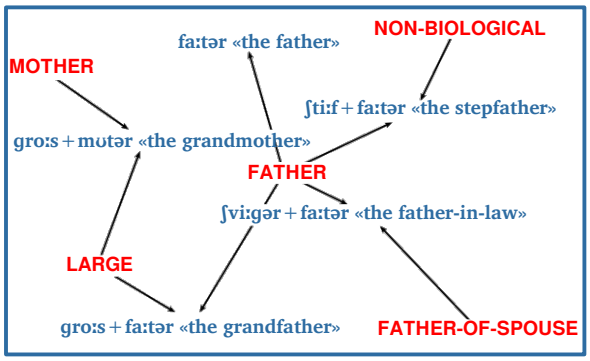


Figure 5: Word family network in the EDICTOR: The morphemes (in red) link the words around German *Großvater* 'grandfather' (in blue).

4.3 Analysing Sound Correspondences

Once cognate sets are identified and aligned, searching for regular sound correspondences in the data is a straightforward task. The Correspondences panel of the EDICTOR allows to analyse sound correspondence patterns across pairs of languages. In addition to a simple frequency count, however, *conditioning context* can be included in the analysis. Context is modeled as a separate string that provides abstract context symbols for each sound segment of a given word. This means essentially that context is handled as an additional *tier* of a sequence. This multi-tiered representation is very flexible and also allows to model suprasegmental context, like tone or stress. If users do not provide their own tiers, the EDICTOR employs a default context model which distinguishes consonants in syllable onsets from consonants in syllable offsets.

5 Customising the EDICTOR

The EDICTOR can be configured in multiple ways, be it while editing a dataset or before loading the data. The latter is handled via URL parameters passed to the URL that loads the application. In order to facilitate the customization procedure, a specific panel for customisation allows the users to define their default settings and creates a URL which users can bookmark to have quick access to their preferred settings.

The EDICTOR can be loaded in read-only mode by specifying a "publish" parameter. Additionally, server-side files can be directly loaded when loading the application. This makes it very simple and straightforward to use the EDICTOR to publish raw etymological datasets in a visually appealing format as can be seen from this exemplary URL: `http://edictor.digling.org?file=Tujia.tsv&publish=true&preview=500`.

6 Conclusion and Outlook

This paper presented a web-based tool for creating, inspecting, editing, and publishing etymological datasets. Although many aspects of the tool are still experimental, and many problems still need to be solved, I am confident that – even in its current form – the tool will be helpful for those working with etymological datasets. In the future, I will develop the tool further, both by adding more useful features and by increasing its consistency.

Acknowledgements

This research was supported by the DFG research fellowship grant 261553824 *Vertical and lateral aspects of Chinese dialect history*. I thank Nathan W. Hill, Guillaume Jacques, and Laurent Sagart for testing the prototype.

References

Russell D. Gray and Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965):435–439.

Gerhard Jäger. 2015. Support for linguistic macrofamilies from weighted alignment. *PNAS*, 112(41):1275212757.

Johann-Mattis List and Robert Forkel. 2016. *LingPy*. Max Planck Institute for the Science of Human History, Jena. URL: `http://lingpy.org`.

Johann-Mattis List, Simon J. Greenhill, and Russell D. Gray. 2017. The potential of automatic word comparison for historical linguistics. *PLOS ONE*, 12(1):1–18, 01.

Johann-Mattis List. 2014. *Sequence comparison in historical linguistics*. Düsseldorf University Press, Düsseldorf.

Guillaume Segerer and S. Flavier. 2015. *RefLex*. Laboratoire DDL, Paris and Lyon. URL: `http://reflex.cnrs.fr`.

Sergej Anatol'evič Starostin. 2000. *STARLING*. RGGU, Moscow. URL: `http://starling.rinet.ru`.

Supplementary Material

Supplements for this paper contain a demo video (`https://youtu.be/IyZuf6SmQM4`), the application website (`http://edictor.digling.org`), the source (v. 0.1, `https://zenodo.org/record/48834`), and the development version (`http://github.com/digling/edictor`).

Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages

Nathan W. Hill[a] and Johann-Mattis List[b]
[a]SOAS, London
[b]Max-Planck-Institute for the Science of Human History, Jena
[b]mattis.list@shh.mpg.de

Abstract

The use of computational methods in comparative linguistics is growing in popularity. The increasing deployment of such methods draws into focus those areas in which they remain inadequate as well as those areas where classical approaches to language comparison are untransparent and inconsistent. In this paper we illustrate specific challenges which both computational and classical approaches encounter when studying South-East Asian languages. With the help of data from the Burmish language family we point to the challenges resulting from missing annotation standards and insufficient methods for analysis and we illustrate how to tackle these problems within a computer-assisted framework in which computational approaches are used to pre-analyse the data while linguists attend to the detailed analyses.

Keywords: historical linguistics, linguistic reconstruction, Burmish languages, annotation, analysis, computer-assisted language comparison

1. Introduction

The quantitative turn in historical linguistics created a gap between "new and innovative" quantitative methods and classical approaches. Classical linguists are often skeptical of the new approaches, partly because the results do not seem to coincide with those of classical methods (Holm 2007), partly because they only confirm well established findings (Campbell 2013: 485f). Computational linguists, on the other hand, complain about inconsistencies in the application of the classical methods (McMahon and McMahon 2005: 26–29).

Both classical and computational approaches have strong and weak points. Steeped in philological learning, classical linguists enjoy extensive knowledge of, and refined intuitions about both common and language-specific processes of language change. Basing their analyses on multiple types of evidence, classical linguists can work out probable solutions even in situations where data are sparse. Their disadvantage is that they have difficulties coping with large amounts of data. The advantage of computational methods is their efficiency and consistency, and thus their ability to handle large amounts of data. The weakness of computational linguists is their tendency to ignore language-specific idiosyncrasies, being accustomed to deal only with homogeneous evidence. For this reason, computational approaches function poorly with sparse data. Since most of the data in historical linguistics are sparse and heterogenous (Sturtevant 1920: 11; Makaev 1977: 88), it is no wonder that the triumphs of computational analyses still lag behind those of classical approaches.

In the following, we concentrate on two specific challenges which both computational and classical historical linguists encounter when working with South-East Asian and specifically Sino-Tibetan (Trans-Himalayan) languages.[1] In particular, we focus on the Burmish languages, a small Sino-Tibetan sub-branch, but the analogous challenges are encountered in South-East Asian languages of other language families. We concentrate on processes of lexical change, pointing to specific challenges of *annotation* (Section 2) and *analysis* (Section 3). We then turn to addressing these problems in the Burmish Etymological Database (BED, https://dighl.github.io/burmish), where we use improved annotation and analysis techniques in order to create an etymological dictionary of the Burmish languages which is amenable to both qualitative and quantitative analyses.

2. Challenges of annotation

In historical linguistics we look back at a tradition of over 200 years of research on language families from around the world. Given this long tradition,

[1] By the term "Sino-Tibetan" we mean that language family of which Chinese, Tibetan, and Burmese are members. We use this term agnostically with regard to the shape of the Stammbaum of this family. Specifically, we see no reason to posit a branch of this family that contains Tibetan and Burmese but not Chinese.

it is surprising that our field still lacks common *annotation guidelines*: a general set of best practices stating how particular findings should be presented. By this, we do not mean the use of certain characters, like the asterisk to indicate that a word is reconstructed and not attested in written or spoken sources (see Koerner 1976 on the history of this practice), but rather a standardized way of how the fundamental findings, such as regular sound correspondences, convincing cognate sets, or shared innovations, are not only presented to the readers in publications, but also handled as data points amenable to statistical analyses. Historical linguistics has always been a data-driven discipline, even in pre-computer times, scholars would develop their individual practice of arranging their data with the help of index cards (see, for example, the detailed description in Gabelentz 1891, as well as his questionnaire for foreign language documentation from 1892, which is discussed in detail in Kürschner 2014) or punch cards (Swadesh 1963). Unfortunately, scholars rarely shared or discussed their practice but instead expected neophytes to learn by doing (Schwink 1994: 29).

The lack of annotation guidelines has immediate consequences both for classical and computational approaches. Computational approaches suffer from ambiguously annotated data which may confuse the algorithms, bound as they are by strict assumptions about the major processes of lexical change. Classical approaches suffer from a lack of transparency in data annotation when it comes to assessing the work of colleagues, especially vis-à-vis proposed regular sound correspondences and cognate sets. Since arguments on cognates and sound correspondences are often presented in an idiosyncratic way that varies not only from subfield to subfield but also among scholars working on the same language family, it is extremely difficult to base discussions on data and conclusions alone. This may be one of the reasons why debates often become personal in historical linguistics: since it is often not entirely clear where two scholars exactly differ, debates drift into polemics with scholars accusing each other of deliberately disregarding major facts.

In the following we quickly point to two major problems of annotation when analysing South-East Asian languages: cognates and sound correspondences. While the former constitutes primarily a problem for computational approaches to phylogenetic reconstruction, the latter is a major drawback for the discussion and evaluation of proposals in classical historical linguistics.

2.1. Partial cognate annotation

Cognacy is not a binary relation and cannot be reduced to a simple yes-no question. Instead, judging whether two words are cognate is both a question of perspective and degree. For example, one can distinguish "root" cognates from "stem" cognates. An example of *root cognates* is French *donner* 'to give' compared to Italian *dare* 'to give'. Both words descend from Proto-Indo-European *$*deh_3$- 'to give', the French indirectly, via a verbalized *no*-stem (PIE *$*deh_3$-*no*- 'that which is given' > Latin *dōnāre* 'to give as present'), the Italian directly (PIE *$*deh_3$ > Latin *dare* 'to give', Meiser 1998). An example of *stem cognates* is the comparison of Italian *dare* and Spanish *dar* 'to give', which both descend directly from Latin *dare*. The relativity of perspective and degree inherent in the notion of cognancy is comparable to the relation of *homology* in evolutionary biology, which denotes a relation of *commen descent* (Koonin 2005: 311). While we can say, for example, that wings in birds and wings in bats are deeply homologous, in so far as both represent the upper limbs of tetrapods, we can also say that they are homoplastic (i.e., independent innovations), in so far as their specific function, allowing tetrapods to fly, has evolved independently (Butler 2000, Morrison 2015).

Even more problematic than the vagaries of root etymology versus stem etymology are cases of *partial cognacy* (List 2015: 42; List 2016). Partial cognacy reflects a situation where words share cognate material only in part, such as French *aujourd'**hui***, which can be seen as partially cognate with Latin *hodiē*, itself a compound of Latin *hic* 'this' and *dies* 'day' (Vaan 2008: 287), of which the latter is again cognate with Ancient Greek Ζεύς [dzeus] (Meier-Bruegger 2002: L303). While partial cognacy generally holds for all root cognates reflected in words with different stems, including the case of French *donner* and Italian *dare*, mentioned above, partial cognacy is most frequently met in languages in which compounding is a frequent and productive process of word formation, such as South-East Asian languages.

As an example from the Burmish languages, consider the translational equivalents for 'yesterday' in Bola, Lashi, Rangoon Burmese, and Xiandao, given in Figure 1. As we have indicated with the aid of font colors, four languages have at least one morpheme in common (Bola [nɛʔ^{31}], Lashi [nap^{31}], Rangoon [ne^{53}] and Xiandao [n̩^{31}] all meaning 'day' in isolation), but only Bola and Lashi share the same compound structure. If we were forced to make a binary cognate decision out of this example, as we must when prepar-

Language	Form	Strict	Loose	Exact
Bola	a^{31} ŋji^{35} nɛʔ^{31}	1	1	1 2 3
Lashi	a^{31} ŋjei^{55} nap^{31}	1	1	1 2 3
Rangoon	mɑ^{53} ne^{53} kɑ^{53}	2	1	0 3 0
Xiandao	n̥^{31} m̥an^{35}	3	1	3 4
Achang	man^{35}	4	1	4

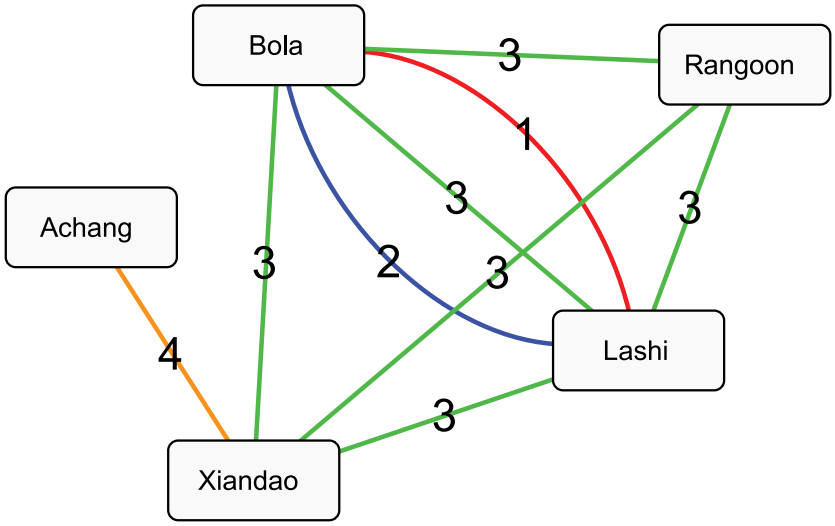


Figure 1. Annotation of cognate relations for words for ‘yesterday’ in five Burmish languages. Four languages share one morpheme, originally meaning ‘day’, marked in green in the table. But while Bola and Lashi show an identical compound structure, Rangoon and Xiandao show different structures, and the mono-morphemic word in Achang could have easily resulted from the loss of the first element of the cognate word in Xiandao. Coding these relationships in a strict fashion (column Strict) will ignore the similarity among all word forms in the morphemes they share, while coding in a loose fashion leads to an exaggeration of the similarities, rendering all words cognate. The same problems are further illustrated in the network on the right, where each edge represents one shared cognate morpheme across the five languages, based on the data in the table on the left. While all words form a connected component in this network, not all connections are equally strong.

ing cognate-coded datasets for the purpose of phylogenetic reconstruction analyses (Atkinson and Gray 2006), we would have a hard time deciding where to draw the boundaries in our cognate judgments. Are only Bola [a^{31} ŋji^{35} nɛʔ^{31}] and Lashi [a^{31} ŋjei^{55} nap^{31}] truely cognate, or should we say that all words are cognate, given that they form a connected component in a network, as illustrated in Figure 1? These decisions are reflected in what List (2016) calls *strict* and *loose partial cognate coding*. In strict cognate coding, only words which share the same compound structure and are cognate in all their parts are assigned to the same cognate set. In loose coding, one shared element is sufficient to assign two words to the same cognate set. For lexicostatistical datasets and phylogenetic reconstruction loose cognate coding necessarily masks important processes of *lexical replacement*: the fact that four of the five Burmish languages have a cognate morpheme in the word for 'yesterday' does not provide any important information for subgrouping. On the other hand, the case of Achang [man^{35}] and Xiandao [n̥^{31} m̥an^{35}] can be easily explained by assuming a recent loss of the first element in Achang, which is further confirmed by the overall closeness of the two languages. These examples illustrate that we should not blindly follow a strict cognate coding, as we may easily loose information relevant for subgrouping.

It seems that the best way to treat partial cognacy would be to follow an exact cognate coding of partial cognates, by annotating the cognacy of each morpheme in each word rather than for each word form. Unfortunately, available tools are not up to the task. Computational methods for automatic cognate detection, which could be used to pre-parse the data for the linguists, usually assume that words are morphologically simple (Steiner et al. 2011; List et al. 2017) and automatic partial cognate detection is still in its infancy (List et al. 2016).

Manual handling of partial cognacy is extremely tedious, since we lack consistent standards and tools for partial cognate annotation. As a result, studies which make use of manually annotated cognate sets usually ignore the problem of partial cognacy, as can be seen when inspecting the current practice of cognate coding in large lexicostatistic databases such as the Austronesian Basic Vocabulary Database (ABVD, Greenhill et al. 2008) or the Indo-European Lexical Cognacy Database (IELex, Dunn et al. 2012). In classical studies, scholars usually content themselves with the extraction of morphemes to establish sound correspondences or etymologies (Mann 1998),

and often even omit the information that the data from which their examples were drawn originally were morphologically complex words (Nishi 1999).

2.2. Sound correspondence annotation

Processes of sound change can be incredibly complex, especially when they involve suprasegmental developments, such as tone change or tono-genesis, which is often triggered by segmental features like the phonation of syllable-initial consonants, or the presence or absence of syllable-final plosives. For scholars who are unfamiliar with a particular language family, it is often impossible to say which sounds correspond when looking at a particular set of cognate words.

But even when ignoring complex sound correspondences, it may be extremely difficult for non-experts to see where two or more cognate sets display correspondences. As an example, consider two words for the comparison concept 'grease/fat', taking from the ABVD (Greenhill et al. 2008), namely Central Amis *simar* vs. Thao *lhimash*. The two words are labelled as cognates in the databases, but for non-experts, it is difficult to see which sounds correspond in the word forms. While it is straightforward to assume non-trivial sound correspondences between Central Amis *s-* and Thao *lh-*, as well as *-r* and *-sh*, it is still impossible for non-experts to assess whether this comparison makes sense or not, as we do not know how regular these correspondences are. Whether the sounds actually correspond or not, is not important for the sake of our example. What *is* important is the fact that we cannot transparently see what the people who annotated the words as being cognate were basing their opinion on.

3. Challenges of analysis

In the preceding section, we mentioned challenges of *annotation*, pointing to cases in South-East Asian languages where both computational and classical approaches have a hard time in achieving transparency. In the following, we show that similar problems arise in *analysing* the processes which pose a challenge for annotation. Having discussed the challenge of partial cognate annotation and sound correspondence annotation above, we here turn to the

problem of the reconstruction of compounds (Section 3.1) and the identification of irregular cognates (Section 3.2).

3.1. Reconstruction of compounds

Compounding is a frequent and vivid process in many languages and language families, not only in South-East Asia, but the world over. Given the prevalence of compounding in some Sino-Tibetan branches like Burmish or Sinitic, it is implausible to assume that the ancestors of the relevant languages had only monomorphemic words. Suprisingly, however, scholars have rarely tried to reconstruct concrete compounds in ancestral languages. Reconstruction systems of Proto-Burmish, for example, only give collections of morphemes with tentative semantic reconstructions (Burling 1967; Nishi 1999), and even where scholars provide reconstructions for tentative compounds in the proto-language (Mann 1998), they fail to provide a transparent account of how they arrived at these conclusions, that is, how they *analysed* the data.

That reconstructions and etymological dictionaries neglect the lexeme level is a general South-East Asian problem, found in etymological analyses of Hmong-Mien (Ratliff 2010), for Austro-Asiatic (Jenny and Sidwell 2015), and Tai-Kadai (Norquest 2007). Furthermore, the problem of treating compound structures consistently in etymological analysis is not unique to South-East Asian linguistics. In 1954, Malkiel criticized the lack of typological investigations on derivation and composition in historical linguistics. What he said by then, namely, that "[one] finds fleeting allusions and casual hints at certain varieties of derivational and compositional hierarchy, but surely no attempt at organized typology" (Malkiel 1954: 266) still holds today.

It is obvious that reconstruction at the lexeme level is more challenging than reconstruction at the morpheme level. True lexical reconstruction may at times even be impossible due to the incompleteness of available data and the complexity of compounding processes. However, scholars often do not even attempt to address these questions and there is little awareness of the inadequacies of the current "morphemes-first" approaches in South East Asian historical linguistics. If we want to advance our knowledge of language change, we cannot stop with sound change but need to try to find regularities and tendencies throughout all levels of language, including processes of word formation.

3.2. Identifying irregular cognate sets

If language contact can be excluded, sound change is a predominantly regular process that affects the whole lexicon of a language (Blevins 2004: 260–268; Kiparsky 1988; Labov 1981). Morphological processes, like suffixation, compounding, or analogy, however, are predominantly *sporadic*. Such morphological processes can mask the regularity of sound change and obstruct the identification of regular sound correspondences.

While the regularity of correspondence is still the major criterion to identify cognate words in different languages, it is by no means the only criterion employed by scholars applying the comparative method. As an example, consider German *fünf* 'five' vs. French *cinq* 'five'. While both words go back to the same Proto-Indo-European root **pénk^{w}e* 'five' (see Meier-Brügger 2002: 265), their phonetic development is highly irregular. While **pénk^{w}e* became *quinque* [k^{w}ink^{w}e] in Latin as a result of an assimilation process replacing the original **p* with **k^{w}* (Meiser 1998), a similar process happened in Proto-Germanic, where the word is reconstructed as **fimfe* (<**pimpe*), reflecting a sporadic change that replaced the **k^{w}* with **p,* which then became *f in Proto-Germanic (Kroonen 2013: 140). Without forms like Classical Greek *πέντε* [pénte] 'five' (with t <**k^{w}*) and Sanskrit *páñca* 'id.' (c < **k^{w}*), it is unlikely that we could identify the French and the German forms as true cognates going back to the same Indo-European root. It is the *cumulative evidence* drawn from regular sound correspondences among Greek, Sanskrit, Latin, and Proto-Germanic that allows us to first identify the Germanic and the Latin forms as irregular and then resolve this irregularity relying on our general knowledge of language-specific and general processes of sound change.

To operationalize such language specific developments when working on concrete language data is difficult. Regularities, at least in shallow language families, can usually be reliably detected when following the general protocol of the comparative method. Even automatic methods for cognate detection are getting more and more reliable and yield convincing results for shallow language families like Germanic or Romance (List et al. 2017). With their help, linguists could preparse the data, and quickly identify the major sound correspondences after manual correction. Finding the irregularities, however, is a much more difficult task, since it not only requires the knowledge of the regularities, but also a general strategy of how to identify cognate material which behaves irregularly in terms of the sound corre-

spondences. Up to today, no heuristics has been proposed for this task, neither in classical nor in computational historical linguistics.

4. Improving annotation and analysis in the Burmish etymological database

Our concerns with annotation and analysis in historical linguistics result from our own efforts in creating an etymological database of the Burmish language family. In this Burmish Etymological Database project (BED, http://dighl.github.io/burmish/), we aim to establish a new type of etymological database which provides data in both human- and machine-readable form, serving both for manual inspection and computational analysis. In the following, we briefly show how we address the aforementioned problems. Since the major part of our endeavour is still a work-in-progress, we are unable to present full-fledged solutions for all the problems mentioned, but we hope that our initial ideas serve future discussions in the field and may inspire new approaches.

4.1. Materials

4.1.1. The Burmish language family

The Burmish languages comprise a small and neatly identifiable group of languages spoken in Southwest China and Northeast Burma. The major languages of the Burmish Family include Burmese, Achang, Xiandao, Maru, Atsi (Zaiwa), Bola, and Lashi, as indicated in the map in the top panel of Figure 2. As can be seen, four of the varieties were recorded in the same city (Máng City 芒市 in China, formerly called Lùxī 路西). When comparing the languages their close proximity must be borne in mind, as we should expect intensive language contact among them. Characteristics of the languages in this family include a generally isolating morphological structure, the use of lexical tone, and tense or creaky phonation.

Nishi (1999: 68) distinguishes two subbranches, Maruic and Burmic, the latter comprising Burmese, Achang, and Xiandao. His classification rests on the observation that the Burmic languages lost tense phonation, replacing it with aspiration of the initial. However, this development does not allow the

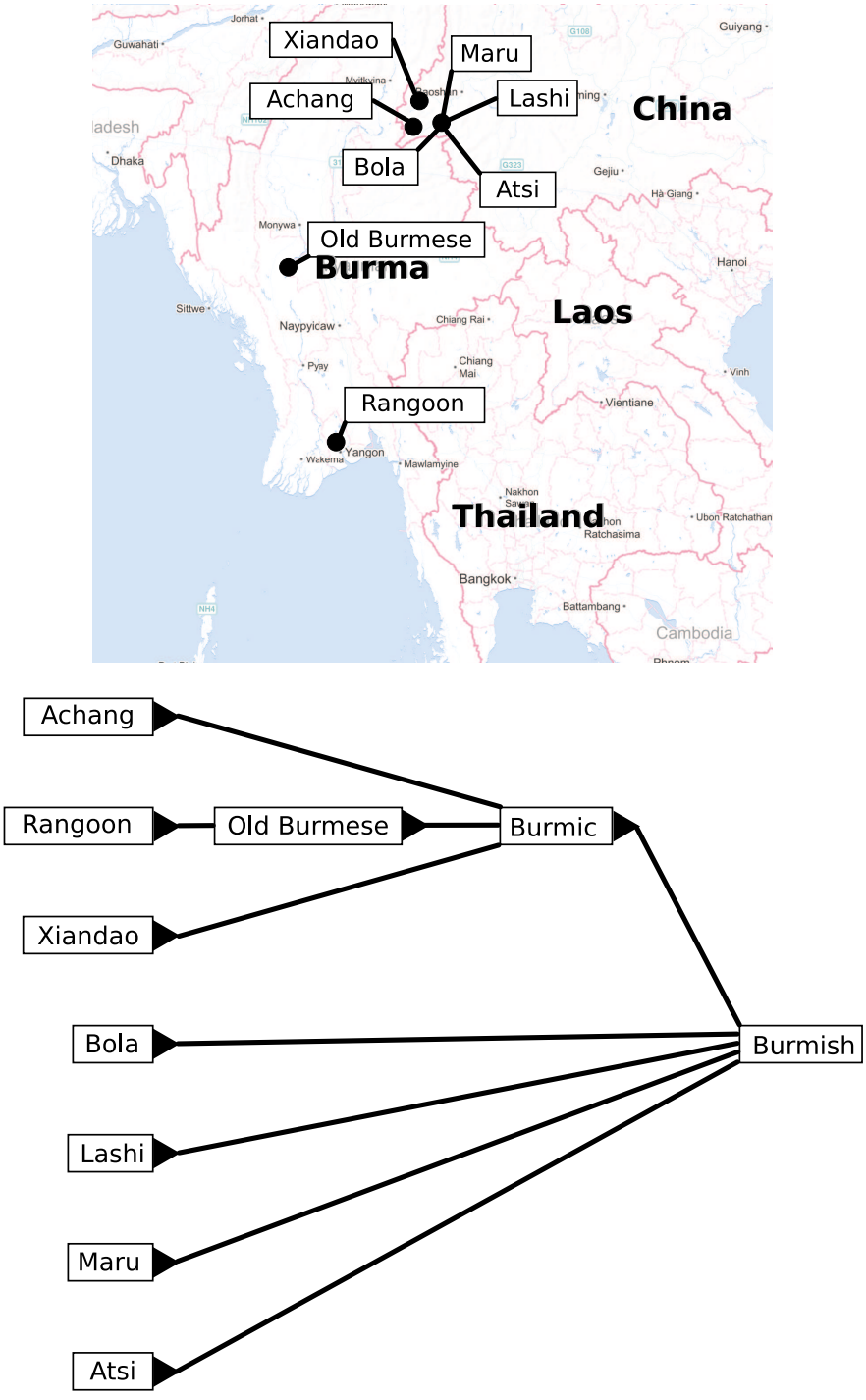


Figure 2. The top panel shows the geographic location of the Burmish varieties in our database (Rangoon is the prestige dialect of modern Burmese), with the location of Old Burmese at Pagan, the capital of the first Burmese dynasty. The bottom panel shows a tentative phylogeny based on sound changes identified as shared innovations, using multi-furcations to indicate uncertainty.

identification of Maruic as a sub-branch, since by keeping tense phonation the languages in question share a retention rather than an innovation. Thus, we propose the preliminary genetic classification seen in the bottom panel of Figure 2, with uncertainties indicated using polytomic (multifurcating) splits. Note that this classification deviates from the one provided in Glottolog (Hammarström et al. 2017), which follows the classification of Mann (1998), one that is not sufficiently substantiated with linguistic evidence.

4.1.2. The Burmish Etymological Database

The Burmish Etymological Database (BED) currently provides data for a basic word list of 240 items translated into the 8 varieties (including Rangoon as the modern prestige dialect of Burmese) given in Figure 2. The data were taken from Huáng et al. (1992) in the digital version provided by the Sino-Tibetan Etymological Dictionary and Thesaurus (STEDT) project (Matisoff 2015), to which we added Old Burmese on the basis of Okell (1971), Luce (1985) and Nishi (1999). The etymologies we arrived at independently of the STEDT project, and the degree of annotation was, as will be further illustrated below, considerably refined.

4.1.3. Availability of data, tools, and code

All data which we used for the following illustrations along with the source code of the software we applied are available in the supplementary material accompanying this paper. In addition to our analyses, we provide explicit links for the languages in the data to Glottolog (Version 3.0, Hammarström et al. 2017), and the concepts in the data to the CLLD Concepticon (Version 1.0, List et al. 2016). All words are further linked to the STEDT database, apart from those for Old Burmese which was not taken from STEDT.

4.2. Methods and tools for annotation and analysis

In order to address the problems mentioned above, several methods and tools were developed, which are presented in more detail below. Computationally

intensive methods for automatic analyses were generally written as plugins for LingPy, a Python software library for quantitative tasks in historical linguistics (Version 2.5.1, http://lingpy.org, List and Forkel 2016), and are available in the supplementary material accompanying this paper. Tools for manual annotation and inspection were implemented as part of the Etymological Dictionary Editor (EDICTOR, http://edictor.digling.org, List 2017), a web-based interactive tool for creating, inspecting, and editing etymological datasets, and are already implemented in the most recent online version of the tool.

4.2.1. Partial cognate annotation

As mentioned above, the manual annotation of partial cognates is tedious. In order to ease the task, a partial cognate editor was included in the most recent version of the EDICTOR tool, which greatly facilitiates the annotation task. All that is required is that the data are morphologically segmented by the user. Once this is done, users can load their data into the EDICTOR tool and indicate which morphemes in a set of pre-defined words (usually translations of the same comparison concept) are cognate. Since this can be done in a simple drag-and-drop fashion, by which the user selects and deselects the words which are grouped into one partial cognate set, the annotation can be carried out quickly and is also less prone to error than the use of spreadsheet software not designed for this task.

In order to identify partial cognates in the BED projects, we first analysed the data automatically, using the algorithm recently proposed by List et al. (2016) for the automatic detection of partial cognates, and then manually corrected the errors in the automatic analysis.

4.2.2. Using alignments for sound-correspondence annotation

To detect regularly recurring sound correspondences linguists usually rely on *alignment analyses* (Prokić et al. 2009; List 2014). Alignments are a formal way to compare sequences. In an alignment analsyis, two or more strings of segments are arranged in a matrix in such a way that corresponding segments are placed in the same column, while placeholders (so-called *gaps*, usually represented by the symbol "-") mark segments lacking a counterpart. In addi-

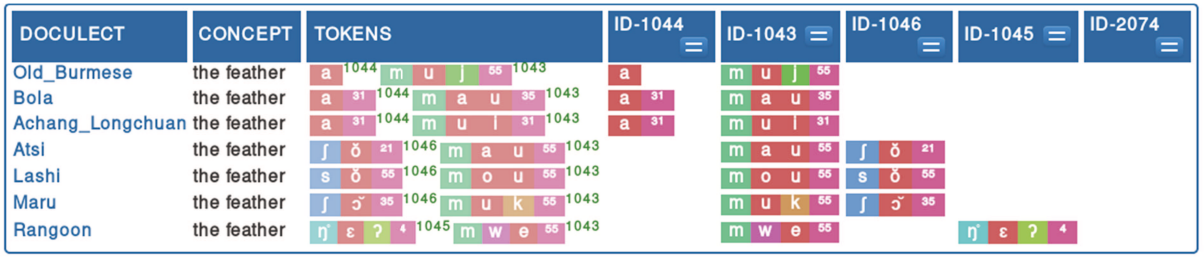


Figure 3. Partial cognate annotation with the EDICTOR tool. Annotation of partial cognates is essentially drag and drop. The user first selects morphemes by clicking on them in order to assign them to a common cognate set in a second step. The figure shows how we cluster translations of the comparison concept 'the feather' in the BED.

tion to identifying partial cognates in the Burmish language data, we also aligned the data, using a computer-assisted work-flow in which we first aligned the partial cognate sets automatically using the SCA algorithm (List 2012) available in the LingPy software package, and then refined them manually, using the alignment module of the EDICTOR tool. An example alignment analysis is illustrated in Figure 4 for translations of the comparison concept 'the man (male human)'.

DOCULECTS	CONCEPTS	ID: 446					ID: 448			
Achang_Longchuan	the man (male human)	-	i	-	-	31	tɕ	i	-	55
Atsi	the man (male human)	j	u	-	ʔ	21	k	e	-	51
Bola	the man (male human)	j	a	u	ʔ	31	k	a	i	55
Lashi	the man (male human)	j	-	u	ʔ	55	k	ɛ	-	31
Maru	the man (male human)	j	a	u	k	31	k	a	i	31
Xiandao	the man (male human)	j	-	u	ʔ	31	ɕ	ɛ	-	55
Rangoon	the man (male human)	j	ɑ	u	ʔ	4	tɕ	ɑ	-	55

Figure 4. Example for the tentative alignment of words for the comparison concept 'the man (male human)' in seven of the eight Burmish languages in our sample.

The use of alignments to annotate sound correspondences is an old technique that goes at least back to the early 20th century (Dixon and Kroeber 1919), long before automatic alignment algorithms were proposed (Covington 1996, Kondrak 2000). Unfortunately, alignments have only sporadically been employed so far (Haas 1969; Fox 1995: 67; Payne 1991). Scholars often consider alignments as too simple to represent the complex relations they see when looking at cognate words. This, however, is not a convincing ground for the rejection of alignments. If alignments are indeed too simple to reflect sound correspondences in all their complexity, scholars should work on enhanced ways to transparently annotate their judgments.

4.2.3. Compound analysis and word family detection

List (2016) presents an initial approach to reconstructing processes of word compounding with the help of a reference phylogeny and ancestral state re-

construction based on weighted parsimony. Given that our data are available in a similar form, we could use the same technique to analyse compound processes in the Burmish languages. However, since this approach requires a good idea of the general phylogeny of the languages, whereas the phylogeny of the Burmish languages remains rather unclear, we base our initial compound analysis on a semi-automated approach that helps to identify the *motivation structure* underlying the formation of specific compounds. Our core idea is to follow Urban (2011) in searching for *partial colexifications* across the words in our data, and to represent them as *bipartite networks*. Following François (2008), we see colexification as a term to cover cases in which a word form is used to denote more than one concept, without distinguishing between homophony or polysemy. Partial colexification therefore points to cases where a specific morpheme is shared across two words denoting distinct concepts.

Given that each syllable usually corresponds to one morpheme in the Burmish languages, it is easy to write a computer application to search for these patterns in our data. In contrast to approaches that are solely interested in the relations between different concepts (List et al. 2013), we wish to investigate both the actual word forms in our data and the concepts which they denote. *Bipartite* networks, which are increasingly used to investigate molecular datasets in evolutionary biology (Corel et al. 2016), provide an intuitive and simple structure for such a computer-assisted investigation. Bipartite networks are networks consisting of two types of nodes. Edges in these networks are only allowed to be drawn from nodes of one type to nodes of another type. In our case the first node type are the *concepts* in the concept list and the second node type are the *word forms* in a given language. We create our network by linking all individual morphemes in our data to the concepts denoted by the words in which they occur. This yields a large graph, which is almost completely connected, but sparse enough to allow interactive search for interesting structures using graph-visualization software, such as Cytoscape (Smoot et al. 2011), and without applying heavy algorithmic machinery. In our supplementary material, we provide the full network created from our data along with the source code as an interactive web-application that works in most web browsers.

In addition, and in order to complement this computational analysis, the EDICTOR tool contains a *morpheme annotation module* that allows one to inspect automatically created bipartite networks for individual languages and to annotate compounds in a meaningful way. The general idea behind this

compound structure analysis is to annotate compounds in a way similar to how linguists annotate sentences in inter-linear glossed text. For each word in the data, we provide a language-internal analysis that reveals the motivation of compound formation. Essentially, this yields a language-internal *word family analysis*, as it allow us to identify cognates within the same language.

ID	COGID	CONCEPT	MORPHEMES	TOKENS
3368	400	the river	water + mo-suffix	v u i 51 + m o 55
3535	425	the sea	water + sea	v u i 51 + m i ŋ 21
4868	409	the water	water	v u i 51
598	619	to buy	buy	v u i 51

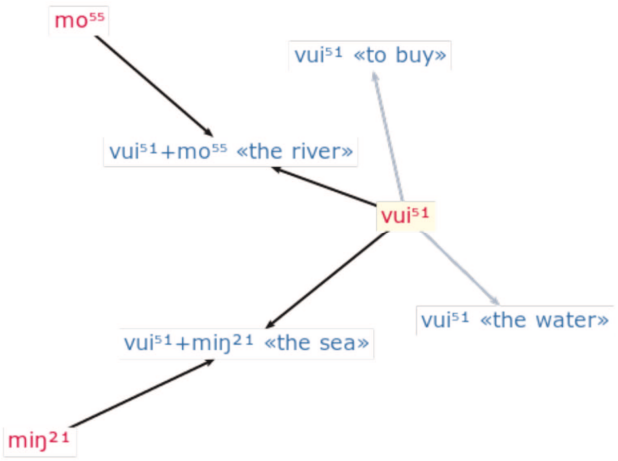


Figure 5. Compound analysis (language-internally) with the help of partial colexification networks. The example shows four words in Atsi (Zaiwa), of which three constitute a word family. The table shows the morpheme analysis and the raw data, while the network below shows the bipartite graph which is automatically created by the EDICTOR tool.

As an example, consider Atsi [vui^{51} mo^{55}] ‘river’, [vui^{51} miŋ^{21}] ‘sea’ and [vui^{51}] ‘water’. When inspecting these words, it is obvious, that [vui^{51}] ‘water’ recurs in the words for ‘sea’ and ‘river’, and it is also easy to identify [mo^{55}] as a suffix, as it recurs in a few other words , such as [lo^{21} mo^{55}] ‘tiger’

and [vam^{51} k^{h}ui^{21} mo^{55}] ‘wolf’.[2] The language-internal bipartite networks drawn from partial colexifications available in the EDICTOR drastically facilitate this task. Scholars can first automatically search for potential word families and then annotate them step by step, eventually distinguishing coincidental cases of homophony, such as Atsi [vui^{51}] ‘to buy’, from the reuse of an etymon in distinct lexemes. Figure 5 shows the user-annotated data and the automatically reconstructed partial colexification network for this example.

5. Results

In the following, we present the results of the analyses described above. We should add that most of these results are anecdotal and not quantitative. There are two reasons for this: first, our general intention in the BED project is to pursue a computer-assisted rather than a computer-based approach to language comparison. This means that we use quantitative analyses to do the bulk of the heavy lifting while we inspect the data manually to find those patterns which cannot be explained with algorithms alone. Second, our methodology comprises preliminary work that to our knowledge has so far not yet been tested on other language families. By pointing to some of our initial findings, we hope we can advertise the tools and approaches discussed here. In this way, we hope that the preliminary approaches presented in this study may in the future bear further fruits, be it in our own work or that of our colleagues working on language families that present similar difficulties.

5.1. Comparison with STEDT

As we assigned the cognate sets independently of the cognates provided by the STEDT project (Matisoff 2015), one can compare the differences between our analysis of the Burmish languages and the analysis provided by the STEDT project. The 240 concepts and 7 languages which were originally taken from STEDT's digitalized version of Huáng et al. (1992) consists of 1611 distinct words and 1002 distinct morphemes. 743 (46%) of the words

[2] By ‘wolf’ we understand ‘dhole’ (*Cuon alpinus*). The grey wolf (*Canis lupus*) is not endemic to the relevant parts of Asia.

are annotated in STEDT, i.e., they are given etymologies; 828 (83%) of the morphemes are assigned to cognate sets in STEDT. Having excluded 23 out of the 743 words for which we found no link between our data and the data in STEDT, we compared the similarity in cognate judgments for the remaining 720 words, using B-Cubed Scores (Bagga and Baldwin 1998) to estimate the differences. These scores are usually measured in *precision*, *recall*, and *harmonic mean* (F-Score), by comparing the results of a cluster analysis A with a cluster analysis B. Precision indicates how often clusters proposed by analysis B are also found in analysis A, recall indicates how often clusters proposed in analysis A are also found in analysis B, and the harmonic mean provides a summary of the two scores. All scores are measured in terms of floating points between 0 and 1, with 1 indicating complete identity and 0 indicating complete difference.

The comparison of our BED analysis with the analysis provided by STEDT (assuming that BED is analysis A and STEDT is analysis B) yielded a precision of 0.88, a recall of 1.0, and an F-Score of 0.94. These results are remarkable, given that the analyses were carried out independently. The high recall means that whenever BED says that two words are cognate, STEDT will also do so. The low precision shows that our analysis is more conservative, having the tendency to refuse cognate judgments rather than to propose them, and as a result, if BED refuses cognacy, STEDT may in quite a few cases still tend to propose it.

5.2. Proving cognacy despite irregularities

Thanks to the alignment analyses carried out on our data, we are able to determine quickly whether the sound correspondence patterns inherent in a given cognate set are regular or not. For convenience, the EDICTOR offers a module in which sound correspondences are automatically counted for each pair of languages in the data. Ideally, this should likewise be offered for the major patterns across all of the languages in the data, but at the moment, this is not feasible, as no algorithms for the detection of general correspondence patterns have been proposed so far.

In order to identify potential cognates independently of regular sound correspondences, we can employ our bipartite partial colexification networks. As an example for this idea, compare the words for ‘good’ across seven Burmish varieties given in Table 3. At first sight, the words all look quite

similar, and no linguist would immediately rule out the possibility that they could be cognate. Based on the sound correspondences we identified, however, the forms in Achang and Xiandao are not regular, as the correspondence among [tɕ] in Achang, [ɕ] in Xiandao and [k] in the other Burmish varieties is only attested in the words for 'good' and the word for 'man', also given in Table 3.

Despite the irregularity of the sound correspondences between Achang and the other varieties, it is still justifiable to regard all words as cognate (except for Rangoon Burmese [kə̃u^{55}] 'good' and Lashi [kɛː^{31}] which has an unpredicted long vowel). We reconstruct the word 'man' in Proto-Burmish as a compound of 'person' and 'good', supported by the fact that the first morpheme of the words for 'man' occurs in the words for 'who' in Bola and Maru (as shown in the same table), and that – except for in Rangoon Burmese – the second morpheme in the words for 'man' is cognate in all languages in the table (we suspect that the vowel length in Lashi is a secondary phenomenon, probably resulting from loss of syllable weight in compounds).

Table 3. Irregular sound correspondences among Achang and Xiandao and five other Burmish languages: Achang [tɕ] and Xiandao [ɕ] in the word for 'good' exhibits an irregular correspondence with [k] in the other Burmish languages. The fact that the compound word 'man' has the word for 'good' as its second part in all Burmish languages apart from Rangoon, and the peculiarity of the motivation of this compound justify assuming cognacy despite irregularity. As a result, we label cognacy among the morphemes in the table by assigning the same color to cognate morphemes, leaving black as the color for words we cannot relate to any other word.

Language	'man'	'good'	'who?'
Achang	i^{31} tɕi^{55}	tɕi^{55}	xau^{55}
Atsi	juʔ^{21} kɐ^{51}	kɐ^{51}	o^{55}
Bola	jauʔ^{31} kai^{55}	kai^{55}	kʰak^{55} jauʔ^{31}
Lashi	juʔ^{55} kɛ^{31}	kɛː^{31}	xaŋ^{55}
Maru	jauk^{31} kai^{31}	kai^{31}	kʰɤ̃^{31} jauk^{31}
Rangoon (Burmese)	jɑuʔ^{4} tɕɑ^{55}	kə̃u^{55}	bɛ^{22} tθu^{22}
Xiandao	juʔ^{31} ɕɛ^{55}	ɕɛ^{55}	xau^{55}

Since this compound is semantically and syntactically peculiar and uniquely occurs in the Burmish languages (we found no similar motivation in the more than 40 other Sino-Tibetan languages in Huáng et al. 1992), it is very likely that this word originated only once in the history of the Burmish languages. No matter what the explanation for the irregular sound correspondences in Achang and Xiandao will be (if it can ever be found), given the overwhelming similarity in the *motivation structure* of the compound for 'man' in the Burmish languages, one cannot resist the conclusion that these words are indeed cognate, and we mark them accordingly in Table 3.

5.3. Compound structure and subgrouping

Compound structure can provide us with initial hints regarding subgrouping. We must be careful, however, since it is obvious that words can easily be borrowed among languages, and closely related languages will also allow for the borrowing of full compounds, as we can see in numerous examples from the Chinese dialects (compare, for example, List et al. 2014). Nevertheless, when such cases can be excluded, compound structure may serve as a proxy for the identification of shared traits between languages and thus help us to identify potential innovations that provide us evidence for subgrouping.

As an example, consider Table 4 which gives words for 'mountain', 'dog', 'thunder', 'wolf', and 'bear (n.)' in the modern languages in our sample along with our comparative analysis of the motivation structure of these words, derived from the bipartite partial colexification networks. First, we find four different motivations for 'wolf' in the sample. Except for the Rangoon word form, all are derived from the word for 'dog', but the first part of the compound differs, and we find 'bear' + 'dog' in Atsi and Lashi, 'thunder' + 'dog' in Bola and Maru, and 'mountain' + 'dog' in Achang and Xiandao. Achang and Xiandao further show the same motivation structure for 'thunder', which can be seen as a further argument that both varieties form a subbranch of the Burmic branch of Burmish.

The situation with Lashi, Bola, and Maru is more complicated and requires further explanation. We find that Maru shares the same motivation structure for 'thunder' with Lashi ('sky' + 'thunderB'), while it also shares the motivation structure for 'wolf' with Bola ('thunder' + 'dog'). Note that our analysis of Maru [mjaŋ^{31} k^{h}a^{35}] as 'thunder' + 'dog' is based only on the similarity with Bola, as the word for 'thunder' in Maru does not contain [mjaŋ^{31}].

Table 4. Compound motivation patterns across the modern Burmish languages. Items with identical color in the annotation of the motivation structure are presumed to be cognate across and inside the four varieties. Black is reserved for items which are not related to any other item in the data.

Language	'mountain'	'dog'	'thunder'	'wolf'	'bear (n.)'
Atsi	pum^{51}	k^{h}ui^{21}	mau^{21} mjiŋ^{51}	vam^{51} k^{h}ui^{21} mo^{55}	vam^{51}
	mountain	dog	sky + thunder	bear + dog + *m-suff.*	bear
Bola	pam^{55}	k^{h}ui^{35}	mau^{31} mjaŋ^{55}	mjaŋ^{55} k^{h}ui^{35}	vɛ̃^{55}
	mountain	dog	sky + thunder	thunder + dog	bear
Lashi	pɔm^{31}	k^{h}ui^{55}	mou^{33} kɔm^{33}	wɔm^{31} k^{h}ui^{55}	wɔm^{31}
	mountain	dog	sky + thunderB	bear + dog	bear
Maru	pam^{31}	lə̆^{31} k^{h}a^{35}	muk^{55} kum^{31}	mjaŋ^{31} k^{h}a^{35}	vɛ̃^{31}
	mountain	? + dog	sky + thunderB	thunder + dog	bear
Achang	pum^{55}	xui^{31}	mau^{31} ʐau^{31}	pum^{55} xui^{31}	ɔm^{55}
	mountain	dog	sky + thunderC	mountain + dog	bear
Xiandao	pum^{55}	fui^{31}	mau^{31} cau^{31}	pum^{55} fui^{31}	om^{55}
	mountain	dog	sky + thunderC	mountain + dog	bear
Rangoon	tãu^{22}	k^{h}we^{55}	mo^{55} tɕ^{h}ẽ^{55}	wũ^{22}pu^{53}lwe^{22}	wũ^{22}
	mountain2	dog	sky + thunderD	bear + ? + ?	bear

Given that the data for Maru, Lashi, Bola, and Atsi were collected in the same area, and close contact among the varieties is therefore expected, we may suspect that the divergence in compound structures results from language contact. Given that 'bear' occurs in the word for 'wolf' in Atsi, Lashi, Achang, Xiandao and particularly in the otherwise untransparent Rangoon Burmese form , we suspect that the 'thunder-dog' in Maru and Bola is a later innovation rather than a retention. This suspicion however gives rise to a further complication. If Maru and Bola together innovated the structure 'thunder-dog' then the Maru word for 'thunder' should be cognate with the form of the word 'thunder' that occurs in the Maru word for 'wolf', which it is not. To explain the Maru word for 'thunder' one can suggest that Maru has borrowed it from Lashi. This proposal is not only confirmed by the irregular vowel correspondence between the two varieties, but also by alternative data in Clerk (1911: 163), who gives *muk myang* as the word for 'thunder' in a Maru variety spoken in the Myitkina area of Burma, far away from Máng

City, where the Maru variety we considered for our database is spoken. The Myitkina form appears to preserve the inherited etymon as opposed to the Máng City form, which is borrowed from Lashi. This explanation is yet further buttressed by the fact that Wannemacher (2011: 37) gives /mou⁴ gøm⁴/ as translation for 'thunder' in a Lashi variety spoken in the Waimaw area of the Kachin State in Burma, again far away from the Lashi variety we considered in our study. The obvious cognancy of the Lashi forms from distinct regions of Burma points to the fact that Lashi here retains an inheritance. In other words, the Lashi word is geographically stable whereas the Maru word is not.

It would go beyond the scope of this paper to resolve the phylogeny of the Burmish languages by listing potential shared innovations or even using phylogenetic methods to arrive at a subgrouping of the language family. We think, however, that our small analysis of the words in Table 4 has shown that compound motivation structures bears substantial potential for linguistic subgrouping, provided they are analysed with care, and borrowing are thoroughly identified. Both the analysis of compound motivation structures and the identification of borrowings cannot be done automatically. Our methods for the reconstruction of bipartite partial colexification networks, however, provide great help for a detailed computer-assisted analysis.

5.4. Compound structure and semantic reconstruction

A compound motivaton structure analysis derived from bipartite partial colexification networks can also serve as a starting point for semantic reconstruction, both from a semasiological perspective, seeking the original meaning of a given morpheme, and form an onomasiological perspective, seeking to identify how a given concept was pronounced in ancestral languages. As an illustration, consider the colexification newtork given in Figure 6. In this example, we find three major semantic complexes: the verbs 'to shoot (arrow)' and 'to throw', the verb 'to hunt', and several concepts denoting body parts ('hair', 'tail', 'bone', etc.). These semantic groups are connected by two form groups, the first one pointing to Proto-Burmish **pak⁴* and the second pointing to **ʃa²* (both in the reconstructions of Mann 1998). The verb 'to shoot' is expressed by single morphemes (reflexes of **pak⁴*) in Atsi, Bola, Maru, and Xiandao, while the verb for 'to hunt' is expressed by two morphemes, the former colexifying with the forms for 'to shoot', and the latter,

reflexes of Mann's **ʃa²*, occurring as one of the elements in the numerous body part terms in our third semantic cluster. Given these patterns, we find it straightforward to reconstruct the rough semantics of Proto-Burmish **pak⁴* as 'to throw/to shoot', and the semantics of **ʃa²* as 'body/flesh', since these meanings (which are admittedly not extremely precise at this stage of the analysis) allow best to explain why reflexes of **ʃa²* occur in compounds denoting body parts, and as the object of verb-object compounds meaning 'to hunt' (lit. 'shoot meat' or 'shoot bodies') in the Burmish varieties.

The pattern in Figure 6 is but a small example of a computer-assisted procedure, but it illustrates the main idea of computer-assisted approaches : the analytical work is still carried out by the linguists who interpret the data and draw their conclusions, but an advanced computational modeling of linguistic problems helps the linguists in identifying patterns deserving explanation. No doubt one could identify the pattern in Figure 6 by simply inspecting the data in a book. The representation as bipartite networks of partial colexifications, however, drastically speeds up this process.

6. Conclusion

With more than 7000 languages currently spoken and numerous other languages now lost, existing in philological records, historical linguistics faces the tremendous task of charting the evolution of these languages into their current shape. Computational approaches offer quick solutions to analyze large amounts of digitally available data. However, they face specific difficulties, resulting from their lack of flexibility which makes them vulnerable in situations of sparse data. Classical approaches handle data sparseness well, but they face efficiency and transparency problems. A combined framework can cope with the shortcomings of both disciplines while at the same time preserving their specific advantages.

In this paper, we have tried to illustrate how computational and classical approaches can be combined, concentrating on specific challenges of annotation and analysis in the Burmish language family. With the help of computational methods and interactive tools for the correction of errors, we consistently annotated partial cognates and regular sound correspondences for eight Burmish varieties. With the help of bipartite partial colexification networks, we further annotated compound motivation structures for a large part of the words in our data. We illustrated the benefit of these new approaches to an-

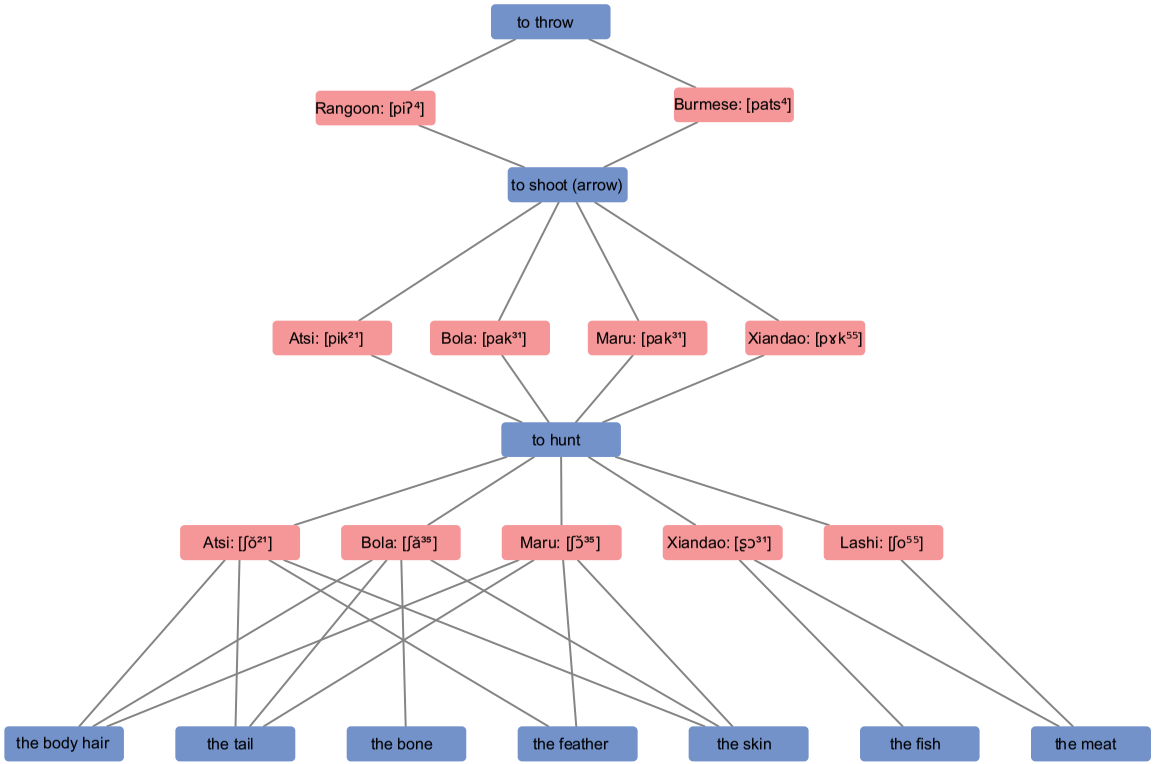


Figure 6. A bipartite word family network indicating the relations between words for ‘to shoot’ and ‘to hunt’ in the Burmish language. The bipartite graph constructed from partial colexifications shows which concepts (blue nodes) are expressed with similar morphemes (red nodes) in the Burmish languages in our sample. As can be seen from the network, the word for ‘to hunt’ is expressed by morphemes in most languages: one which also means ‘to shoot’ in isolation, and one which occurs as a further element in words like ‘body hair’, ‘bone’, ‘skin’, and ‘meat’. This leads us to conclude that the Proto-Burmish word for ‘to hunt’ was a compound motivated as {shoot} + {body/flesh}. Our evidence is two-fold, drawing from information regarding the regularity of the sound correspondences in the languages under investigation as well as the structural information exhibited in the bipartite word family network.

notation and analysis, by showing how cognate words can be identified even when sound correspondences are irregular, how shared innovations can be detected by searching for similar compound structures, and how compound structure comparison allows us to make initial steps towards semantic reconstruction. The proposed methods and techniques are preliminary and need to be further developed. We are, however, confident that they provide new insights not only into the Burmish languages but also into South-East Asian languages in general, since they offer not only a more complete perspective on linguistic reconstruction, but also deliver additional evidence for subgrouping, hidden cognates, and semantic reconstruction.

7. Acknowledgements

This research would not have been possible without the LFK Young Scholars Symposium (University of Washington, Seattle, 2013), generously hosted by the Li Fang-Kuei Society for Chinese Linguistics (http://lfksociety.org/), during which both authors made first acquaintance and began their collaboration. We would like to acknowledge the generous support of the European Research Council for supporting this research under the auspices of 'Beyond Boundaries: Religion, Region, Language and the State' (ERC Synergy Project 609823 ASIA, NWH) and 'Computer-Assisted Language Comparison' (ERC Starting Grant, JML), and the German Research Foundation (DFG) for supporting JML from 2015 to 2016 with a research scholarship on 'Vertical and Lateral Aspects of Chinese Dialect History' (Grant No. 261553824). We would further like to thank Guillaume Jacques and Harald Hammarström for providing helpful comments on an earlier version of this paper, as well as Doug Cooper and Mark Miyake for providing invaluable help with linguistic data.

8. Supplementary material

The supplementary material accompanying this paper contains the source code with which we created our sample bipartite network application, as well as the Burmish data set which we analyzed and prepared with help of the Etymological Dictionary Editor (http://edictor.digling.org). All material can be downloaded from https://zenodo.org/record/886179. The code was curated

on GitHub at http://github.com/digling/challenges-of-annotation-paper. The data is additionally shared in CLDF (http://cldf.clld.org), following the most recent specifications.

References

Atkinson, Q. and R. Gray. 2006. "How old is the Indo-European language family? Illumination or more moths to the flame?" In: Forster, P. and C. Renfrew (eds.), *Phylogenetic methods and the prehistory of languages*. Cambridge, Oxford and Oakville: McDonald Institute for Archaeological Research. 91–109.

Bagga, A. and B. Baldwin. 1998. "Entity-based cross-document coreferencing using the vector space model". In: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*. Association of Computational Linguistics. 79–85.

Blevins, J. 2004. *Evolutionary phonology*. The emergence of sound patterns. Cambridge: Cambridge University Press.

Burling, R. 1967. *Proto-Lolo-Burmese*. Bloomington: Indiana University Press.

Butler, A. and W. Saidel. 2000. "Defining sameness: Historical, biological, and generative homology". *BioEssays* 22. 846–853.

Campbell, L. 2013. *Historical linguistics*. Edinburgh: Edinburgh University Press.

Clerk, F. 1911. *A manual of the Lawngwaw or Maru language, containing: the grammatical principles of the language, glossaries of special terms, colloquial exercises, and Maru–English and English–Maru vocabularies*. Rangoon: American Baptist mission Press.

Corel, E., P. Lopez, R. Méheust and E. Bapteste. 2016. "Network-thinking: Graphs to analyze microbial complexity and evolution". *Trends in Microbiology* 24(3). 224–237.

Covington, M. 1996. "An algorithm to align words for historical comparison". *Computational Linguistics* 22(4). 481–496.

Dixon, R. and A. Kroeber. 1919. *Linguistic families of California*. Berkeley: University of California Press.

Dunn, M. (ed.). 2012. Indo-European lexical cognacy database (IELex). http://ielex.mpi.nl/.

Fox, A. 1995. *Linguistic reconstruction. An introduction to theory and method.* Oxford; Oxford University Press.

François, A. 2008. "Semantic maps and the typology of colexification: Intertwining polysemous networks across languages". In: Vanhove, M. (ed.), *From polysemy to semantic change*.Amsterdam: Benjamins. 163–215.

Gabelentz, G. v. d. 1891. *Die Sprachwissenschaft. Ihre Aufgaben, Methoden und bisherigen Ergebnisse*. Leipzig: T. O. Weigel.

Gabelentz, G. v. d. 1892. *Handbuch zur Aufnahme fremder Sprachen* [Handbook for the description of foreign languages]. Berlin: Ernst Siegfried Mittler & Sohn.

Greenhill, S., R. Blust and R. Gray. 2008. “The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics”. *Evolutionary Bioinformatics* 4. 271–283.

Haas, M. 1969. *The prehistory of languages*. Mouton: The Hague and Paris.

Hammarström, H., R. Forkel and M. Haspelmath. 2017. *Glottolog*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Holm, H. 2007. “The new arboretum of Indo-European ‘trees’. Can new algorithms reveal the phylogeny and even prehistory of Indo-European?” *Journal of Quantitative Linguistics* 14(2–3). 167–214.

Huáng Bùfán 黃布凡 .1992. *Zàngmiǎn yǔzú yǔyán cíhuì* [A Tibeto-Burman lexicon]. Zhōngyāng Mínzú Dàxué 中央民族大学 [Central Institute of Minorities]: Běijīng 北京.

Jenny, M. and P. Sidwell (eds.). 2015. *The handbook of Austroasiatic languages*. Leiden and Boston: Brill.

Kiparsky, P. 1988. “Phonological change”. In: Newmeyer, F. (ed.), *The Cambridge Survey of Linguistics* (vol. 1). Cambridge: Cambridge University Press. 363–415.

Koerner, E. 1976. “Zu Ursprung und Geschichte der Besternung in der historischen Sprachwissenschaft. Eine historiographische Notiz”. *Zeitschrift für vergleichende Sprachforschung* 89(2). 185–190.

Kondrak, G. 2000. “A new algorithm for the alignment of phonetic sequences”. In: *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*. 288–295.

Koonin, E. 2005. “Orthologs, paralogs, and evolutionary genomics”. *Annual Review of Genetics* 39. 309–338.

Kroonen, G. 2013. *Etymological dictionary of Proto-Germanic*. Leiden and Boston: Brill.

Kürschner, W. 2014. “Georg von der Gabelentz’ *Handbuch zur Aufnahme fremder Sprachen* (1892). Entstehung, Ziele, Arbeitsweise, Wirkung“. In: Ezawa, K., F. Hundsnurscher and A. Vogel (eds.), *Beiträge zur Gabelentz-Forschung*. Tübingen: Narr. 239–259.

Labov, W. 1981. “Resolving the Neogrammarian Controversy”. *Language* 57(2). 267–308.

List, J.-M. 2012. “LexStat. Automatic detection of cognates in multilingual wordlists”. In: *Proceedings of the EACL 2012 Joint Workshop of Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources*. 117–125.

List, J.-M., A. Terhalle and M. Urban. 2013. “Using network approaches to enhance the analysis of cross-linguistic polysemies”. In: *Proceedings of the 10th International Conference on Computational Semantics – Short Papers*. Association for Computational Linguistics. 347–353.

List, J.-M., S. Nelson-Sathi, W. Martin and H. Geisler. 2014. “Using phylogenetic networks to model Chinese dialect history”. *Language Dynamics and Change* 4(2). 222–252.

List, J.-M. 2014. *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.

List, J.-M. 2015. “Network perspectives on Chinese dialect history”. *Bulletin of Chinese Linguistics* 8. 42–67.
List, J.-M., M. Cysouw and R. Forkel. 2016. “Concepticon. A resource for the linking of concept lists”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. 2393–2400.
List, J.-M. and R. Forkel. 2016. *LingPy. A Python library for historical linguistics*. Jena: Max Planck Institute for the Science of Human History.
List, J.-M. 2016. “Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction”. *Journal of Language Evolution* 1(2). 119–136.
List, J.-M., P. Lopez and E. Bapteste. 2016. “Using sequence similarity networks to identify partial cognates in multilingual wordlists”. In: *Proceedings of the Association of Computational Linguistics 2016*. (Volume 2: *Short Papers*.) Association of Computational Linguistics. 599–605.
List, J.-M., S. Greenhill and R. Gray. 2017. “The potential of automatic word comparison for historical linguistics”. *PLOS ONE* 12(1). 1–18.
List, J.-M. 2017. “A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. *System Demonstrations*. 9-12.
Luce, G.H. 1985. *Phases of Pre-Pagán Burma: Languages and history*. Oxford: Oxford University Press.
Makaev, E. 1977. *Obščaja teorija sravnitel’nogo jazykoznanija* [General theory of comparative linguistics]. Moscow: Nauka.
Malkiel, Y. 1954. “Etymology and the structure of word families”. *Word* 10(2–3). 265–274.
Mann, N. 1998. A phonological reconstruction of Proto Northern Burmic. (MA thesis, the University of Texas at Arlington.)
Matisoff, J. 2015. *The Sino-Tibetan Etymological Dictionary and Thesaurus project*. Berkeley: University of California.
McMahon, A. and R. McMahon. 2005. *Language classification by numbers*. Oxford: Oxford University Press.
Meier-Brügger, M. 2002. *Indogermanische Sprachwissenschaft*. Berlin: de Gruyter.
Meiser, G. 1998. *Historische Laut- und Formenlehre der lateinischen Sprache*. Darmstadt: Wissenschaftliche Buchgesellschaft.
Morrison, D. 2015. “Molecular homology and multiple-sequence alignment: an analysis of concepts and practice”. *Australian Systematic Botany* 28. 46–62.
Nishi, Y. 1999. *Four papers on Burmese: Toward the history of Burmese (the Myanmar language)*. Tokyo: Institute for the study of languages and cultures of Asia and Africa, Tokyo University of Foreign Studies.
Norquest, P. 2007. A phonological reconstruction of Proto-Hlai. (PhD dissertation, The University of Arizona.)
Okell, J. 1971. “K Clusters in Proto-Burmese”. Paper presented at the Sino-Tibetan Conference, October 8–9, 1971. Bloomington, IN.

Payne, D. 1991. “A classification of Maipuran (Arawakan) languages based on shared lexical retentions”. In: Derbyshire, D. and G. Pullum (eds.), *Handbook of Amazonian languages* (vol. 3). Berlin: Mouton de Gruyter. 355–499.

Prokić, J., M. Wieling and J. Nerbonne. 2009. “Multiple sequence alignments in linguistics”. In: *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*. 18–25.

Ratliff, M. 2010. *Hmong-Mien language history*. Canberra: Pacific Linguistics.

Schwink, F. 1994. *Linguistic typology, universality and the realism of reconstruction*. Washington: Institute for the Study of Man.

Smoot, M., K. Ono, J. Ruscheinski, P. Wang and T. Ideker. 2011. “Cytoscape 2.8. New features for data integration and network visualization”. *Bioinformatics* 27(3). 431–432.

Steiner, L., P. Stadler and M. Cysouw. 2011. “A pipeline for computational historical linguistics”. *Language Dynamics and Change* 1(1). 89–127.

Sturtevant, E. 1920. *The pronunciation of Greek and Latin*. Chicago: University of Chicago Press.

Swadesh, M. 1963. “A punchcard system of cognate hunting”. *International Journal of American Linguistics* 29(3). 283–288.

Urban, M. 2011. “Asymmetries in overt marking and directionality in semantic change”. *Journal of Historical Linguistics* 1(1). 3–47.

Vaan, M. 2008. *Etymological dictionary of Latin and the other Italic languages*. Leiden: Brill.

Wannemacher, M. 2011. *A phonological overview of the Lacid language*. Chiang Mai: Linguistics Institute, Payap University.

Address for correspondence:
Johann-Mattis List
Kahlaische Str. 10
07754 Jena
Germany
mattis.list@shh.mpg.de

4 Advances in Automatic Sequence Comparison

4.1 Advanced Cognate Detection

The task of *automated cognate detection* is of particular interest for historical linguistics, since the identification of cognates, that is, words that are assumed to have descended from a common form, serves multiple purposes, both in traditional and computational approaches to historical language comparison. In traditional historical linguistics, identifying cognate word forms is not only important for the initial proof of genetic language relationship, but also for the identification of regular sound correspondences, or – ultimately – for the compilation of etymological dictionaries in which cognate sets for a particular language family or subgroup are systematically assembled. In computational historical linguistics, cognate sets are particularly important for the reconstruction of language phylogenies.

While initial methods for the automated identification of cognates have been discussed among linguists already for a longer time, with proposals appearing already in the 1960s (Kay 1964), the increased availability of computers along with scripting languages with powerful third-party libraries, such as Python and Perl, have made it much more convenient for scholars of different backgrounds to experiment with their own approaches to the problem. Already during my dissertation, I have worked intensively on the task of automated cognate detection and was able to propose a first algorithm that took inspiration from sequence comparison approaches in evolutionary biology along with newly developed models for the representation of phonetic sequences within a computational frameworks (List 2014). This algorithm, however, still had one serious drawback, since it is only available to detect fully cognate words, while partial cognates could not be readily handled. While working on annotation frameworks for partial cognate relations (as presented in Section 3), I also started to experiment with methods for the automated detection of partial cognates, using *sequence similarity networks*, an approach that has been successfully used in biological applications as the main methodological tool. The thoughts behind this first algorithm for partial cognate detection along with an evaluation of its performance are discussed in detail in the first of the two following studies (List et al. 2016b).

While the cognate detection method developed as part of my dissertation could show by than to have a rather satisfying performance in comparison with alternative methods when applying it to a gold standard of six datasets from four language families, it was important to get a clearer impression of the performance of the approach when applied to more datasets from additional language families. For this reason, we conducted a follow-up study for which we created six additional gold standard datasets from 5 different language families, which were similar in size and diversity to the gold standard datasets used in the previous study. In addition, we tested a new approach for the *flat clustering* of words into cognate sets, based on similarity networks as already applied in the previous study on partial cognate detection and algorithms for the detection of communities in social networks. The results confirm the overall satisfying performance of the cognate detection approaches, reaching about 89% of accuracy compared to the annotations of cognacy done by experts, and show also that community detection algorithms applied to similarity networks further improve the performance of the method (List et al. 2017).

Using Sequence Similarity Networks to Identify Partial Cognates in Multilingual Wordlists

Johann-Mattis List
CRLAO/UPMC
2 rue de Lille
75007 Paris
mattis.list@lingpy.org

Philippe Lopez
UPMC
9 quai de Bernard
75005 Paris
philippe.lopez@upmc.fr

Eric Bapteste
UPMC
9 quai de Bernard
75005 Paris
eric.bapteste@upmc.fr

Abstract

Increasing amounts of digital data in historical linguistics necessitate the development of automatic methods for the detection of cognate words across languages. Recently developed methods work well on language families with moderate time depths, but they are not capable of identifying cognate morphemes in words which are only partially related. Partial cognacy, however, is a frequently recurring phenomenon, especially in language families with productive derivational morphology. This paper presents a pilot approach for partial cognate detection in which networks are used to represent similarities between word parts and cognate morphemes are identified with help of state-of-the-art algorithms for network partitioning. The approach is tested on a newly created benchmark dataset with data from three sub-branches of Sino-Tibetan and yields very promising results, outperforming all algorithms which are not sensible to partial cognacy.

1 Introduction

In a very general notion, cognacy is similar to the concept of *homology* in biology (Haggerty et al. 2014), denoting a relation between words which share a common history (List 2014b). In classical linguistics, borrowings are often excluded from this notion (Trask 2000). Quantitative approaches additionally distinguish cognates which have retained, and cognates which have shifted their meaning (Starostin 2013b). Further aspects of cognacy are rarely distinguished, although they are obvious and common. Words which go back to the same ancestor form can for example have been morphologically modified, such as French *soleil* which does not go directly back to Latin *sōl* `sun' but to *sōliculus* `small sun' which is itself a derivation of *sōl* (Meyer-Lübke 1911).

Variety	Form	Character	Cognacy
Fúzhōu	ŋuoʔ^{5}	月	1
Měixiàn	ŋiat^{5} kuoŋ^{44}	月光	1 2
Wēnzhōu	ɳy^{21} kuɔ^{35} vai^{13}	月光佛	1 2 3
Běijīng	yɛ^{51} liɑŋ^{1}	月亮	1 4

Table 1: Partial cognacy in Chinese dialects.

Another problem are words which have been created from two or more morphemes via processes of *compounding*. While these cases are rather rare in the core vocabulary of Indo-European languages, they are very frequent in South-East Asian language families like Sino-Tibetan or Austro-Asiatic. In 200 basic words across 23 Chinese dialects (Ben Hamed and Wang 2006), for example, almost 50% of the nouns and more than 30% of all words consist of two or more morphemes (see the Sup. Material for details).

The presence of words consisting of more than one morpheme challenges the notion that words can either be cognate or not. It poses problems for phylogenetic approaches which require binary presence-absence matrices as input and model language evolution as cognate gain and cognate loss (Atkinson and Gray 2006). This is illustrated in Table 1 where words for `moon' in four Chinese dialects (Hóu 2004) are compared, with cognate elements being given the same color. If we assign cognacy *strictly*, only matching those words which are identical in all their elements (Ben Hamed and Wang 2006), we would have to label all words as being not cognate. If we assign cognacy *loosely* (Satterthwaite-Phillips 2011), labeling all words as cognate when only they share a common morpheme, we would have to label all

words as cognate. No matter how we code in phylogenetic analyses, as long as we use binary states, we will loose information (List 2016).

Partial cognacy is also a problem for current cognate detection algorithms which compare words in their entirety (List 2014b, Turchin et al. 2010). Given the frequency of compound words in South-East Asian languages, it is not surprising that the algorithms perform much worse on diverse South-East Asian language families, than they perform on other language families where compounding is less frequent (List 2014b:197f).

This paper presents a new algorithm for cognate detection which does not identify cognate *words* but instead searches for cognate *elements* in words. The algorithm takes multilingual word lists as input and outputs statements regarding the cognacy of morphemes, just as the ones shown in the last column of Table 1, where identical numerical IDs are given for all morphemes identified as cognate.

Dataset	**Bai**	**Chinese**	**Tujia**
Languages	9	18	5
Words	1028	3653	513
Concepts	110	180	109
Strict Cogn.	285	1231	247
Partial Cogn.	309	1408	348
Sounds	94	122	57
Source	Wang, 2006	Běijīng Dàxué, 1964	Starostin, 2013b

Table 2: Partial cognate detection gold standard

2 Materials

Three gold standard datasets from different branches of Sino-Tibetan with different degrees of diversity were prepared, including Bai dialects, Chinese dialects, and Tujia dialects. All datasets were taken from existing datasets with cognate codings provided independently. To facilitate further use of the data, all languages were linked to Glottolog (Hammarström et al. 2015) and all concepts were linked to the Concepticon (List et al. 2016a). Furthermore, phonetic transcriptions were cleaned by segmenting phonetic entries into meaningful sound units and unifying phonetic variants representing the same pronunciation. Morphological segmentation was not required, since all languages in our sample (and the majority of all South-East Asian languages) have a morpheme-syllabic structure in which each syllable denotes one morpheme. Partial cognate judgments are displayed with help of multiple integer IDs assigned to a word in the order of its morphemes, as displayed above in Table 1. For the Chinese dataset, partial cognate information was provided in the source itself, for Bai and Tujia, it was manually derived from the cognate judgments in the sources. Detailed information regarding the datasets is given in Table 2, and the full dataset along with further information is given in the Sup. Material.

3 Methods

The workflow for partial cognate detection consists of three major steps. (1) In a first step, pairwise sequence similarities are determined between all morphemes of all words in the same meaning slot in a word list. (2) These similarities are then used to create a similarity network in which nodes represent morphemes and edges between the nodes represent similarities between the morphemes. (3) In a third step, an algorithm for network partitioning is used to cluster the nodes of the network into groups of cognate morphemes.

3.1 Sequence Similarity

There are various ways to determine the similarity or distance between words and morphemes. A general distinction can be made between *language-independent* and *language-specific* approaches. The former determine the word similarity independently of the languages to which the words belong. As a result, the scores only depend on the substantial and structural differences between words. Examples for language-independent similarity measures are SCA distances, as produced by the Sound-Class-Based Phonetic Alignment algorithm (List 2012b), or PMI similarities as produced by the Weighted String Alignment algorithm (Jäger 2013). Language-specific approaches, on the other hand, are based on previously identified recurring correspondences between the languages from which the words are taken (List 2014b: 48-50) and may differ across languages.[1] An example for language-specific similarity measures is the LexStat algorithm, first proposed in List (2012a) and later refined in List

[1]Comparing, for example, German *Kuckuck* with French *coucou* and English *cuckoo* may yield quite different scores, although the English and the French words are almost identical in pronunciation.

		A						
Fúzhōu	ŋuoʔ^{5}							
Měixiàn	ŋiat^{5}	0.44						
	kuoŋ^{44}	0.78	0.78					
Wēnzhōu	ɳy^{21}	0.30	0.35	0.67				
	kuɔ^{35}	0.80	0.85	0.27	0.67			
	vai^{13}	0.85	0.85	0.82	0.73	0.73		
Běijīng	yɛ^{51}	0.77	0.84	0.73	0.56	0.56	0.66	
	liɑŋ^{1}	0.78	0.78	0.44	0.67	0.82	0.82	0.80

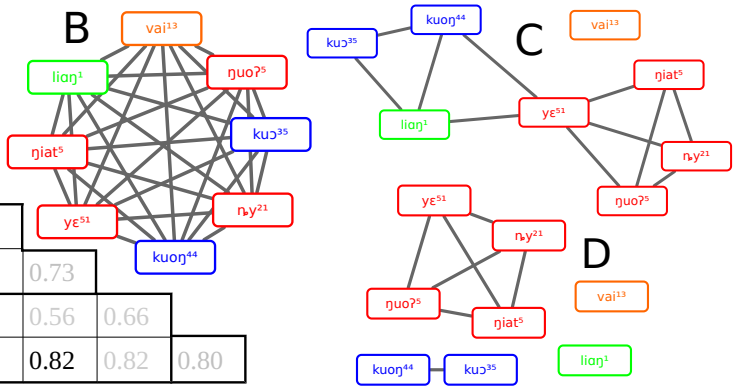


Figure 1: Similarity networks for partial cognate detection. A shows pairwise SCA distances computed between all morphemes of Chinese dialect words for `moon'. Values shaded in gray are excluded following filtering rules 1 and 2 (see text). B shows the initial similarity network with all nodes connected. C shows the network after filtering, and D shows the network after applying the partitioning algorithm.

(2014b). As a general rule, language-specific approaches outperform language-independent ones, provided the sample size is large enough (List 2014a).

Two similarity measures are used in this paper, one language-independent, and one language-specific one. The above-mentioned SCA method for phonetic alignments (List 2012b, 2014b) reduces the phonetic space of sound sequences to 28 sound classes. Based on a scoring function which defines transition scores between the sound classes, phonetic sequences are aligned and similarity and distance scores can be determined. The LexStat approach List (2012a, 2014b) also uses sound classes, but instead of using a pre-defined scoring function, transition scores between sound classes are determined with help of a permutation test. In this test, words drawn from a randomized sample are repeatedly aligned with each other in order to create a distribution of sound transitions for unrelated languages. This distribution is then compared with the actual distribution retrieved from aligned words in the word list, and a language-specific scoring function is created List (2014b). SCA is very fast in computation, but LexStat has a much higher accuracy. Both approaches are freely available as part of the LingPy software package (List and Forkel 2016).

3.2 Sequence Similarity Networks

Sequence similarity networks are tools for exploratory data analysis. In evolutionary biology they are used to study complex evolutionary processes (Méheust et al. 2016, Corel et al. 2016). They represent sequences as nodes and connections between nodes represent similarities which are usually determined from similarity scores exceeding a certain threshold (Alvarez-Ponce et al. 2013). Since evolutionary processes leave specific traces in the network topology, they can be identified by applying techniques for network analysis. In linguistics, sequence similarity networks have been rarely applied (Lopez et al. 2013), although they are applicable, provided that one uses informed measures for phonetic similarity.

For the application of sequence similarity networks it is essential to decide when to draw an edge between two nodes and when not. For the new approach to partial cognate detection, three filtering criteria are applied. (1) No edges are drawn between morphemes which occur in the same word. (2) No morpheme in one word is linked to two morphemes in another word, with the preference given to morpheme pairs with the lowest phonetic distance applying a greedy strategy. (3) Edges are only drawn when the phonetic distance between the morphemes is beyond a certain threshold. The application of the filtering criteria is illustrated in Fig. 1 for the exemplary words shown in Table 1.

3.3 Network Partitioning

Cognate morphemes in a similarity network can be found by partitioning the network into groups. Many algorithms are available for this purpose, as can be seen from evolutionary biology, where homology detection is frequently approached from a network perspective (Vlasblom and Wodak 2009). Three different algorithms were tested for this purpose. A flat version of the UPGMA algorithm for hierarchical clustering (Sokal and Mich-

ener 1958), which terminates when a certain user-defined threshold is reached is originally underlying the LexStat algorithm and was therefore also included in this study. Markov Clustering (van Dongen 2000) uses techniques for matrix multiplication to inflate and expand the edge weights in a given network until weak edges have disappeared and a few clusters of connected nodes remain. Markov Clustering is very popular in biology and was shown to outperform the popular Affinity Propagation algorithm (Frey and Dueck 2007) in the task of homolog detection in biology (Vlasblom and Wodak 2009). As a third method, we follow List et al. (2016b) in testing Infomap (Rosvall and Bergstrom 2008), a method that was originally designed to detect *communities* in complex networks. Communities are groups that share more links with each other than outside the group (Newman and Girvan 2004). Infomap uses random walks to find the best partition of a network into communities. Infomap is not a classical partitioning algorithm, and we do not know of any studies which tested its suitability for the task of homolog detection in evolutionary biology, but according to List et al. (2016b), Infomap shows a better performance than UPGMA in automatic cognate detection.

3.4 Analyses and Evaluation

All methods, be it classical or partial cognate detection, require a user-defined threshold. Since our gold standard data was too small to split it into training and tests sets, we carried out an exhaustive comparison of all methods on different thresholds varying between 0.05 and 0.95 in steps of 0.05. B-cubed scores were chosen as an evaluation measure for cognate detection (Bagga and Baldwin 1998), since they have been shown to yield sensible results (Hauer and Kondrak 2011).

With SCA and LexStat, two classical methods for cognate detection were tested List (2014b), and their underlying models for phonetic similarity (see Sec. 3.1) were used as basis for the partial cognate detection algorithm. All in all, this yielded four different methods: LexStat, LexStat-Partial, SCA, and SCA-Partial. Since our new algorithms yield partial cognates, while LexStat and SCA yield ``complete'' cognates, it is not possible to compare them directly. In order to allow for a direct comparison, partial cognate sets were converted into ``complete'' cognate sets using the above-mentioned strict coding approach proposed by Ben Hamed and Wang (2006): only those words in which *all* morphemes are cognate were assigned to the cognate same set. With a total of three different clustering algorithms (UPGMA, Markov Clustering, and Infomap), we thus carried out twelve tests on complete cognacy (three for each of our four approaches), and six additional tests on pure partial cognate detection, in which we compared the suitability of SCA and LexStat as string similarity measures.

LexStat				
Cluster-Method	**T**	**P**	**R**	**FS**
UPGMA	0.60	0.9030	0.8743	0.8878
Markov	0.50	0.9123	0.8752	0.8933
Infomap	0.50	0.9131	0.8866	0.8995
SCA				
Cluster-Method	**T**	**P**	**R**	**FS**
UPGMA	0.45	0.8595	0.8707	0.8648
Markov	0.45	0.8049	0.8097	0.8031
Infomap	0.35	0.8901	0.8573	0.8734
LexStat-Partial Complete Cognacy				
Cluster-Method	**T**	**P**	**R**	**FS**
UPGMA	0.90	0.9193	0.9638	0.9399
Markov	0.70	0.9275	0.9342	0.9298
Infomap	0.65	0.9453	0.9363	0.9404
SCA-Partial Complete Cognacy				
Cluster-Method	**T**	**P**	**R**	**FS**
UPGMA	0.60	0.9304	0.9045	0.9172
Markov	0.95	0.8153	0.8949	0.8446
Infomap	0.55	0.9104	0.9366	0.9223
LexStat-Partial Partial Cognacy				
Cluster-Method	**T**	**P**	**R**	**FS**
UPGMA	0.75	0.8920	0.8820	0.8867
Markov	0.60	0.8858	0.8724	0.8782
Infomap	0.60	0.8876	0.8844	0.8856
SCA-Partial Partial Cognacy				
Cluster-Method	**T**	**P**	**R**	**FS**
UPGMA	0.50	0.8597	0.8509	0.8552
Markov	0.50	0.8074	0.7621	0.7755
Infomap	0.35	0.8676	0.8439	0.8553

Table 3: General performance of the algorithms on all datasets. The table shows for each of the 18 different methods the threshold (T) for which the best B-Cubed F-Score was determined, as well as the B-Cubed precision (P), recall (R), and F-score (FS). The best result in each block is shaded in gray.

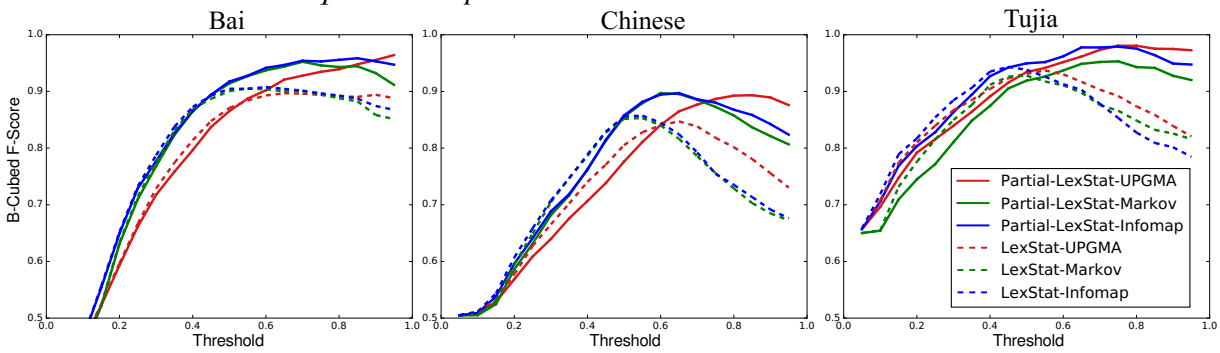


Figure 2: Comparing the results for the LexStat sequences similarities

3.5 Implementation

The code was implemented in Python, as part of the LingPy library (Version 2.5, List and Forkel (2016), `http://lingpy.org`). The Igraph software package (Csárdi and Nepusz 2006) is needed to apply the Infomap algorithm.

4 Results

The aggregated results of the test (thresholds, precision, recall, and F-scores) are given in Table 3, specific results for the comparison of LexStat with LexStat-Partial are given in Table 3. In general, one can clearly see that the partial cognate detection algorithms outperform their non-partial counterparts when applying the complete cognacy measure. The differences are very striking, with LexStat-Partial outperforming its non-partial counterpart by up to four points, and SCA-Partial outperforming the classical SCA variant by almost five points.[2] In contrast, we do not find strong differences in the performance of the cluster algorithms. Infomap outperforms the other cluster algorithms in almost all tests (all other aspects being equal), but the differences are not high enough to make any further conclusions at this point.

When comparing the aggregated results for the true evaluation of partial cognate detection (the last two blocks in Figure 2), the scores are less high than in the complete cognate analyses. Given that we cannot detect any striking tendency, like a drastic drop of precision or recall, this suggests that the algorithms generally loose accuracy in the task of ``true'' partial cognate detection. This is surely not surprising, since the task of detecting exactly which morphemes in the data are historically related is much more complex than the task of detecting which words are completely cognate.

[2]By one point, we mean 0.01 on the B-Cube scale.

In Figure 2, detailed analyses for the LexStat analyses with complete cognate evaluation (the first and the third block in Table 3) are shown for each of the datasets, and throughout all thresholds we tested. The superior performance of the partial cognate detection variants is reflected in all datasets. That the internal diversity of the Chinese languages largely exceeds Bai and Tujia can be seen from the generally lower scores which all algorithms achieve for the datasets.

5 Discussion

This paper has presented a pilot approach for the detection of partial cognates in multilingual word lists. Although the results are very promising at this stage, we can think of many points where improvement is needed, and further studies are needed to fully assess the potential of the current approach. First, it should be tested on additional datasets, and ideally also on language families other than Sino-Tibetan. Second, since our approach is very general, it can easily be adjusted to employ different string similarity measures or different partitioning algorithms, and it would be interesting to see whether alternative measures can improve upon our current version.

Acknowledgments

This research was supported by the DFG research fellowship grant 261553824 *Vertical and lateral aspects of Chinese dialect history* (JML). EB is supported by by the ERC under the European Community's Seventh Framework Programme, FP7/2007-2013 Grant Agreement # 615274.

Supplementary Material

The Sup. Material contains results, benchmark datasets, and code, downloadable at: `https://zenodo.org/record/51328`.

References

David Alvarez-Ponce, Philippe Lopez, Eric Bapteste, and James O. McInerney. 2013. Gene similarity networks provide tools for understanding eukaryote origins and evolution. *Proceedings of the National Academy of Sciences of the United States of America* 110(17):E1594--1603.

Quentin D. Atkinson and Russell D. Gray. 2006. How old is the Indo-European language family? Illumination or more moths to the flame? In Peter Forster and Colin Renfrew, editors, *Phylogenetic methods and the prehistory of languages*, McDonald Institute for Archaeological Research, Cambridge, pages 91--109.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 36th Annual Meeting of the ACL*. pages 79--85.

Mahe Ben Hamed and Feng Wang. 2006. Stuck in the forest: Trees, networks and Chinese dialects. *Diachronica* 23:29--60.

Běijīng Dàxué 北京大学, editor. 1964. *Hànyǔ fāngyán cíhuì* 汉语方言词汇[Chinese dialect vocabularies]. Wénzì Gǎigé 文字改革, Běijing 北京.

Eduardo Corel, Philippe Lopez, Raphaël Méheust, and Eric Bapteste. 2016. Network-thinking: Graphs to analyze microbial complexity and evolution. *Trends Microbiology* 24(3):224--237.

Gábor Csárdi and Tamás Nepusz. 2006. The igraph software package for complex network research. *InterJournal Complex Systems* page 1695.

Brendan J. Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *Science* 315:972--976.

Leanne S. Haggerty, Pierre-Alain A. Jachiet, William P. Hanage, David A. Fitzpatrick, Philippe Lopez, Mary J. O'Connell, Davide Pisani, Mark Wilkinson, Eric Bapteste, and James O. McInerney. 2014. A pluralistic account of homology: adapting the models to the data. *Mol. Biol. Evol.* 31(3):501--516.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2015. *Glottolog*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Bradley Hauer and Grzegorz Kondrak. 2011. Clustering semantically equivalent words into cognate sets in multilingual lists. In *Proceedings of the 5th International Joint NLP conference*. pages 865--873.

Hóu, Jīngyī 侯精一, editor. 2004. *Xiàndài Hànyǔ fāngyán yīnkù* 现代汉语方言音库[Phonological database of Chinese dialects]. Shànghǎi Jiàoyù 上海教育, Shànghǎi 上海.

Gerhard Jäger. 2013. Phylogenetic inference from word lists using weighted alignment with empirical determined weights. *Language Dynamics and Change* 3(2):245--291.

Johann-Mattis List. 2012a. Lexstat. automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS and UNCLH*. Stroudsburg, pages 117--125.

Johann-Mattis List. 2012b. SCA. phonetic alignment based on sound classes. In Marija Slavkovik and Dan Lassiter, editors, *New directions in logic, language, and computation*, Springer, Berlin and Heidelberg, pages 32--51.

Johann-Mattis List. 2014a. Investigating the impact of sample size on cognate detection. *Journal of Language Relationship* 11:91--101.

Johann-Mattis List. 2014b. *Sequence comparison in historical linguistics*. Düsseldorf University Press, Düsseldorf. URL: `http://sequencecomparison.github.io`.

Johann-Mattis List. 2016. Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction. *Journal of Language Evolution* 1(2). Published online before print.

Johann-Mattis List, Michael Cysouw, and Robert Forkel. 2016a. *Concepticon: A resource for the linking of concept lists*. Max Planck Institute for the Science of Human History, Jena. Version: 1.0, URL: `http://concepticon.clld.org`.

Johann-Mattis List and Robert Forkel. 2016. *LingPy. A Python library for historical linguistics*. Max Planck Institute for the Science of Human History, Jena. Version 2.5. URL: `http://lingpy.org`. With contributions by Steven Moran, Peter Bouda, Johannes Dellert, Taraka Rama, Frank Nagel, and Simon Greenhill.

Johann-Mattis List, Simon Greenhill, and Russell Gray. 2016b. The potential of automatic cognate

detection for historical linguistics. Manuscript in preparation.

Philippe Lopez, Johann-Mattis List, and Eric Bapteste. 2013. A preliminary case for exploratory networks in biology and linguistics. In Heiner Fangerau, Hans Geisler, Thorsten Halling, and William Martin, editors, *Classification and evolution in biology, linguistics and the history of science*, Franz Steiner Verlag, Stuttgart, pages 181--196.

Wilhelm Meyer-Lübke. 1911. *Romanisches etymologisches Wörterbuch*. Winter, Heidelberg.

Raphaël Méheust, Ehud Zelzion, Debashish Bhattacharya, Philippe Lopez, and Eric Bapteste. 2016. Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis. *Proceedings of the National Academy of Sciences of the United States of America* 113(3): 3579--3584.

M. E. J. Newman and M. Girvan. 2004. Finding and evaluating community structure in networks. *Physical Review E* 69(2):026113+.

Martin Rosvall and Carl T. Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America* 105(4):1118--1123.

Damian Satterthwaite-Phillips. 2011. *Phylogenetic inference of the Tibeto-Burman languages*. PhD Thesis, Stanford University, Stanford.

Robert. R. Sokal and Charles. D. Michener. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin* 28:1409--1438.

George S. Starostin. 2013a. Annotated Swadesh wordlists for the Tujia group. In George Starostin, editor, *The Global Lexicostatistical Database*, RGGU, Moscow. URL: `http://starling.rinet.ru/new100/tuj.xls`.

George S. Starostin. 2013b. Lexicostatistics as a basis for language classification. In Heiner Fangerau, Hans Geisler, Thorsten Halling, and William Martin, editors, *Classification and evolution in biology, linguistics and the history of science*, Franz Steiner Verlag, Stuttgart, pages 125--146.

Robert L. Trask. 2000. *The dictionary of historical and comparative linguistics*. Edinburgh University Press, Edinburgh.

Peter Turchin, Ilja Peiros, and Murray Gell-Mann. 2010. Analyzing genetic connections between languages by matching consonant classes. *Journal of Language Relationship* 3:117--126.

Stijn M. van Dongen. 2000. *Graph clustering by flow simulation*. PhD Thesis, University of Utrecht.

James Vlasblom and Shoshana J. Wodak. 2009. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC Bioinformatics* 10:99.

Feng Wang. 2006. *Comparison of languages in contact*. Academia Sinica, Taipei.

RESEARCH ARTICLE

The Potential of Automatic Word Comparison for Historical Linguistics

Johann-Mattis List[1]*, **Simon J. Greenhill**[2,3], **Russell D. Gray**[2]

1 Centre des Recherches Linguistiques sur l'Asie Orientale, École des Hautes Études en Sciences Sociales, 2 Rue de Lille, 75007 Paris, France, **2** Department for Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Kahlaische Straße 10, 07743, Jena, Germany, **3** ARC Centre of Excellence for the Dynamics of Language, Australian National University, Canberra, 2600, Australia

* mattis.list@lingpy.org

OPEN ACCESS

Citation: List J-M, Greenhill SJ, Gray RD (2017) The Potential of Automatic Word Comparison for Historical Linguistics. PLoS ONE 12(1): e0170046. doi:10.1371/journal.pone.0170046

Editor: Robert C Berwick, Massachusetts Institute of Technology, UNITED STATES

Received: October 18, 2016

Accepted: December 28, 2016

Published: January 27, 2017

Data Availability Statement: The Supplementary Material contains additional results, as well as data and code to replicate the analyses. You can download it from: https://zenodo.org/badge/latestdoi/75610836 (DOI:10.5281/zenodo.192607).

Funding: As part of the GlottoBank Project, this work was supported by the Max Planck Institute for the Science of Human History and the Royal Society of New Zealand Marsden Fund grant 13¬UOA-121. This paper was further supported by the DFG research fellowship grant 261553824 "Vertical and lateral aspects of Chinese dialect history" (JML), and the Australian Research

Abstract

The amount of data from languages spoken all over the world is rapidly increasing. Traditional manual methods in historical linguistics need to face the challenges brought by this influx of data. Automatic approaches to word comparison could provide invaluable help to pre-analyze data which can be later enhanced by experts. In this way, computational approaches can take care of the repetitive and schematic tasks leaving experts to concentrate on answering interesting questions. Here we test the potential of automatic methods to detect etymologically related words (cognates) in cross-linguistic data. Using a newly compiled database of expert cognate judgments across five different language families, we compare how well different automatic approaches distinguish related from unrelated words. Our results show that automatic methods can identify cognates with a very high degree of accuracy, reaching 89% for the best-performing method *Infomap*. We identify the specific strengths and weaknesses of these different methods and point to major challenges for future approaches. Current automatic approaches for cognate detection—although not perfect—could become an important component of future research in historical linguistics.

Introduction

Historical linguistics is currently facing a dramatic increase in digitally available datasets [1–5]. The availability of data for more and more languages and language families challenges the ways in which we traditionally compare them. The comparative method has been the core method for linguistic reconstruction for the past 200 years [6], and is based on manually identifying systematic phonetic correspondences between many words in pairs of languages. However, there are too few expert historical linguists to analyse the world's more than 7500 languages [7] and, consequently, only a small percentage of these languages have been thoroughly investigated leaving us in the dark about their history and relationships. This becomes especially evident in largely understudied linguistic areas like New Guinea, parts of South America, or the Himalayan region, and our lack of knowledge about these languages has immediate implications for our understanding of human prehistory.

Council's Discovery Projects funding scheme (project number DE120101954, SJG).

Competing Interests: The authors have declared that no competing interests exist.

Over the last two decades computational methods have been become more prevalent in historical linguistics. Advocates of computational methods emphasize the speed and replicability as the main advantage of computational techniques [8, 9]. However, sceptics criticise the validity and accuracy of these methods as lagging far behind those achieved by human experts. [10]. One approach in computational historical linguistics is to design fully-automated methods to identify language relationships with no input from researchers [11, 12]. Although these methods may provide interesting insights into linguistic macroareas [13], their "black-box" character makes it difficult to evaluate the results, as judgements about sound correspondences and decisions of cognacy are hidden. This opacity makes it difficult to improve the algorithms. More problematically, however, it limits the scientific value of these methods, as we do not just want to know how languages are related, but why and which pieces of evidence support this conclusion. As a result, there is much suspicion about these methods in historical linguistics [14–16].

Another approach—the one we take here—is to opt for a computer-assisted framework. In contrast to fully automated frameworks, computer-assisted frameworks seek to support and facilitate the task of language comparison by using human expertise where available to correct errors and improve the quality of the results. One of the core tasks of the comparative method is the identification of cognate words in multiple languages. If two words are cognate, this means that they are genetically related, and have descended from a common ancestor [17]. Cognate identification, along with the identification of regular sound correspondences, is the basis for proving that two or more languages are genetically related. It is also the basis for the reconstruction of ancestral word forms in historically unattested languages, and for the genetic classification of language families. In practice, cognate identification is a time-consuming process that is based on an iterative manual procedure where cognate sets are proposed, evaluated, and either kept or rejected [18].

This process of manual cognate identification should be an ideal candidate for computer-assisted tasks. As a possible workflow, scholars could first run an automatic cognate detection analysis and then edit the algorithmic findings. Even an iterative workflow in which the data is passed between computers and experts would be fruitful. An important question which arises in this context concerns the quality of automatic methods for cognate detection: Are these methods really good enough to provide concrete help to a highly trained expert? In order to find an answer to this question, we tested four publicly available methods and one newly proposed method for automatic cognate detection on six test sets covering five different language families, evaluated the performance of these methods, and determined their shortcomings.

Materials and Methods

Materials

There are few datasets available for testing the potential of cognate detection methods on language data, As such, testing algorithms run the risk of *over-fitting*. When developing an algorithm, one usually *trains* it on some datasets. If those datasets are afterwards used to also test the algorithm, the accuracy should be quite high, but we cannot tell whether the method will work on datasets apart from the ones on which the algorithm was trained. For this reason, it is important to split the available data into a training set and a test set. In our case, the training set will be used to determine the best parameters for each of the algorithms we test, while the test set will be used to carry out the actual test of cognate recovery.

For this study, we took training data from existing sources [19], while a new test dataset was compiled from scratch. The new test set consists of six datasets from five language families. These data were collected from different sources, including published datasets [3, 20–23],

Table 1. Test data used in our study.

Dataset	Words	Conc.	Lang.	Cog.	Div.
Bahnaric (Sidwell, 2015) [20]	4546	200	24	1055	0.20
Chinese (Běijīng Dàxué, 1964) [24]	3653	180	18	1231	0.30
Huon (McElhanon, 1967) [22]	1668	139	14	855	0.47
Romance (Saenko, 2015) [21]	4853	110	43	465	0.07
Tujia (Starostin, 2013) [23]	513	109	5	179	0.17
Uralic (Syrjänen et al, 2013) [25]	1401	173	7	870	0.57
TOTAL	16634	911	111	4655	0.30

doi:10.1371/journal.pone.0170046.t001

books [24], and ongoing research by scholars who allowed us to use parts of their data in advance (Uralex project, [25]). All datasets were formatted to tabular format and semi-automatically cleaned for various kinds of errors, like misspelled phonetic transcriptions, empty word slots, or obviously erroneous cognate judgments. We further linked all languages to Glottolog [7], and all wordlist concepts to the Concepticon [26].

Table 1 lists all datasets along with additional details, such as the number of words, concepts, languages, and cognate sets in the data. The diversity index given in the last column of the table is calculated by dividing the difference between cognate sets and meanings with the difference between words and meanings [19]. This score, which ranges between 0 and 1, indicates whether large numbers of words in a given dataset are unrelated (high index) or are cognate (low index). As can be seen from the diversity indices listed in the table, our test sets have varying degrees of diversity, ranging from 0.07 (Romance, Saenko, 2015) to 0.57 (Uralic).

As mentioned above, training data is needed for parameter estimation. The key parameter we need to estimate is the *best thresholds* for cognate identification in some of the methods. As training data we employed the collection of benchmark datasets for automatic cognate detection by List [19], which also covers six datasets from five language families. Details for this dataset (number of words, concepts, languages, cognate sets, and the diversity index) are given in Table 2. This dataset is available online at http://dx.doi.org/10.5281/zenodo.11877.

Methods

Automatic Cognate Detection. Many methods for automatic cognate detection have been proposed in the past (see Table 3 below). Unfortunately, only a few of these methods qualify as candidate methods for computer-assisted language comparison, since the majority are either (a) not able to analyse multiple languages at once, (b) have further requirements making their use more complicated [31, 32] e.g. require a user-specified reference phylogeny (and therefore assume that language groupings are already known), or need extensive training sets, or (c) are not freely available (see Table 3).

Table 2. Training data used in our study.

Dataset	Words	Conc.	Lang.	Cog.	Div.
Austronesian (Greenhill et al., 2008) [1]	4358	210	20	2864	0.64
Bai (Wang, 2006) [27]	1028	110	9	285	0.19
Chinese (Hóu, 2004) [28]	2789	140	15	1189	0.40
IndoEuropean (Dunn, 2012) [2]	4393	207	20	1777	0.38
Japanese (Hattori, 1973) [29]	1986	200	10	460	0.15
ObUgrian (Zhivlov, 2011) [30]	2055	110	21	242	0.07
TOTAL	16609	977	95	6817	0.30

doi:10.1371/journal.pone.0170046.t002

Table 3. Recent approaches to cognate detection. A plus "+" indicates that the algorithm meets the requirement, a minus "-" indicates that its failure. ML (multilingual) refers to the ability of an algorithm to identify cognate words across more than two languages at the same time. RQ (requirements) refers to additional requirements apart from the raw word list data, such as needing reference phylogenies or extensive training data. FA (free availability) means that the method has a useable public implementation.

Cognate Detection Approach	ML?	RQ?	FA?
Mackay and Kondrak, 2005, [34]	-	+	-
Bergsma and Kondrak, 2007, [35]	+	+	-
Turchin et al., 2010, [44]	+	+	+
Berg-Kirkpatrick and Klein, 2011, [36]	-	+	-
Hauer and Kondrak, 2011, [37]	+	+	-
Steiner et al., 2011, [38]	+	+	-
List, 2014, [19]	+	+	+
Beinborn et al., 2013, [31]	-	-	-
Bouchard-Côté, et al. 2013, [32]	+	-	-
Rama, 2013, [39]	-	+	-
Ciobanu and Dinu, 2014, [40]	-	+	-
Jäger and Sofroniev 2016, [41]	+	-	-

doi:10.1371/journal.pone.0170046.t003

We decided to take four publicly available methods as the basis of our test study, the Turchin Method, the Edit Distance Method, the SCA Method, and the LexStat Method. Additionally, we tested a modified version of the LexStat method which we call Infomap. In this modified version of LexStat we introduced an improved partitioning method based on the Infomap algorithm for community detection [33]. All methods are presented in more detail below.

The four publicly available methods are all implemented as part of the same software package (LingPy, http://lingpy.org, [42]), and represent different degrees of algorithmic sophistication and closeness to linguistic theory, with the Turchin Method being very simple and computationally extremely fast, and the LexStat Method being rather complex and time-consuming. For the usage of the fifth method, we wrote a small LingPy plugin which builds on the python-igraph package (http://igraph.org/python-igraph/, [43], see details below) and is provided along with our supplementary material.

Cognate Detection following Turchin et al [44]. The Turchin method (also called *Consonant Class Matching* approach) was proposed by Turchin et al. [44]. In this method, the consonants of the words are converted to one of 10 possible consonant classes. The idea of consonant classes (also called sound classes) was proposed by Dolgopolsky [45], who stated that certain sounds occur more frequently in correspondence relation than others and could therefore be clustered into classes of high historical similarity. In the approach by Turchin et al., two words are judged to be cognate, if they match in their first two consonant classes.

Cognate Detection using the Edit Distance approach. A second method provided by LingPy, the Edit Distance approach, takes the normalized Levenshtein distance [46], between all word pairs in the same meaning slot and clusters these words into potential cognate sets using a flat version of the UPGMA algorithm [47] which terminates once a certain threshold of average distances between all words is reached. This general procedure of flat clustering, which is also employed for the two remaining cognate detection methods provided by LingPy, is illustrated in Fig 1A and 1B.

Cognate Detection using the Sound Class Algorithm. A third method available in the LingPy package, the SCA method, uses the same threshold-based clustering algorithm as the Edit Distance but employs distance scores derived from the Sound-Class Based Alignment

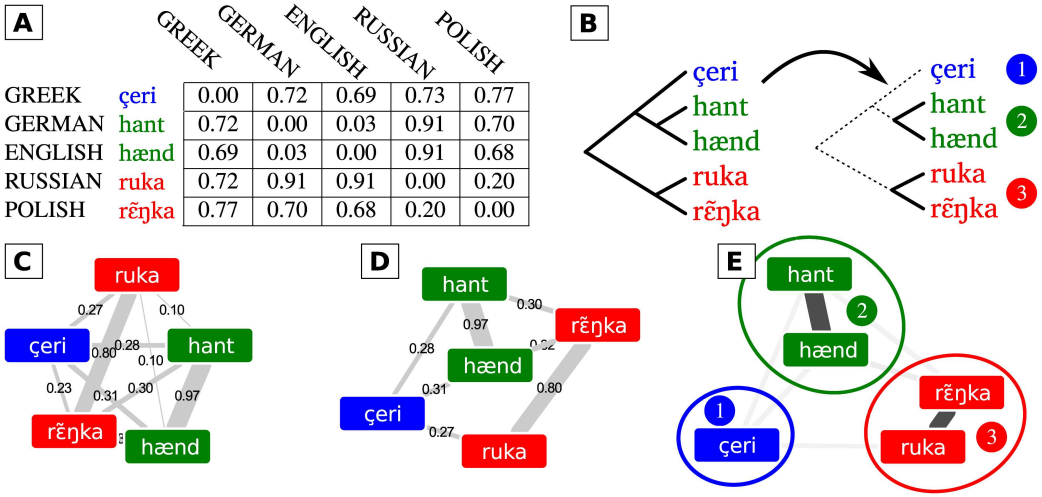


Fig 1. Workflows for automatic cognate detection. In LingPy, cognate detection is treated as a hierarchical clustering task. After distances or similarities between word pairs have been determined (A), a hierarchical clustering algorithm is applied to the matrix and terminates when a certain threshold is reached (B). Similarity networks start from a graph-representation of the similarity or distance matrix (C). In a first step, edges whose score exceeds a certain threshold are removed from the graph (D). In a second step, state-of-the-art algorithms for community detection are used to partition the graph into groups of cognate words (E).

doi:10.1371/journal.pone.0170046.g001

(SCA) method [19]. This method for pairwise and multiple alignment analyses uses expanded sound class models along with detailed scoring functions as its basis. In contrast to previous alignment algorithms [48], the SCA algorithm takes prosodic aspects of the words into account and is also capable of aligning within morpheme boundaries, if morpheme information is available in the input data [19].

Cognate Detection using the LexStat method. The last publicly available method we tested, the LexStat method, is again based on flat UPGMA clustering, but in contrast to both the Edit-Distance method and the SCA method, it uses language-specific scoring schemes which are derived from a Monte-Carlo permutation of the data [19]. This permutation, by which the wordlists of all language pairs are shuffled in such a way that words denoting different meanings are aligned and scored, is used to derive a distribution of sound-correspondence frequencies under the assumption that both languages are not related. The permuted distribution is then compared with the attested distribution, and converted into a language-specific scoring scheme for all language pairs. Using this scoring scheme, the words in the data are aligned again, and distance scores are derived which are then used as the basis for the flat cluster algorithm.

Differences between algorithms. In order to illustrate the differences between these four algorithms, we analysed the test set by Kessler [49]. This dataset is particularly interesting for the task of cognate detection, since the sample of languages contains not only four Indo-European languages with different degrees of genetic affiliation, but also unrelated languages from different language families. When running the algorithm with default thresholds as proposed in List [19], LexStat performs best, showing the smallest amount of false positives and false negatives, followed by SCA, Edit-Distance, and Turchin. When looking at specific results of this analysis, like the cognate judgments for the concept 'there', given in Table 4, for example, we can immediately see the shortcomings of the language-independent methods. The Turchin method (T), for example, links Albanian [aty] and Navajo [ʔaːdi] as cognate, where these are a

Table 4. Cognate detection algorithms in LingPy. Columns show the performance of cognate identification for the given wordforms in the International Phonetic Alphabet (IPA). The algorithms are the **T**urchin, **E**dit distance, **S**ound Class Algorithm, and **L**exStat methods. Italic numbers indicate false positives (forms incorrectly identified as cognate) and bold numbers indicate false negatives (forms incorrectly identified as not cognate) in comparison with the **G**old Standard.

Language	Word	IPA	T	E	S	L	G
Albanian	aty	aty	*1*	1	1	1	1
English	there	ðɛr	**2**	**2**	**2**	2	2
French	là	la	3	*3*	3	3	3
German	da	daː	**4**	**4**	**4**	2	2
Hawaiian	laila	laila	5	*3*	5	4	4
Navajo	’áadi	ʔaːdi	*1*	5	6	5	5
Turkish	orada	ora	6	6	7	6	6

doi:10.1371/journal.pone.0170046.t004

clear chance resemblance in the consonant class structure. Note that initial vowel is treated identical with initial glottal stop in the Turchin method, following the original sound class proposal by [45].

The Edit Distance (E) method also identifies a chance resemblance by proposing that French [la] and Hawaian [laila] are cognate. The Edit-Distance method is especially prone to identifying chance similarity as cognacy, and this risk increases as languages get more and more different [15]. The threshold of the SCA method (S) is too low to identify any cognate set for the concept ‘there’. Only the LexStat method (L) correctly identifies English [ðɛr] and German [daː] as cognates, but not due to the phonetic similarity of the words, but due to the fact that matches of English [ð] and German [d] recur frequently in the dataset.

Similarity Networks. All the above cognate detection methods currently use a rather simple flat clustering procedure. The basis of this procedure is a clustering algorithm which terminates when average distances among sequences exceed a certain threshold. In evolutionary biology, the task of homolog detection is often approached from a *network perspective*. In *similarity networks*, for example, gene or protein sequences are modeled as the nodes of a network, and edges between the nodes are drawn with weights representing the pairwise similarities [50, 51]. Homolog detection is then modeled as a network partitioning task by which the network is divided into subgraphs with some objective criterion being used to define the best partition of the original network. While originally developed for the application in evolutionary biology, sequence similarity networks are now increasingly being tested on linguistics data [52, 53] and it was proposed that they might not only help to detect both genetically related words as well as words which have been borrowed [54]. Many strategies for network partitioning exist. The most common methods used in biology are Markov Clustering [55], *k*-means [56], and Affinity Propagation [57]. *k*-means has the strong disadvantage that it requires that the number of clusters into which the data shall be partitioned needs to be specified in advance. Tests in evolutionary biology have further shown that Markov Clustering outperforms Affinity Propagation [58]. This finding suggests that Markov Clustering would be an ideal choice for linguistic applications. However, when testing the approach on our training data, the results were inconclusive, and no real improvement compared with the default clustering algorithm used in LingPy could be observed.

For this study, we followed List et al. [53] in testing a partitioning approach which was originally developed for the task of community detection in social network analysis [59] and has shown to perform with a high accuracy: The Infomap algorithm [33] uses random walks to identify the best way to assign the nodes in a network to distinct communities. In order to convert the matrix of pairwise distances between words into a graph, we first define a threshold,

and then add edges between all words whose pairwise distance is below the threshold. The edge weight is the distance score converted to a similarity score by subtracting it from 1. We use the pairwise distance matrices produced by the LexStat method, since this was shown to outperform the other three methods implemented in LingPy [19]. How cognate detection is modeled as a graph partitioning problem applied to similarity networks is displayed in more detail in Fig 1C and 1D.

Evaluation. It is not necessarily an easy task to compare how well an algorithm for automatic cognate detection performs in comparison with a "gold" standard. In our study, our gold standard are the expert cognate decisions by historical linguists using the comparative method. Scholars often use pairwise scores [32] for evaluation. In these scores, all words in a concept slot are assembled into pairs. The pair score is then calculated by comparing how many pairs in the gold standard are identically clustered by the algorithm, and vice versa. This is simple and straightforward, since, for pairs, there are only two possible decisions, namely whether they are cognate or not. We can then simply count how many pairs in the gold standard are also judged to be cognate by the algorithm, or how many pairs proposed to be cognate by the algorithm are also cognate according to the gold standard. The advantage of this score is that we can directly convert it into an intuitive notion of *false positives* and *false negatives* versus *true positives* and *true negatives*.

Breaking down the comparison of two different clusters into pairs is, however, problematic, since it has a strong bias in favoring datasets containing large amounts of non-cognate words [19]. In order to avoid these problems, we used B-Cubed scores as our primary evaluation method [37, 60, 61]. For the calculation of B-Cubed scores, we need to determine for each of the words the intersection of words between its cognate set in the gold standard and its cognate set proposed by the algorithm, as well as the size of the respective cognate sets. This is illustrated in Table 5 for a fictive test analysis of the five words in Fig 1, which wrongly clusters the Greek word with the English and the German word. For the B-Cubed precision we then average the size of the intersection divided by the size of the cognate set proposed by the algorithm for each of the words in our sample:

$$P = \frac{\frac{1}{3} + \frac{2}{3} + \frac{2}{3} + \frac{2}{2} + \frac{2}{2}}{5} = 0.7\overline{3} \quad (1)$$

For the B-Cubed recall we average the intersection size divided by the cognate set size in the gold standard:

$$R = \frac{\frac{1}{1} + \frac{2}{2} + \frac{2}{2} + \frac{2}{2} + \frac{2}{2}}{5} = 1.0 \quad (2)$$

Table 5. Preliminaries for B-Cubed score calculation. Cognate clusters, cluster size and cluster intersection for a fictive test analysis of the five words from Fig 1 compared to a gold standard.

Word	Cogn. Clusters		Cluster Size		Intersection
	Gold	Test	Gold	Test	
çeri	1	1	1	3	1
hant	2	1	2	3	2
hænd	2	1	2	3	2
ruka	3	2	2	2	2
rɛ̃ŋka	3	2	2	2	2

doi:10.1371/journal.pone.0170046.t005

The B-Cubed F-Score is then computed as usual:

$$F = 2 \times \frac{P \times R}{P + R} = 2 \times \frac{0.7\overline{3} \times 1}{0.7\overline{3} + 1} = 0.\overline{846153} \quad (3)$$

Threshold and Parameter Selection

Apart from the Turchin method, all analyses require a threshold which ranges between 0 and 1, denoting the amount of similarity needed to judge two items as cognate. In order to find the most suitable threshold for each of the three methods, we used the expert cognate decisions in our training set and ran the analyses on these data with varying thresholds starting from 0.05 up to 0.95. Fig 2 shows box-plots of the training analyses for the four methods, depending on

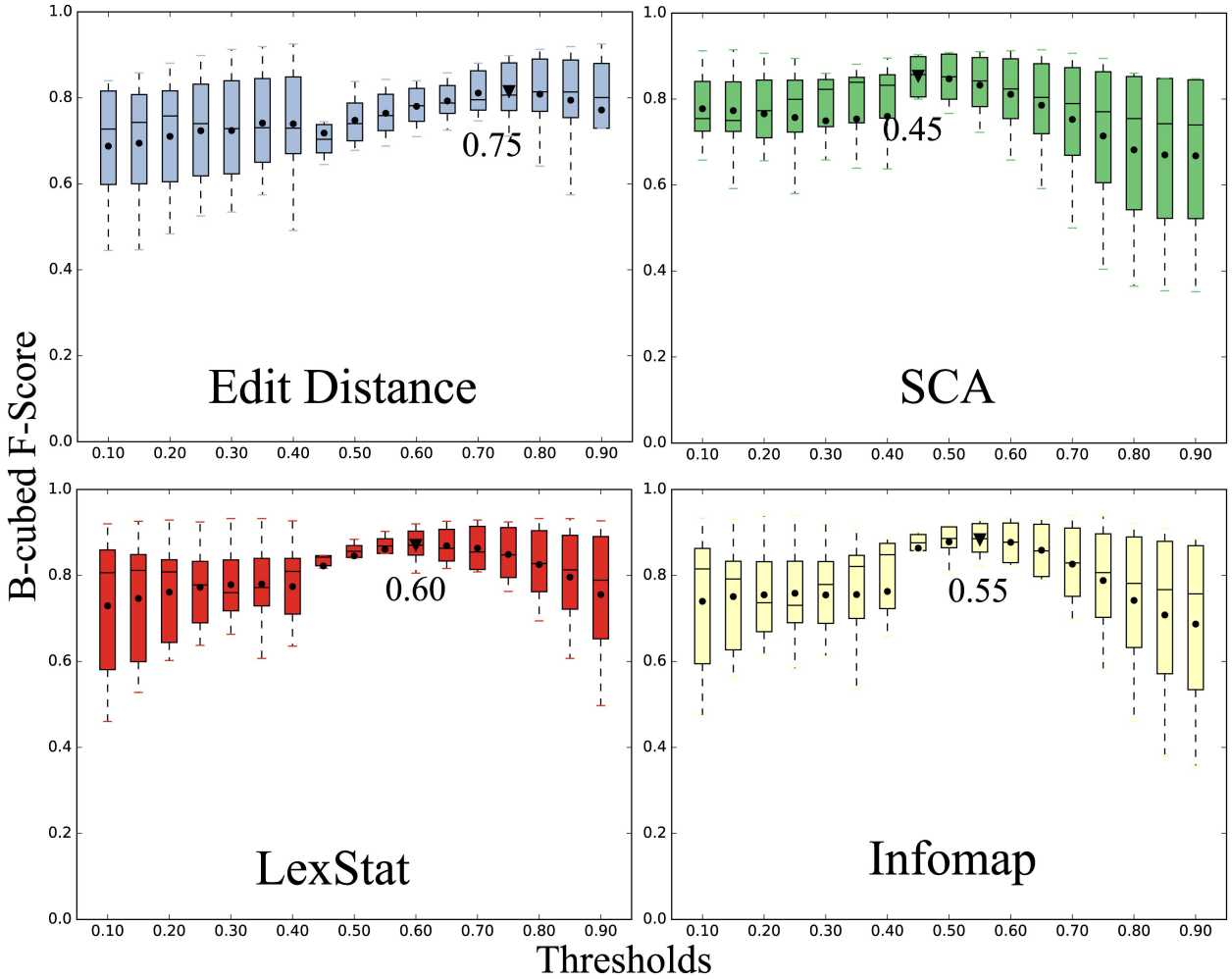


Fig 2. Determining the best thresholds for the methods. The *y*-axis shows the B-Cubed F-scores averaged over all training sets, and the *x*-axis shows the threshold for the 5 methods we tested. Infomap shows the best results on average, Edit Distance performs worst. Dots in the plots indicate the mean for each sample, with triangular symbols indicating the peak.

doi:10.1371/journal.pone.0170046.g002

Table 6. Results of the training analysis to identify the best thresholds. Bold numbers indicate best values.

Method	Thr.	Prec.	Recall	F-Score
Turchin	-	0.8953	0.7276	0.8006
Edit Distance	0.75	0.8341	0.8101	0.8144
SCA	0.45	0.8650	0.8449	0.8529
LexStat	0.60	**0.9204**	0.8287	0.8700
Infomap	0.55	0.9012	**0.8712**	**0.8830**

doi:10.1371/journal.pone.0170046.t006

the threshold. As can be seen from this figure, all methods show a definite peak where they yield the best results for all datasets. In order to select the best threshold for each of the four methods, we selected the threshold which showed the best average B-Cubed F-Score (i.e. the best accuracy at recovering the known cognate sets). For the Edit Distance Method, the threshold was thus set to 0.75, for the SCA Method it was set to 0.45, for the LexStat Method, it was set to 0.60, and for the Infomap method, it was set to 0.55. The B-Cubed scores for these analyses are given in Table 6. These results indicate that the Infomap method performs best, followed by LexStat and SCA. Of the two worst-performing methods, the Turchin method performs worst in terms of F-Scores, but shows a much higher precision than the Edit-Distance method.

Results

We analyzed the datasets with each of the five methods described above, using the individual thresholds for each method, setting the number of permutations to 10,000, and using the default parameters in LingPy. For each analysis, we further calculated the B-Cubed scores to evaluate the performance of each method on each dataset.

Table 7 shows the averaged results of our experiments. While the LexStat method shows the highest precision, the Infomap method shows the highest recall and also the best general performance. The results are generally consistent with those reported by List [19] for the performance of Turchin, Edit Distance, SCA, and LexStat: The Turchin method is very conservative with a low amount of false positives as reflected by the high precision, but a very large amount of undetected cognate relations as reflected by the low recall. The Edit Distance method shows a much higher cognate detection rate, but at the cost of a high rate of false positives. The SCA method outperforms the Edit Distance, thus showing that refined distance scores can make a certain difference in automatic cognate detection.

However, as the performance of LexStat and Infomap shows: Language-specific approaches for cognate detection clearly outperform language-independent approaches. The reason for this can be found in the specific similarity measure that is employed by the methods: the better performing methods are not based on surface similarities, but on similarities derived from previously inferred probability scores for sound correspondences. These methods are therefore

Table 7. General results of the test analysis.

Method	Prec.	Recall	F-Score
Turchin	0.9108	0.7501	0.8175
Edit Distance	0.8397	0.8484	0.8396
SCA	0.8826	0.8492	0.8632
LexStat	**0.9227**	0.8488	0.8831
Infomap	0.9031	**0.8898**	**0.8942**

doi:10.1371/journal.pone.0170046.t007

much closer to the traditional comparative method than methods which employ simple surface similarities between sounds. Our experiment with the Infomap algorithm shows that a shift from simple agglomerative clustering approaches to a network perspective may further strengthen the results. Similarity networks have been successfully employed in evolutionary biology for some time now and should now become a fruitful topic of research in computational historical linguistics as well.

Dataset Specific Results

There are interesting differences between method performance across language datasets, with marked variation in cognate identification accuracy between different languages. Fig 3 shows the performance of the methods on the individual test sets, indicating which method performed best and which method performed worst. These results confirm the high accuracy of the LexStat method and the even better accuracy of the Infomap approach. All methods apart from the Turchin method perform the worst on the Chinese data. Since compounding is very frequent in Chinese, it is difficult to clearly decide which words to assign to the same cognate set. Often, words show some overlap of cognate material without being entirely cognate. This is illustrated in Fig 4, where cognates and partial cognates for Germanic and Sinitic languages are compared. We followed a strict procedure by which only words in which all morphemes are cognate are labelled as cognate [62], rather than loosely placing all words sharing a single

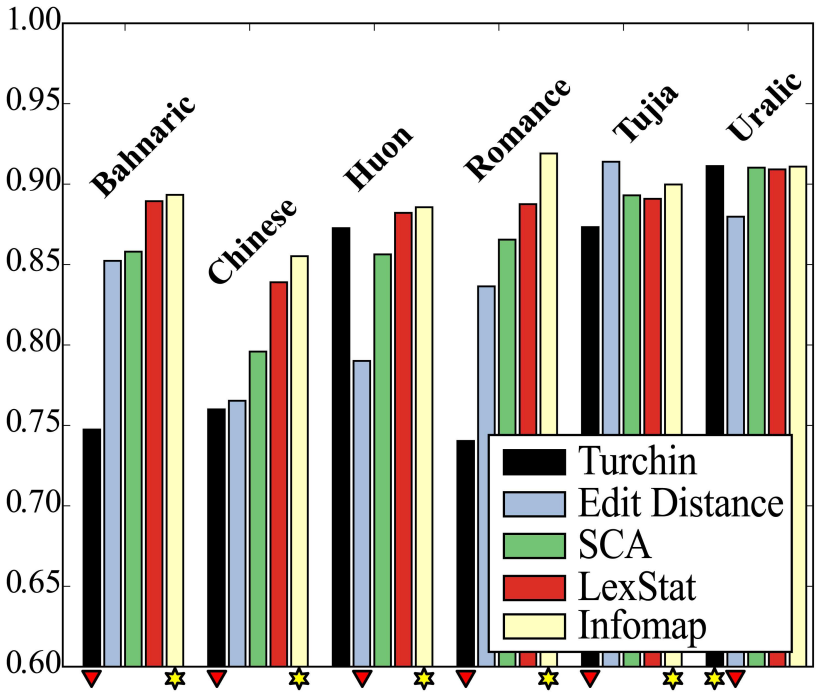


Fig 3. Individual test results (B-Cubed F-Scores). The figure shows the individual results of all algorithms based on B-Cubed F-Scores for each of the datasets. Results marked by a red triangle point to the worst result for a given subset, and results marked by a yellow star point to the best result. Apart from Uralic, our new Infomap approach always performs best, while the Turchin approach performs worst in four out of six tests.

doi:10.1371/journal.pone.0170046.g003

"moon" in Germanic languages		"moon" in Chinese dialects	
English	moon	Fúzhōu	ŋuoʔ^{5}
German	Mond	Měixiàn	ŋiat^{5} kuoŋ^{44}
Dutch	maan	Wēnzhōu	n̩y^{21} kuɔ^{35} vai^{13}
Swedish	måne	Běijīng	yɛ^{51} liɑŋ^{1}

Fig 4. Partial and non-partial cognate relations. The word for "moon" in Germanic and Sinitic languages is mono-morphemic in Germanic languages, while it is usually compounded in Chinese dialects, with the first element in the compound meaning "moon" proper, while the second often originally meant "shine" or "glance". The different cognate relations among the morphemes in the Chinese words make it impossible to give a binary assessment regarding the cognacy of the four words.

doi:10.1371/journal.pone.0170046.g004

cognate morpheme in the same cognate set [63]. Since neither of the algorithms we tested is specifically sensitive for partial cognate relations (for a recent proposal for this task, see [53]), they all show a very low precision, because they tend to classify only partially related words as fully cognate.

The Turchin method has three extreme outliers in which it lags far behind the other methods: Chinese, Bahnaric and Romance. There are two major reasons for this. First, the Turchin method only compares the first two consonants and will be seriously affected by the problem of partial cognates discussed above. These partial cognates are especially prevalent in Chinese and Bahnaric where compounding is an important linguistic process. Second, a specific weakness of the Turchin method is the lack of an alignment and words are not exhaustively compared for structural similarities but simply mapped in their first two initial consonants. When there is substantial sound change, as is evident in both Bahnaric and some branches of Romance, this may lead to an increased amount of false negatives. Since the Turchin method only distinguishes 10 different sound classes and only compares the first two consonant classes in each word in the data, it is very likely to miss obvious cognates. The main problem here is that the method does not allow for any transition probabilities between sound classes, but treats them as discrete units. As a result, it is likely that the Turchin method often misses valid cognate relations which are easily picked up by the other methods. This shortcoming of the Turchin approach is illustrated in Fig 5, where the amount of true positives and negatives is contrasted with the amount of false positives and negatives in each dataset and for each of the five methods. This figure indicates that the Turchin method shows exceptionally high amounts of false negatives in Bahnaric and Romance. The clear advantage of the Turchin method is its speed, as it can be computed in linear time. Its clear disadvantage is its simplicity which may under certain circumstances lead to a high amount of false negatives.

The Edit-Distance method also performs very poorly. While, on average, it performs better than the Turchin approach, it performs considerably worse on the Chinese and Huon test sets. The reason for this poor performance can be found in a high amount of false positives as shown in Fig 5. While the Turchin method suffers from not finding valid cognates, the Edit-Distance method suffers from the opposite problem—identifying high amounts of false cognates. Since false positives are more deleterious for language comparison, as they might lead to false conclusions about genetic relationship [15], the Edit-Distance method should be used with very great care. Given that the SCA method performs better while being similarly fast, there is no particular need to use the Edit-Distance method at all.

In Fig 6, we further illustrate the difference between the worst and the best approaches in our study by comparing false positives and false negatives in Turchin and Infomap across all language pairs in the Chinese data. As can be seen from Fig 5, the Turchin approach has about

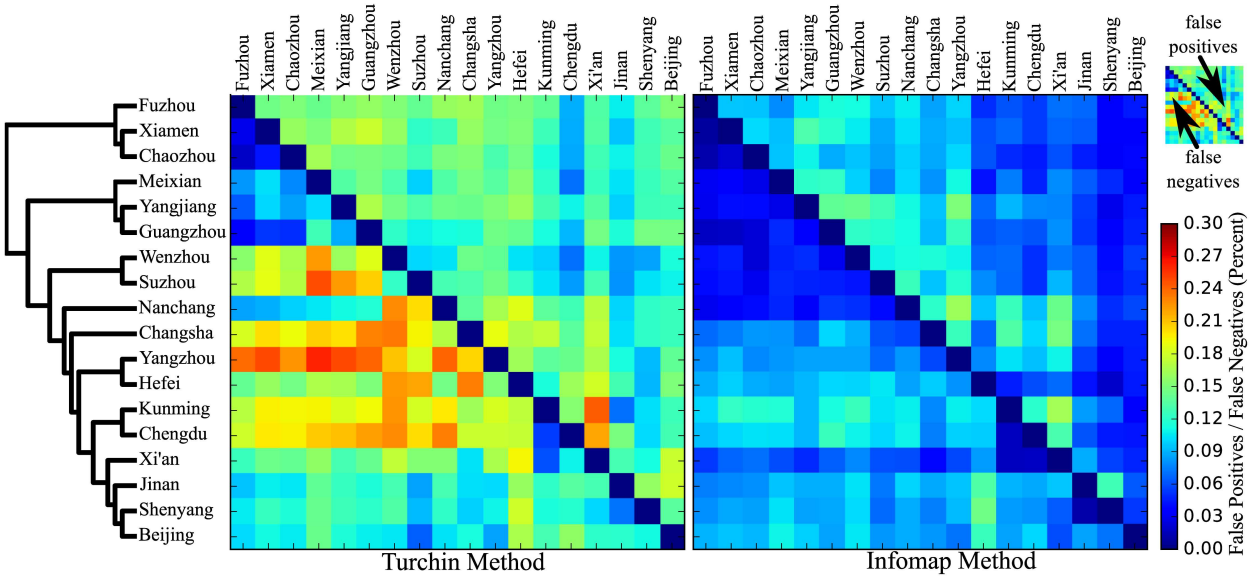


Fig 5. Distribution of true and false positives and true and false negatives.

doi:10.1371/journal.pone.0170046.g005

as many false positives as false negatives. The Infomap approach shows slightly more false positives than false negatives. This general picture, however, changes when looking at the detailed data plotted in Fig 6. Here, we can see that false positives in the Turchin approach occur in almost all dialect pairings, while the major number of cognates is missed in the mainland dialects (bottom of the *y*-axis). Infomap, on the other hand, shows drastically fewer false positives and false negatives, but while false negatives can be mostly observed in the Northern dialects (bottom of *y*-axis), false positives seem to center around the highly diverse Southern dialects (top of the *y*-axis). This reflects the internal diversity in Northern and Southern Chinese dialects, and the challenges resulting from it for automatic cognate detection. While word compounding is very frequent in the North of China, where almost all words are bisyllabic and bimorphemic, the Southern dialects often preserve monosyllabic words. While Northern dialects are rather homogeneous, showing similar sound systems and a rather large consonant inventories, Southern dialects have undergone many consonant mergers in their development, and are highly diverse. The unique threshold for cognate word detection overestimates similarities among the Southern dialects (upper triangle, left quarter), while it underestimates similarities among Northern dialects compared to Southern dialects (lower triangle, left quarter). What further contributes to this problem is also the limited size of the word lists in our sample, which make it difficult for the language-specific algorithms to acquire enough deep signal.

Discussion

In this study we have applied four published methods and one new method for automated cognate detection to a set of six different test sets from five different language families. By training our data on an already published dataset of similar size, we identified the best thresholds to obtain a high accuracy for detecting truly related words for four out of the five methods (EditDistance: 0.75, SCA: 0.45, LexStat: 0.6, Infomap: 0.55). Using these thresholds, we tested the

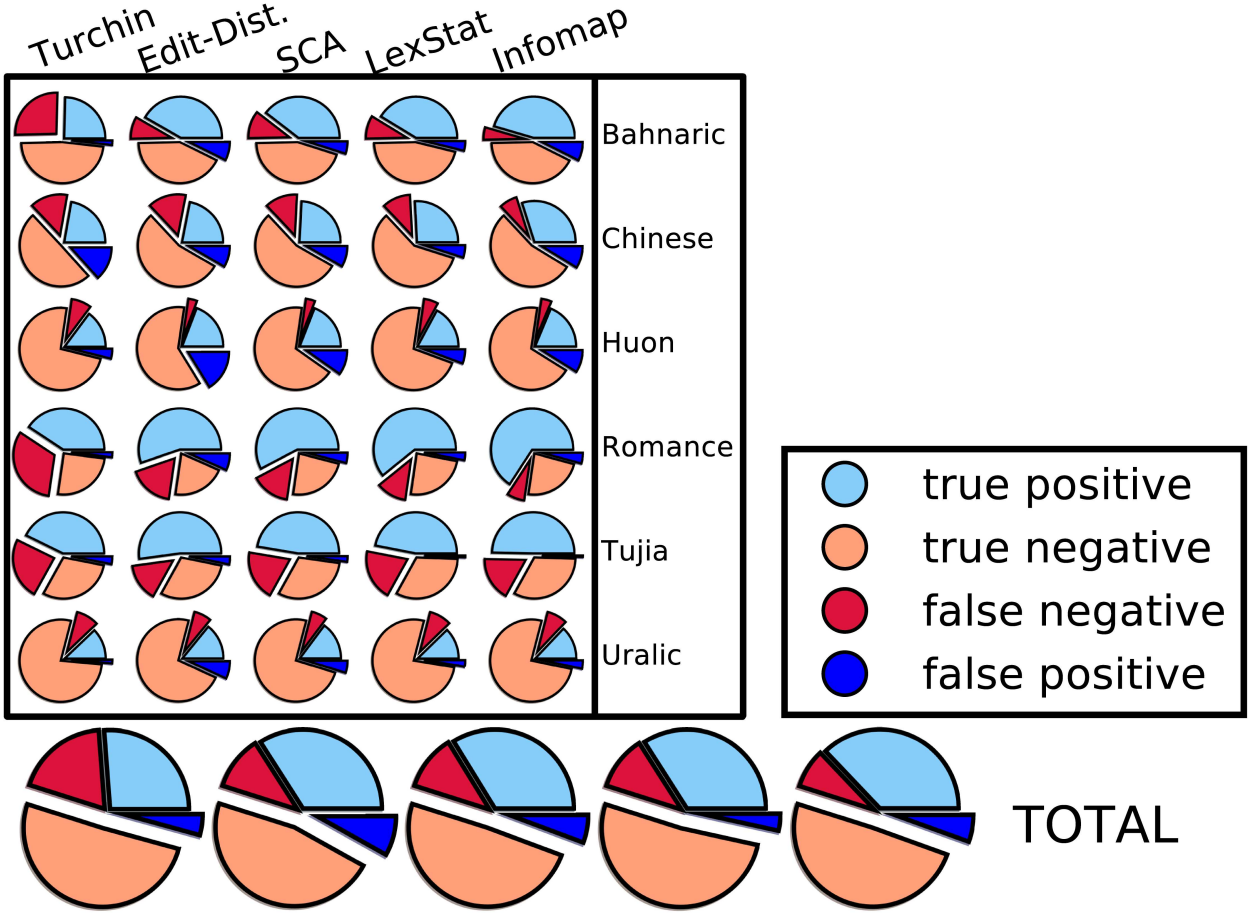


Fig 6. Comparing false positives and false negatives in the Chinese data. The figure compares the amount of false positives and false negatives, as measured in pairwise scores for the Turchin method and our Infomap approach for all pairs of language varieties in the Chinese test set. The upper triangle of the heatmaps shows the amount of false positives, while the lower triangle shows the amount of false negatives.

doi:10.1371/journal.pone.0170046.g006

methods on our new gold standard, and found that most methods identified cognates with a considerable amount of accuracy ranging from 0.82 (Tuchin) to 0.89 (Infomap). Our new method, which builds on the LexStat method but employs the Infomap algorithm for community detection to partition words into cognate sets, outperforms all other methods in almost all regards, slightly followed by the LexStat approach. Given that the LexStat method and our Infomap approach are based on *language-specific* language comparison, searching for similar patterns in individual language pairs, our results confirm the superiority of cognate detection approaches which are closer to the theoretical foundation of the classical comparative method in historical linguistics. The Consonant Class Matching method by Turchin et al. confirmed worst in our experiment, followed by the Edit-Distance approach, which was criticized in earlier work [15]. While the major drawback of the Turchin approach is a rather large amount of false negatives, the Edit-Distance approach shows the highest amount of false positives in our test.

The method of choice may well depend on the task to which cognate detection is to be applied. If the task is to simply identify some potential cognates for future inspection and annotation, then a fast algorithm like the one by Turchin et al. should provide enough help to get started. This practice, which is already applied by some scholars [64], is further justified by the rather small amount of false positives. While the use of the Turchin method may be justified in computer-assisted workflows, the use of the Edit-Distance approach should be discouraged, since it lacks the speed advantages and is very prone to false positives.

When searching for deeper signals in larger datasets, however, we recommend using the more advanced methods, like SCA, LexStat or our new Infomap approach. LexStat and Infomap have the great advantage of taking regular sound correspondences into account. As a result, these methods tend to refuse chance resemblances and borrowings. Their drawback is the number of words needed to carry out the analysis. As we know from earlier tests [65], language-specific methods require at least 200 words for moderately closely related languages. When applied to datasets with higher diversity among the languages, the number of words should be even higher. Thus, when searching for cognates in very short word lists, we recommend using the SCA method to achieve the greatest accuracy. However, as demonstrated by the poorer performance of all methods on the Chinese language data where compounding has played a major role in word formation, language family specific considerations about the methods and processes need to be taken into consideration.

Our results show that the performance of computer-assisted automatic cognate detection methods has advanced substantially, both with respect to the applicability of the methods and the accuracy of the results. Moreover, given that the simple change we made from agglomerative to network-based clustering could further increase the accuracy of the results, shows that we have still not exhausted the full potential of cognate detection methods. Future algorithms may bring us even closer to expert's judgments, and it seems worthwhile to invest time to increase the performance of our algorithms. Essential tasks for the future include (a) the work on parameter-free methods which do not require user-defined thresholds and state the results as probabilities rather as binary decisions, (b) the further development of methods for *partial cognate detection* [53], (c) approaches that search for cognates not only in the same meaning slot but across different meanings [66], and (d) approaches that integrate expert annotations to allow for a true iterative workflow for computer-assisted language comparison. A key problem to solve is the performance of these methods on larger datasets that trace language relationships to a greater depth. Most of our test cases in this paper are shallow families or subgroups of larger families. Deeper relationships between languages spoken in more complicated language situations are where the real challenge lies.

Currently automatic cognate detection algorithms are highly accurate at detecting a substantial proportion of the cognates in a lexical dataset. Tools like LingPy are already at a stage where they can act as a computer-assisted framework for language comparison. These tools therefore provide a powerful way of supplementing the historical linguistics toolkit by enabling linguists to rapidly identify the cognate sets which can then be checked, corrected, and augmented as necessary by experts. In regions where there has been an absence of detailed historical comparative work, these automated cognate assignments can provide a way to pre-process linguistic data from less well studied languages and speed up the process by which experts apply the comparative method. Additionally, these tools can be employed for exploratory data analysis of larger datasets, or to arrive at preliminary classifications for language families which have not yet been studied with help of the classical methods.

Acknowledgments

We thank the anonymous reviewers for helpful advice. We thank Outi Vesakoski and Jury Lehtinen (Uralic data from the Uralex project), Paul Sidwell (Bahnaric), George Starostin (Tujia), and M. Saenko (Romance) for sharing their data with us by either exchanging them directly or making them accessible online.

Author Contributions

Conceptualization: JML SJG RDG.

Data curation: JML.

Formal analysis: JML SJG.

Funding acquisition: SJG RDG.

Investigation: JML SJG.

Methodology: JML SJG.

Project administration: RDG.

Software: JML.

Validation: JML SJG RDG.

Visualization: JML SJG.

Writing – original draft: JML.

Writing – review & editing: JML SJG RDG.

References

1. Greenhill SJ, Blust R, Gray RD. The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics. Evolutionary Bioinformatics. 2008; 4:271–283.
2. Dunn M. Indo-European lexical cognacy database (IELex). Nijmegen: Max Planck Institute for Psycholinguistics; 2012. URL: http://ielex.mpi.nl/.
3. Greenhill SJ. TransNewGuinea.org: An online database of New Guinea languages. PLoS ONE. 2015; 10(10):e0141563. doi: 10.1371/journal.pone.0141563 PMID: 26506615
4. Kitchen A, Ehret C, Assefa S, Mulligan CJ. Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. Proc Biol Sci. 2009 Aug; 276(1668):2703–2710. doi: 10.1098/rspb.2009.0408 PMID: 19403539
5. Bowern C. Chirila: Contemporary and historical resources for the indigenous languages of Australia. Language Documentation and Conservation. 2016; 10:1–44. Available from: http://nflrc.hawaii.edu/ldc/?p=1002.
6. Fox A. Linguistic reconstruction. Oxford: Oxford University Press; 1995.
7. Hammarström H, Forkel R, Haspelmath M, Bank S. Glottolog. Leipzig: Max Planck Institute for Evolutionary Anthropology; 2015. URL: http://glottolog.org.
8. McMahon A, McMahon R. Language classification by numbers. Oxford: Oxford University Press; 2005.
9. Embleton S. Lexicostatistics/glottochronology. From Swadesh to Sankoff to Starostin to future horizons. In: Renfrew C, McMahon A, Trask L, editors. Time depth in historical linguistics. Cambridge: The McDonald Institute for Archaeological Research; 2000. p. 143–165.
10. Holm HJ. The new arboretum of Indo-European "trees". Journal of Quantitative Linguistics. 2007; 14(2–3):167–214. doi: 10.1080/09296170701378916
11. Holman EW, Wichmann S, Brown CH, Velupillai V, Müller A, Bakker D. Explorations in automated lexicostatistics. Folia Linguistica. 2008; 20(3):116–121.

12. Wheeler WC, Whiteley PM. Historical linguistics as a sequence optimization problem: the evolution and biogeography of Uto-Aztecan languages. Cladistics. 2015; 31:113–125. doi: 10.1111/cla.12078
13. Jäger G. Support for linguistic macrofamilies from weighted alignment. PNAS. 2015; 112(41):12752–12757. doi: 10.1073/pnas.1500331112 PMID: 26403857
14. Campbell L. Comment on: Automated dating of the world's language families based on lexical similarity. Current Anthropology. 2011; 52:866–867.
15. Greenhill SJ. Levenshtein distances fail to identify language relationships accurately. Computational Linguistics. 2011; 37(4):689–698. doi: 10.1162/COLI_a_00073
16. Sidwell P. Comment on: Automated Dating of the World's Language Families Based on Lexical Similarity. Current Anthropology. 2011; 52:869–870.
17. Trask RL. The dictionary of historical and comparative linguistics. Edinburgh: Edinburgh University Press; 2000.
18. Ross MD, Durie M. Introduction. In: Durie M, editor. The comparative method reviewed. New York: Oxford University Press; 1996. p. 3–38.
19. List JM. Sequence comparison in historical linguistics. Düsseldorf: Düsseldorf University Press; 2014.
20. Sidwell P. Austroasiatic dataset for phylogenetic analysis: 2015 version. Mon-Khmer Studies (Notes, Reviews, Data-Papers). 2015; 44:lxviii–ccclvii.
21. Saenko M. Annotated Swadesh wordlists for the Romance group (Indo-European family). In: Starostin GS, editor. The Global Lexicostatistical Database. RGU; 2015. Http://starling.rinet.ru/new100/tuj.xls.
22. McElhanon KA. Preliminary Observations on Huon Peninsula Languages. Oceanic Linguistics. 1967; 6 (1):1–45. doi: 10.2307/3622923
23. Starostin GS. Annotated Swadesh wordlists for the Tujia group. In: Starostin GS, editor. The Global Lexicostatistical Database. Moscow: RGU; 2013. URL: http://starling.rinet.ru/new100/tuj.xls.
24. Běijīng Dàxué. Hányǔ fāngyán cíhuì 漢語方言詞匯 [Chinese dialect vocabularies]. Beijing: Wénzì Gǎigé; 1964.
25. Syrjänen K, Honkola T, Korhonen K, Lehtinen J, Vesakoski O, Wahlber N. Shedding more light on language classification using basic vocabularies and phylogenetic methods. Diachronica. 2013; 30 (3):323–352. doi: 10.1075/dia.30.3.02syr
26. List JM, Cysouw M, Forkel R. Concepticon: A resource for the linking of concept lists. Leipzig: Max Planck Institute for Evolutionary Anthropology; 2016. URL: http://concepticon.clld.org.
27. Wang F. Comparison of languages in contact. The distillation method and the case of Bai. Taipei: Institute of Linguistics Academia Sinica; 2006.
28. Hóu J. Xiàndài Hànyǔ fāngyán yīnkù 現代漢語方言音庫 [Phonological database of Chinese dialects]. Shànghǎi: Shànghǎi Jiàoyù; 2004.
29. Hattori S. Japanese dialects. In: Hoenigswald HM, Langacre RH, editors. Diachronic, areal and typological linguistics. The Hague and Paris: Mouton; 1973. p. 368–400.
30. Zhivlov M. Annotated Swadesh wordlists for the Ob-Ugrian group (Uralic family). In: Starostin GS, editor. The Global Lexicostatistical Database. RGU; 2011. URL: http://starling.rinet.ru/new100/oug.xls.
31. Beinborn L, Zesch T, Gurevych I. Cognate production using Character-based Machine Translation. In: Mitkov R, Park JC, editors. Proceedings of the Sixth International NLP Conference; 2013. p. 883–891.
32. Bouchard-Côté A, Hall D, Griffiths TL, Klein D. Automated reconstruction of ancient languages using probabilistic models of sound change. PNAS. 2013; 110(11):4224–4229. doi: 10.1073/pnas.1204678110 PMID: 23401532
33. Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal community structure. PNAS. 2008; 105(4):1118–1123. doi: 10.1073/pnas.0706851105 PMID: 18216267
34. Mackay W, Kondrak G. Computing word similarity and identifying cognates with pair hidden markov models. In: Proceedings of the Ninth Conference on Computational Natural Language Learning; 2005. p. 40–47.
35. Bergsma S, Kondrak G. Multilingual cognate identification using integer linear programming. In: Proceedings of the RANLP Workshop; 2007. p. 656–663.
36. Berg-Kirkpatrick T, Klein D. Simple effective decipherment via combinatorial optimization. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing; 2011. p. 313–321.
37. Hauer B, Kondrak G. Clustering semantically equivalent words into cognate sets in multilingual lists. In: Proceedings of the 5th International Joint NLP conference; 2011. p. 865–873.
38. Steiner L, Stadler PF, Cysouw M. A pipeline for computational historical linguistics. Language Dynamics and Change. 2011; 1(1):89–127. doi: 10.1163/221058211X570358

39. Rama T, Kolachina P, Kolachina S. Two methods for automatic identification of cognates. In: Wielfaert T, Heylen K, Speelman D, editors. Proceedings of the 5th QITL Conference; 2013. p. 76–80.
40. Ciobanu AM, Dinu LP. Automatic detection of cognates using orthographic alignment. In: Proceedings of the 52nd Annual Meeting of the ACL (Short Papers); 2013. p. 99–105.
41. Jäger G, Sofroniev P. Automatic cognate classification with a Support Vector Machine. In: Proceedings of the 13th Conference on Natural Language Processing; 2016. p. 128–133.
42. List JM, Moran S. An open source toolkit for quantitative historical linguistics. In: Proceedings of the ACL 2013 System Demonstrations. Stroudsburg: Association for Computational Linguistics; 2013. p. 13–18.
43. Csárdi G, Nepusz T. The igraph software package for complex network research. InterJournal Complex Systems. 2006;p. 1695.
44. Turchin P, Peiros I, Gell-Mann M. Analyzing genetic connections between languages by matching consonant classes. Journal of Language Relationship. 2010; 3:117–126.
45. Dolgopolsky AB. Gipoteza drevnejšego rodstva jazykovych semej Severnoj Evrazii s verojatnostej točky zrenija [A probabilistic hypothesis concerning the oldest relationships among the language families of Northern Eurasia]. Voprosy Jazykoznanija [Linguistic Inquiries]. 1964; 2:53–63.
46. Levenshtein VI. Dvoičnye kody s ispravleniem vypadenij, vstavok i zameščenij simvolov [Binary codes with correction of deletions, insertions and replacements]. Doklady Akademij Nauk SSSR. 1965; 163 (4):845–848.
47. Sokal RR, Michener CD. A statistical method for evaluating systematic relationships. University of Kansas Scientific Bulletin. 1958; 28:1409–1438.
48. Kondrak G. A new algorithm for the alignment of phonetic sequences. In: Proceedings of the 1st North American chapter of the ACL conference; 2000. p. 288–295.
49. Kessler B. The significance of word lists. Stanford: CSLI Publications; 2001.
50. Méheust R, Zelzion E, Bhattacharya D, Lopez P, Bapteste E. Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis. PNAS. 2016;In press.
51. Corel E, Lopez P, Méheust R, Bapteste E. Network-thinking: Graphs to analyze microbial complexity and evolution. Trends Microbiol. 2016; 24(3):224–237. doi: 10.1016/j.tim.2015.12.003 PMID: 26774999
52. Lopez P, List JM, Bapteste E. A preliminary case for exploratory networks in biology and linguistics. In: Fangerau H, Geisler H, Halling T, Martin W, editors. Classification and evolution in biology, linguistics and the history of science. Stuttgart: Franz Steiner Verlag; 2013. p. 181–196.
53. List JM, Lopez P, Bapteste E. Using sequence similarity networks to identify partial cognates in multilingual wordlists. In: Proceedings of the Association of Computational Linguistics 2016 (Volume 2: Short Papers). Berlin: Association of Computational Linguistics; 2016. p. 599–605.
54. List JM, Pathmanathan JS, Lopez P, Bapteste E. Unity and disunity in evolutionary sciences: process-based analogies open common research avenues for biology and linguistics. Biology Direct. 2016; 11 (39):1–17.
55. van Dongen SM. Graph clustering by flow simulation [PhD Thesis]. University of Utrecht; 2000.
56. MacQueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. vol. 1. Berkeley: University of California Press; 1967. p. 281–297.
57. Frey BJ, Dueck D. Clustering by passing messages between data points. Science. 2007; 315:972–976. Available from: www.psi.toronto.edu/affinitypropagation. doi: 10.1126/science.1136800 PMID: 17218491
58. Vlasblom J, Wodak SJ. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. BMC Bioinformatics. 2009; 10:99. doi: 10.1186/1471-2105-10-99 PMID: 19331680
59. Girvan M, Newman ME. Community structure in social and biological networks. PNAS. 2002; 99 (12):7821–7826. doi: 10.1073/pnas.122653799 PMID: 12060727
60. Bagga A, Baldwin B. Entity-based cross-document coreferencing using the vector space model. In: Proceedings of the 36th Annual Meeting of the ACL; 1998. p. 79–85.
61. Amigó E, Gonzalo J, Artiles J, Verdejo F. A comparison of extrinsic clustering evaluation metrics based on formal constraints. Information Retrieval. 2009; 12(4):461–486. doi: 10.1007/s10791-008-9066-8
62. Ben Hamed M, Wang F. Stuck in the forest: Trees, networks and Chinese dialects. Diachronica. 2006; 23:29–60. doi: 10.1075/dia.23.1.04ham
63. Satterthwaite-Phillips D. Phylogenetic inference of the Tibeto-Burman languages [PhD Thesis]. Stanford University. Stanford; 2011.

64. Starostin G, Krylov P. The Global Lexicostatistical Database. Compiling, clarifying, connecting basic vocabulary around the world: From free-form to tree-form. Moscow: RGGU; 2011. URL: http://starling.rinet.ru/new100/main.htm.

65. List JM. Investigating the impact of sample size on cognate detection. Journal of Language Relationship. 2014; 11:91–101.

66. Wahle J. An approach to cross-concept cognacy identification. In: Bentz C, Jäger G, Yanovich I, editors. Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics. Tübingen; 2016. Available from: http://dx.doi.org/10.15496/publikation-10060.

4.2 Phonetic Alignments and Sound Correspondences

Although the past research has illustrated that methods for automated sequence comparison are already in a state where they could provide real help to historical linguists, be it by helping them to speed up the process of data annotation for the purpose of studies on phylogenetic reconstruction, or in assembling data for etymological dictionaries, the methods are still only rarely applied. There are different reasons for this situation. First, not all linguists are experienced in the application of software. They do not know how to follow code examples and how to apply them on their own data. Second, scholars are often also quite skeptical regarding the new methods and see them as potential enemies rather than useful helpers.

In order to address the first problem, I decided to write a detailed tutorial, together with colleagues with a historical and a computational background. The major idea was to present the *state-of-the-art* of computational sequence comparison approaches in historical linguistics and to provide colleagues with a detailed tutorial presenting major caveats (data handling, parameter settings, evaluation) when applying the new techniques. The tutorial was presented in form of an article describing the background and providing literature and key example and in form of an accompanying online tutorial which would lead interested users step by step through the analysis of a newly created dataset of Polynesian languages (List 2018).

In order to address the second problem, I realized that it would be important to show that automated sequence comparison methods could go beyond the rather straightforward identification of cognate words. Since practitioners of historical language comparison often assume that they do not need much help in identifying cognate sets in their data, it was important to develop an approach that could immediately show a use-case for which there was no traditional counterpart. It turned out that this use-case was the identification of *sound correspondence patterns*. Although linguists often list regular sound correspondences for two and sometimes also for more languages, they barely do so in a systematic way that would trace all positions in known cognate sets to a given pattern. The reason why this had never been done in historical linguistics so far, seems to lie in the complexity of the problem. Scholars have been assembling cognate sets in etymological dictionaries for a long time. This resulted in specific styles and unspoken rules in which data are usually assembled. A detailed listing of sound correspondence patterns, however, drastically exceeds the complexity of etymological data, requiring the use of alignments and cross-references across multiple tables. There is no straightforward way to represent these inferences in a book-style resource, but books have been for a long time the typical resource in which data were assembled in historical linguistics.

When working on automated methods to infer sound correspondences from aligned cognate sets, I realized at some point that the problem of correspondence pattern inference could be treated as a very specific *network partitioning problem*. The goal was to assemble correspondence patterns from the individual columns of aligned cognate sets and to assign those individual columns to the same pattern which are identical or *compatible* with each other. As discussed in detail in (List 2019b), it turned out that the partitioning problem could be modeled as the *minimum clique cover problem* (Bhasker and Samad 1991: 2), which is a well-known problem in graph theory and computer science, for which approximate solutions exist. Having identified the problem, it was straightforward to show how it could be applied to concrete linguistic datasets. However, since linguists never tried to solve the problem themselves exhaustively on a given dataset, it was only possible to test the inferences indirectly. Here, it turned out that the correspondence patterns, once inferred, can be easily use to *predict* possible cognate words, even if they are not present in the data. By deleting words artificially from test datasets, it was possible to show that the predictive capacity of the automatically inferred correspondence patterns is rather high, ranking from 50% to more than 90%, depending on the age and diversity of the language family.

Sequence comparison in computational historical linguistics

Johann-Mattis List,[1,*] Mary Walworth,[1] Simon J. Greenhill,[1,2] Tiago Tresoldi,[1] and Robert Forkel[1]

[1]Department of Linguistic and Cultural Evolution, MPI-SHH, Jena and and [2]ARC Centre of Excellence for the Dynamics of Language, Australian National University, Canberra

*Corresponding author: mattis.list@shh.mpg.de

Abstract

With increasing amounts of digitally available data from all over the world, manual annotation of cognates in multi-lingual word lists becomes more and more time-consuming in historical linguistics. Using available software packages to pre-process the data prior to manual analysis can drastically speed-up the process of cognate detection. Furthermore, it allows us to get a quick overview on data which have not yet been intensively studied by experts. LingPy is a Python library which provides a large arsenal of routines for sequence comparison in historical linguistics. With LingPy, linguists can not only automatically search for cognates in lexical data, but they can also align the automatically identified words, and output them in various forms, which aim at facilitating manual inspection. In this tutorial, we will briefly introduce the basic concepts behind the algorithms employed by LingPy and then illustrate in concrete workflows how automatic sequence comparison can be applied to multi-lingual word lists. The goal is to provide the readers with all information they need to (1) carry out cognate detection and alignment analyses in LingPy, (2) select the appropriate algorithms for the appropriate task, (3) evaluate how well automatic cognate detection algorithms perform compared to experts, and (4) export their data into various formats useful for additional analyses or data sharing. While basic knowledge of the Python language is useful for all analyses, our tutorial is structured in such a way that scholars with basic knowledge of computing can follow through all steps as well.

Key words: historical linguistics; computer-assisted language comparison; Polynesian languages; cognate detection; phonetic alignment

1. Introduction

Sequence comparison is one of the key tasks in historical linguistics. By comparing words or morphemes across languages, linguists can identify which words have sprung from a common source in genetically related languages, or which words have been borrowed from one language to another. By comparing words within a language, linguists can identify grammatical and lexical morphemes, cluster words into families, and shed light on the internal history of languages. So far the majority of this work has been carried out manually. Linguists sift through dictionaries and fieldwork notes, trying to identify those words which reflect a shared history across languages. All etymological dictionaries available today have been based on manual word comparison and their results fill thousands of pages. Even the largest databases which offer cognate judgments, such as the *Austronesian Basic Vocabulary Database* (ABVD,

Greenhill et al., 2008) or the *Indo-European Lexical Cognacy Database* (Dunn, 2012) are based on manual assessments of cognacy.

With the increasing amounts of digitally available data it becomes harder for linguists to keep up. For example, the *Sino-Tibetan Etymological and Thesaurus database* (Matisoff, 2015), contains more than 500,000 words, but only a small amount of words have been compared etymologically (see Hill and List, 2017: 64f). We need to take advantage of increasing amounts of data, refining work on well-established languages, and fostering work on the world's understudied languages. To do this, however, we will have to rethink the way we compare languages.

Historical linguists are skeptical about automating the methods for cognate identification (see Holman et al. (2011) and commentaries, as well as List et al. (2017b)). First, the *accuracy* of automated methods is often low, failing to reproduce the analyses of linguistic experts. Especially, the use of the *edit distance* (Levenshtein, 1965) has been criticized for being linguistically too nave, conflating sound correspondences and lexical replacement, to be useful for subgrouping or cognate detection (Campbell, 2011; Greenhill, 2011). Second, it is hard to *verify* many algorithms as they are seen as black-boxes which hide the crucial decisions leading to cognate judgments and subgroupings, making it difficult for scholars to determine whether similarities are due to inheritance or contact (Jäger, 2015; List et al., 2017b). The nontransparency of automatic methods is highly problematic for computational historical linguistics: if we do not know what evidence decisions are based on, we cannot criticize and improve them.

However, methods for automatic sequence comparison in historical linguistics have dramatically improved during the last two decades. Starting with the pioneering work on pairwise and multiple phonetic alignment (Kondrak, 2000; Prokić et al., 2009), new methods for phonetic alignment and automatic cognate detection solve both the problems of verification and accuracy (List et al., 2017b; Jäger et al., 2017). First, these algorithms are based on phonetically informed metrics on sound similarities. Importantly, any algorithmically identified correspondences are logged and can be inspected by researchers. Second, in a wide-ranging test of these methods, they have been found to be highly accurate and able to correctly identify cognates in almost 90% of the cases (List et al., 2017b).

LingPy (List et al., 2017a) provides these algorithms as part of a stable open-source software package that works on all major platforms. Given the complexity human judgments any time soon, but with the recent advancements, the methods are definitely good enough to provide substantial help for classical historical linguists to *pre-analyze* the data to be later corrected by experts, or to check the consistency of human cognate judgments. Over the long run, computational methods can also contribute to the bigger questions of language evolution, be it indirectly, by increasing the amount of digitally available high-quality annotated data, or directly, by providing scholars' access to data too large to be processed by humans alone.

In the following, we will give a concise overview on how automatic sequence comparison can be carried out. After discussing general aspects of sequence comparison (Section 2), we will introduce basic ideas on the data needed (Section 3). We will then turn to the core tasks of automatic sequence comparison, namely automatic phonetic alignment (Section 4) and automatic cognate detection (Section 5). We conclude by showing how automatic approaches for cognate detection can be evaluated (Section 6), and how results can be exported to various formats (Section 7).

This article is supplemented by a detailed interactive tutorial in form of an IPython Notebook (Pérez and Granger, 2007) which illustrates how all methods discussed here can be practically applied (see the Supplementary material for more information). Having installed the necessary software (Tutorial: 1), readers can follow the tutorial step by step and investigate how the algorithms work in practise. Our data is based on a small sample of Polynesian languages taken from the ABVD, which we substantially revised, both with respect to the phonetic transcriptions and the expert cognate judgments. All data needed to replicate the analyses discussed here are supplemented. We give more information in the interactive tutorial (Tutorial: 2.1).

2. Basic aspects of sequence comparison

The words and morphemes which constitute a language are best modeled as *sequences of sounds*. Sequences have information content not only from their elements (*segments*, whether these are phonemes, graphemes, or morphemes) but also via the *order* of the elements, a consistent comparison of sequences should account for both order and content. *Alignments* are a very general way to model differences between sequences. The major idea is to arrange two or more sequences in a matrix in such a way that similar or identical segments which occur in similar positions are placed in the same column

represented by a *gap character*, usually the *dash*-symbol (List, 2014b).

Sequence alignments are crucial in biology, where they are used to compare protein and DNA sequences (Durbin et al., 2002). In historical linguistics, however, they are usually only *implicitly* employed, and initial attempts to arrange cognate words in a matrix go back to the early 20th century, as one can see from an early example based on Dixon and Kroeber (1919: 61) given in Fig. 1. The authors themselves describe this way of representing sequence similarities as a 'columnar form' with the goal to 'bring out parallelisms that otherwise might fail to impress without detailed analysis and discussion' (Dixon and Kroeber, 1919: 55). The figure further shows how the data would look if they were rendered in contemporary alignment editors for historical linguistics (List, 2017). Dixon and Kroeber's wording nicely expresses one of the major advantages of alignments: the transparency of homology assessments. Scholars often list long lists of cognate sets in the literature, claiming that all words are somehow related to each other, but if they do not list the alignments, it is often impossible, even for experts in the same language family, to understand *where exactly* the authors think that certain segments are similar.

Given that the inference of historically related words is *not* based on superficial word similarities but on recurrent systematic similarities, known as *regular sound correspondences* (Lass, 1997: 130), all judgments regarding the relatedness of words across languages directly rely on previously established sequence alignments (Fox, 1995: 67f). Alignment analyses not only increase the transparency of cognate judgments, but they also play a crucial role in substantiating these judgments in a first place. As can be seen from Table 1, similarities in cognate words in Sikaiana and Tahitian (data taken from Greenhill et al., 2008) are *not* based on the identity of sounds, but rather in the regularity of occurrence: whenever Sikaiana has a [k] and a [l], Tahitian has a [ʔ] and a [r], respectively. Without alignments, we could not identify this similarity. Alignments are also at the core of all automatic sequence comparison approaches in historical linguistics, as we will see throughout this tutorial.

3. Data preparation

When searching for cognates across languages, we usually assume that our data are given in some kind of *wordlist*, a list in which a number of concepts is translated into various languages. How many concepts we select depends on the research question, and various concept lists and questionnaires, ranging from 40 (Brown et al., 2008) up to more than 1,000 concepts (Haspelmath and Tadmor, 2009) have been proposed so far (see the overview in List et al. (2016a)). Our data example for this tutorial is based on the questionnaire of the ABVD project (Greenhill et al., 2008), consisting of 210 concepts, which were translated into 31 different Polynesian languages. For closely related languages, such as those in the Polynesian family, this gives us enough information to infer regular correspondences automatically, although it is clear that for analyses of

Table 1. Recurring similarities in Sikaiana and Tahitian.

Cognate list		Alignment				Correspondences		
Sikaiana	*louse*	k	u	t	u	Sik.	Tah.	Freq.
Tahitian	*louse*	ʔ	u	t	u	k	ʔ	3 x
Sikaiana	*dog*	k	u	1	iː	u	u	3 x
Tahitian	*dog*	ʔ	u	r	iː	t	t	1 x
Sikaiana	*skin*	k	i	1	i	r	1	2 x
Tahitian	*skin*	ʔ	i	r	i	i(ː)	i(ː)	3 x

Group	Variety	Pref.		Suff.	Postf.	Comment
W	N		kaha	i		
	C		kʼa	i		
	SE		tcʼa	i		
	SW		tca	i		
Md	NW	tsiʼ	bi			
	NE, S	bi	tsi			(metathesis)
Y	B	go	tco	yi	-c	
Mw	P, L		ti			
	CO	pi	tci			(Borrowed?)
C	J, CR		tu	r		
	CL		tu	r	-em	
	B		tu	r	-is	

ID	DOCULECT	CONCEPT	VALUE	IPA	TOKENS	VARIANTS	COGID
188	Emae	Eight	βaru	βaru	β a r u		750
447	Rennell_Bellona	Eight	baŋgu	baŋgu	b a ⁿg u		750
703	Tuvalu	Eight	valu	valu	v a l u		750
927	Sikaiana	Eight	valu	valu	v a l u		750
1135	Penrhyn	Eight	varu	varu	v a r u		750
6114	Kapingamarangi	Eight	waru,walu	waru	w a r u	walu	750
6115	Kapingamarangi	Eight	waru,walu	walu	w a l u	waru	750

numeric identifier, language name, concept label, value in source, phonetic transcription, space-segmented transc, variants of word, if given, cognate identifier

Figure 2. Input format required by the LingPy package. The last two entries show how synonyms can be handled by placing different variants of one concept in one language variety into different rows with a separate ID each.

more distant language relationship the number of words per language may not be enough.

The basic format used by LingPy is a tab-separated input file in which the first row serves as a header and defines the content of the rest of the rows. The very first column is reserved for numerical identifiers (which all need to be unique), while the order of the other columns is arbitrary, with specific columns being required, and others being optional. Essential columns which always must be provided are the *language name* (DOCULECT), the *comparison concept* (CONCEPT), the *original transcription* (International Phonetic Alphabet (IPA), FORM, or VALUE), and a *space-segmented form* of the transcription (TOKENS). Multiple synonyms for the same comparison concept in the same language should be written in separate rows and given a separate ID each. The data in the TOKENS-column should supply the transcriptions in *space-segmented form*, that is, instead of transcribing the Fila word for 'all' as [eutʃi], the software expects [e u tʃ i], which is internally interpreted as a sequence of five segments, namely [e], [u], [tʃ] and [i], with [tʃ] representing a voiceless post-alveolar affricate. If the TOKENS are not supplied to the algorithm, it will try to segment the data automatically, provided it can find the column IPA, which is otherwise not necessarily required to appear in the data. This however, may lead to various problems and unexpected behavior. We therefore urge all users of LingPy to make sure that they supply segmented data to the algorithm, making furthermore sure that they adhere to the general standards of transcription as they are represented in the IPA (IPA, 1999).[1] The format can be created manually by degree, this input format is compatible with the one advocated by the Cross-Linguistic Data Formats (CLDF) initiative (Forkel et al., 2017), the main difference being that LingPy requires a flat single file with tab-stop as separators, while CLDF supports multiple files. CLDF furthermore encourages the use of reference catalogs, such as Glottolog (Hammarström et al., 2017) or Concepticon (List et al., 2018), in order to increase the comparability of linguistic data across datasets, while LingPy is indifferent regarding the overall comparability as long as the data is internally consistent. As of version 2.6, LingPy offers routines to convert to and from CLDF (see Tutorial: 6.3). Figure 2 provides a basic summary on LingPy's input formats. More information on the format, and how it can be loaded into LingPy can be found in the supplemented interactive tutorial (Tutorial: 2.2-3).

Data quality and consistency plays a crucial role in the outcome of an automatic sequence comparison. As a general rule of thumb, we recommend all linguists who apply LingPy or other software to carry out automatic sequence comparison, to pay careful attention to what we call the SANE rules for data sanity: users should pay close attention to providing a sensible *segmentation* of their data, they should *aim* for high coverage, there should be *no* mixing of data from different sources (as this usually leads to inconsistent transcriptions and may also increase the number of synonyms), and synonyms should be *evaded*.[2] These rules are summarized in Table 2. If the original data does not provide reliable phonetic transcriptions, as it was the case with the Polynesian data we use in this tutorial, *orthography pro-*

Table 2. SANE rules for data sanity.

Segmentation matters		
Consistent phonetic transcription and segmentation are of crucial importance for automatic sequence comparison. Computers cannot guess whether multiple graphemes represent separate or single sound segments.	**NOT:**	Fila [eutʃi] 'all'
	BUT:	Fila [e u tʃ i] 'all'
Aim for high coverage		
Each language should have about the same number of words recorded across the wordlist. A high mutual coverage is important to allow algorithms to find enough information to determine the major signal.	**NOT:**	L_1 150, L_2 50
	BUT:	L_1 200, L_2 200
No mixing of data from different sources		
Mixing data for the same language from various sources can lead to inconsistencies in the phonetic representation of words, even if they are all given in plain phonetic transcriptions. This will weaken the evidence for regular sound correspondences.	**NOT:**	L_1=Source$_1$+Source$_2$
	BUT:	L_1=Source$_1$, L_2=Source$_2$
Evade synonyms		
Languages often have multiple words for a given meaning. However, these can cause problems for sequence comparison and further downstream analyses like phylogenetic reconstruction. Having abundant synonyms in the data (e.g. 40 words for *snow*) will necessarily blur this signal.	**NOT:**	Tahitian [tai] 'sea', [moana] 'ocean'
	BUT:	Tahitian 'sea'

the data, and the EDICTOR tool (List, 2017) offers convenient ways to check phonological inventories of all varieties (Tutorial: 2.4). Various coverage statistics can be computed in LingPy (see Tutorial: 2.5). Synonym statistics can also be easily computed (see Tutorial: 2.6). Users should always keep in mind that the quality of automatic sequence comparison crucially depends on the quality of the data submitted to the algorithms.

4. Automatic phonetic alignment

Alignments are crucial for historical language comparison to search for *regular sound correspondence* patterns, *layers of borrowed words*, or even use them as the starting point for *linguistic reconstruction* (Fox, 1995). A further important advantage is that they can be easily quantified, as we will see in Section 5. Since phonetic alignment is heavily influenced by bioinformatics, linguists using phonetic alignments should have some basic understanding of original algorithms and terminology. In this context, it is not necessarily important to understand *how* the algorithms work in detail. Instead, we think it is more important to learn (also by testing the algorithms with different data and parameters) how the different options from which users can choose influence the results. In the following, we will quickly introduce basic algorithms and concepts involving alignments in historical linguistics, and how they relate to alignments in bioinformatics. We will follow the traditional division and introduce the most important concepts and parameters that users should know when applying the methods.

4.1 Pairwise alignment analyses

Pairwise alignment analyses in biology and computer science date back to the 1970s when scholars like Needleman and Wunsch (1970), and Wagner and Fischer (1974) proposed algorithms based on the *dynamic programming paradigm* (Eddy, 2004b) which drastically reduced the computation time for the task of aligning two sequences with each other. The basic idea of the algorithms by Needleman and Wunsch and Wager and Fischer was to split the problem of finding one optimal alignment between two sequences into subparts and building the general solution from optimal alignments of smaller subsequences (Durbin et al., 2002: 19).[3]

The major parameters of pairwise alignment algorithms are the *scoring function*, the *gap function,* and the *alignment mode*. The scoring function (Fig. 3A, Tutorial: 3.1.1) determines how the matching of segments is penalized (or favored). In biology, it is well known that amino acid mutations follow certain transition preferences. The scoring function defines transition probabilities for each segment pair, and biologists make use of a large number of empirically derived scoring functions (Eddy, 2004a). In linguistics, on the other hand, we know well that certain sounds are more likely to occur in correspondence relations with each other (Dolgopolsky, 1964; Brown et al., 2013), and this

A Scoring Function

Determines how segments are compared with each other. Most generally represented as a symmetric matrix of transition probability scores. Scores between segments are usually given in logarithmic scale, with unexpected matches being smaller 0 and expected ones being greater 0. On the right, a short scoring matrix is shown, in which matches between vowels (a) and consonants (p, f, h) are not allowed and therefore assigned high negative values.

	p	**f**	**h**	**a**
p	10	6	2	-10
f	6	10	6	-10
h	2	6	10	-10
a	-10	-10	-10	10

B Sound Classes

To reduce the huge number of sounds in phonetic alignment analyses, sounds are clustered into classes, assuming that correspondences inside a class are more likely than outside a class. Scoring functions are defined for sound classes, and sound sequences are converted to sound-class sequences before aligning them. On the right, it is shown, how Austral "you" can be rendered in different sound-class alphabets.

IPA	ʔ	oː	ɢ	u	a
Dolgo	H	V	K	V	V
SCA	H	U	K	Y	A
ASJP	7	o	q	u	a

C Gap Function

Gaps are introduced in alignments if a given element cannot be matched with any other element. Gap penalties can be defined globally, by assigning the same penalty to any segment of a given sequence, or individually, based on the context in which the segment occurs. In phonetic alignment analyses, individual gap penalties can be derived from *prosodic profiles* which reflect the relative sonority of each segment in a sequence. On the right, it is shown, how individual gap penalties are derived for the Austral "you".

IPA	ʔ	oː	ɢ	u	a
Sonority	1	7	1	7	7
Prostring	#	V	C	V	>
Weights	2.0	1.5	1.75	1.3	0.8
Gap-Scores	-4	-3	-3.5	-2.6	-1.6

D Alignment Mode

Alignment modes determine how sequences are compared. Global alignment compares the sequences entirely, assuming that they are indeed comparable in all their parts. Local alignment seeks to find the most similar subsequence of an alignment and deliberately strips off prefixes or suffixes. Semi-global alignment can only ignore prefixes or suffixes in one of the sequences. If the other sequence also has a suffix, it needs to strip it off. On the right, it is shown, how the different alignment modes account for the phonetic alignment of "you (dual)" in Ra'ivavae and Mangareva. While the local and the semi-global alignment fail to identify the regular sound correspondence of glottal stop and [k], the global alignment correctly matches words. This does not mean, however, that global alignment always yields the best solutions. Unaligned parts are shaded gray.

Global Alignment

Ra'ivavae	ʔ	oː	ɢ	u	a	
Mangareva	k	oː	r	u	a	

Local Alignment

Ra'ivavae	(ʔoː)	ɢ	-	-	u	a
Mangareva	-	k	oː	r	u	a

Semi-Global Alignment

Ra'ivavae	ʔ	oː	ɢ	-	-	u	a
Mangareva	-	-	k	oː	r	u	a

Figure 3. Basic parameters and concepts in pairwise alignment analyses: (A) Scoring function, (B) Sound classes, (C) Gap function and (D) Alignment mode.

small alphabets, in linguistics, the numbers of possible sounds in the languages of the world amounts to the thousands (Moran et al., 2014). It is not practical to design a matrix containing and confronting all sounds with each other, and most algorithms reduce the size of the alphabet by lumping similar sounds into a set of predefined sound classes (Fig. 3B, Tutorial: 3.1.2), for which transition probabilities can be efficiently defined, and which are then given as input for the alignment algorithm (List, 2012a; Holman et al., 2008).

The introduction of gaps in an alignment (Fig. 3C, Tutorial: 3.1.3) can be seen as a special case of a scoring function. Instead of comparing two segments, the algo-penalty, independent of the element with which they were compared, later studies showed that they could even be individually adjusted for each position in a sequence (Thompson et al., 1994). In linguistics, we know that sounds in certain positions (like initial consonants) are less likely to be lost and that new sounds tend to appear in specific contexts as well. In LingPy, position-specific gap penalties are derived from the *prosodic profiles* of sequences (List, 2012a). Prosodic profiles essentially reflect for each segment of a word whether it occurs in weak or strong prosodic positions, and the user-defined gap penalty is modified accordingly.

The alignment mode (Fig. 3D, Tutorial: 3.1.4) basic-

entirely. Instead, we compare only certain parts of which we know that they are cognate, ignoring parts of which we know they are not. Since the same problem occurs when comparing the genes of diverse species in bioinformatics, biologists have long since been working on solutions, reflected in local alignment analyses (Smith and Waterman, 1981) in which only the most similar parts of sequences are compared (see Fig. 3), while the rest is ignored, or *semi-global alignments* (Durbin et al., 2002: 26f).

What should users keep in mind when carrying out pairwise alignment analyses? As a rule of thumb, we recommend caution with local alignment analyses, since these can show unexpected behavior. We also recommend care with custom changes applied to the scoring or the gap function. Users often naively think by just 'telling' the computers which sound changes, this would automatically lead to excellent alignments and at times complain that LingPy's standard algorithms fail to 'detect certain obvious changes'. However, alignments are no way to determine sound changes, they are at best a first step for linguistic reconstruction, and none of the algorithms which have been proposed so far models any kind of change. What is modeled instead are *correspondences* of sounds. It is difficult, if not impossible, to design an algorithm that aligns sequences of all kinds of diversity without proposing certain analyses which look awkward to a trained linguist. But remember, automatic sequence comparison is not there to replace the experts, but to help them.

4.2 Multiple alignment analyses in linguistics

Pairwise alignments are crucial for most automatic cognate detection methods (List, 2014b; Jäger et al., 2017). In order to visualize cognate judgments, or to reconstruct proto-forms, however, pairwise alignments are not of great help, as most linguistic research applies to at least three if not more language varieties. It may sound counterintuitive for readers not familiar with the major workflows for automatic cognate detection that pairwise alignments are mainly used to detect cognates across multiple languages, while multiple alignments are only later computed from existing cognate sets. Why not compute multiple alignments right from the beginning, as for example, proposed by Wheeler and Whiteley (2015)? The reason for this workflow is that alignments only make sense when representing cognate words—aligning unrelated words just leads to chance similarities.

For reasons of algorithmic complexity, pairwise alignments for an arbitrary number of sequences. In order to address this problem, early approaches used heuristics that approximate optimal multiple alignments (Feng and Doolittle, 1987; Thompson et al., 1994). Most of these algorithms compute pairwise alignments in a first step and then combine the data in a pairwise fashion until all alignments are merged into one multiple alignment. The easiest way to do so is with help of a *guide tree*, a clustering of all sequences, which determines in which order sequences are merged with each other. This procedure is illustrated in Fig. 4 for the alignment of four words for 'dog' in four Polynesian languages (Tutorial: 3.2).

Many extensions of the classical guide-tree heuristics have been proposed in the biological literature (Notredame et al., 2000; Morgenstern et al., 1998) and also adapted in linguistic applications (List, 2012a; Jäger and List, 2015; Hruschka et al., 2015). While the fine-tuning of the algorithms may have a solid impact on multiple alignment analyses involving large sets of language varieties, as we often encounter in dialectology (compare the results of Prokić et al., 2009 with List, 2012a), the problem of erroneous alignments is much less pronounced when using smaller datasets and working in workflows which start from cognate detection and compute multiple alignments in a later stage. For these reasons, we refrain from giving more detailed descriptions of multiple sequence alignment here, but instead refer the readers to the literature that we quoted in this section and the examples in the interactive tutorial (Tutorial: 3.2).

5. Automatic cognate detection

As mentioned in the previous section, we can only meaningfully align words if we know they are historically related. In order to identify which words *are* related, however, we still need to compare them, and most automatic approaches, including the core methods available in LingPy, make use of pairwise sequence comparison techniques in order to find historically related words in linguistic datasets.

The basic workflow of most automatic cognate detection methods can be divided into two major steps. In the first step, pairwise alignment is used to align all words to retrieve distance scores for each pair of words in the data which occur in the same concept slot. If normalized, distance scores typically rank between 0 and 1, with 0 indicating the identity of the objects under comparison, and 1 indicating the maximal difference that can be encountered for the objects. In a second step,

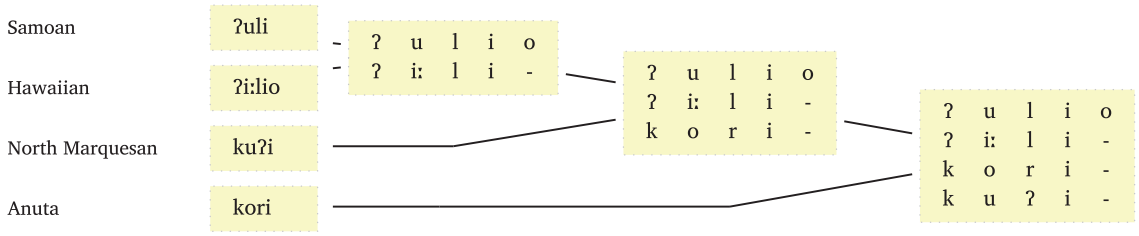


Figure 4. Combining words for 'dog' in Samoan, Hawaiian, North Marquesan, and Anuta into a multiple alignment with help of a guide tree.

presumable cognate sets using tree- or network-based partitioning algorithms. If we take five words for 'neck' from our Polynesian data, Ra'ivavae [ʔaɡapoʔa], Hawaiian [ʔaːʔiː], Mangareva [kaki], Maori [ua], and Rapanui [ŋao], for example, we can use the *normalized edit distance* (NED) to compare all four words with each other and write the results into a matrix, as shown in Table 5A.[4]

In Table 5B, we have carried out the same pairwise comparison, but this time with a different sequence comparison measure, following the sound-class-based alignment method (SCA, List 2012a), in which the idea of sound classes is combined with sequence alignment methods. Table 5C shows the results retrieved from the LexStat method (List, 2012b) which derives distances from a previous search for regular sound correspondences. As can be seen, when comparing only the matrices, the methods generally differ in the way they handle sequence similarities. While NED has rather high scores which do not vary much from each other, SCA has consistently smaller scores with more variation, and LexStat has higher scores but more variation than NED.

In the second step, the matrix of word pair distances is used to partition the words into cognate sets. For this, partitioning algorithms are used which split the words into cognate sets by trying to account as closely as possible for the pairwise distances of all words in a given meaning slot. Early approaches were based on a flat version of the well-known UPGMA algorithm (Sokal and Michener, 1958), which is an agglomerative cluster algorithm that returns the data in the form of a tree. The flat variant of UPGMA stops merging words into bigger subgroups once a user-defined threshold of average pairwise distances among the words in each cluster has been reached (List, 2012b). In order to show how algorithms arrive from pairwise distance scores in a matrix at cognate set partitions, we provide a concrete example in Fig. 5. First, we have marked all cells in which the dis-

added guide trees (reflecting the clustering proposed when applying the UPGMA algorithm without stopping it earlier) below each matrix, which show how the flat clustering algorithm proceeds. If the algorithm stops grouping words into a given cluster, because the average threshold has been reached, this is indicated by a dashed line, which indicates how the clustering would have proceeded if the algorithm had not stopped. Given that we know that of these five words in the figure, only Hawaiian [ʔaːʔiː] and Mangareva [kaki] *are* cognate, we can immediately see that the LexStat algorithm is proposing the correct cognates in this example.

The performance of LexStat is not surprising, if we take its more sophisticated working procedure into account. LexStat uses global and local pairwise alignments to pre-analyze the data, computing *language-specific* scoring functions (List, 2012b), in which the similarity of the segments in a given language pair depends on the overall number of matches that could be found in the preprocessing stage.[6] In these scoring functions, sound segments for all languages in the data are represented as sound-class strings in a certain prosodic environment. This representation is useful to handle sound correspondences in different contexts (word-initial, word-final, etc.). For each language pair in the data, *LexStat* creates an *attested* and an *expected* distribution of sound correspondences. The attested distribution is computed for words with the same meaning and whose SCA score is beyond a user-defined threshold. The expected distribution is computed by shuffling the word lists in such a way that words with different meanings are aligned and compared, with the users defining how often word lists should be shuffled. This *permutation test* following suggestions by Kessler (2001) makes sure that the sound correspondences identified are unlikely to have arisen by chance. The distributions resulting from this permutation test are then combined in *log-odds* scores (see Fig. 3 above) which can then in turn be used to realign all

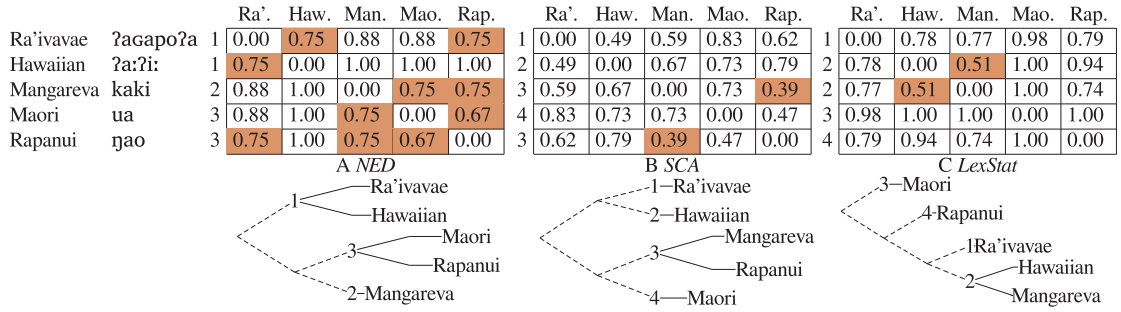


Figure 5. Contrasting distances retrieved from three different alignment approaches for Polynesian words for 'neck'. Cells highlighted indicate that distances are smaller than the default threshold for the algorithms. The first column of each table indicates the cognate decisions resulting from the matrix and the threshold. How these cognate decisions are determined is further illustrated in the trees below each matrix. They show how a flat cluster algorithm which stops once a certain threshold is reached can be used to partition the words into cognate sets.

distances as shown in Fig. 5. Our interactive tutorial shows how input data can be quickly checked before carrying out the (at times time-consuming) computation (Tutorial: 4.1) and provides additional information regarding the differences between the cognate detection methods available in LingPy (Tutorial: 4.2) and illustrates in detail how each of them can be applied (Tutorial: 4.3).

More recent approaches for cognate set partitioning use Infomap (Rosvall and Bergstrom, 2008), a *community detection* algorithm which uses random walks in a graph representation of the data to identify those clusters in which significantly more edges can be found inside a group than outside (Newman, 2006). In order to model the data as a graph, words are represented as nodes and distances between words are represented as edges which are drawn between all nodes whose pairwise distance is beyond a user-defined threshold (List et al., 2017b). Recent studies have shown that the graph-based partitioning approaches slightly outperform the flat agglomerative clustering procedures (List et al., 2016b, 2017b; Jäger et al., 2017).

The advantage of *LexStat* and similar algorithms is that the algorithm infers a lot of information from the data itself. Instead of assuming language-independent distance scores which would be the same for all languages in the world, it essentially infers potential sound correspondences for each language pair in separation and uses this information to determine language-specific distance scores. The disadvantages of *LexStat* are the computation time and the dependency of data with high mutual coverage. It was designed in such a way that it refuses to cluster words into cognate sets if sufficient information is lacking. As a rule of thumb, derived from consists of at least 200 words, and if the mutual coverage of the data exceeds 150 word pairs. If the data is too sparse, such as, for example, in the ASJP database (Wichmann et al., 2016) which gives maximally 40 concepts per language, we recommend to use either the SCA approach, or to turn to more sophisticated machine learning approaches (Jäger et al., 2017), which have been designed and trained in such a way that they yield their best scores on smaller datasets. In all cases, users should be aware that the algorithms may fail to detect certain cognates. The reasons range from rare sound correspondences which can trigger problematic alignments, via sparseness of data (especially when dealing with divergent languages), up to problems of morphological change which may easily confuse the algorithms as they may yield partial cognates and produce words that cannot be fully aligned anymore (List et al., 2017b). In Table 3, we summarize some basic differences between the four methods mentioned so far.

Once the words have been clustered into cognate sets, it is advisable to align all cognate words with each other, using a multiple alignment algorithm (Tutorial: 4.4). Alignments are useful in multiple ways. First, users can easily inspect them with web-based tools (Tutorial: 4.5). Second, they can be used to statistically investigate the identified sound correspondence patterns in the data (see Tutorial: 4.6). Both the manual and the automatic check of the results provided by automatic cognate detection methods are essential for a successful application of the methods. Only in this way can users either convince themselves that the results come close to their expectations or that something weird is going on. In the latter situation, we recommend that users thoroughly check to which degree they have conformed to

Table 3. Comparing different algorithms for cognate detection implemented in LingPy with respect to some fundamental parameters of sequence comparison.

Method	Scoring function	Sound classes	Gap function	Alignment mode	Partitioning
NED	identity	–	–	global	flat UPGMA
SCA	language-independent	SCA-model	prosodic profiles	global	flat UPGMA
LexStat	language-specific	SCA-model	prosodic profiles	semi-global	flat UPGMA
LexStat-Infomap	language-specific	SCA-model	prosodic profiles	semi-global	Infomap

the different parameters too much, especially when applying LingPy the first time. Instead of trying to fix minor errors (such as obvious cognates missed or look-alikes marked as cognates) by changing parameters, it is often more efficient to correct errors manually. Although Rama et al. (2018) report promising results on fully automated workflows, we do not recommend relying entirely on automatic cognate detection when it comes to phylogenetic reconstruction, since the algorithms tend to be too conservative, often missing valid cognates (List et al., 2017b), but we are confident enough to recommend it for initial data exploration, and for the preparsing of data in order to increase the efficiency of cognate annotation.

6. Evaluation

We have claimed above that automatic cognate detection had made great progress of late. We make this claim based on tests in which the performance of automatic cognate detection algorithms was compared with expert cognate judgments (List et al., 2017b). There are different ways to compare expert cognate judgments with algorithmic ones. A very simple but nevertheless important one is to compare different cognate judgments manually, by eyeballing the data. Even if one lacks expert cognate judgments for a given dataset, this may be useful, as it helps to get a quick impression on potential weaknesses of the algorithm used for a given analysis. Comparing cognate judgments in concrete, however, can be quite tedious, especially if the data are not presented in any ordered fashion. For this reason, LingPy offers a specific format that helps to compare different cognate judgments in a rather convenient way. How this comparison can be carried out is illustrated in Table 4, where we use the numeric annotation for cognate clusters as described in Fig. 6 to compare expert cognate judgments for 'to turn' in eight East Polynesian languages with those produced by edit distance, the seen from the table, NED lumps all words into one cluster, obviously being confused by the similarity of the vowels across all words. SCA comes close to the expert annotation, but wrongly separates Hawaiian [wili] from the first cluster, obviously being confused by the dissimilarity of the sound classes. LexStat correctly identifies all cognates, obviously thanks to its initial search for language-specific similarities between sound classes. In the interactive tutorial, we show how users can compute similar overviews on differences in cognate detection analyses and conveniently compare them (Tutorial: 5.1).

While manual inspection is important, it is also crucial to have an independent and objective score that tells us how well algorithms perform on a given dataset. Knowing the approximate performance may, for example, be useful when working with large datasets which would take too long to be analyzed manually. If we annotate part of the data and see that the automatic methods perform well enough, we could then use the automatic approaches to carry out our analyses and report the expected accuracy in the study. Our recommended evaluation measures are B-Cubed scores (Bagga and Baldwin, 1998; Amigó et al., 2009), which Hauer and Kondrak (2011) first introduced as a measure to assess the quality of cognate detection algorithms compared to expert judgments.

The details of how B-Cubed scores are computed are explained elsewhere in detail (List et al., 2017b), and it would go beyond the scope of this tutorial to introduce them here again. For users interested in automatic cognate detection, but reluctant in learning in depth about evaluation measures in computational linguistics, it is sufficient to know how the B-Cubed scores should be interpreted. Usually the scores are given in three forms, which all rank between 0 and 1: *precision*, *recall*, and *F-Score*. Precision comes closest to the notion of *true positives* in historical linguistics. Recall is close to the notion of *true negatives*, accordingly, and the F-Score, the harmonic mean of precision and recall, can be seen as a

$2\frac{P \times R}{P+R}$, where P is the precision and R is the recall. If the scores are high, this means the algorithms come close to the judgment of the experts, a score of 1.0 in precision and recall (and therefore also the F-Score) means that the results are 100% identical.

In Table 5, we report the results achieved by four automatic cognate detection methods on a small subset of ten East Polynesian languages which we retrieved from our Polynesian dataset for illustrative purposes.[8] In addition to the three methods reported already in Table 4, we added a random cognate detector which was sampled from 100 trials, and the Infomap version of the LexStat algorithm (LS-Infomap), in which the cognate set partitioning is carried out with the Infomap algorithm instead of the flat version of UPGMA (see Section 5 above).[9] NED shows a rather low precision compared to the other nonrandom approaches, indicating that it proposes many false positives (as we could see above in Table 4). On the other hand, its recall is very high, indicating that it does not miss many cognate sets. SCA obviously has a lot of problems with the data, performing worse than NED in general, with a rather low precision and recall. Both LexStat approaches largely outperform the other approaches in general, and especially the very high precision is very comforting, since it indicates that the algorithms do not propose too many false positives. That the Infomap version of LexStat

Table 4. Comparing automatic cognate detection methods with expert cognate judgments for words for 'to turn' in East Polynesian languages.

Doculect	Form	Expert	NED	SCA	LexStat
Ra'ivavae	ta: viGi [4580]	1	1	1	1
Hawaiian	wili [5835]	1	1	4	1
North-Marquesan	kavi?i [3575]	1	1	1	1
Rapanui	taviri [1838]	1	1	1	1
Hawaiian	huli [5834]	2	1	2	2
Maori	huri [936]	2	1	2	2
Sikaiana	tahuli [3283]	2	1	2	2
Mangareva	ti: rori [2101]	3	1	3	3

Highlighted cells indicate where the respective algorithms fail compared to the expert judgment.

Table 5. B-Cubed scores for different cognate detection algorithms compared against a test set of East Polynesian languages.

	RANDOM	NED	SCA	LexStat	LS-Infomap
Precision	0.47	0.81	0.88	0.95	0.94
Recall	0.73	0.96	0.84	0.92	0.93
F-score	0.57	0.88	0.86	0.93	0.94

Highlighted cells indicate the best scores for a given measure.

A Pairwise Distance Calculation

In order to calculate distances between words, different methods can be used. A simple way is to count, how often two two aligned strings show different characters (*edit distance*). By dividing this number by the length of the alignment, we can normalize this distance so that it ranges between 0 and 1. More sophisticated ways consist in computing the alignment scores (which are rendered as similarities) and converting these to distance scores using a formula for normalization (Downey et al. 2008).

Hawaiian	ʔ	a:	ʔ	i:	
Mangareva	k	a	k	i	
	1 +	1 +	1 +	1	= 4 / 4 = 1.0
Mangareva	k	a	k	i	
Rapanui	ŋ	a	-	o	
	1 +	0 +	1 +	1	= 3 / 4 = 0.75

B Flat Clustering (Partitioning)

Algorithms for flat clustering (also called partitioning algorithms) take a matrix of pairwise distances between objects as input and return a partition of the objects in which each object is assigned to exactly one group. As a common notation, we can assign each object a numeric ID. In this way we can easily compare to which degree different partition proposals for the same data differ.

Objects	Part. 1	Part. 2	Part. 3
A	1	1	1
B	1	2	1
C	2	2	1

C Permutation Test

It is not possible to use simple combinatorics to predict how many sound correspondences we would expect if all languages in our data were unrelated, since we do not know the underlying phonotactics of our languages. But we can compute the sound correspondences by shuffling our data multiple times, comparing words expressing different concepts, and counting how many correspondences we find for presumably unrelated word pairs. The table on the right shows correspondence distributions and the resulting LexStat score for some sounds in Sikaiana and Tuamotuan.

Sikaiana		Tuamotuan		Distributions		
IPA	Class	IPA	Class	Att.	Exp.	Score
f, v	B	f, v	B	10.0	1.1	7.32
f, v	B	k	K	0	1.8	-3.33
h	H	h, ʔ	H	28.0	8.4	5.69
h	H	k	K	1.0	4.2	-3.33

performs better than LexStat with UPGMA is also shown in this comparison, although the differences are much lower than reported in List et al. (2017b). It would be very interesting to compare the scores we achieved with general scores of levels of agreement among human experts. Unfortunately, no systematic study has been carried out so far.[10] The interactive tutorial gives a detailed introduction into the computation of B-Cubed scores with LingPy (Tutorial: 5.2).

Given the differences in the results regarding precision, recall, and generalized F-scores, it is obvious that the choice of the algorithm to use depends on the task at hand. If users plan to invest much time into manual data correction, having an algorithm with high recall that identifies most of the cognates in the data while proposing a couple of erroneous ones is probably the best choice. Users can achieve this by choosing a high threshold or an algorithm such as NED, which yields a rather high recall in form of the B-Cubed scores, at least for the Polynesian data in our sample. In other cases, however, when user-correction is not feasible because of the size of the dataset, it is useful to choose low thresholds or generally conservative algorithms with high B-Cubed precision in order to minimize the amount of false positives.

7. Data export

LingPy provides direct export of the cognate judgments to the Nexus format (Maddison et al., 1997), allowing users to analyze automated cognate judgments with popular packages for phylogenetic reconstruction, such as SplitsTree (Huson, 1998), MrBayes (Ronquist et al., 2009), or BEAST 2 (Bouckaert et al., 2014, see Tutorial: 6.1). If phylogenetic trees are computed from distance matrices, both matrices and trees can be written to file and further imported in software packages for tree manipulation and visualization (Tutorial: 6.2). In addition, data can be exported (and also be imported) to the wordlist format proposed by the CLDF initiative (Forkel et al., 2017), which is intended to serve as a generic format for data sharing in cross-linguistic studies (Tutorial: 6.3).

8. Concluding remarks

In this tutorial we have tried to show how automatic sequence comparison in LingPy can be carried out. Given the scope of this article, it is clear that we could not cover all aspects of alignments and cognate detection in all due detail. We hope, however, that we could help readers understand what they should keep in mind if interactive tutorial supplemented with this article, and for deeper questions going beyond the pure application of sequence comparison algorithms—such as additional analyses (e.g. the *minimal lateral network* method for borrowing detection, List et al., 2014, or an algorithm for the detection of partial cognates, List et al., 2016b), routines for plotting and data visualization, or customization routines for user-defined sound-class models—we recommend the readers to turn to the extensive online documentation of the LingPy package (http://lingpy.org). We have emphasized multiple times throughout this article that the algorithms cannot and should not be used to replace trained linguists. Instead, they should be seen as a useful complement to the large arsenal of methods for historical language comparison which can help experts to derive initial hypotheses on cognacy, speed up tedious annotation of cognate sets, and increase their efficiency and consistency.

Supplementary data

Supplementary data is available at *Journal of Language Evolution* online. Stable updates of this material with the latest version are also available at Zenodo (https://doi.org/10.5281/zenodo.1252230).

Funding

This research was supported by the European Research Council Starting Grant 'Computer-Assisted Language Comparison' (Grant CALC 715618, J.M.L., T.T.) and the Australian Research Council's Centre of Excellence for the Dynamics of Language (Australian National University, Grant CE140100041, S.J.G.). As part of the GlottoBank project (http://glottobank.org), this work was further supported by the Department of Linguistic and Cultural Evolution of the Max Planck Institute for the Science of Human History (Jena) and the Royal Society of New Zealand (Marsden Fund, Grant 13-UOA-121).

Notes

1. Linguists are often skeptical when they hear that LingPy requires explicit phonetic transcriptions, and often, they are even reluctant to interpret their data along the lines of the IPA. But in order to give the algorithms a fair chance to interpret the data in the same way in which they would be interpreted by linguists, a general practice for phonetic transcriptions is indispensable, and the IPA is the most widely employed transcription system.
2. We know well how difficult it is to conform to the lat-

Gudschinsky (1956), will have a deleterious impact on any analysis (List, 2018). In order to avoid synonyms in qualitative work, we recommend to thoroughly review the guidelines in Kassian et al. (2010).

3. It would go beyond the scope of this tutorial to explain these famous algorithms in all detail. Instead, we refer the readers to Kondrak (2002: 20–65) as well as to an interactive demo of the Wagner–Fischer algorithm in List (2016).
4. In the *normalized edit distance* (NED), the *edit distance* between two strings is further normalized by dividing it by the length of the longer string. In this way, we can control for the length of the compared sequences.
5. The threshold for the algorithms are: NED: 0.75, SCA: 0.45, LexStat: 0.6.
6. For an example, consider the matches between Sikaiana and Tahitian shown in Table 1. Although Sikaiana [k] is different from [ʔ], they are similar from a language-specific perspective, since they recur across many aligned cognate sets between both languages. When comparing [k] in English with [ʔ] in German, however, they are not similar, as we will not find a cognate set in which those two sounds correspond.
7. As alignment algorithms yield similarity scores as a default, the similarity scores are converted to distance scores with help of the formula proposed by Downey et al. (2008).
8. We have not fully explored the practical limitations in terms of number of languages or number of concepts when comparing languages with LingPy. Jäger et al. (2017) and Rama et al. (2017) report successful applications of LingPy's cognate detection algorithms for as many as 100 languages. Although we think that the number might in fact be even higher, based on tests we carried out ourselves on 150 and more languages, we recommend to be careful when analyzing too many languages, as algorithmic performance may drastically drop when investigation samples are too large
9. The threshold for LexStat-Infomap was set to 0.55, following List et al. (2017b). The random cognate annotation algorithm was designed in such a way that it has the tendency to lump cognates to larger clusters.
10. The only study known to us addressing these problems is Geisler and List (2010), but it has, unfortunately, not been sufficiently quantified.

Acknowledgements

We are grateful to three anonymous reviewers for challenging and Christoph Rzymski for helpful comments on an earlier version of this draft.

References

Amigó, E. et al. (2009) 'A Comparison of Extrinsic Clustering Evaluation Metrics Based on Formal Constraints', *Information Retrieval*, 12/4: 461–86.

Bagga, A. and Baldwin, B. (1998) 'Entity-based cross-document coreferencing using the vector space model' in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pp. 79–85. Móntreal: Association of Computational Linguistics.

Bouckaert, R. et al. (2014) 'BEAST 2: A Software Platform for Bayesian Evolutionary Analysis', *PLoS Computational Biology*, 10/4: e1003537.

Brown, C. H. et al. (2008) 'Automated Classification of the World's Languages', *Sprachtypologie und Universalienforschung*, 61/4: 285–308.

——, Holman, E. W., and Wichmann, S. (2013) 'Sound Correspondences in the World's Languages', *Language*, 89/1: 4–29.

Campbell, L. (2011) 'Comment On: Automated Dating of the World's Language Families Based on Lexical Similarity', *Current Anthropology*, 52: 866–7.

Dixon, R. B. and Kroeber, A. L. (1919) *Linguistic Families of California*. Berkeley: University of California Press.

Dolgopolsky, A. B. (1964) 'Gipoteza drevnejego rodstva jazykovych semej Severnoj Evrazii s verojatnostej toky zrenija', *Voprosy Jazykoznanija*, 2: 53–63.

Downey, S. S. et al. (2008) 'Computational Feature-sensitive Reconstruction of Language Relationships: Developing the ALINE Distance for Comparative Historical Linguistic Reconstruction', *Journal of Quantitative Linguistics*, 15/4: 340–69.

Dunn, M. (2012) *Indo-European lexical cognacy database (IELex)*. Nijmegen: Max Planck Institute for Psycholinguistics <http://ielex.mpi.nl>.

Durbin, R. et al. (2002) *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids*, 7th edn. Cambridge: Cambridge University Press.

Eddy, S. R. (2004a) 'Where Did the BLOSUM62 Alignment Score Matrix Come From?' *Nature Biotechnology*, 22/8: 1035–6.

—— (2004b) 'What is Dynamic Programming?' *Nature Biotechnology*, 22/7: 909–10.

Feng, D. F. and Doolittle, R. F. (1987) 'Progressive Sequence Alignment as a Prerequisite to Correct Phylogenetic Trees. *Journal of Molecular Evolution*, 25/4: 351–60.

Forkel, R. et al. *CLDF. Cross-Linguistic Data Formats. Version 1.0*. Max Planck Institute for the Science of Human History, Jena, 2017. doi: 10.5281/zenodo.1117644. <https://doi.org/10.5281/zenodo.1117644>.

Fox, A. (1995) *Linguistic Reconstruction*. Oxford: Oxford University Press. ISBN 0-19-870000-8.

Hettrich Heinrich (ed.) *Die Ausbreitung des Indogermanischen. Thesen aus Sprachwissenschaft, Archologie und Genetik. Wiesbaden: Reichert. In press*.

Greenhill, S. J. (2011) 'Levenshtein Distances Fail to Identify Language Relationships Accurately', *Computational Linguistics*, 37/4: 689–98.

——, Blust, R., and Gray, R. D. (2008) 'The Austronesian Basic Vocabulary Database: From Bioinformatics to Lexomics', *Evolutionary Bioinformatics*, 4: 271–83.

Gudschinsky S. C. (1956) 'The ABC's of Lexicostatistics (glottochronology)', *Word*, 12/2:175–210.

Hammarström, H., Forkel, R., and Haspelmath, M. (2017) *Glottolog*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://glottolog.org>.

Haspelmath, M. and Tadmor, U. (2009) 'The Loanword Typology Project and the World Loanword Database', in Haspelmath Martin and Tadmor Uri (eds) *Loanwords in the World's Languages, pp. 1–34*. Berlin and New York: de Gruyter.

Hauer, B. and Kondrak, G. (2011) 'Clustering semantically equivalent words into cognate sets in multilingual lists'. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pp. 865–73. Chiang Mai: AFNLP.

Hill, N. W. and List, J. -M. (2017) 'Challenges of Annotation and Analysis in Computer-assisted Language Comparison: A Case Study on Burmish Languages', *Yearbook of the Pozna Linguistic Meeting*, 3/1: 4776.

Holman, E. W. et al. (2008) 'Explorations in Automated Lexicostatistics', *Folia Linguistica*, 20/3: 116–21.

—— et al. (2011) 'Automated Dating of the World's Language Families Based on Lexical Similarity', *Current Anthropology*, 52/6: 841–75.

Hruschka, D. J. et al. (2015) 'Detecting Regular Sound Changes in Linguistics as Events of Concerted Evolution', *Current Biology*, 25/1: 1–9.

Huson, D. H. (1998) 'SplitsTree: Analyzing and Visualizing Evolutionary Data', *Bioinformatics*, 14/1: 68–73.

International Phonetic Association (IPA) (1999) *Handbook of the International Phonetic Association. A Guide to the Use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press.

Jäger, G. (2015) 'Support for Linguistic Macrofamilies from Weighted Alignment', *Proceedings of the National Academy of Sciences of the United States of America*, 112/41: 12752–7.

—— and List, J. -M. (2015) 'Factoring Lexical and Phonetic Phylogenetic Characters from Word Lists', in Baayen H. et al. (eds) *Proceedings of the 6th Conference on Quantitative Investigations in Theoretical Linguistics*. Eberhard-Karls University: Tübingen.

——, List, J.-M., and Sofroniev, P. (2017) 'Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists'. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*.

Kassian Alexei et al. (2010) 'The Swadesh Wordlist. An Attempt at Semantic Specification', *Journal of Language Relationships*, 4: 46–89.

Kessler, B. (2001) *The Significance of Word Lists*. Stanford: CSLI Publications.

Kondrak, G. (2000) 'A new algorithm for the alignment of phonetic sequences'. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pp. 288–95. Seattle: Association of Computational Linguistics.

—— (2002) '*Algorithms for Language Reconstruction', Dissertation*, University of Toronto, Toronto.

Lass, R. (1997) *Historical Linguistics and Language Change*. Cambridge: Cambridge University Press.

Levenshtein V. I. (1965) 'Dvoičnye kody s ispravleniem vypadenij, vstavok i zameščenij simvolov', *Doklady Akademij Nauk SSSR*, 163/4: 845–8.

List, J. -M. (2012a) 'SCA. Phonetic Alignment Based on Sound Classes', in Slavkovik Marija and Lassiter Dan(eds) *New Directions in Logic, Language, and Computation, pp. 32–51*. Berlin and Heidelberg: Springer.

—— (2012b). 'LexStat. Automatic detection of cognates in multilingual wordlists'. In *Proceedings of the EACL 2012 Joint Workshop of Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources*, pp. 117–25, Stroudsburg.

—— (2014a). 'Investigating the Impact of Sample Size on Cognate Detection', *Journal of Language Relationship*, 11: 91–101.

—— (2014b) *Sequence Comparison in Historical Linguistics*. Düsseldorf: Düsseldorf University Press.

——. Wagner-Fischer Demo. *figshare*, 2016. doi: http://dx.doi.org/10.6084/m9.figshare.3158836.v1. <https://figshare.com/articles/Wagner_Fischer_Demo/3158836>

—— (2017) 'A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets'. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*, pp. 9–12, Valencia: Association for Computational Linguistics.

—— (2018) 'Tossing Coins: Linguistic Phylogenies and Extensive Synonymy', *The Genealogical World of Phylogenetic Networks*, 7/2. <http://phylonetworks.blogspot.de/2018/02/tossing-coins-linguistic-phylogenies.html>.

—— et al. (2014) 'Networks of Lexical Borrowing and Lateral Gene Transfer in Language and Genome Evolution', *Bioessays*, 36/2: 141–50.

——, Cysouw, M., and Forkel, R. (2016a) 'Concepticon. A resource for the linking of concept lists'. In N. Calzolari et al. (eds) *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pp. 2393–400. Portorož: European Language Resources Association (ELRA).

——, Lopez, P., and Bapteste, E. (2016b) 'Using sequence similarity networks to identify partial cognates in multilingual wordlists'. In *Proceedings of the Association of Computational Linguistics*

——, Greenhill, S. J., and Forkel, R. (2017a) *LingPy. A Python Library for Historical Linguistics*. Jena: Max Planck Institute for the Science of Human History, doi: https://zenodo.org/badge/latestdoi/5137/lingpy/lingpy. <http://lingpy.org>.

——, ——, and Gray, R. D. (2017b). 'The Potential of Automatic Word Comparison for Historical Linguistics', *PLOS One*, 12/1: 1–18.

—— et al., eds. (2018) *Concepticon. A Resource for the Linking of Concept List*. Jena: Max Planck Institute for the Science of Human History. <http://concepticon.clld.org/>.

Maddison D. R., Swofford, D. L., and Maddison W. P. (1997) 'NEXUS: An Extensible File Format for Systematic Information', *Systematic Biology*, 46/4: 590–621.

Matisoff, J. A. (2015) *The Sino-Tibetan Etymological Dictionary and Thesaurus Project*. Berkeley: University of California.

Moran, S. and Cysouw, M. (2017) *The Unicode Cookbook for Linguists: Managing Writing Systems Using Orthography Profiles*. Zürich: Zenodo. doi: 10.5281/zenodo.290662. <https://doi.org/10.5281/zenodo.290662>.

——, McCloy, D. and Wright, R., eds. (2014) *PHOIBLE Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://phoible.org/>.

Morgenstern, B. et al. (1998) 'Segment-based scores for pairwise and multiple sequence alignments'. In J. Glasgow (eds), *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology*, pp. 115–21. AAAI Press: Menlo Park.

Needleman, S. B. and Wunsch, C. D. (1970) 'A Gene Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins', *Journal of Molecular Biology*, 48: 443–53.

Newman, M. E. J. (2006) 'Finding Community Structure in Networks using the Eigenvectors of Matrices', *Physical Review E*, 74/3: 1–19.

Notredame, C., Higgins, D. G., and Heringa, J. (2000) 'T-Coffee', *Journal of Molecular Biology*, 302: 205–17.

Pérez, F. and Granger, B. E. (2007) 'IPython: A System for Interactive Scientific Computing', *Computing in Science and Engineering*, 9/3: 21–9.

Prokić, J., Wieling, M., and Nerbonne, J. (2009) 'Multiple sequence alignments in linguistics'. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, pp. 18–25. Athens: Association for Computational Linguistics.

Rama, T. et al. (2017) 'Fast and unsupervised methods for multilingual cognate clustering'. *CoRR*, abs/1702.04938. <http://arxiv.org/abs/1702.04938>.

—— et al. (2018) 'Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics?' In *Proceedings of the North American Chapter of the Association of Computational Linguistics*. New Orleans: Association for Computational Linguistics.

Ronquist, F., van der Mark, P., and Huelsenbeck, J. P. (2009) 'Bayesian Phylogenetic Analysis Using MrBayes', in Lemey P., Salemi M., and Vandamme A. -M. (eds) *The Phylogenetic Handbook. A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*, 2nd edn, pp. 210–66. Cambridge: Cambridge University Press.

Rosvall, M. and Bergstrom, C. T. (2008) 'Maps of Random Walks on Complex Networks Reveal Community Structure', *Proceedings of the National Academy of Sciences of the United States of America*, 105/4: 1118–23.

Smith T. F. and Waterman M. S. (1981) 'Identification of Common Molecular Subsequences', *Journal of Molecular Biology*, 1: 195–7.

Sokal, R. R. and Michener, C. D. (1958) 'A Statistical Method for Evaluating Systematic Relationships', *University of Kansas Scientific Bulletin*, 28: 1409–38.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) 'CLUSTAL W', *Nucleic Acids Research*, 22/22: 4673–80.

Wagner, R. A. and Fischer, M. J. (1974) 'The String-to-string Correction Problem', *Journal of the Association for Computing Machinery*, 21/1: 168–73.

Wheeler, W. C. and Whiteley, P. M. (2015) 'Historical Linguistics as a Sequence Optimization Problem: The Evolution and Biogeography of Uto-Aztecan Languages', *Cladistics*, 31/2: 113–25, 2015.

Wichmann, S., Holman, E. W., and Brown, C. H. (2016) *The ASJP Database*. Jena: Max Planck Institute for the Science of Human History. <http://asjp.clld.org>.

Automatic Inference of Sound Correspondence Patterns across Multiple Languages

Johann-Mattis List
Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Jena

Sound correspondence patterns play a crucial role for linguistic reconstruction. Linguists use them to prove language relationship, to reconstruct proto-forms, and for classical phylogenetic reconstruction based on shared innovations. Cognate words that fail to conform with expected patterns can further point to various kinds of exceptions in sound change, such as analogy or assimilation of frequent words. Here I present an automatic method for the inference of sound correspondence patterns across multiple languages based on a network approach. The core idea is to represent all columns in aligned cognate sets as nodes in a network with edges representing the degree of compatibility between the nodes. The task of inferring all compatible correspondence sets can then be handled as the well-known minimum clique cover problem in graph theory, which essentially seeks to split the graph into the smallest number of cliques in which each node is represented by exactly one clique. The resulting partitions represent all correspondence patterns that can be inferred for a given data set. By excluding those patterns that occur in only a few cognate sets, the core of regularly recurring sound correspondences can be inferred. Based on this idea, the article presents a method for automatic correspondence pattern recognition, which is implemented as part of a Python library which supplements the article. To illustrate the usefulness of the method, I present how the inferred patterns can be used to predict words that have not been observed before.

1. Introduction

By comparing the languages of the world, we gain invaluable insights into human prehistory, predating the appearance of written records by thousands of years. The classical methods for historical language comparison, a collection of different techniques summarized under the term **comparative method** (Meillet 1954; Weiss 2015), date back to the early 19th century and have since then been constantly refined and improved (see Ross and Durie 1996 for details on the practical workflow). Thanks to the comparative method, linguists have made groundbreaking insights into language change in general and into the history of many specific language families (Campbell and Poser 2008) and external evidence has often confirmed the validity of the findings (McMahon and

Submission received: 4 April 2018; revised version received: 3 October 2018; accepted for publication: 21 November 2018.

doi:10.1162/COLI_a_00344

McMahon 2005, pages 10–14). With increasing amounts of data, however, the methods, which are largely manually applied, reach their practical limits. As a result, scholars are now increasingly trying to automate different aspects of the classical comparative methods (Kondrak 2000; Prokić, Wieling, and Nerbonne 2009; List 2014).

One of the fundamental insights of early historical linguistic research was that—as a result of systemic changes in the sound system of languages—genetically related languages exhibit structural similarities in those parts of their lexicon that were commonly inherited from their ancestral languages. These similarities surface in the form of **correspondence relations** between sounds from different languages in cognate words. English *th* [θ], for example, is usually reflected as *d* in German, as we can see from cognate pairs like English *think* versus German *denken*, or English *thorn* and German *Dorn*. English *t*, on the other hand, is usually reflected as *z* **[ts]** in German, as we can see from pairs like English *toe* versus German *Zeh*, or English *ten* versus German *zehn*. The identification of these **regular sound correspondences** plays a crucial role in historical language comparison, serving not only as the basis for the proof of genetic relationship (Dybo and Starostin 2008; Campbell and Poser 2008) or the **reconstruction of proto-forms** (Hoenigswald 1960, pages 72–85; Anttila 1972, pages 229–263), but (indirectly) also for classical subgrouping based on shared innovations (which would not be possible without identified correspondence patterns).

With the beginning of this millennium, historical linguistics has witnessed an increased number of attempts to quantify specific tasks of the traditional comparative method. Since then, scholars have repeatedly attempted to either directly infer regular sound correspondences across genetically related languages (Kay 1964; Brown, Holman, and Wichmann 2013; Kondrak 2003, 2009) or integrated the inference into workflows for automatic cognate detection (Guy 1994; List 2012, 2014; List, Greenhill, and Gray 2017). What is interesting in this context, however, is that almost all approaches dealing with regular sound correspondences, be it early formal—but classically grounded—accounts (Grimes and Agard 1959; Hoenigswald 1960) or computer-based methods (Kondrak 2002, 2003; List 2014) only consider sound correspondences between *pairs* of languages.

A rare exception can be found in the work of Antilla (1972, pages 229–263) who presents the search for regular sound correspondences across multiple languages as the basic technique underlying the comparative method for historical language comparison. Anttila's description starts from a set of cognate word forms (or morphemes) across the languages under investigation. These words are then arranged in such a way that corresponding sounds in all words are placed into the same column of a matrix. The extraction of regularly recurring sound correspondences in the languages under investigation is then based on the identification of similar patterns recurring across different columns within the cognate sets. The procedure is illustrated in Figure 1, where four cognate sets in Sanskrit, Ancient Greek, Latin, and Gothic are shown, two taken from Anttila (1972, page 246) and two added by me.

Two points are remarkable about Anttila's approach. First, it builds heavily on the **phonetic alignment** of sound sequences,[1] by which the sound sequences of words are arranged in a matrix in such a way that all corresponding sounds are placed in the same cell (List 2014). Second, it reflects a concrete technique by which regular sound

1 This concept was only recently adapted in linguistics (Covington 1996; Kondrak 2000; List 2014), building heavily on approaches in bioinformatics and computer science (Needleman and Wunsch 1970; Wagner and Fischer 1974), although it was implicitly always an integral part of the methodology of historical language comparison (compare Dixon and Kroeber 1919, Fox 1995, 67f).

	A			B		C			D		E				F				
Sanskrit	y	u	g	a	m	dh	u	h	i	(tar)	s	n	u	ṣ	(ā)	-	r	u	dh	(iras)
Greek	z	u	g	o	n	th	u	g	a	(ter-)	-	n	u	-	(os)	e	r	u	th	(rós)
Latin	i	u	g	u	m	Ø	Ø	Ø	Ø	(Ø)	-	n	u	r	(us)	-	r	u	b	(er)
Gothic	j	u	k	-	-	d	au	h	-	(tar)	Ø	Ø	Ø	Ø	(Ø)	Ø	Ø	Ø	Ø	(Ø)
Gloss	'yoke'					'daughter'					'daughter-in-law'					'red'				

Figure 1
Regular sound correspondences across four Indo-European languages, illustrated with help of alignments along the lines of Anttila (1972, page 246). In contrast to the original illustration, lost sounds are displayed with help of the dash "-" as a gap symbol, while missing words (where no reflex in Gothic or Latin could be found) are represented by the "∅" symbol.

correspondences for multiple languages can be detected and employed as a starting point for linguistic reconstruction. If we look at the framed columns in the four examples in Figure 1, which are further labeled alphabetically, we can easily see that the patterns A, E, and F are remarkably similar. The only difference is that we miss data for Gothic in the patterns E and F, and, as a result, we don't have **reflex sounds** (sounds in a given alignment column as reflected in a cognate word) for the full sound correspondence patterns in the respective columns. The same holds, however, for columns C, E, and F. Since A and C differ regarding the reflex sound of Gothic (*u* vs. *au*), they cannot be assigned to the same correspondence set at this stage, and if we want to solve the problem of finding the regular sound correspondences for the words in the figure, we need to decide which columns in the alignments we assign to the same correspondence set, thereby "imputing" missing sounds where we miss a reflex. Assuming that the "regular" pattern in our case is reflected by the group of C, E, and F, we can make *predictions* about the sounds missing in Gothic in E and F, concluding that, if ever we find the missing reflex in so far unrecognized sources of Gothic in the future, we would expect a *-au-* in the words for "daughter-in-law" and "red".[2]

We can easily see how patterns of sound correspondences across multiple languages can serve as the basis for multiple tasks in historical linguistics. First, we could use them to guess how a word that is missing in a given alignment would sound in that language, if it could be found. Since the task of identifying cognate words across multiple languages is very complex, and words may have drastically shifted their meanings, we could use the predictions to search for missing cognate forms in those areas of the lexicon that we have not considered before.[3] Second, if two alignment columns are identical, they must reflect the same proto-sound, if alternative processes like borrowing can be excluded. Thus, similarly to the prediction of missing words in our cognate sets, we could use correspondence patterns to infer proto-forms, provided that parts of the data are already annotated.[4] Third, we could use them to check linguistic claims

2 As pointed out by the anonymous reviewer, Gothic *ráups* is a reflex of 'red' (Wright 1910, page 340), but as mentioned by Eugen Hill (personal communication), the Gothic form reflects a derivationally different formation and was therefore correctly not listed in Anttila's examples.

3 Consider cases of shifted meanings like English *hound* vs. German *Hund* 'dog,' or English *-thorp* as a prefix in place names compared to German *Dorf* 'village.'

4 But even if correspondence patterns are not identical, they could be assigned to the same proto-sound, provided that one can show that the differences are conditioned by phonetic context. This is the case for Gothic *au* [o] in pattern C, which has been shown to go back to *u* when preceding *h* (Meier-Brügger 2002, page 210f). As a result, scholars usually reconstruct Proto-Indo-European **u* for A, C, E, and F.

about cognate words themselves: If it turns out that the aligned cognate sets proposed by linguists do not pattern into recurring correspondences across the languages under consideration, we can directly criticize both individual claims regarding word relations and general claims about the genetic relation of languages.

While it seems trivial to identify sound correspondences across multiple languages from the few examples provided in Figure 1, the problem can become quite complicated if we add more cognate sets and languages to the comparative sample. Especially the handling of **missing reflexes** for a given cognate set becomes a problem here, as missing data makes it difficult for linguists to decide which alignment columns to group with each other. This can already be seen from the examples given in Figure 1, where we have two possibilities to group the patterns A, C, E, and F, if we base our judgments only on these four patterns: E and F could be grouped with either A or C, and it may even be possible that one should be grouped with A and one with C. The "true" solution here depends on the history of the languages, but if the data that would allow us to reconstruct this history is lost, we can never infer the historically correct grouping with full confidence.

The goal of this article is to illustrate how a manual analysis in the spirit of Anttila can be automated and fruitfully applied—not only in purely computational approaches to historical linguistics, but also in computer-assisted frameworks that help linguists to explore their data before they start carrying out painstaking qualitative comparisons (List 2016). In order to illustrate how this problem can be solved computationally, the article will first discuss some important general aspects of sound correspondences and sound correspondence patterns in Section 2, introducing specific terminology that will be needed in the remainder. In Section 3, we will see that the problem of finding sound correspondences across multiple languages can be modeled as the well-known *clique-cover problem* in an undirected network (Bhasker and Samad 1991). While this problem is *hard* to solve in an exact way computationally,[5] fast approximate solutions exist (Welsh and Powell 1967) and can be easily applied. Based on these findings, the article will introduce a fully automated method for the recognition of sound correspondence patterns across multiple languages (Section 4). This method is implemented in the form of a Python library and can be readily applied to multilingual wordlist data as it is also required by software packages such as LingPy (List, Greenhill, and Forkel 2017) or software tools such as EDICTOR (List 2017). Section 5 will then illustrate how the method can be applied by testing how it performs in the task of predicting missing cognate words and missing proto-forms.

2. Preliminaries on Sound Correspondence Patterns

In the introduction, it was emphasized that the traditional comparative method is itself less concerned with regular sound correspondences attested for language pairs, but for all languages under consideration. In the following, this claim will be further substantiated, while at the same time introducing some major methodological considerations and ideas that are important for the development of the new method for sound correspondence pattern recognition.

5 Both the clique-cover problem and its inverse problem, the graph coloring problem, have been shown to be *np-complete* (Bhasker and Samad 1991).

Table 1
Comparing correspondence patterns for Proto-Germanic reflexes of **d-*, **þ-*, and **t-* in German, English, and Dutch (Germanic proto-forms follow Kroonen [2013]).

Gloss	Proto-Germanic		German		English		Dutch	
'dead'	**daudaz*	daudaz	*tot*	toːt	*dead*	dɛd	*dood*	doːt
'deed'	**dēdiz*	deːdiz	*Tat*	taːt	*deed*	diːd	*daad*	daːt
'blood'	**blōdan*	bloːdan	*Blut*	bluːt	*blood*	blʌd	*bloed*	*blut*
'courage'	**mōdaz*	moːdaz	*Mut*	muːt	*mood*	muːd	*moed*	muːt
'thick'	**þekuz*	θekuz	*dick*	dɪk	*thick*	θɪk	*dik*	dɪk
'thorn'	**þurnuz*	θurnuz	*Dorn*	dɔrn	*thorn*	θɔːn	*doorn*	doːrn
'field'	**felþuz*	felθuz	*Feld*	fɛlt	*field*	fiːld	*veld*	vɛlt
'gold'	**gulþan*	gulθan	*Gold*	gɔlt	*gold*	gəʊld	*goud*	xɑut
'tongue'	**tuŋgōn*	tuŋgoːn	*Zunge*	tsʊŋə	*tongue*	tʌŋ	*tong*	tɔŋ
'tooth'	**tanþs*	tanθs	*Zahn*	tsaːn	*tooth*	tuːθ	*tand*	tɑnt
'become stiff'	**sterbanan*	sterbanan	*sterben*	ʃtɛrbən	*to starve*	staːv	*sterven*	stɛrvə
'thirst	**þurstuz*	θurstuz	*Durst*	dʊrst	*thirst*	θɜːst	*dorst*	dɔrst

2.1 From Sound Correspondences to Sound Correspondence Patterns

Sound correspondences are most easily defined for pairs of languages. Thus, it is straightforward to state that German **[d]** regularly corresponds to English **[θ]** (or [ð]), that German **[ts]** regularly corresponds to English **[t]**, and that German **[t]** corresponds to English **[d]**. We can likewise expand this view to multiple languages by adding another Germanic language, such as, for example, Dutch to our comparison, which has **[d]** in the case of German **[d]** and English [θ], **[t]** in the case of German **[ts]** and English **[t]**, and **[d]** in the case of German **[t]** and English **[d]**.

The more languages and examples we add to the sample, however, the more complex the picture becomes, and while we can state three (basic) patterns for the case of English, German, and Dutch, given in our example, we may get easily more patterns, due to secondary sound changes in the different languages, although we would still reconstruct only three sounds in the proto-language (**[θ, t, d]**). This is illustrated in Table 1, where Proto-Germanic forms containing ***p[þθ]**, ***t**, and ***d** in different phonetic environments are contrasted with their descendant forms in German, English, and Dutch. The example shows that there is a one-to-*n* relationship between what we interpret as a proto-sound of the proto-language, and the regular correspondence patterns that we may find in our data. While we will reserve the term **sound correspondence** for pairwise language comparison, we will use the term **sound correspondence pattern** (or simply **correspondence pattern**) for the abstract notion of regular sound correspondences across a set of languages that we can find in the data.

2.2 Correspondence Patterns in the Classical Literature

Scholars like Meillet (1908, page 23) have stated that the core of historical linguistics is not linguistic reconstruction, but the inference of correspondence patterns, emphasizing that "reconstructions are nothing else but the signs by which one points to the

Table 2
Sound correspondence patterns for Indo-European stops, following Clackson (2007, page 37) .

PIE	Hittite	Sanskrit	Greek	Latin	Gothic	...
*p	p	p	p	p	f b	...
*b	b p	b	b	b	p	...
*b^h	b p	$b^ɦ$/bh	p^h/ph	f b	b	...
*t	t	t	t	t	θ/þ d	...
*d	d t	d	d	d	t	...
*d^h	d t	d^h/dh h	t^h/th	f d b	d	...
...	...	...	...	...	...	...
*k^w	k^w/ku	k c	k p t	k^w/qu	h^w/hw g	...
*g^w	k^w/u	g j	g b d	g^w/gu u	q	...
*g^{wh}	k^w/ku g^w/gu	$g^ɦ$/gh h	p^h/ph t^h/th k^h/kh	f g^w/gu u	g b	...

correspondences in short form".[6] However, given the one-to-*n* relation between proto-sounds and correspondence patterns, it is clear that this is not quite correct. Having inferred regular correspondence patterns in our data, our reconstructions will add a different level of analysis by further *clustering* these patterns into groups that we believe to reflect one single sound in the ancestral language.

That there are usually more than just one correspondence pattern for a reconstructed proto-sound is nothing new to most practitioners of linguistic reconstruction. Unfortunately, however, linguists rarely list all possible correspondence patterns exhaustively when presenting their reconstructions, but instead select the most frequent ones, leaving the explanation of weird or unexpected patterns to comments written in prose. A first and important step of making a linguistic reconstruction system transparent, however, should start from an exhaustive listing of all correspondence patterns, including irregular patterns that occur very infrequently but would still be accepted by the scholars as reflecting true cognate words.

What scholars do instead is provide tables that summarize the correspondence patterns in a rough form, for example, by showing the reflexes of a given proto-sound in the descendant languages in a table, where multiple reflexes for one and the same language are put in the same cell. An example, taken with modifications[7] from Clackson (2007, page 37), is given in Table 2. In this table, the major reflexes of Proto-Indo-European stops in 11 languages representing the oldest attestations and major branches of Indo-European are listed. This table is a very typical example for the way in which scholars discuss, propose, and present correspondence patterns in linguistic reconstruction (Beekes 1995; Brown et al. 2011; Holton et al. 2012; Jacques 2017). The shortcomings of this representation become immediately transparent. Neither are we told about the frequency by which a given reflex is attested to occur in the descendant languages, nor are we told about the specific phonetic conditions that have been proposed to trigger the change where we have two reflexes for the same proto-sound.

6 My translation, original text: 'Les «restitutions» ne sont rien autre chose que les signes par lesquels on exprime en abrégé les correspondances.'
7 We added phonetic transcriptions, preceding the original sound given by the author, separated by a slash.

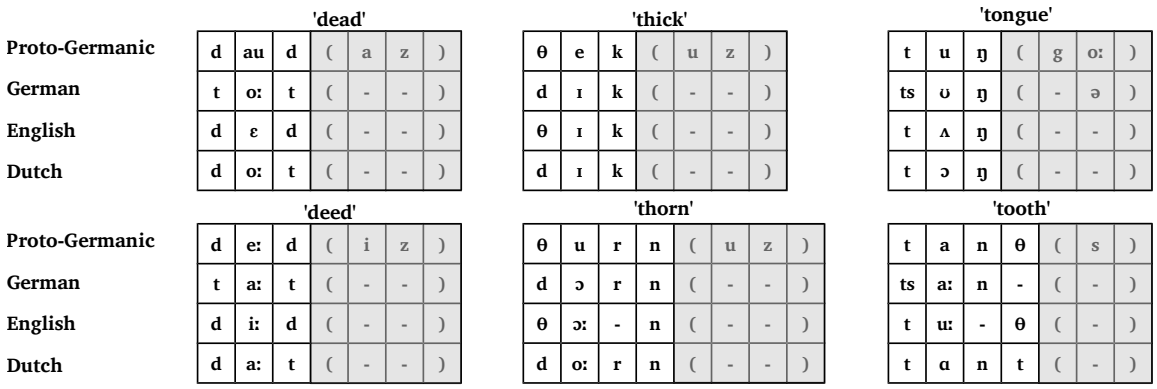


Figure 2
Alignment analyses of the six cognate sets from Table 1. Brackets around subsequences indicate that the alignments cannot be fully resolved due to secondary morphological changes.

While scholars of Indo-European tend to know these conditions by heart, it is perfectly understandable why they would not list them. However, when presenting the results to outsiders to their field in this form, it makes it quite difficult for them to correctly evaluate the findings. A sound correspondence table may look impressive, but it is of no use to people who have not studied the data themselves.

A further problem in the field of linguistic reconstruction is that scholars barely discuss workflows or procedures by which sound correspondence patterns can be *inferred*. For well-investigated language families like Indo-European or Austronesian, which have been thoroughly studied for more than one hundred years (Blust 1990), it is clear that there is no direct need to propose a heuristic procedure, given that the major patterns have been identified long ago and the research has reached a stage where scholarly discussions circle around individual etymologies or higher levels of linguistic reconstruction, like semantics, morphology, and syntax.[8] For languages whose history is less well known and where historical language reconstruction has not even reached a stage of reconstruction where a majority of scholars agree, however, a procedure that helps to identify the major correspondence patterns underlying a given data set would surely be incredibly valuable.

2.3 Correspondence Patterns and Alignments

In order to infer correspondence patterns, the data must be available in **aligned form** (see Section 1), that is, we must know which of the sound segments that we compare across cognate sets are assumed to go back to the same ancestral segment. This is illustrated in Figure 2 where the cognate sets from Table 1 are presented in *aligned form*, with zero-matches (**gaps**) being represented as a dash ("-"), and with brackets indicating **unalignable parts** in the sequences, that is, parts that cannot be aligned, since the differences are not due to regular sound change.[9] Although alignments are never explicitly mentioned in Clackson (2007), they are implied by the provided

8 For examples, compare the very detailed etymological discussions by Meier-Brügger (2002, pages 173–187).

9 Scholars at times object to this claim, but it should be evident, also from reading the account by Anttila (1972) mentioned above, that without alignment analyses, albeit implicit ones that are never provided in concrete, no correspondence patterns could be proposed.

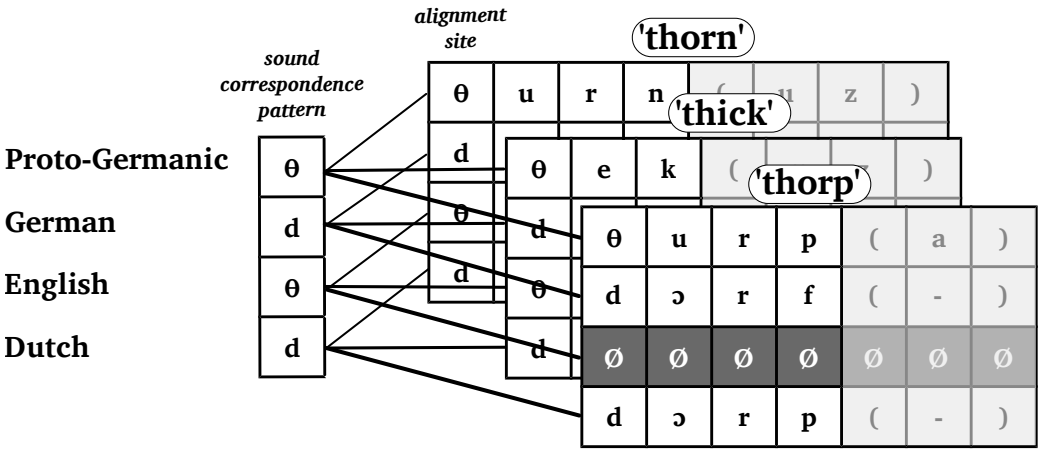


Figure 3
Alignment sites and correspondence patterns: While alignment sites are concrete representations of the presumed relations among cognate words, correspondence patterns are a further stage of abstraction.

correspondence patterns, which are presumably derived from the alignment of reflexes in each of the daughter languages. These assumed alignments are given in Table 2.

Following evolutionary biology, a given column of an alignment is called an **alignment site** (or simply a **site**). An alignment site may reflect the same values as we find in a correspondence pattern, and correspondence patterns are usually derived from alignment sites, but in contrast to a correspondence pattern, an alignment site may reflect a correspondence pattern only incompletely, due to missing data in one or more of the languages under investigation. For example, when comparing German *Dorf* [dɔrf] "village" with Dutch *dorp* [dɔrp], it is immediately clear that the initial sounds of both words represent the same correspondence pattern as we find for the cognate sets for "thick" and "thorn" given in Figure 2, although no reflex of their Proto-Germanic ancestor form *þurpa-* (originally meaning "crowd," see Kroonen [2013, 553]) has survived in Modern English.[10] Thanks to the correspondence patterns in Table 1, however, we know that—if we project the word back to Proto-Germanic—we must reconstruct the initial with **þ-* '[θ], since the match of German *d-* and Dutch *d-* occurs—if we ignore recent borrowings—only in correspondence patterns in which English has *th-*.

These "gaps" due to missing reflexes of a given cognate set are not the same as the gaps inside an alignment, since the latter are due to the (regular) loss or gain of a sound segment in a given alignment site, while gaps due to missing reflexes may either reflect processes of **lexical replacement** (List 2014, page 37f), or a preliminary stage of research resulting from insufficient data collections or insufficient search for potential reflexes. While we use the dash as a symbol for gaps in alignment sites, we will use the character ∅ (denoting the empty set) to represent missing data in correspondence patterns and alignment sites. The relation between correspondence patterns in the sense developed here and alignment sites is illustrated in Figure 3, where the initial alignment sites of three alignments corresponding to Proto-Germanic **þ* [θ] are assembled to form one correspondence pattern.

10 Old English still has the word *þorp*, but in Modern English, we only find *thorp* in names.

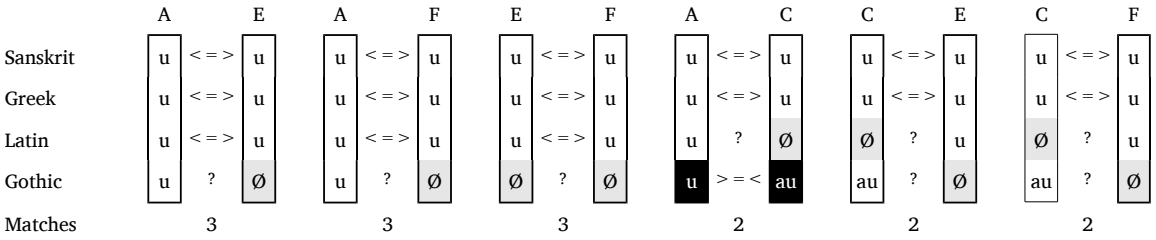


Figure 4
Assessing the compatibility of the four alignment sites from Figure 1.

3. Preliminary Thoughts on Correspondence Pattern Recognition

If we recall the problem we had in grouping the alignment sites E and F from Figure 1 with either A or C, we can see that the general problem of grouping alignment sites to correspondence patterns is their **compatibility**. If we had reflexes for all languages under investigation in all cognate sets, the compatibility would not be a problem, since we could simply group all identical sites with each other, and the task could be considered as solved. However, since it is rather the exception than the norm to have all reflexes for all cognate sets in all languages, we will always find possible alternative groupings for the alignment sites.

In the following, we will assume that two alignment sites are compatible, if they (a) share at least one sound that is not a gap symbol, and (b) do not have any conflicting sounds. This is illustrated in Figure 4 for our four alignment sites A, C, E, and F from Figure 1. As we can see from the figure, only two sites are incompatible, namely A and C, as they show different sounds for the reflexes in Gothic. Given that the reflex for Latin is missing in site C, we can further see that C shares only two sounds with E and F.

Having established the notion of **alignment site compatibility**, it is straightforward to go a step further and model alignment sites in the form of a *network*. Here, all sites in the data represent nodes (or vertices), and edges are only drawn between those nodes that are *compatible*, following the criterion of compatibility outlined in the previous section.[11]

Having shown how the data can be modeled in the form of a network, we can rephrase the task of identifying correspondence patterns as a **network partitioning task** with the goal of splitting the network into non-overlapping sets of nodes. Given that our main criterion for a valid correspondence pattern is full compatibility among all alignment sites of a given partition, we can further specify the task as a **clique partitioning task**. A **clique** in a network is "a maximal subset of the vertices [nodes] in an undirected network such that every member of the set is connected by an edge to every other" (Newman 2010, page 193). Demanding that sound correspondence patterns should form a clique of compatible nodes in the network of alignment sites directly reflects the basic practice of historical language comparison as outlined in Anttila (1972). Any further grouping would require us to identify complementary phonetic environments for the incompatible alignment sites.

11 We can further weight the edges in the alignment site network, for example, by using the number of matching sounds (where no missing data is encountered) to represent the strength of the connection (but we will disregard weighting in the approach presented here).

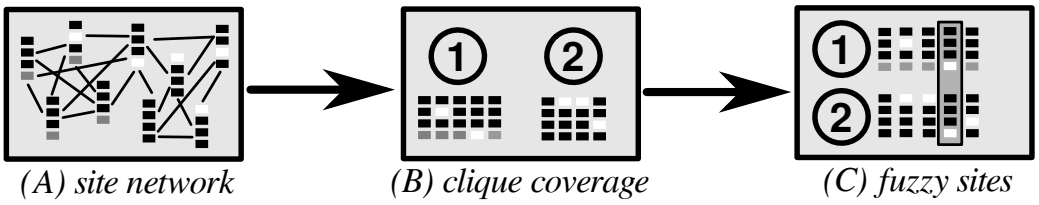


Figure 5
General workflow of the method for automatic correspondence pattern recognition.

Parsimony dictates that—when partitioning our alignment site graph—we should try to minimize the number of cliques to which the different nodes are assigned. This is the **minimum clique cover problem** (Bhasker and Samad 1991, page 2). The minimum clique cover problem is a well-known problem in graph theory and computer science, although it is usually more prominently discussed in the form of its inverse problem,[12] the **graph coloring problem**. In the graph coloring problem, one tries to assign all those nodes in a graph to different clusters (i.e., to "color" them in different colors) which are directly connected (Hetland 2010, page 276). While the problem is generally known to be *NP-hard* (Hetland 2010, page 276), fast approximate solutions like the Welsh-Powell algorithm (Welsh and Powell 1967) are available. Using approximate solutions seems to be appropriate for the task of correspondence pattern recognition, given that we do not (yet) have formal linguistic criteria to favor one clique cover over another.[13]

4. An Automatic Method for Correspondence Pattern Recognition

The method for automatic correspondence pattern recognition requires that the data be coded for cognacy, and that all cognate sets be phonetically aligned. Thanks to recently proposed algorithms, these tasks can be carried out automatically,[14] but to guarantee reliable results, it is useful to provide manually annotated data, or to manually correct data that was automatically analyzed in a first step.[15]

The general workflow of the method consists of three basic steps (see Figure 5). In a first step, the alignments in the data are used to construct an *alignment site network* in which edges are drawn between compatible sites (A). The alignment sites are then partitioned into distinct non-overlapping subsets using an approximate algorithm for the minimum clique cover problem (B). In the final step (C), alternate correspondence sets are considered for each individual alignment site. Any existing partitions with which the site is compatible are added as potential correspondents. In the following sections, I will provide more detailed explanations on the different stages.

12 The inverse problem of a given problem in graph theory provides a solution to the original problem for a graph in which the original edges are deleted and nodes formerly unconnected are connected.

13 We should furthermore bear in mind that an optimal resolution of sound correspondence patterns for linguistic purposes would additionally allow for uncertainty when it comes to assigning a given alignment site to a given sound correspondence pattern. If we decided, for example, that the pattern C in Figure 1 could by no means cluster with E and F, this may well be premature before we have figured out whether the two patterns (**u-u-u-u** vs. **u-u-u-au**) are *complementary* and what phonetic environments explain their complementarity.

14 For automatic cognate detection, compare for example List (2014), List, Greenhill, and Gray (2017), Arnaud, Beck, and Kondrak (2017), and Jäger, List, and Sofroniev (2017), and for automatic phonetic alignment, compare Prokić, Wieling, and Nerbonne (2009) and List (2014).

15 For manual annotation of cognates and alignments, compare List (2017).

Table 3
Input format with the basic values needed to apply the method for automatic correspondence pattern recognition.

ID	DOCULECT	CONCEPT	FORM	TOKENS	STRUCTURE	COGID	ALIGNMENT
1	German	tongue	Zunge	ts ʊ ŋ ə	**c v c** v	1	ts ʊ ŋ (ə)
2	English	tongue	tongue	t ʌ ŋ	c v c	1	t ʌ ŋ (-)
3	Dutch	tongue	tong	t ɔ ŋ	c v c	1	t ɔ ŋ (-)
4	German	tooth	Zahn	ts aː n	c v c	2	ts aː n -
5	English	tooth	tooth	t uː θ	c v c	2	t uː - θ
6	Dutch	tooth	tand	t ɑ n t	**c v c** c	2	t ɑ n t
7	German	thick	dick	d ɪ k	c v c	3	d ɪ k
...	...	...	...	...	...	...	...

4.1 Implementation, Input Format, and Output Format

The method has been implemented as a Python package that can be used as a plug-in for the LingPy library for quantitative tasks in historical linguistics (List, Greenhill, and Forkel 2017). The supplementary material offers precise instructions on how the software package can be installed and how the experiments can be replicated.

The input format for the method described here generally follows the input format employed by LingPy. In general, this format is a tab-separated text file with the first row being reserved for the header, and the first column reserved for a unique numerical identifier. The header specifies the entry types in the data. Table 3 provides an example of the minimal data that needs to be provided to our method for automatic correspondence pattern recognition. In addition to the generally needed information on the identifier of each word (ID), on the language (DOCULECT), the concept or elicitation gloss (CONCEPT), the (not necessarily required) orthographic form (FORM), and the phonetic transcription provided in space-segmented form (TOKENS), the method requires information on the type of sound (consonant or vowel, STRUCTURE),[16] the cognate set (COGID), and the alignment (ALIGNMENT).

The method offers different output formats, ranging from the LingPy wordlist format in which additional columns added to the original wordlist provide information on the inferred patterns, or in the form of tab-separated text files, in which the patterns are explicitly listed. The wordlist output can also be directly inspected in the EDICTOR tool, allowing for a convenient manual inspection of the inferred patterns.

16 The values passed to the STRUCTURE column can be arbitrarily filled. When running the analysis, they are used to identify those positions in the alignments that should be analyzed separately, that is, they will be considered as a useful pre-partitioning of the alignment sites.

4.2 Detailed Description of the Algorithm

As mentioned above, the method for correspondence pattern recognition consists of three stages. It starts with the reconstruction of an alignment site network in which each node represents a unique alignment site, and links between alignment sites are drawn if the sites are compatible, following the criterion for site compatibility outlined in Section 3 (A). It then uses a greedy algorithm to compute an approximate minimal clique cover of the network (B). All partitions proposed in stage (B) qualify as potentially valid correspondence patterns of our data. But the individual alignment sites in a given data set may as well be compatible with more than one correspondence pattern.[17] For this reason, the method iterates again over all alignment sites in the data, checking whether each is compatible with any other existing partition. This procedure assigns each alignment site to at least one but potentially more different sound correspondence patterns (C).[18]

The clique cover algorithm (A) is an inverse version of the Welsh-Powell algorithm for graph coloring (Welsh and Powell 1967). It starts with k cliques of size 1, which are sorted in increasing order by the amount of missing data they contain. The algorithm then picks the first pattern and compares it with the set of all other patterns. If this first pattern is compatible with one of the other patterns, the two patterns will be merged into a new pattern that is then further compared with the remaining ones. After the iteration, the first pattern is added to the set of results, and the same procedure is repeated with the remaining patterns that have not yet been merged and remain in the queue until no patterns are left.

Since alignment sites may suffer from missing data, their assignment to particular correspondence patterns is not always unambiguous. The example alignment from Figure 1, for example, would yield two general correspondence patterns, namely **u-u-u-au** versus **u-u-u-u**. While the assignment of alignment sites A and C in the figure would be unambiguous, sites E and F could be assigned to either partition, since they are missing the disambiguating data. In order to reflect the fuzziness of the partition assignment, the method therefore requires an additional step. In addition to the partition from stage (B), alternative partitions are found for E and F during stage (C). The patterns, to which a given alignment site is assigned, can further be ranked by counting the total amount of alignment sites with which they are compatible, thus allowing us to prefer only those site-to-pattern assignments that have a reasonable number of examples.

Figure 6 gives an artificial example that illustrates how the basic method infers the clique cover. Starting from the data in (A), the method assembles patterns A and B in (B) and computes their pattern, thereby retaining the non-missing data for each language in the pattern as the representative value. Having added C and D in this fashion in steps (C) and (D), the remaining three alignment sites, E–G, are merged to form a new partition, accordingly, in steps (E) and (F). Step (G) reflects the reassignment of individual alignment sites to the previously inferred patterns. In this example, all sites are only assigned to one pattern, but it is possible, depending on the amount of missing data, that one site can be assigned to more than one pattern.

17 Compare, for example, site E in Figure 1, which is both compatible with the pattern *u-u-u-u* reflected by the site A, and the pattern *u-u-u-au*, reflected by site B.

18 By further weighting and sorting the fuzzy patterns to which a given site has been assigned, the number of fuzzy alignment sites can be further reduced.

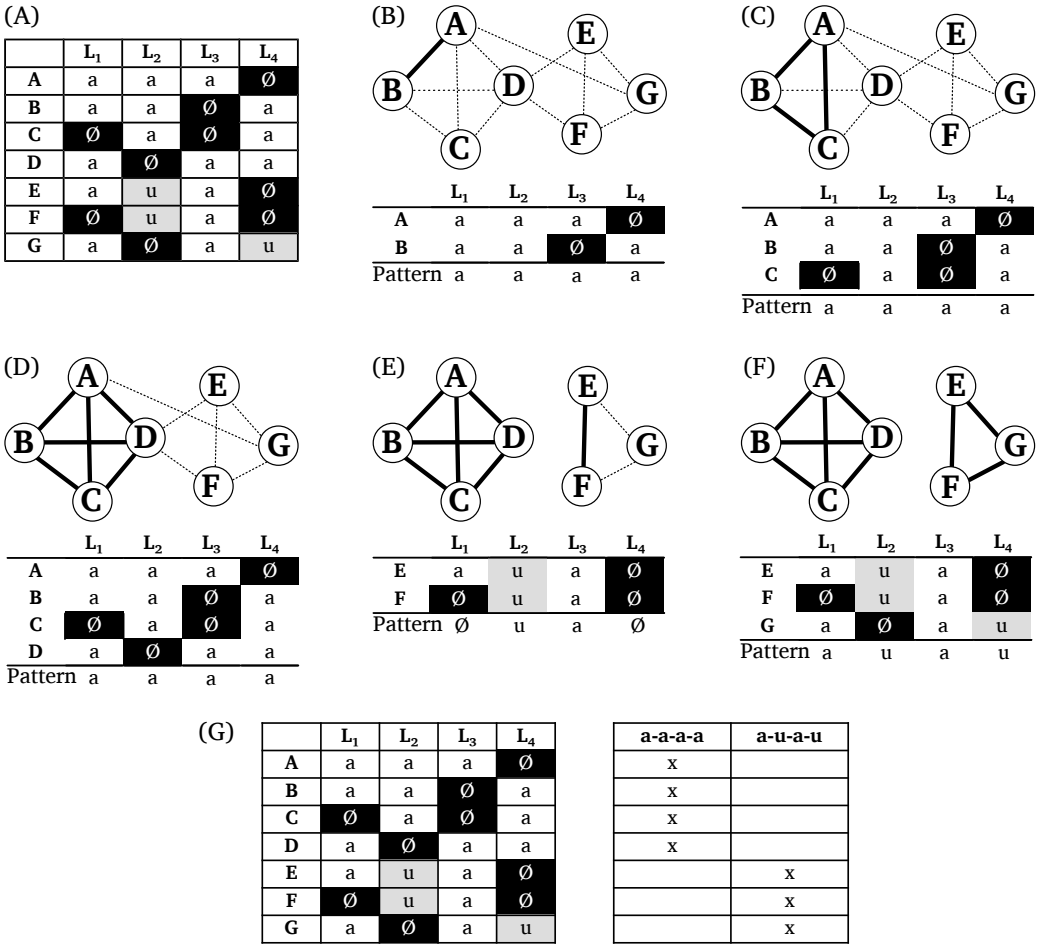


Figure 6
Example for the basic method to compute the clique cover of the data. (A) shows all alignment sites in the data. (B–D) show how the algorithm selects potential edges step by step in order to arrive at a first larger clique cover. (E–F) show how the second cover is inferred. In each step during which one new alignment site is added to a given pattern, the pattern is updated, filling empty spots. While there are two missing data points in (E), where only alignment sites E and F are merged, these are filled after adding G. (G) shows how patterns are reassigned to individual alignment sites.

It is important to note that the originally selected pattern may change during the merge procedure, since missing spots can be filled by merging the pattern with a new alignment site (as also shown in Figure 6). For this reason, it is possible that this procedure, when only carried out one time, may not result in a true clique cover (in which all compatible alignment sites are merged). For this reason, at the end of the iteration, the algorithm checks if patterns exist that could be further combined, and repeats the procedure with the existing patterns until the resulting partitioning represents a true clique cover.

Pseudocode is in Algorithm 1 for the core function of the method for correspondence pattern detection. In a worst-case scenario in which all alignment sites will be assigned to distinct correspondence patterns, the algorithm requires $\sum_{k=1}^{n} = \frac{n(n+1)}{2}$ iterations in the while-loop, where k represents the number of alignment sites in the data, so the general complexity of the algorithm is $\mathcal{O}(n^2)$. In applications to real-world-data,

Algorithm 1 Main part of the correspondence pattern detection method in pseudocode.

```
1:  function CORRPATTERNS(almSites)
2:      sort(almSites, sortKey:=countMissing);                ▷ sort the sites
3:      patterns := [];                                       ▷ stores the patterns
4:      while length(almSites)≠0 do
5:          first, rest := almSites[0], almSites[1:];         ▷ compare first site against rest
6:          almSites := [];                                   ▷ fill with unmerged sites during for-loop
7:          for i := 0; i < length(rest); i++ do
8:              if compatible(rest[i], first) then
9:                  first :=merge(first, rest[i]);
10:             else
11:                 append(almSites, rest[i]);
12:             end if
13:         end for
14:         append(patterns, first);
15:     end while
        return patterns;
16: end function
```

however, this worst-case scenario is never reached, and the method converges rather fast.[19]

5. Testing the Method for Correspondence Pattern Recognition

The quantitative treatment of sound correspondence patterns presented in this study is novel. As a result, no expert-annotated data listing all observable correspondence patterns for a certain language family exhaustively is available.[20] and it is not possible to compare the suitability of this novel approach with expert-annotated gold standards, as it is usually done in similar studies in computational historical linguistics.

The lack of suitable gold-standard data, however, does not mean that we cannot test the method for its suitability. Since the core service the method provides is to impute missing values in alignment sites resulting from cognate sets that are not reflected in all languages in a given data set, we can easily design tests in which we test the power of the method to *predict* those missing values in controlled settings.

5.1 Data for Testing

Three different data sets were selected to test the method proposed in this study. The data sets were chosen with great care, since only a few of the many data sets offering manually coded cognate sets also provide the cognate sets in aligned form. Apart from the data by Hill and List (2017) on Burmish languages (original data based on Huáng 1992), Walworth (2018) on East Polynesian languages (original data based

19 This can easily be seen when assuming clique cover that includes all nodes in a given network: Here, the algorithm would need only one iteration, as it would consecutively merge each next node visited in the first iteration into the same partition.

20 As we have seen in Figure 2, scholars list major sound correspondences across multiple languages, but they do not show individual patterns for aligned cognate sets.

Table 4
Three test sets used in this study.

Data set	Source	Languages	Concepts	Words	Cognates
Burmish	Hill and List (2017)	8	240	1,819	798
Chinese	Hóu (2004)	14	623	8,371	623
Polynesian	Walworth (2018)	10	210	2,427	1,187
Japanese	Hattori (1973)	10	200	1,986	454

on Greenhill, Blust, and Gray 2008), and Hattori (1973) on Japanese dialect (data in electronic form supplemented in List 2014), an additional data set of 14 Chinese varieties originally published by Hóu (2004) was specifically modified and manually aligned for this study. While the two former data sets are classical wordlists that are further coded for cognacy and alignments,[21] the Chinese data is based on a collection of 623 morphemes (reflected by a Chinese character each) whose pronunciation across the 14 dialects used in our sample was elicited by field workers. As a result, the amount of cognate sets with missing reflexes in this data set is extremely low.

An overview of the data sets, along with additional information regarding the data sources, the number of cognate sets, language varieties, and words in the data, is given in Table 4. Needless to say that all data sets are provided in the supplementary material accompanying this article.

5.2 General Characteristics

As a first illustrative test, the method was applied to the four data sets, and some basic statistics were calculated. These include the original number of alignment sites in the data, the number of patterns into which these sites were partitioned by the method, and the number of singleton patterns, that is, patterns that are reflected by only one alignment site in the data. By dividing the number of alignment sites assigned to non-unique patterns by the number of all sites, we can further determine the proportion of "regular" correspondence patterns in a given data set, assuming that a pattern is regular if it recurs in at least two different alignment sites.

The results of this analysis are summarized in Table 5. As we can see from this table, the number of correspondence patterns inferred by the algorithm is much lower than the number of alignment sites. This is, of course, not surprising, if we assume that the hypothesis that sound change is an overwhelmingly regular process holds. However, across the data sets, we can find rather large differences with respect to the amount of singleton patterns, that is, patterns reflecting only one alignment site. That an alignment site is not compatible with any other site in the data can have different reasons. First, there can be idiosyncratic sound changes, resulting, for example, from taboo, or from the assimilation of frequently used words. Second, there can be errors in the data, resulting from incorrectly assigned cognates, alignments, or undetected borrowings. It is also possible that the data sample is too small, and that additional samples could be found, but have not been included in the data.

21 All data sets are coded for partial cognates and across semantic categories.

Table 5
Basic statistics after applying the correspondence pattern recognition method to the four data sets.

Data set	Align. Sites	Corr. Patterns	Singletons	Reg. Patterns
Burmish	1,833	432	173	0.91
Chinese	2,891	1,341	966	0.67
Polynesian	1,863	243	64	0.97
Japanese	1,590	556	311	0.80

Table 6
Examples for idiosyncratic correspondence patterns in the Chinese dialects reflecting the major groups (Běijīng, Sūzhōu, Chángshā, Nánchāng, Méixiàn, Táoyuán, Guǎngzhōu, Fúzhōu, and Táiběi [Mǐn dialect spoken in Táiběi]).

#	Cognates	BJ	SZ	CS	NC	MX	TY	GZ	FZ	TB
73	5	t	d	t	t^h	t^h	t^h	t	t	t
484	1	t	d	t	l	t^h	t^h	t	t	t
654	1	t	d	t	t^h	t	t^h	t	t	t
239	6	n	n̥	n̥	n̥	ŋ	ŋ	n	n	l
679	1	n	n̥	n̥	n̥	t	ŋ	n	t͡s	t͡s
699	1	n	n̥	n̥	n̥	ŋ	ŋ	ŋ	g	g

When comparing the proportion of "regular" patterns that are reflected by at least two alignment sites in the data across the data sets, the Chinese data shows the lowest proportion, with only 67% of all alignment sites being assigned to patterns that recur in the data. Given the intertwined history of the Chinese dialects, in which language contact among the dialect varieties played an important role, it is not necessarily surprising that the data looks less regular in general: If languages borrow from each other, and borrowing is sporadic, rather than systematic, this will lead to an increase in irregular correspondence patterns and therefore impact on the regularity we can observe. As a manual inspection of the inferences reveals, the majority of the singleton alignment sites in the Chinese data could be assigned to one of the regular patterns if one of the reflexes would be ignored. Examples for these patterns are given in Table 6. On the other hand, we find some patterns that are largely irregular for specific reasons like taboo. An example is given in the same table with Chinese *niǎo* "bird" (pattern 679), which is reflected by nasal and dental initials across the Chinese dialects. As we know from older readings, the original reading had the initial **[t]**, but it was later replaced by a nasal in some Chinese varieties to avoid homophony with the word for "penis," which was most likely metaphorically shifted from "bird."

What we can see from the individual analyses of the different data sets is that the overall regularity of correspondence patterns does not necessarily reflect the time depth of the languages in a given data set. Instead, correspondence patterns reflect different aspects of the data, which have so far not been thoroughly investigated by researchers. The overwhelming regularity of the patterns in the Polynesian data set, for example, is probably also due to the fact that the languages contain very small phoneme inventories,

with no more than 17 different sounds on average (compared to Chinese dialects with about 35 sounds), while many idiosyncratic patterns in the Japanese data result from morphological differences, which are difficult to handle in phonetic alignments.

5.3 Tests on Word Prediction

As mentioned briefly already in Section 1, correspondence patterns—once readily inferred—can provide hints regarding the potential pronunciation of missing cognates in an alignment. Since the method for correspondence pattern recognition imputes missing data in its core, it can also be used to predict how a given word should look in a given language if the reflex of the corresponding cognate set is missing. An example for the prediction of forms has been given above for the cognate set Dutch *dorp* and German *Dorf*. Since we know from Table 1 that the correspondence pattern of *d* in Dutch and German usually points to Proto-Germanic **þ*, we can propose that the English reflex (which is missing in Modern English, apart from place names) would start with *th*, if it was still preserved.[22] Since the method for correspondence pattern recognition assigns one or more correspondence patterns to each alignment site, even if the site has missing data for a certain number of languages, all that needs to be done in order to predict a missing entry is to look up the alignment pattern and check the value that is proposed for the given language variety.

The test on word prediction was designed as follows: from each of the data sets, a certain number of cognate sets was randomly deleted, and the resulting data was then analyzed with the help of the correspondence pattern recognition algorithm. In a second step, these inferred patterns were used to predict the cognate words which were deleted before. For the prediction, only the largest correspondence pattern was considered for the imputation, in order to avoid that multiple proposals for one sound could be made by the algorithm. For each data set, three different proportions of words to be deleted were tested (25%, 50%, and 75%).[23] For each proportion and data set, 1,000 trials were tested and the results were averaged. To assess the accuracy of a predicted word, the proportion of correctly predicted sounds in the given word was estimated and divided by the total length of the word. The individual accuracies of predicted words were then averaged by dividing the number of individual prediction scores by the number of predicted words for each trial.

The results of this experiment are given in Table 7. In general, we can note that the prediction experiment works very well across all data sets for wordlists reduced by 25% and 50% of their words appearing in cognate sets, while the accuracy of prediction drastically drops in all data sets when removing up to 75% of the data. The only exception is the Polynesian data set, where the difference in accuracy across the three experiments is only small, with a rather large standard deviation.

What may come as a surprise is that the reduction of the data by 25% and 50% does not seem to influence the accuracy of prediction in all data sets. On the contrary, in the Chinese and the Polynesian data sets, we find even slightly higher accuracy scores for the larger data reduction. At least in the Chinese data, the reason for this can be found in the large number of singleton patterns that deviate only in one reflex from regularly

22 We ignore deliberately in this context that the alternative of the correspondence in Dutch and German is a borrowing from Dutch, Frisian, or English to German.

23 Depending on the specific distribution of cognates in the individual data, these proportions could vary in each run.

Table 7
Results of the test on word prediction, based on 1,000 random samples for each subset of the data. The column Proportion reflects the different proportions of the data that was deleted during the experiments. Patterns refers to the average number of correspondence patterns inferred in each trial, and Reg. Patterns points to the proportion of alignment sites covered by patterns recurring at least twice.

Data set	Proportion	Patterns	Reg. Patterns	Accuracy
Burmish	25%	231.68 ± 6.86	0.94 ± 0.01	0.59 ± 0.02
	50%	165.55 ± 6.08	0.94 ± 0.01	0.53 ± 0.02
	75%	99.33 ± 5.71	0.89 ± 0.02	0.37 ± 0.03
Chinese	25%	1,040.62 ± 11.88	0.81 ± 0.01	0.69 ± 0.01
	50%	672.35 ± 10.53	0.95 ± 0.00	0.70 ± 0.01
	75%	373.23 ± 7.67	0.97 ± 0.00	0.64 ± 0.01
Japanese	25%	399.82 ± 10.04	0.89 ± 0.01	0.64 ± 0.01
	50%	259.71 ± 9.39	0.93 ± 0.01	0.62 ± 0.01
	75%	142.65 ± 7.35	0.92 ± 0.01	0.52 ± 0.02
Polynesian	25%	127.30 ± 5.38	0.97 ± 0.00	0.81 ± 0.01
	50%	89.10 ± 5.51	0.97 ± 0.01	0.82 ± 0.01
	75%	51.37 ± 4.69	0.95 ± 0.01	0.80 ± 0.03

recurring patterns. If the data is reduced by 50%, the number of idiosyncratic patterns also drops, as we can see from the proportion of regularly recurring patterns given in the table. While these cover 95% of all alignment sites in the data set reduced by 50%, their proportion drops to 81% when being reduced by only 25%, and is (as we have seen in Table 5) even lower when analyzing the whole data set. If enough words are deleted from singleton patterns, like the ones shown in the examples in Table 6, the method for correspondence pattern recognition will assign them to the same clusters. As a result, the words whose pronunciation deviates will still be wrongly predicted, but the words that are not affected by individual sound changes will be predicted correctly, and since there are more regular words in the data, the overall prediction accuracy will increase.

When comparing the differences in the scores across the four data sets, we can also see that the overall "regularity" of the data, as measured by the number of patterns that recur more than one time, is not a good predictor of the success of the prediction quality. The Burmish data, for example, has rather high rates of pattern regularity, but performs worse in prediction than the other data sets. It is clear that the number of singleton patterns that only reflect one alignment site in a data set will have a direct impact on the word prediction quality, since only patterns that recur at least two times in the data can be used for prediction. But this is not the only factor influencing the prediction quality. Ambiguous alignment sites that can be assigned to more than one pattern may, for example, likewise produce erroneous predictions. For the time being, we cannot offer a full account of all the different factors that might influence prediction quality. More studies on different data sets will be needed to increase our knowledge in the future.

The fact that the prediction accuracy does not seem to improve or may even drop when more data is retained in our experiments is important for further applications of

the method (for example, when carrying out field work or when searching for missing cognate sets), as it shows that we can reduce the amount of time spent on manual annotation substantially when annotating data sets for historical linguistics. Linguists could, for example, annotate half of their data manually and then use our method to impute potentially missing cognates in their data. If actual words that were not annotated in the first run turn out to have the same form as words predicted by our algorithm, this would be a very strong argument that they are really cognate. Another example would be guided field work for the purpose of historical language comparison. If insufficient amounts of data have been collected, scholars can use the prediction method to predict the most likely forms for certain cognate sets and use them to ease the elicitation of the relevant forms when asking new informants.

5.4 Examples

Table 8 gives some examples illustrating the scoring procedure and typical failures of the method, again illustrated for the Chinese data set.[24] In cognate set 687, we find one correctly predicted form for Chángshā, and one incorrectly predicted tone for the Jínán form. As we can see from the frequencies of alignment sites supporting the proposed pattern given in the column "Frequency" in the table, the inferred pattern clusters only two alignment sites. As a result, it is not surprising that a wrong tone is proposed. The wrong form for Měixiān in cognate set 319 is due to a wrong clustering of the cognate set with the irregular cognate set 654, listed earlier in Table 6. Since the Měixiān word was deleted in the experiment, the whole pattern is compatible with pattern # 73 in the table, which predicts that the Měixiān form should start with **t^{h}**. In cognate set 518, we can see that the method fails to propose a valid sound for the form in Wēnzhōu for the second and the third site in the alignment, given that these sites are assigned to singleton patterns (of one alignment site only) in which no sound for Wēnzhōu could be imputed.

While the success or failure of the prediction experiments can help us to improve the method in the future, we can also illustrate how the analysis can aid in practical work on linguistic reconstruction. This example will again be based on the Chinese data, since it has the advantage of offering quick access to Middle-Chinese reconstructions. Because Middle Chinese is only partially reconstructed on the basis of historical language comparison, and mostly based on written sources, such as ancient rhyme books and rhyme tables (Baxter 1992), the reconstructions are not entirely dependent on the modern dialect readings, which is a great advantage for testing the consequences of the correspondence pattern analysis.

In Table 9, patterns inferred by the method for correspondence pattern recognition for a reduced number of dialects (one of each major subgroup) have been listed. The examples can all be reconstructed to a dental stop in Middle Chinese (*__t__, *__t^{h}__, or *__d__). If we inspect only reflexes of Middle Chinese *__d__ in the data, we can see that the initial consonant is reflected in seven different patterns in our data. Four of these patterns, however, occur only one time (# 719, # 1096, # 484, and # 654), as reflected in the column Cogn. (pointing to supporting *cognate sets*), and if we exclude the reflexes for Méixiàn

24 This data set and the detailed predictions are available from the supplementary material as files `<chinese25.tsv>` (wordlist) and `predictions-chinese25.txt`.

Table 8
Examples for the word prediction experiment for the Chinese data. The column Frequency lists the size of the inferred patterns for each position of the predicted word form. The score is calculated by dividing the number of correctly predicted sounds by the total number of sounds.

Cogn. Set	Dialect	Frequency	Predicted	Attested	Score
687	Chángshā	3 4 2 2	ʂ ə n ^{13}	ʂ ə n ^{13}	1.00
	Jínán		t͡ʂʰ ə ŋ ^{55}	t͡ʂʰ ə ŋ ^{42}	0.75
319	Běijīng	4 2 2 45	t a i ^{51}	t a i ^{51}	1.00
	Táiběi		t a i ^{33}	t a i ^{33}	1.00
	Měixiān		t o - ^{53}	tʰ o - ^{53}	0.75
	Wēnzhōu		d e i ^{22}	d e - ^{22}	0.75
518	Běijīng	6 3 3 24 45	t͡ɕ i a n ^{51}	t͡ɕ i a n ^{51}	1.00
	Wēnzhōu		dʑ Ø Ø - ^{22}	dʑ - i - ^{35}	0.40

(# 719, # 654), Táiběi (# 1096), and Nánchāng (# 484), respectively, we can assign # 719 and # 1096 to # 718 and # 484 and # 654 to # 73. In patterns # 718 and # 747, only Fúzhōu shows a different reflex. Since we have forms that are homophones in Middle Chinese in both correspondence patterns (糖 in # 747 and 堂 in # 718 were both pronounced as ***dam** in Middle Chinese), we cannot find a conditioning context that would explain this difference from the perspective of Middle Chinese alone. We know, however, that the Mǐn dialects (to which Fúzhōu belongs) reflect features that are more archaic than Middle Chinese. In this case, the difference between the patterns is regularly reflecting the difference between plain voiced and breathy voiced initials in the ancestor of the Mǐn dialects, with the latter going back to complex onsets in Old Chinese, the predecessor of all Chinese dialects (Baxter and Sagart 2014, page 171f). Furthermore, if we compare the patterns # 747 and # 73 directly, we can see that, although only Fúzhōu has a direct reflex of the original voiced sound in Middle Chinese, we can still find its traces in the different correspondence patterns, since Běijīng and Guǎngzhōu have contrastive outcomes in both patterns (**[tʰ]** versus **[t]**). When inspecting the tones that are reconstructed for the different words in Middle Chinese, we can easily find a conditioning context where the reflexes differ. The *píng* (flat) tone category in Middle Chinese correlates with aspiration, while the other tone categories correlate with devoicing in the three dialects.[25] If we had no knowledge of Middle Chinese, it would be harder to understand that both patterns correspond to the same proto-sound, but once assembled in such a way, it would still be much easier for scholars to search for a conditioning context that allows them to assign the same proto-sound to the two patterns in questions.

The example shows that, as far as the Middle Chinese dental stops are concerned, we do not find explicit exceptions in our data, but can rather see that multiple correspondence patterns for the same proto-sound may easily evolve. We can also see that a careful alignment and cognate annotation is crucial for the success of the method, but

25 This phenomenon most likely goes back to an earlier phonation contrast between the first (*píng*) tone in Middle Chinese and the other tones.

Table 9
Contrasting inferred correspondence patterns with Middle Chinese reconstructions (MC) and tone patterns (MC Tones: P: píng (flat), S: shǎng (rising), Q: qù (falling), R: rù (stop coda)) for representative dialects of the major groups (Běijīng, Sūzhōu, Chángshā, Nánchāng, Méixiàn, Táoyuán, Guǎngzhōu, Fúzhōu, Táiběi).

#	Cogn.	MC	MC Tones	BJ	SZ	CS	NC	MX	TY	GZ	FZ	TB
30	13	*t	PSQR	t	t	t	t	t	t	t	t	t
41	9	*th	PSQR	tʰ	tʰ	tʰ	tʰ	tʰ	tʰ	tʰ	tʰ	tʰ
747	3	*d	P	tʰ	d	t	tʰ	tʰ	tʰ	tʰ	tʰ	tʰ
718	2	*d	P	tʰ	d	t	tʰ	tʰ	tʰ	tʰ	t	tʰ
719	1	*d	P	tʰ	d	t	tʰ	t	tʰ	tʰ	t	tʰ
1,096	1	*d	P	tʰ	Ø	t	tʰ	tʰ	tʰ	tʰ	t	t
73	5	*d	QR	t	d	t	tʰ	tʰ	tʰ	t	t	tʰ
484	1	*d	R	t	d	t	l	tʰ	tʰ	t	t	tʰ
654	1	*d	S	t	d	t	tʰ	t	tʰ	t	t	tʰ

even if the cognate judgments are fine, but the data are sparse, the method may propose erroneous groupings.

In contrast to manual work on linguistic reconstruction, where correspondence patterns are never regarded in the detail in which they are presented here, the method has the potential to drastically increase both the transparency and the quality of linguistic data sets, especially in combination with tools for cognate annotation, like EDICTOR, to which we added a convenient way to inspect inferred correspondence patterns interactively (see the example in Appendix A). Because linguists can run the new method on their data and then directly inspect the consequences by browsing all correspondence patterns conveniently in the EDICTOR, the method makes it a lot easier for linguists to come up with first reconstructions or to identify problems in the data.

6. Conclusion and Outlook

This study has presented a new method for the inference of sound correspondence patterns in multilingual wordlists. Thanks to its integration with popular software packages, the method can be easily applied, both within automated, or computer-assisted workflows. The usefulness of the method was illustrated by showing how it can be used to predict missing words in linguistic data sets. The method, however, has much additional potential. Since the method can impute words not attested in existing languages, it could likewise be used for the automatic reconstruction of proto-forms, the identification of cognates, or the assessment of the general regularity of a given data set. In addition to revealing potential correspondence structures underlying a given data set, the method can additionally help to assess how well a given data set has been analyzed before. By helping to improve the quality and transparency of existing and future data sets in historical linguistics in this way, we hope that the method will in the long run also contribute to new and important findings about the past of our world's languages.

Supplementary Material

The supplementary material accompanying this article contains the code and all instructions needed to repeat the experiments described in this article. The original package for correspondence pattern detection is publicly available from GitHub under `https://github.com/lingpy/lingrex` (Version 0.1.0). The package providing the supplementary material with results and instructions for running the code is also available via GitHub under `https://github.com/lingpy/correspondence-pattern-paper` (Version 1.1.1) and has been archived with Zenodo at `https://doi.org/10.5281/zenodo.1544949`.

Appendix A: Inspecting Correspondence Patterns in EDICTOR

The following screenshots show how the modified version of the EDICTOR allows for an enhanced inspection of sound correspondence patterns inferred by the method.

Investigate correspondence patterns in the data

Select Sets | THR. 4 | PREV. 54 | OK | ← | 1-25 of 25 Sites | →

COGNATES	INDEX	PATTERN	CONCEPTS	Bei	Suz	Cha	Nan	Mei	Gua	Fuz	SIZE
646	1	tʰ / 85	the body hair (hair or fur)	tʰ	d	t	tʰ	Ø	tʰ	tʰ	5.14 / 7
649	1	tʰ / 85	the hair (of the head)	tʰ	d	t	tʰ	Ø	tʰ	tʰ	5.14 / 7
740	1	tʰ / 85	the big frog	tʰ	d	t	tʰ	tʰ	tʰ	Ø	5.14 / 7
189	5	tʰ / 85	the fish (one piece of fish)	tʰ	d	t	Ø	tʰ	tʰ	Ø	5.14 / 7
607	5	tʰ / 85	the wood (material)	tʰ	d	t	tʰ	tʰ	Ø	Ø	5.14 / 7
948	5	tʰ / 85	the upper level (above)	tʰ	d	t	tʰ	Ø	Ø	Ø	5.14 / 7
965	5	tʰ / 85	the lower level (below)	tʰ	d	t	tʰ	Ø	Ø	Ø	5.14 / 7
1038	1	tʰ / 76	to hear	tʰ	tʰ	tʰ	tʰ	tʰ	tʰ	tʰ	4.00 / 5
1097	1	tʰ / 76	to pull	tʰ	tʰ	tʰ	tʰ	tʰ	tʰ	tʰ	4.00 / 5
1058	1	tʰ / 76	to lick	tʰ	tʰ	tʰ	tʰ	Ø	tʰ	Ø	4.00 / 5
1085	1	tʰ / 76	to push	tʰ	tʰ	tʰ	tʰ	Ø	tʰ	Ø	4.00 / 5
380	1	tʰ / 76	to make a journey	Ø	tʰ	tʰ	tʰ	Ø	tʰ	Ø	4.00 / 5
598	6	t / 177	the sickle	t	t	t	t	Ø	t	Ø	2.57 / 4
667	5	t / 177	the flower (one piece of flower)	t	t	t	t	Ø	t	Ø	2.57 / 4
673	4	t / 177	the ear	t	t	t	t	Ø	Ø	Ø	2.57 / 4
432	1	t / 177	all	t	Ø	Ø	t	t	Ø	t	2.57 / 4

Acknowledgments

This research was funded by the DFG research fellowship grant 261553824 "Vertical and lateral aspects of Chinese dialect history" (2015–2016), and by the ERC Starting Grant 715618 "Computer-Assisted Language Comparison" (http://calc.digling.org, 2017-2018). Originally, the approach presented here was inspired by a novel (so far still unpublished) biological technique presented to me by Eric Bapteste and Philippe Lopez, which later turned out to be completely different from the one presented here, as I misunderstood the original intention of the draft. This misunderstanding, which helped me to address a problem that had been following me for a long time, reflects how inspiring my collaboration with Eric and Philippe was. I am particularly indebted to Nathan W. Hill for supporting this project from the beginning, by discussing the findings, the methods, and their potential improvement. I am also extremely thankful to Taraka Rama for commenting on many previous versions of this draft and the code, discussing details and recommending enhancements, as well as to Simon J. Greenhill for his support after I received the first reviews, and Mary Walworth for helping with data. Timotheus Bodt also deserves special thanks for being an early tester of the methods. In addition, many people provided helpful comments on an earlier version(s) of this article, including Adam Powell, David A. S. Moslehi, Eugen Hill, Juho Pystynen, Martin Kümmel, Rémy Viredaz, Tiago Tresoldi, and Yoram Meroz, to whom I would also like to express my gratitude.

References

Anttila, Raimo. 1972. *An Introduction to Historical and Comparative Linguistics*, Macmillan, New York.

Arnaud, Adam S., David Beck, and Grzegorz Kondrak. 2017. Identifying cognate sets across dictionaries of related languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2509–2518, Association for Computational Linguistics.

Baxter, William H. 1992. *A Handbook of Old Chinese Phonology*. de Gruyter, Berlin.

Baxter, William H., and Laurent Sagart. 2014. *Old Chinese. A New Reconstruction*. Oxford University Press, Oxford.

Beekes, Robert S. P. 1995. *Comparative Indo-European Linguistics. An Introduction*. John Benjamins, Amsterdam and Philadelphia.

Bhasker, J., and Tariq Samad. 1991. The clique-partitioning problem. *Computers & Mathematics with Applications*, 22(6):1–11.

Blust, Robert. 1990. Patterns of sound change in the Austronesian languages. In Philip Baldi, editor, *Linguistic Change and Reconstruction Methodology*. Mouton de Gruyter, Berlin; New York, pages 231–270.

Brown, Cecil H., David Beck, Grzegorz Kondrak, James K. Watters, and Søren Wichmann. 2011. Totozoquean. *International Journal of American Linguistics*, 77(3):323–372.

Brown, Cecil H., Eric W. Holman, and Søren Wichmann. 2013. Sound correspondences in the world's languages. *Language*, 89(1):4–29.

Campbell, Lyle, and William John Poser. 2008. *Language Classification: History and Method*. Cambridge University Press, Cambridge.

Clackson, James. 2007. *Indo-European Linguistics*. Cambridge University Press, Cambridge.

Covington, Michael A. 1996. An algorithm to align words for historical comparison. *Computational Linguistics*, 22(4):481–496.

Dixon, R. B., and A. L. Kroeber. 1919. *Linguistic Families of California*. University of California Press, Berkeley.

Dybo, Anna, and George S. Starostin. 2008. In defense of the comparative method, or the end of the Vovin controversy. In I. S. Smirnov, editor, *Aspekty komparativistiki*, Volume 3. RGGU, Moscow, pages 119–258.

Fox, Anthony. 1995. *Linguistic Reconstruction*. Oxford University Press, Oxford.

Greenhill, Simon J., Robert Blust, and Russell D. Gray. 2008. The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics. *Evolutionary Bioinformatics*, 4271–283.

Grimes, Joseph E., and Frederick B. Agard. 1959. Linguistic divergence in romance. *Language*, 35(4):598–604.

Guy, Jacques B. M. 1994. An algorithm for identifying cognates in bilingual wordlists and its applicability to machine translation. *Journal of Quantitative Linguistics*, 1(1):35–42.

Hattori, Shirō. 1973. Japanese dialects. In Henry M. Hoenigswald and Robert H. Langacre, editors. *Diachronic, Areal and Typological Linguistics,* Number 11 in *Current Trends in Linguistics.* Mouton, The Hague and Paris, pages 368–400.

Hetland, Magnus Lie. 2010. *Python Algorithms. Mastering Basic Algorithms in the Python Language*. Apress, New York.

Hill, Nathan W., and Johann-Mattis List. 2017. Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages. *Yearbook of the Poznań Linguistic Meeting*, 3(1):47–76.

Hoenigswald, Henry Max. 1960. *Language Change and Linguistic Reconstruction*, 4. aufl. 1966 edition. The University of Chicago Press, Chicago.

Holton, Gary, Marian Klamer, František Kratochvíl, Laura C. Robinson, and Antoinette Schapper. 2012. The historical relations of the Papuan languages of Alor and Pantar. *Oceanic Linguistics*, 51(1):86–122.

Hóu, Jīngī. 2004. *Xiàndài Hànyǔ fāngyán yīnkù [Phonological Database of Chinese Dialects]*, Shànghǎi Jiàoyù, Shànghǎi.

Huáng, Bùfán. 1992. *Zàngmiǎn yǔzú yǔyán cíhuì*, Zhōngyāng Mínzú Dàxué [Central Institute of Minorities], Beijīng.

Jacques, Guillaume. 2017. A reconstruction of Proto-Kiranti verb roots. *Folia Linguistica Historica*, 38(1):177–215.

Jäger, Gerhard, Johann-Mattis List, and Pavel Sofroniev. 2017. Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Long Papers*. pages 1204–1215.

Kay, Martin. 1964. *The Logic of Cognate Recognition in Historical Linguistics*. The RAND Corporation, Santa Monica.

Kondrak, Grzegorz. 2000. A new algorithm for the alignment of phonetic sequences. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, pages 288–295.

Kondrak, Grzegorz. 2002. Determining recurrent sound correspondences by inducing translation models. In *Nineteenth International Conference on Computational Linguistics*, pages 488–494, Taipei.

Kondrak, Grzegorz. 2003. Identifying complex sound correspondences in bilingual wordlists. Alexander Gelbukh, editor. *Computational Linguistics and Intelligent Text Processing*. Springer, Berlin, pages 432–443.

Kondrak, Grzegorz. 2009. Identification of cognates and recurrent sound correspondences in word lists. *Traitement Automatique des Langues*, 50(2):201–235.

Kroonen, Guus. 2013. *Etymological Dictionary of Proto-Germanic.* Number 11 in *Leiden Indo-European Etymological Dictionary Series*. Brill, Leiden and Boston.

List, Johann-Mattis. 2012. LexStat. Automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources*, pages 117–125, Stroudsburg.

List, Johann-Mattis. 2014. *Sequence Comparison in Historical Linguistics*. Düsseldorf University Press, Düsseldorf.

List, Johann-Mattis. 2016. Computer-assisted language comparison: Reconciling computational and classical approaches in historical linguistics. Technical Report, Max Planck Institute for the Science of Human History, Jena.

List, Johann-Mattis. 2017. A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*. pages 9–12, Association for Computational Linguistics, Valencia.

List, Johann-Mattis, Simon Greenhill, and Robert Forkel. 2017. *LingPy. A Python Library for Quantitative Tasks in Historical Linguistics*. Max Planck Institute for the Science of Human History, Jena.

List, Johann-Mattis, Simon J. Greenhill, and Russell D. Gray. 2017. The potential of automatic word comparison for historical linguistics. *PLOS ONE*, 12(1):1–18.

McMahon, April, and Robert McMahon. 2005. *Language Classification by Numbers*. Oxford University Press, Oxford.

Meier-Brügger, Michael. 2002. *Indogermanische Sprachwissenschaft*, 8th edition. de Gruyter, Berlin and New York.

Meillet, Antoine. 1908. *Les dialectes Indo-Européens*, Librairie Ancienne Honoré Champion, Paris.

Meillet, Antoine. 1954. *La méthode comparative en linguistique historique, reprint edition*. Honoré Champion, Paris.

Needleman, Saul B., and Christan D. Wunsch. 1970. A gene method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453.

Newman, M. E. J. 2010. *Networks. An Introduction*. Oxford University Press, Oxford.

Prokić, Jelena, Martijn Wieling, and John Nerbonne. 2009. Multiple sequence alignments in linguistics. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural*

Heritage, Social Sciences, Humanities, and Education, pages 18–25.

Ross, Malcolm, and Mark Durie. 1996. Introduction. In Durie, Mark, editor. *The Comparative Method Reviewed. Regularity and Irregularity in Language Change*. Oxford University Press, New York, pages 3–38.

Wagner, Robert A., and Michael J. Fischer. 1974. The string-to-string correction problem. *Journal of the Association for Computing Machinery*, 21(1):168–173.

Walworth, Mary. 2018. Polynesian segmented data. Version 1. Zenodo. http://doi.org/10.5281/zenodo.1689909.

Weiss, Michael. 2015. The comparative method. In Claire Bowern and Nicholas Evans, editors. *The Routledge Handbook of Historical Linguistics*, 1st edition, Routledge Handbooks in Linguistics. Routledge, New York, pages 127–145.

Welsh, D. J. A., and M. B. Powell. 1967. An upper bound for the chromatic number of a graph and its application to timetabling problems. *The Computer Journal*, 10(1):85–86.

Wright, Joseph. 1910. *Grammar of the Gothic Language*, 2 edition. Clarendon Press, Oxford.

5 Conclusion and Outlook

The popularity of computational approaches in the fields of historical linguistics and linguistic typology is constantly increasing. In contrast to the early attempts by Morris Swadesh, whose methods for the automated dating of language divergence times based on the amount of shared items in selected lists of basic vocabulary (Swadesh 1952, Swadesh 1955) items were soon abandoned after they had first been proposed (Bergsland and Vogt 1962, Hoijer 1956), it seems that the more recent computational approaches, which were established during the past two decades, have entered the realm of the language sciences without the intention to leave. Nevertheless, the recent intruders are still viewed skeptically by many practitioners of historical linguistics and linguistic typology, and one can observe a divide in the fields, with computational linguists promising novel revolutionary techniques on the one hand, and traditional linguists who look at the new methods with suspicion, emphasizing the past success of their manual and qualitative approaches.

In order to overcome this divide between quantitative and qualitative approaches, I have tried to work on a combined framework for *computer-assisted language comparison* that would help qualitative linguists to increase the efficiency of their work while helping quantitative linguists to increase the accuracy of their approaches. Obviously, this framework cannot solve all problems once and for all times. In order to reconcile quantitative and qualitative approaches, detailed and concrete proposals for concrete problems are needed.

In this study, I have tried to show how the framework for computer-assisted language comparison can be filled with life and how combined approaches which allow for a quantitative and a qualitative treatment at the same time can be advanced. Starting from novel approaches to the reconstruction of phylogenetic networks and ancestral character states in Chapter 2, I have presented new approaches towards data formats and annotation frameworks in historical linguistics in Chapter 3, and finally presented in Chapter 4, how the concrete task of *sequence comparison* in historical linguistics can be further refined with help of improved approaches to phonetic sequence modeling and new algorithms for the inference of correspondence patterns. In this way, I have tried to illustrate how computer-assisted research can be carried out and advanced in concrete.

Looking back at this research, which began pursuing more than six years ago, three aspects have proven to be crucial for any work on computer-assisted language comparison and possibly even for computer-assisted approaches in the humanities in general. These aspects are the *data*, the the frameworks for *annotation* (the interfaces), and the *computational approaches* (the software). All three are crucial for a successful application of computer-assisted approaches to historical and typological language comparison.

Since both qualitative and quantitative research in historical linguistics usually deals with data (albeit in different form), a first step towards a reconciliation of the two perspectives is to guarantee that data can be both qualitatively and quantitatively processed. Since quantitative processing is based on computers, one can also say that data should be not only human- but also machine-readable. All too often, scholars have neglected this aspect in the past. Computational approaches have used numerical data representations and restricted themselves to presenting only aggregated results to the experts. Experts on the other hand, have created large datasets that can only be digested manually by eyeballing the collections page by page. In order to overcome this problem, I have presented concrete standards in Section 3.1, with the CLDF

initiative as a first example on general standards for a human- and machine-readable treatment of cross-linguistic datasets (Forkel et al. 2018), and the CLTS initiative as a concrete example of how standards for phonetic transcription systems can be established and practically used (Anderson et al. 2018).

In addition to the establishment of standards, computer-assisted language comparison also needs to allow experts to conveniently curate their data in concordance with the proposed standards. For this reason, I have invested a considerable amount of time to develop specific *interfaces* and annotation frameworks, as presented in Section 3.2. The EDICTOR (List 2017), for example, allows for an efficient annotation of etymological data in concordance with basic aspects of linguistic standards. The guidelines for the annotation of morphological relations and partial cognate relations, as presented in the study by (Hill and List 2017), furthermore, illustrate how important it is to work towards a strict formalization of those relations which qualitative linguists use in a mostly implicit manner.

As a last aspect, computer-assisted language comparison has to develop new methods which help to solve concrete problems in the field of historical language comparison. Here, I have presented two kinds of methods which I have been developing during the past years. First, in Chapter 2, I have presented novel approaches to phylogenetic reconstruction, concentrating on phylogenetic networks (List et al. 2014a, List et al. 2014b) and on the reconstruction of ancestral character states (Jäger and List 2018, List 2016). Although these approaches are more computational in their nature than qualitative, they represent concrete advancements over past approaches. Thus, they deal explicitly with problems that have been so far ignored in the majority of quantitative approaches, such as lexical borrowing or partial cognacy. Furthermore, their usefulness has been explored qualitatively. Second, in Chapter 4, I have presented novel methods for automated sequence comparison, which are supposed to help qualitative researchers to increase the efficiency of annotation, by allowing them to pre-process their data with the new approaches and later use the annotation tools presented in Section 3.2 to correct the errors. That the methods are already good enough to be used in this way could be illustrated with the help of the two studies presented in Section 3.1, which deal with partial cognate detection in particular (List et al. 2016b) and the evaluation of the performance of different cognate detection algorithms in general (List et al. 2017). In Section 4.2, finally, I first presented a tutorial helping scholars to get started in making active use of the new methods developed during the past years (List et al. 2018), and finally showed the improved data formats and annotation frameworks along with the novel algorithms could be combined to address the problem of sound correspondence pattern detection within a computer-assisted framework, a problem which had not yet been addressed neither in quantitative, nor in qualitative approaches (List 2019b).

While I consider the work that I have presented here as a success, at least in parts, there remain many challenges that I have not been able to tackle so far, and which may well remain challenges for quite some time in the future. Among these is the problem of borrowing detection, for which I presented initial ideas by using phylogenetic network approaches, but which remains one of the major obstacles that prevent us from finding deeper genetic affiliations among the languages of the world (List 2019a). A further problem, which became evident in the algorithm for partial cognate detection is that we are still unable to identify morpheme boundaries from linguistic data without relying on experts. Methods which have been proposed by colleagues in the past all fail when being applied to data in historical linguistics, because the word lists which are available are all too small for the data-hungry approaches traditionally used in Natural Language Processing. A third major challenge, which could not be addressed in this context relates to the reconstruction of proto-sounds. While the work presented on ancestral state reconstruction may seem applicable to tackle this problem (at least in part), it is obvious that current ancestral state reconstruction methods cannot cope with the postulation of ancestral states that one cannot find in the data, while linguists often do so in linguistic reconstruction (Fox 1995).

In addition to these inference problems, we have also numerous problems of modeling, where we lack the proper models to describe the processes of language change. Among these, we find – again – sound change as a chief candidate that is difficult to model consistently, also since general language-independent principles of phonotactics are still purely understood. A second candidate for an open problem is the modeling of lexical change, which is still modeled as a simple process of word gain and word loss in most phylogenetic approaches. As a third problem, we still lack clear-cut approaches to estimate language relatedness. Although there have been some attempts by scholars to find a simple formula that assesses the probability of two languages being related (Baxter and Manaster Ramer 2000, Kessler 2001), none of the approaches proposed so far could really do justice to the complexity of language change, and only a few approaches have been tested on sufficient amounts of data.

A third group of problems relates to the establishment of typological accounts on the classical phenomena of language change. Here, the problems of modeling sound change make it also difficult to arrive at a first dataset that would allow us to compare the major sound change processes which happened so far in the languages of the world. While pioneering work has been done in this respect (Kümmel 2008), it seems still impossible to come up with a general solution of collecting which sound changes occurred in which language families, that would be applicable to all languages of the world. A similar problem can be encountered when dealing with semantic change. While datasets that try to show the major tendencies of semantic change across different languages and times have been published already (Zalizniak 2018), their applicability is still heavily exacerbated by the fact that the elicitation of specific meanings has not been done in a systematic manner. Some specific typological problems have even never been really asked for, although they should be interesting for historical linguists. An example in this regard is a consistent typology of what Blank (1997) calls *Attraktion* and *Expansion*. The fact that the words expressing certain concepts are repeatedly subject to change, while some concepts keep a rather stable connection with word forms expressing them, has still not been systematically investigated by linguists, although we would expect to find numerous interesting factors that could contribute to the attractivity or the expansivity of a given concept, such as, among others, its frequency of use, its concreteness, or how important the concept is to its speakers. Such a systematic investigation of "semantic promiscuity" (as inspired by the concept of "domain promiscuity" in biology, see List et al. 2016c) in word formation processes documented in the languages of the world might therefore provide us with many interesting new insights into human perception and cognition. However, given the difficulty of making linguistic data cross-linguistically comparable, it is not likely that this problem will be tackled any time soon.

All in all, we can say that there is, beyond doubt, still a lot to do in order to reconcile quantitative and qualitative approaches in historical linguistics and linguistic typology within a unifying framework of computer-assisted language comparison. With this study, I hope, nevertheless to have shown that these attempts bear a lot of potential, and that it is therefore worthwhile to pursue them further.

Bibliography

Anderson, C., T. Tresoldi, T. C. Chacon, A.-M. Fehn, M. Walworth, R. Forkel, and J.-M. List (2018). "A Cross-Linguistic Database of Phonetic Transcription Systems." *Yearbook of the Poznań Linguistic Meeting* 4.1, 21–53.

Anthony, D. W. and D. Ringe (2015). "The Indo-European homeland from linguistic and Archaeological perspectives." *Annual Review of Linguistics* 1, 199–219.

Barrachina, S. et al. (2008). "Statistical approaches to computer-assisted translation." *Computational Linguistics* 35.1, 3–28.

Baxter, W. H. and A. Manaster Ramer (2000). "Beyond lumping and splitting: Probabilistic issues in historical linguistics." In: *Time depth in historical linguistics*. Ed. by C. Renfrew, A. McMahon, and L. Trask. Cambridge: McDonald Institute for Archaeological Research, 167–188.

Bergsland, K. and H. Vogt (1962). "On the validity of glottochronology." *Current Anthropology* 3.2, 115–153. JSTOR: 2739527.

Bhasker, J. and T. Samad (1991). "The clique-partitioning problem." *Computers & Mathematics with Applications* 22.6, 1 –11.

Blank, A. (1997). *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*. Beihefte zur Zeitschrift für romanische Philologie 285. Tübingen: Niemeyer.

Dagan, T. and W. Martin (2009). "Getting a better picture of microbial evolution en route to a network of genomes." *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 364.1527, 2187–2196.

Forkel, R., J.-M. List, S. J. Greenhill, C. Rzymski, S. Bank, M. Cysouw, H. Hammarström, M. Haspelmath, G. A. Kaiping, and R. D. Gray (2018). "Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics." *Scientific Data* 5.180205, 1–10.

Fox, A. (1995). *Linguistic reconstruction. An introduction to theory and method*. Oxford: Oxford University Press.

Hammarström, H., M. Haspelmath, R. Forkel, and S. Bank (2020). *Glottolog. Version 4.3*. Jena: Max Planck Institute for the Science of Human History. URL: https://glottolog.org.

Hill, N. W. and J.-M. List (2017). "Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages." *Yearbook of the Poznań Linguistic Meeting* 3.1, 47–76.

Hoijer, H. (1956). "Lexicostatistics. A critique." *Language* 32.1, 49–60. JSTOR: 410652.

Holm, H. J. (2007). "The new arboretum of Indo-European "trees□. Can new algorithms reveal the phylogeny and even prehistory of Indo-European?" *Journal of Quantitative Linguistics* 14.2-3, 167–214.

IPA, ed. (1999). *Handbook of the International Phonetic Association. A guide to the use of the international phonetic alphabet*. Cambridge: Cambridge University Press.

Jäger, G. (2018). "Global-scale phylogenetic linguistic inference from lexical resources." *Scientific Data* 5.180189, 1–16.

Jäger, G. and J.-M. List (2018). "Using ancestral state reconstruction methods for onomasiological reconstruction in multilingual word lists." *Language Dynamics and Change* 8.1, 22–54.

Kay, M. (1964). *The logic of cognate recognition in historical linguistics*. Santa Monica: The RAND Corporation.

Kessler, B. (2001). *The significance of word lists. Statistical tests for investigating historical connections between languages*. Stanford: CSLI Publications.

Kümmel, M. J. (2008). *Konsonantenwandel* [Consonant change]. Wiesbaden: Reichert.

List, J.-M. (2014). *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.

– (2016). "Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction." *Journal of Language Evolution* 1.2, 119–136.

– (2017). "A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets." In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*. Valencia: Association for Computational Linguistics, 9–12.

– (2018). "Ho well do automatic methods for language comparison work?" *Latest Thinking* 4.3, LT-PUB10576.

– (2019a). "Automated methods for the investigation of language contact situations, with a focus on lexical borrowing." *Language and Linguistics Compass* 13.e12355, 1–16.

– (2019b). "Automatic inference of sound correspondence patterns across multiple languages." *Computational Linguistics* 1.45, 137–161.

List, J.-M., M. Cysouw, and R. Forkel (2016a). "Concepticon. A resource for the linking of concept lists." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. "LREC 2016" (Portorož, 05/23–05/28/2016). Ed. by N. C. C. Chair), K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis. Luxembourg: European Language Resources Association (ELRA), 2393–2400.

List, J.-M., S. Greenhill, T. Tresoldi, and R. Forkel (2019). *LingPy. A Python library for quantitative tasks in historical linguistics*. Version 2.6.5. URL: `http://lingpy.org`.

List, J.-M., S. J. Greenhill, and R. D. Gray (2017). "The potential of automatic word comparison for historical linguistics." *PLOS ONE* 12.1, 1–18.

List, J.-M., P. Lopez, and E. Bapteste (2016b). "Using sequence similarity networks to identify partial cognates in multilingual wordlists." In: *Proceedings of the Association of Computational Linguistics 2016 (Volume 2: Short Papers)*. Association of Computational Linguistics. Berlin, 599–605.

List, J.-M., S. Nelson-Sathi, H. Geisler, and W. Martin (2014a). "Networks of lexical borrowing and lateral gene transfer in language and genome evolution." *Bioessays* 36.2, 141–150.

List, J.-M., S. Nelson-Sathi, W. Martin, and H. Geisler (2014b). "Using phylogenetic networks to model Chinese dialect history." *Language Dynamics and Change* 4.2, 222–252.

List, J.-M., J. S. Pathmanathan, P. Lopez, and E. Bapteste (2016c). "Unity and disunity in evolutionary sciences: process-based analogies open common research avenues for biology and linguistics." *Biology Direct* 11.39, 1–17.

List, J. M., C. Rzymski, S. Greenhill, N. Schweikhard, K. Pianykh, A. Tjuka, M.-S. Wu, C. Hundt, T. Tresoldi, and R. Forkel (2020). *Concepticon. A resource for the linking of concept lists. Version 2.4.0*. Jena: Max Planck Institute for the Science of Human History. URL: `https://concepticon.clld.org/`.

List, J.-M., M. Walworth, S. J. Greenhill, T. Tresoldi, and R. Forkel (2018). "Sequence comparison in computational historical linguistics." *Journal of Language Evolution* 3.2, 130–144.

Maddieson, I., S. Flavier, E. Marsico, C. Coupé, and F. Pellegrino. (2013). "LAPSyD: Lyon-Albuquerque Phonological Systems Database." In: *Proceedings of Interspeech*. (Lyon, 08/25–08/29/2013).

Moran, S., D. McCloy, and R. Wright, eds. (2014). *PHOIBLE Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Nelson-Sathi, S., J.-M. List, H. Geisler, H. Fangerau, R. D. Gray, W. Martin, and T. Dagan (2011). "Networks uncover hidden lexical borrowing in Indo-European language evolution." *Proceedings of the Royal Society of London B: Biological Sciences* 278.1713, 1794–1803.

Pullum, G. K. and W. A. Ladusaw (1996). *Phonetic symbol guide*. 2nd ed. Chicago: University of Chicago Press.

Swadesh, M. (1952). "Lexico-statistic dating of prehistoric ethnic contacts. With special reference to North American Indians and Eskimos." *Proceedings of the American Philosophical Society* 96.4, 452–463.

– (1955). "Towards greater accuracy in lexicostatistic dating." *International Journal of American Linguistics* 21.2, 121–137. JSTOR: 1263939.

Zalizniak, A. A. (2018). "The Catalogue of Semantic Shifts: 20 years later." *Russian Journal of Linguistics* 22.4, 770–787.