This is a preprint that still needs to be peer reviewed and has been submitted for inclusion in the second edition of the Routledge Handbook of Historical Linguistics, edited by Bowern and Evans. Please quote as:

List, Johann-Mattis (2022): Computational approaches to historical language comparison. [Preprint not peer reviewed]. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Computational Approaches to Historical Language Comparison

Johann-Mattis List

The chapter discusses recently developed computational techniques providing concrete help in addressing various tasks in historical language comparison, focusing specifically on those tasks which are typically subsumed under the framework of the comparative method. These include the proof of relationship, cognate and correspondence detection, phonological reconstruction and sound law induction, and the reconstruction of evolutionary scenarios.

1 Introduction

There are numerous techniques which scholars use to reconstruct those parts of the history of languages which are not reflected in any kind of sources. Scholars usually label these techniques as *comparative method* (Meillet [1925] 1954; Weiss 2015), which is given the status of some kind of an *overarching framework* to study language history (Fox 1995; Jarceva 1990; Klimov 1990; Ross and Durie 1996). While the term *comparative method* implies a certain degree of coherence, the linguistic literature largely differs both with respect to the concrete techniques which should be subsumed under the term and with respect to the order in which they should be applied. For almost twenty years, scholars have tried to automatize certain aspects of the traditional comparative method. These new techniques rarely propose a complete, full-fledged computational counter-part of the traditional comparative method (Steiner, Stadler, and Cysouw 2011). Instead, they start from certain individual problems – such as the inference of cognate sets (Hauer and Kondrak 2011) or the identification or regular sound correspondences (Kondrak 2002) – and then try to solve them in a computational setting. In the following, I will refer to these techniques as *computational historical language comparison*, and I will try to provide an overview on the most recent developments in this relatively young field of computational historical linguistics.

As I have emphasized, the comparative method is not a coherent workflow which scholars apply in a schematic fashion but rather a bunch of techniques whose scope and character varies from scholar to scholar. In order to compare classical techniques with their computational counterparts, however, it is nevertheless useful to follow some general workflow. I therefore adapt the workflow presented by Ross and Durie (1996) with some modifications in order to discuss which steps of the comparative method have so far been addressed in the field of computational historical language comparison.

Figure 1 shows a visualization of the workflow by Ross and Durie (1996). The workflow starts from the – at times rudimentary – proof that a certain number of languages are genetically related and thus go back to the same ancestor language (1). It is followed by the identification of cognate words and morphemes (2) and the recognition of regular sound correspondences (3). In order to solve these two problems, scholars often operate in an iterative fashion by identifying a preliminary set of cognates from which a preliminary set of sound correspondences is inferred, which is later refined by updating the list of cognate sets and the list of correspondence patterns. Once cognate sets and sound correspondence patterns have been identified, scholars try to reconstruct the phonology and morphology of the ancestral language by projecting sound correspondence patterns to individual sounds in a fictitious proto-language from which the languages in the sample are assumed to have descended (4). Patterns which point to an abrupt divergence of some languages in the sample from the others are then collected in in order to prove that they are indeed innovations which are indicative of subgrouping (5), from which in a further stage the phylogeny of the language family could be reconstructed (6). The final stage of the comparative method in this notion consists in the compilation of an etymological dictionary in which both regular and irregular patterns are explained for the history of each word in the data.

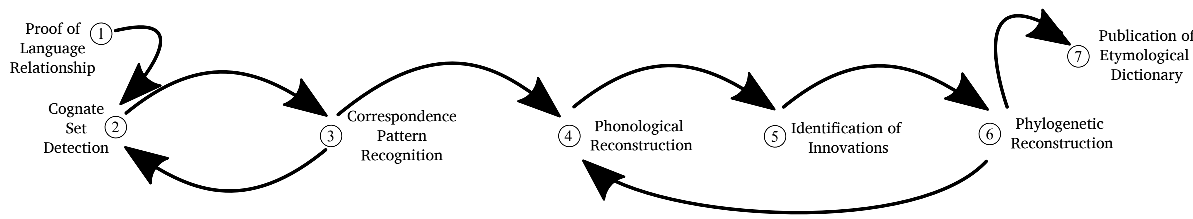

Figure 1: Workflow of the Comparative Method by Ross and Durie (1996)

Since computational approaches rarely correspond directly to the steps marked by this workflow, it is useful to streamline the workflow slightly for the purpose of reviewing recent computational approaches. As a result, I will in the following first discuss statistical techniques by which scholars have tried to provide a *proof of relationship* (§2, task of the original workflow). I will then discuss *cognate and correspondence pattern detection* (§3, summarizing tasks 2 and three of the workflow) and *phonological reconstruction and sound law induction* (§4, task 4 of the workflow) and finish the overview by looking at the *reconstruction of etymological scenarios* (§5, task 7 of the workflow) and discussing *open problems* (§6). Tasks 5 and 6 of the original workflow will be ignored in this overview, since they are typically dealt with in the field of phylogenetic reconstruction, which also emerged during the past 20 years and no longer corresponds to the traditional workflow depicted by Ross and Durie [+++REF to INTERNAL CHAPTER+++].

In this context, I will maintain a broad perspective on *computational solutions* which not only regards full-fledged unsupervised or supervised workflows applied to data in a machine-learning setting, but also includes *computer-assisted* approaches which deal with the formal representation and annotation of data, even if no complete computational solutions to solve the annotation problems can be provided (List 2016b).

2 Proof of Relationship

In the workflow of the comparative method by Ross and Durie (1996), the first stage consists in the identification of proof that the languages one wants to compare are indeed genetically related. This seems to contrasts with views that see the successful application of the comparative method itself as the proof for language relationship (Dybo and Starostin 2008). However, since Ross and Durie (1996) explicitly point to proof in the form of very convincing comparanda, labeled "individual-identify evidence" by Nichols (1996) in the same volume, one does not need to draw a sharp devide betweeen authors . Instead, one can emphasize that the initial stage serves to identify a sample of languages which warrant a closer investigation.

The reason why this *proof* (or "initial justification") is needed when comparing languages historically is because – in contrast to species in evolutionary biology – linguists cannot assume that all languages have sprung from a common source. Despite numerous attempts to push the boundaries of historical language comparison, the statement by the Society of Linguistics in Paris from 1866, which denied research on the origin of language to be published in its journals (Statuts 1871), is still supported by most historical linguists, and most scholars agree that even if language only originated once, it is impossible to trace the development of the worlds' languages back to such a time depth. While biologists can compare the genome of fruit flies with the genome of humans and try to identify commonalities, linguists interested in the history of languages must justify why they chose their language sample.

Since it has never been made clear, what determines whether a set of shared features is convincing evidence of language relationship or not, scholars have repeatedly tried to produce statistical tests that would help them to prove language relationship in a more objective manner. The basic idea of most of these tests is to show that it is highly unlikely that a certain similarity pattern, as it can be observed between two or more languages, has evolved by chance. Starting with the work by Ringe (1992), there have been quite a few attempts to design an ultimate *test* for genetic relatedness (Baxter and Manaster Ramer 1996, 2000; Mortarino 2004, 2009; Kessler 2001; Turchin et al. 2010; Blevins and Sproat 2021), but so far, none of these tests has gained acceptance among practitioners of historical language comparison.

There are different reasons for the reluctance of scholars to accept genetic relationship tests. On the one hand, few of the tests have been published in the form of an implemented software solution. As a result, replicating the studies or applying the studies to new language families is considered a tedious enterprise that would not justify the efforts, since most scholars work on language families which have been established a long time ago. On the other hand, there is a "cultural" problem with significance tests. Classical historical linguistic scholarship treats genetic relationship in a manner similar to mathematicians who use proofs to mark a problem as solved. If the proof for genetic relationship has been identified (which may require a lot of genius but at times it may be fairly trivial), the problem is considered to be settled, and the reconstruction can begin. The idea of designing a *test*, however, is fundamentally different in this regard, since tests only offer approximations to problems, and they are always accompanied by rates of false positives and false negatives. Designing a test to *prove* something is therefore an enterprise which is problematic in itself. The idea of testing and the idea of finding a proof are fundamentally different.

The typical test for language relationship starts from a list of words in a pair of languages (1). In order to faciliate the comparison of word pairs, scholars tend to reduce the words to be compared by the meaning they express, demanding either identity, or similarity (2). The word pairs are then compared one by one and potential cognates are identified, using various criteria by which words are matched (3). In order to check whether the number of obtained matches is significantly different from the number of matches one would obtain when comparing unrelated langugaes, scholars either *calculate* significance with the help of formulas from combinatorics that describe how matches combine, or they *estimate* the significance by using specific techniques, such as, for example, *permutation tests* (4). In a permutation test, the word lists are shuffled, so that words no longer need to express similar meanings, and matches are counted and later compared with the matches obtained when controlling for meaning. In all cases, scholars obtain some that is supposed to tell how likely it would be to obtain these results when comparing unrelated languages.

The various tests proposed in the past differ with respect to four steps of the workflow. Ringes' test (Ringe 1992), later refined by Baxter and Manaster Ramer (1996), for example, uses word lists built from the Swadesh list of 100 items (Swadesh 1955) and only choses one translation equivalent per concept. Matches are determined by considering only the first consonant in all word pairs, listing all potential correspondences between consonants in the two wordlists regardless if these are identical or similar. The probability of obtaining individual matches is then calculated by assuming a binomial distribution (critizied later by Baxter and Manaster Ramer 1996 who propose a hypergeometric distribution instead) and counting those matches which are significant (with $p < 0.01$).

In contrast, Turchin et al. (2010) and Kassian et al. (2015) do not count initial consonant correspondences, but follow Dolgopolsky (1964) in converting their word forms to 10 consonant classes, which are supposed to reflect the most frequently recurring sound correspondences between consonants, and counting those word pairs as a matchin which coincide in their first consonant classes. The expected distribution of matches is then calculated with the help of a permutation method and compared with the attested distribution.

The method by Blevins and Sproat (2021) uses techniques for phonetic alignment analysis (Kondrak 2000; List 2014) for the matching of word forms and estimates the expected distribution of matches by creating lists of artificial words using *lexical language models*. Lexical language models are a specific type of statistical language models which assign probabilities to sequences (Bender and Koller 2020). While language models typically combine words to sentences, lexical language models combine sounds to word forms (Miller et al. 2020). For their main experiment involving a comparison of Proto-Basque with Proto-Indo-European – which the authors think are genetically related – the authors employ a fuzzy matching procedure for the identification of possible comparanda, using the colexifications provided by the CLICS³ database (Rzymski et al. 2020, https://clics.clld.org) as a criterion to assign two different concepts to the same semantic slot.

In order to guarantee the validity of tests for genetic relationship, it is important to test them on a significantly large amount of languages in order to guarantee that they are *conservative*, avoiding large amounts of false positives, while at the same time being successful at identifying distant cases of genetic relationship. Additionally, tests should be applicable in a uniform way to a large number of languages, without requiring that scholars interfere much with the data. Finally, given that language contact can easily obscure the results, tests should be resistant to potential influences from language contact and sound symbolism.

As of now, none of the tests which have been proposed so far fulfils all of these criteria. Given its improved handling of sound correspondences and lexical cognates, the test by Blevins and Sproat (2021) shows many improvements over previous tests, and the authors even test it on a larger sample of language pairs (taken from List 2014). However, the authors use slightly modified workflows for the comparison of Basque and Indo-European and the comparison on the larger sample of languages, since the latter needs to be fully automated in order to be feasible. As a result, it would be premature to assume that the relationship of Indo-European and Basque is proven by their study.

In order to account to improve future tests of genetic relationship, it would be desirable to test and fine-tune them on a normed dataset of language pairs from various language families and different time depths. Unless full-fledged error statistics have been calculated, genetic relationship tests should not be applied to any suspected cases of deep language relations. Instead, they should first be tested on a common gold standard and common error statistics should be calculated. When trying to investigate deep language relations, a range of different tests should be used instead of employing one novel test alone. This would require that scholars share their test applications by publishing their source code. In this way, testing for genetic language relationship could leave the realm of suspicion and develop into a powerful tool that could accompany and inform the traditional work on historical language comparison.

3 Cognate and Correspondence Detection

The second and third stage of our exemplary workflow for historical language comparison consist in the identification of cognate words and regular sound correspondences. Both problems were for a long time regarded as considerably hard, resisting computational solutions. The past two decades, however, have seen great progress with respect to the automated detection of cognates words (Rama 2016; List 2014; Ciobanu and Dinu 2013; Hauer and Kondrak 2011; Hall and Klein 2010; Dellert 2018; Arnaud et al. 2017; Jäger et al. 2017), while methods for the identification of regular sound correspondence patterns have only been addressed sporadically (Kondrak 2002; List 2019b; Turchin et al. 2010).

While there is some work on *pairwise cognate detection*, where two wordslists are compared and an algorithm needs to find the word pairs which are cognate (Rama et al. 2013), the most typical cognate detection task consists in the identification of cognates in *multilingual wordlists*. The typical setting for the cognate detection task starts from a muwordlist in which an initial list of concepts has been translated into a certain number of languages. An algorithm for automatic cognate detection then needs to identify all words in the wordlist which share a common origin. The base task can be further varied by specifying if the detection of cognate words should be restricted to words with the same meaning, or whether cognates should also be detected *across* concepts (*cross-semantic cognate detection*, see Wu et al. 2020), or by specifying whether cognacy should be assigned to words as a whole, or whether only cognate *morphemes* should be identified (*partial cognate detection*, see List et al. 2016).

Most workflows for cognate detection in multilingual wordlists consist of two major stages. In a first stage, all words in a given semantic slot are compared with each other and phonetic distances are computed for all word pairs (1). In a second stage, the distances are analyzed in order to provide a partitioning (or *flat clustering*) of all words into cognate sets (List et al. 2017, see Figure 2). Variants of this workflow consist in the *pre-* and *post-processing* of the data. Preprocessing allows – for example – for the identification of regular sound correspondences, upon which phonetic similarity measures can be based (List 2012), or for the aggregation of several word similarity measures via classification techniques from machine learning (Jäger et al. 2017). Post-processing allows to either refine a given analysis, or to analyze the data further, for example by searching for cases in which the same word form or morpheme recurs across different concepts in order to identify cross-semantic cognates (Wu et al. 2020). Methods also differ with respect to the methods employed for the partitioning of words into cognate sets. Here, the most straightforward approaches consist in flat variants of agglomerative clustering approaches (e.g. UPGMA, Sokal and Michener 1958) which stop merging word forms into larger and larger clusters when a certain threshold is reached (see List 2014). Alternatively, methods for *community detection*, originally designed for social network analysis, have also proven useful (List et al. 2017).

Methods which derive cognate sets through partitioning based on pairwise word comparisons typically need a user-defined threshold that determines when two or more words are judged to be cognate. An alternative family of approaches, originally going back to Dolgopolsky (1964), uses *sound classes* (List 2014) or *consonant classes* to identify potential cognates without using an explicit threshold. These sound-class-based cognate detection approaches preprocess the data by converting the first two consonants in all words in a given semantic slot to their corresponding consonant classes and assigning all words with identical consonant classes to the same cognate set (Turchin et al. 2010). Sound-class-based cognate detection approaches have the advantage of being very fast to apply, since no computationally intensive pairwise comparisons need to be carried out (Rama and List 2019). On the other hand, they lack accuracy in comparison with cognate detection approaches based on phonetic similarity (List et al. 2017).

	Swedish *kvinna*	English *woman*	Danish *kvinde*	Norwegian *kvine*	Dutch *vrouw*	German *Frau*
Swedish `kvina`	0.00	0.69	0.07	0.12	0.71	0.78
English `wumin`	0.69	0.00	0.66	0.57	0.68	0.87
Danish `kveni`	0.07	0.66	0.00	0.08	0.67	0.71
Norwegian `kwini`	0.12	0.57	0.08	0.00	0.75	0.74
Dutch `frou`	0.71	0.68	0.67	0.75	0.00	0.17
German `frau`	0.78	0.87	0.71	0.74	0.17	0.00

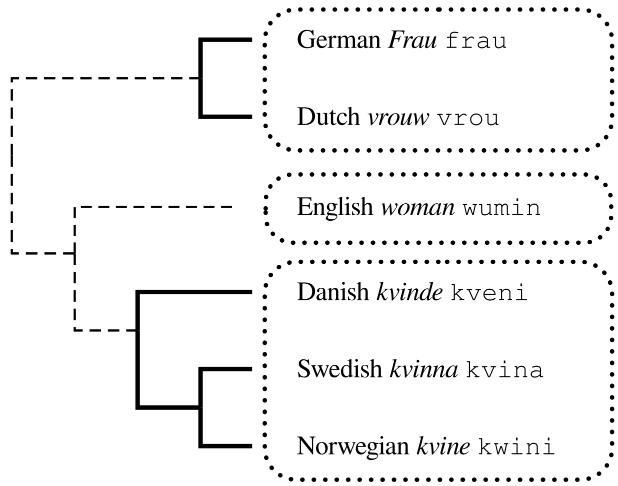

Figure 2: Flat clustering of words into cognate sets, based on a distance matrix inferred with the help of the SCA phonetic alignment algorithm.

In the traditional workflow of the comparative method, the identification of cognates needs to be substantiated by the identification of regular sound correspondences. The identification of cognates and regular sound correspondences is typically done in an iterative fashion in which scholars start from a list of potentially cognate words which is then analyzed in order to identify major sound correspondence patterns. Once major correspondence patterns have been identified, scholars go over the list of potential cognates and check if the correspondences hold. This procedure is typically illustrated for language pairs only (see Figure 3), but in concrete applications of the comparative method, scholars tend to work with more than two languages. Adding more languages to the comparison has advantages and disadvantages. The disadvantage is that comparisons become more and more complex, the more languages one compares. The advantuage, however, is that the resulting *correspondence patterns* are much more coherent and allow – if languages are indeed genetically related – for a much clearer identification of instances of regular sound change and their deviation in individual words from individual languages.

Figure 3: Iterative identification of cognate words and regular sound correspondences.

Although the identification of correspondence patterns across multiple languages constitutes one of the core objectives of the comparative method, the problem was for a long time ignored in computational approaches. Early approaches concentrate almost exclusively on sound correspondences between language pairs. Thus, Kondrak (2002) uses techniques for machine translation to identify regular sound correspondences in several language pairs, and Prokić and Nerbonne (2013) try to identify how sound change processes diffuse over a dialect area based on the investigation of sound correspondences between Bulgarian dialect pairs,

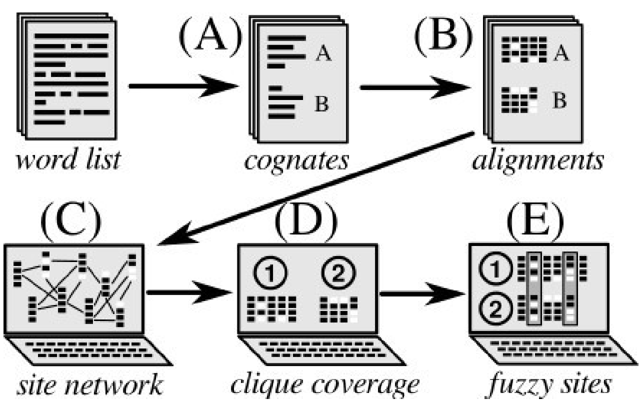

Figure 4: Workflow for the automated recognition of correspondence patterns in multilingual wordlists.

A first algorithm for the recognition of sound correspondence patterns in multilingual wordlists was proposed by List (2019b). The algorithm starts by assembling individual columns of multiple phonetic alignments from comparative wordlists. Each column (also called *site*) of an alignment reflects a potential sound correspondence pattern, but since cognate sets are not always reflected in all languages in a given sample, the alignments have gaps resulting from missing data. In order to identify regular sound correspondence patterns, the algorithm therefore compares all individual alignment sites by constructing an alignment site network in which links are drawn between those sites which are *compatible* with each other. Correspondence patterns are then inferred from this network by identifying the largest groups of compatible alignment sites in the network using an algorithm that computes the *clique cover*, that is, the partition of the network into the smallest number of partitions where each partition constitutes a clique (Bhasker and Samad 1991). Figure 4 illustrates the workflow.

Strictly speaking, there is no reason to assume that the solution of the clique cover problem provides an optimal solution to the problem of identify regularly recurring sound correspondence patterns. We know that a correspondence pattern must be a clique in an alignment site network, but we do not know how the cliques in an alignment site network are typically distributed in genetically related languages. As a result, the approach by List (2019b) needs to be tested further and applied to more language families. What speaks in favor of the method is that it works quite well at predicting missing reflexes in cross-linguistic data (Bodt and List 2022). Thus, even if it does not provide an optimal solution to the problem, it can serve as an important starting point, specifically for those approaches which try to combine computational and classical historical language comparison as part of a computer-assisted (rather than a purely computer-based) framework (List 2016b).

4 Phonological Reconstruction and Sound Law Induction

As a fourth step in the workflow adapted from Ross and Durie (1996), we have listed *phonological reconstruction*, the technique, by which scholars try to determine the most likely proto-sounds for each of the distinct correspondence patterns in the data (Anttila 1972) from which in turn individual proto-forms can be reconstructed for each distinct cognate set (Fox 1995). What was not mentioned in the workflow, but what should be considered as similarly important, going hand in hand with phonological reconstruction, is the *induction of sound laws* which explain under which conditions the sounds in the ancestral languages turn into the sounds in the descendant languages.

As an example for the classical application of the technique of phonological reconstruction, compare cognate words for “to eat” in Romance languages, like Spanish [θenaɾ], Romanian [tʃina], and Portuguese [sjaɹ]. If we align these words, and add more data to the comparison, we can infer correspondences patterns, such as [θ tʃ s] for the initial sound or [a a a] for the second syllable vowel in all words. In order to reconstruct the ancestral form in Proto-Romance, which we know ultimately goes back to Latin [keːnaːrɛ], we would try to find the most plausible ancestral sound for each of the correspondence patterns. For the initial pattern, for example, we could argue that the distribution of reflexes most likely points to a **k*, since this sound can neatly develop into the target sounds under conditions of palatalization. As a result, we would then reconstruct each alignment that shows the correspondence pattern [θ tʃ s] for Spanish, Romanian, and Portuguese with **k*, and we would most likely also posit a *sound law* by which original **k* becomes [tʃ], and consecutively turns into [s] and [θ] when followed by a front vowel, such as **e* or **i*. What we can see from this example is that linguists not only tend to infer proto-forms in phonological reconstruction, but also induce sound laws at the same time when they propose their proto-forms. Similar to the iterative procedure by which correspondence patterns and cognate sets are detected, phonological reconstruction goes hand in hand with *sound law induction*.

In computational historical language comparison, the classical iterative technique for phonological reconstruction and sound law induction has been largely ignored so far. What scholars have instead concentrated on were techniques for *supervised phonological reconstruction*, and techniques for *ancestral state reconstruction*. The former approaches start from a training set of cognate words along with known proto-forms and then *train* a machine learning approach to *learn* how to predict proto-forms when encountering data that had so far not been observed. The latter techniques start from a reference phylogeny that depicts the evolution of the language family under consideration and then climb the tree up from the leaves to the root in order to find the proto-form that best explains how the observed word forms have developed into their observed shape.

Strictly speaking, the task of supervised phonological reconstruction is identical with the task termed *reflex prediction* by Bodt and List (2022). In this task, one tries to predict word unelicited word forms or morphemes in descendant language from the information available from cognate words in related languages. While reflex prediction deals with a set of extand languages which are assumed to be genetically related, supervised phonological reconstruction predicts only the words in the proto-language but typically uses the same techniques which would be used in order to predict missing reflexes. Quite a few attempts to provide methods for supervised phonological reconstruction or reflex prediction can be found in recent approches to Natural Language Processing and machine learning. However, since the task was never really discussed much among historical linguists, it may be difficult to identify related work, since scholars often differ quite substantially in the terminology they use to describe their approaches. Thus, we find terms like *cognate production* (Beinborn et al. 2013), *cognate prediction* (Fourrier 2021), or *word prediction* (Dekker and Zuidema 2021). Additionally, scholars also differ in the use of the term *cognacy*, with some scholars adopting a definition of cognacy by which only words with similar meanings are judged to be cognates (Beinborn et al. 2013), while other scholars follow the classical definition of *cognacy* as referring to etymologically related words which have been inherited by vertial descent, without borrowings (Fourrier, Bawden, and Sagot 2021). While the former definition is commonly used in research on bilingualism, where scholars are interested in the processing of similar words from different languages by the human brain (Gradoville et al. 2021), it goes contrary to the typical definition of cognacy in historical linguistics, which tends to ignore semantic differences as log as words can be shown to have a common ancestor (Szemerényi 1970; List 2016a).

Beinborn et al. (2013) "produce" cognates from one language by training language models with bilingual lists of cognate words and then applying methods for machine translation to the individual characters by which words are represented in orthography. Ciobanu and Dinu (2018) reconstruct Latin words from multiple Romance descendant languages by applying a complex workflow which employs pairwise alignments for the comparison of each descendant language with Latin in order to produce a language specific classifier that predicts from an input word in the descendant language the output word in Latin. In a second step, multiple predictions for individual Romance languages are then ensembled, using different weighting techniques, in order to yield a combined Latin form. Meloni et al. (2021) train recurrent neural networks to predict Latin proto-forms from an extended datasets on Romance languages, reporting very promising evaluation results that seem to outperform previous methods. Dekker and Zuidema

(2021) use different techniques, including recurrent neural network techniques for pairwise word prediction, and show how they can be used to tackle additional tasks, such as automated cognate detection, or phylogenetic reconstruction. They test their approach on the NorthEuralex database which provides large wordlists for about 100 languages from North Eurasia (Dellert et al. 2020). Fourrier et al. (2021) model *cognate prediction* as a machine translation task, testing various pairwise and multilingual settings, based on bilingual collections of cognate sets from three Romance languages. Bodt and List (2022) predict reflexes from correspondence patterns for a small dataset of 8 Western Kho-Bwa languages, which form a subgroup of the Sino-Tibetan language family. In contrast to other approaches, Bodt and List really tried to predict word forms, since the forms they predicted had not yet been elicited in field work. Using the method for correspondence pattern recognition by List (2019b), and refining the automated predictions manually, Bodt and List first registered the predictions online (Bodt and List 2019), in order to verify them through additional fieldwork later. Their results show that manually refined predictions were superior to automatic predictions, reaching an average accuracy of 76%. List et al. (2022) extend their methodology by providing a framework for supervised phonological reconstruction and reflex prediction that allows to make use of classification methods from machine learning in order to assign a given (proto-)sound to a given correspondence pattern.

While supervised phonological reconstruction can be very useful when it comes to the creation and curation of large etymological datasets, the methodology still heavily relies on expert's previous assessment of cognates and proto-forms. Unfortunately, however, only a few attempts have been made so far to infer proto-forms from cognate sets without relying on previously annotated data. All attempts which have been made so far employ phylogenetic information in order to reconstruct cognate word forms in a set of descendant languages step-by-step back to the proto-language. Bouchard-Côté et al. (2013) apply stochastic transducers to basic words in Oceanic languages (Greenhill et al. 2008) in order to reconstruct word forms in Proto-Oceanic. Comparing their results with expert reconstructions shows that their workflow works very well – at least on Austronesian languages. Jäger (2019) applies ancestral state reconstruction techniques (see Jäger and List 2018) to automatically aligned words from Romance languages taken from the ASJP database (Wichmann et al. 2013). In contrast to the results obtained by Bouchard-Côté et al. (2013), the results are, however, rather disappointing, which may specifically also result from the fact that the results were compared against Latin, rather than Proto-Romance, the direct ancestor of Romance languages. As of now, none of the proposed methods for automated unsupervised phonological reconstruction can deal with the fact that the sounds reconstructed for proto-forms may well not be attested in the descendant languages. While linguists routinely propose sounds that may be lacking in the descendant languages, based on their implicit knowledge of correspondence patterns and sound change processes, computational methods that employ varying techniques of ancestral state reconstruction cannot "invent" new sounds that they have not seen before in the data. In order to overcome this problem in the future, scholars would either have to design methods which learn common sound change processes from training datasets, or they would have to turn to feature representations of sounds.

An additional task relevant for computational historical language comparison, which is typically not mentioned in teh canonical literature on the comparative method, is the task of *inducing sound laws* that explain how proto-forms in the proto-language turn into forms in the descendant languages. As many tasks of the comparative method, *sound law induction* usually goes hand in hand with *phonological reconstruction*: when scholars propose first proto-sounds for the major correspondence patterns which they have inferred from their data, they will usually also check how to explain how these proto-sounds would turn into the target sounds in the descendant languages. Sound laws are typically represented in the form of a replacement rule, consisting of the source sound, the target sound, and the conditioning context, quite similar in appearance and formally identical with the replacement rules used in synchronic phonology (Hall and Klein 2011). Thus a rule like (1) describes that voiceless plosives [p], [t], and [k] become voiced in intervocalic context, with `x > y` marking the change of the source into the target sound, the part after the slash symbol marking the conditioning context, and brackets allowing to group sounds into classes.

```
(1) [p t k] > [b d g] / [a e i o u] _ [a e i o u]
```

Although sound law induction plays a crucial role in historical language comparison, no computational methods which could aid linguists working on the phonological reconstruction of a proto-language have been proposed so far. Phonological reconstruction and sound law induction are still mainly seen as an exclusively manual task that only humans can accomplish in a satisfying manner. It is probably also for this reason that there have also been very few attempts to create systems where scholars could encode their sound laws and apply them directly to their data (Hartmann 2003). As a result, most phonological reconstruction systems, although extremely formal by nature, have never been formally tested on concrete data whether they actually hold what they promise.

While concrete methods for sound law induction have not been developed so far, there have been some attempts to learn sound laws from datasets consisting of proto-forms and their reflexes in the descendant languages with the help of machine learning approaches (Cathcart 2020; Cathcart and Wandl 2020), which are quite similar to the approaches used for supervised phonological reconstruction (including specifically neural network models which process sequences) but try to predict the descendant word forms from the proto-forms instead of predicting proto-forms from descendant forms. While these approaches do not allow to induce concrete sound laws as the one shown above in (1), they can be used to identify word forms whose sound change is hard to model automatically.

5 Reconstruction of Etymological Scenarios

As a final step in the workflow, Ross and Durie (1996) mention the publication of an etymological dictionary. An etymological dictionary does not only consist in the publication of proto-forms and their reflexes in the descendant languages, but should rather be understood as the aggregation of all information on individual word histories which scholars could derive from the investigation of the language family at hand and from external sources. This information can best be labeled as an *evolutionary scenario* or an *etymological scenario*, which one may think of as an individual phylogeny of the word family underlying an ancestral word reconstructed for the proto-language, which we would call an *etymon*.

The etymological scenario would ideally include information on the phylogeny of the language family at hand and combine this information with information on the morphological development of the proto-form and also point to sound change processes which do not follow strict sound laws. As a formal schema used to represent etymological scenarios, we can think of *word family trees and networks* (Schweikhard and List forthcoming), which encode information on morphological processes, idiosyncratic sound change processes, processes of semantic change, and processes of lateral transfer (borrowing). An etymological scenario in this sense is close to the model of lexical filiation proposed by Gévaudan (2007), but while Gévaudan (2007) deliberately excludes sound change, sound change idiosyncracies would still be seen as an important aspect of etymological scenarios in the sense defined here (cf. Schweikhard and List 2020).

While no attempts have been made so far to infer etymological scenarios automatically, we can find attempts to provide automated solutions for subtasks, such as the automated identification of borrowings, or the reconciliation of word trees with language trees in order to find cases in which the overall evolution of a language family would be inconflict with the evolution of individual word families. Methods for automatic borrowing detection have received some attention in the past, mostly because of the parallel with lateral gene transfer in biology, which made the topic interesting for a broader reange of scholars from disciplines like bioinformatics, computational linguistics, and natural language processing.

Automated methods for borrowing detection can be roughly divided into two kinds of approaches, phylogeny-based approaches, and sequence-based approaches (List 2019a). Phylogeny-based approaches try to identify borrowings by detecting conflicts between a set of phylogenetic characters and a given reference phylogeny. They have been tested in a larger range of approaches in different variatns (cf. Nelson-Sathi et al. 2011; List et al. 2014a; List et al. 2014b; Willems et al. 2016), but so far, only mixed results regarding their performance could be reported (List 2019a). Sequence-based approaches on the other hand try to identify borrowings from the direct comparison of words. Thus, Ark et al. (2007), Mennecier et al. (2016), and Zhang et al. (2021) use methods for automatic phonetic alignment on language data from different families, assuming that highly similar words from different language families reflect borrowings. List and Forkel (2022) expand this method by providing a workflow that first searches for cognate sets inside language families and then tries to identify sets of borrowed words across families. Although quite promising, these methods work only when data from different language families are being compared. Hantgan and List (forthcoming) propose a workflow that contrasts different cognate detection methods with each other, one method that tends to search for deep cognates, taking regular sound correspondences into account, and a method that searches for similar words, ignoring regular sound correspondences, but they do not test the method on a gold standard but apply it to investigate the origin of Bangime, a language isolate.

Not all cases in which individual word family evolution conflicts with the overall evolution of a language family are due to borrowing and language contact. Due to different forms of linguistic variation, individual patterns of lexical or morphological change may at times look as if they conflict with the overall branching patterns of a language family, although they are in fact perfectly compatible with it. This phenomenon, called *incomplete lineage sorting* in evolutionary biology, has only recently gained the attention of scholars in historical linguistics (List et al. 2016; Jacques and List 2019), and no quantitative approaches to handle or detect the phenomenon in linguistic data has been proposed so far. One potential way to address phenomena of incomplete lineage sorting or – in a broader sense – various forms in which individual patterns of word family evolution conflict with the overall evolution of a language family, would consist in the application of methods by which word family histories are compared with language histories. In evolutionary biology, scholars have for quite some time applied methods by which a *gene tree* can be reconciled with a *species tree* (Nakhleh 2013). While scholars have discussed the importance of applying similar methods in linguistics (Gray et al. 2007), and some attempts have been made to test the suitability of reconciling *word trees* with *language trees* (Willems et al. 2016), the application of tree reconciliation methods in linguistics is still in its infancy, and results have so far been largely disappointing (Köllner 2021). One of the major problems lies in the analogy drawn between word trees in linguistics and gene trees in evolutionary biology. While current approaches construct word trees with the help of techniques for automatic sequence comparison, assigning cognate and non-cognate words to the same word tree, it seems more fruitful to discuss improved ways to model word *family* evolution, even if these cannot yet be automatically inferred (Schweikhard and List forthcoming).

6 Open Problems

This overview of current computational methods that aid historical language comparison has pointed to quite a few open problems which we hope could be tackled in the future. With respect to the proof of genetic relatedness (§ 2), the major challenge for future research would consist in the concrete comparison of various approaches and techniques which have been proposed in the past. This would require to implement methods proposed in the past in software code and to test the methods on unified data collections which offer a large number of related and unrelated languages of different time depths. With such asystem in place, one could systematically test and compare existing solutions by carrying out a detailed error analysis.

While cognate and sound correspondence pattern detection (§ 3) are already quite developed tasks by now, which have made huge progress over the past two decades, additional challenges relate to the detection of partial cognates even in those cases where no morpheme boundaries are available, and the identification of cognate sets across concept slots. While initial methods for partial cognate detection have been proposed (List 2016a), as well as methods for cross-semantic cognate detection (Arnaud et al. 2017; Wu et al. 2020), the former require morpheme boundaries to be known in advance, while the latter have not yet been rigorously tested.

Many challenges remain to enhance current approaches to phonological reconstruction and sound law induction (§ 4). While supervised methods for phonological reconstruction seem to work fine, unsupervised methods have so far only been applied to language families with reduced sound inventories, and none of the approaches proposed so far is able to reconstruct sounds which are not attested in at least one of the descendant languages. While implicitly used in a couple of approaches that test how well descendant forms can be predicted from proto-forms, sound law induction has so far been largely ignored by the field.

In order to improve the reconstruction of etymological scenarios (§ 5), not only methods for the automated detection of borrowings need to be enhanced, but also new approaches to tree reconciliation in historical linguistics need to be developed. With respect to methods for automated borrowing detection, it would be very interesting to see if it is possible to identify borrowings based on irregular sound correspondences. While this technique is widely used in classical historical language comparison, it has not yet been applied in computational framewords. With respect to tree reconciliation methods, it seems that the first crucial step that linguists would have to make is to find a good linguistic analogon of *gene trees* or *gene family trees* in biology. Only if word trees or word family trees can be defined in a meaningful way, one can start to discuss the application of tree reconciliation methods in comparative linguistics.

7 Conclusion

In this overview, we have tried to briefly outline some of the major aspects of the classical comparative method for historical language comparison, which scholars have tried to automatize in the past years and decades. While many problems are still unsolved, and many challenges remain, it is probably fair to say that quite some progress has been made during the past two decades. Given the speed with which computational methods in historical linguistics advance at the moment, we can expect that quite a few interesting new approaches will be developed in the nearer future.

References

Anttila, Raimo. 1972. *An Introduction to Historical and Comparative Linguistics*. New York: Macmillan.

Ark, René van der, Philippe Mennecier, John Nerbonne, and Franz Manni. 2007. “Preliminary Identification of Language Groups and Loan Words in Central Asia.” In *Proceedings of the RANLP Workshop on Acquisition and Management of Multilingual Lexicons*, 13–20.

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*, 2509–18. Association for Computational Linguistics.

Baxter, William H., and Alexis Manaster Ramer. 1996. “Review: On Calculating the Factor of Chance in Language Comparison. By Donald A. Ringe, Jr. Philadelphia: The American Philosophical Society, 1992. Pp. 110.” *Diachronica* 8 (2): 371–84.

———. 2000. “Beyond Lumping and Splitting: Probabilistic Issues in Historical Linguistics.” In *Time Depth in Historical Linguistics*, edited by Colin Renfrew, April McMahon, and Larry Trask, 167–88. Cambridge: McDonald Institute for Archaeological Research.

Beinborn, Lisa, Torsten Zesch, and Iryna Gurevych. 2013. “Cognate Production Using Character-Based Machine Translation.” In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, edited by Ruslan Mitkov and Jong C. Park, 883–91.

Bender, Emily, and Alexander Koller. 2020. “Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–98. Stroudsberg: Association for Computational Linguistics.

Bhasker, J., and Tariq Samad. 1991. “The Clique-Partitioning Problem.” *Computers & Mathematics with Applications* 22 (6): 1–11. https://doi.org/http://dx.doi.org/10.1016/0898-1221(91)90001-K.

Blevins, Juliette, and Richard Sproat. 2021. "Statistical Evidence for the Proto-Indo- European-Euskarian Hypothesis a Word-List Approach Integrating Phonotactics." *Diachronica* 0 (0): 1–59. https://doi.org/10.1075/dia.19014.ble.

Bodt, Timotheus A., and Johann-Mattis List. 2022. "Reflex Prediction. A Case Study of Western Kho-Bwa." *Diachronica* 39 (1): 1–38.

Bodt, Timotheus A., and Johann-Mattis List. 2019. "Testing the Predictive Strength of the Comparative Method: An Ongoing Experiment on Unattested Words in Western Kho-Bwa Languages." *Papers in Historical Phonology* 4 (1): 22–44.

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 United States of America* 110 (11): 4224–29.

Cathcart, Chundra A. 2020. "A Probabilistic Assessment of the Indo-Aryan Inner–Outer Hypothesis." *Journal of Historical Linguistics* 10 (1): 42086.

Cathcart, Chundra, and Florian Wandl. 2020. "In Search of Isoglosses: Continuous and Discrete Language Embeddings in Slavic Historical Phonology." In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 233–44. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.sigmorphon-1.28.

Ciobanu, Alina Maria, and Liviu P. Dinu. 2013. "Automatic Detection of Cognates Using Orthographic Alignment." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, 99–105.

———. 2018. "Simulating Language Evolution: A Tool for Historical Linguistics." In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 68–72. Association of Computational Linguistics.

Dekker, Peter, and Willem Zuidema. 2021. "Word Prediction in Computational Historical Linguistics." *Journal of Language Modelling* 8 (2): 295–336. https://doi.org/10.15398/jlm.v8i2.268.

Dellert, Johannes. 2018. "Combining Information-Weighted Sequence Alignment and Sound Correspondence Models for Improved Cognate Detection." In *Proceedings of the 27th International Conference on Computational Linguistics*, 3123–33.

Dellert, Johannes, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, et al. 2020. "NorthEuraLex: A Wide-Coverage Lexical Database of Northern Eurasia." *Language Resources & Evaluation* 54: 273–301. https://doi.org/10.1007/s10579-019-09480-6.

Dolgopolsky, Aron B. 1964. "Gipoteza Drevnejšego Rodstva Jazykovych Semej Severnoj Evrazii s Verojatnostej Točky Zrenija." *Voprosy Jazykoznanija* 2: 53–63.

Dybo, Anna, and George S Starostin. 2008. "In Defense of the Comparative Method, or the End of the Vovin Controversy." In *Aspekty Komparativistiki*, edited by I. S. Smirnov, 3:119–258. Moscow: RGGU.

Fourrier, Clémentine, Rachel Bawden, and Benoît Sagot. 2021. "Can Cognate Prediction Be Modelled as a Low-Resource Machine Translation Task?" In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 847–61. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.75.

Fox, Anthony. 1995. *Linguistic Reconstruction: An Introduction to Theory and Method*. Oxford: Oxford University Press.

Gévaudan, Paul. 2007. *Typologie Des Lexikalischen Wandels: Bedeutungswandel, Wortbildung Und Entlehnung Am Beispiel Der Romanischen Sprachen*. Stauffenburg-Linguistik ; 45. Tübingen: Stauffenburg.

Gradoville, Michael, Mark Waltermire, and Avizia Long. 2021. “Cognate Similarity and Intervocalic /d/ Production in Riverense Spanish.” *International Journal of Bilingualism* 25 (3): 727–46. https://doi.org/10.1177/1367006921996807.

Gray, Russell D., Simon J. Greenhill, and Robert M. Ross. 2007. “The Pleasures and Perils of Darwinzing Culture (with Phylogenies).” *Biological Theory* 2 (4): 360–75.

Greenhill, Simon J., Robert Blust, and Russell D. Gray. 2008. “The Austronesian Basic Vocabulary Database: From Bioinformatics to Lexomics.” *Evolutionary Bioinformatics* 4: 271–83.

Hall, David, and Dan Klein. 2010. “Finding Cognate Groups Using Phylogenies.” In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1030–39. Stroudsburg: Association for Computational Linguistics.

———. 2011. “Large-Scale Cognate Recovery.” In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 344–54.

Hantgan, Abbie, and Johann-Mattis List. forthcoming. “Bangime: Secret Language, Language Isolate, or Language Island?” *Journal of Language Contact* 0 (0): 1–27.

Hartmann, Lee. 2003. “Phono. Software for Modeling Regular Historical Sound Change.” In *Actas VIII Simposio Internacional de Comunicación Social*, 606–9. Southern Illinois University. http://mypage.siu.edu/lhartman/phono/ver40.htm.

Hauer, Bradley, and Grzegorz Kondrak. 2011. “Clustering Semantically Equivalent Words into Cognate Sets in Multilingual Lists.” In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, 865–73. AFNLP.

Jacques, Guillaume, and Johann-Mattis List. 2019. “Save the Trees: Why We Need Tree Models in Linguistic Reconstruction (and When We Should Apply Them).” *Journal of Historical Linguistics* 9 (1): 128–66. https://doi.org/https://doi.org/10.1075/jhl.17008.mat.

Jäger, Gerhard. 2019. “Computational Historical Linguistics.” *Theoretical Linguistics* 45 (3-4): 151–82. https://doi.org/10.1515/tl-2019-0011.

Jäger, Gerhard, and Johann-Mattis List. 2018. “Using Ancestral State Reconstruction Methods for Onomasiological Reconstruction in Multilingual Word Lists.” *Language Dynamics and Change* 8 (1): 22–54.

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*, 1204–15. Valencia: Association for Computational Linguistics.

Jarceva, V. N., ed. 1990. *Lingvističeskij Ėnciklopedičeskij Slovar (Linguistical Encyclopedical Dictionary)*. Moscow: Sovetskaja Enciklopedija.

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): 301–47.

Kessler, Brett. 2001. *The Significance of Word Lists: Statistical Tests for Investigating Historical Connections Between Languages*. Stanford: CSLI Publications.

Klimov, Georgij Andreevic. 1990. *Osnovy Lingvističeskoj Komparativistiki*. Moscow: Nauka.

Köllner, Marisa. 2021. “Automatic Loanword Identification Using Tree Reconciliation.” Tübingen: Eberhard-Karls-Universität; Eberhard-Karls-Universität Tüpingen. https://repositorium.ixtheo.de/xmlui/bitstream/handle/10900/118123/dissertation_köllner.pdf.

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*, 288–95.

———. 2002. “Determining Recurrent Sound Correspondences by Inducing Translation Models.” In *Nineteenth International Conference on Computational Linguistics (COLING 2002)*, 488–94. Taipei. http://www.cs.ualberta.ca/~kondrak/papers/cic03.pdf.

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*, 117–25. Stroudsburg.

———. 2014. *Sequence Comparison in Historical Linguistics*. Düsseldorf: Düsseldorf University Press.

———. 2016a. “Beyond Cognacy: Historical Relations Between Words and Their Implication for Phylogenetic Reconstruction.” *Journal of Language Evolution* 1 (2): 119–36. https://doi.org/10.1093/jole/lzw006.

———. 2016b. “Computer-Assisted Language Comparison: Reconciling Computational and Classical Approaches in Historical Linguistics.” Jena: Max Planck Institute for the Science of Human History. https://doi.org/10.5281/zenodo.842734.

———. 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* 45 (1): 137–61. https://doi.org/10.1162/coli_a_00344.

List, Johann-Mattis, and Robert Forkel. 2022. “Automated Identification of Borrowings in Multilingual Wordlists [Version 3; Peer Review: 4 Approved].” *Open Research Europe* 1 (79): 1–11. https://doi.org/10.12688/openreseurope.13843.3.

List, J.-M., N. Hill, and R. Forkel (2022): “A new framework for fast automated phonological reconstruction using trimmed alignments and sound correspondence patterns.” In: *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*. Association of Computational Linguistics.

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.

List, Johann-Mattis, Philippe Lopez, and Eric 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)*, 599–605. Berlin: Association of Computational Linguistics.

List, Johann-Mattis, Shijulal Nelson-Sathi, Hans Geisler, and William Martin. 2014a. “Networks of Lexical Borrowing and Lateral Gene Transfer in Language and Genome Evolution.” *Bioessays* 36 (2): 141–50.

List, Johann-Mattis, Shijulal Nelson-Sathi, William Martin, and Hans Geisler. 2014b. “Using Phylogenetic Networks to Model Chinese Dialect History.” *Language Dynamics and Change* 4 (2): 222–52.

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.

Meillet, Antoine. (1925) 1954. *La Méthode Comparative En Linguistique Historique*. Reprint. Paris: Honoré Champion.

Meloni, Carlo, Shauli Ravfogel, and Yoav Goldberg. 2021. “Ab Antiquo: Neural Proto-Language Reconstruction.” In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4460–73. Online: Association for Computational Linguistics.

https://www.aclweb.org/anthology/2021.naacl-main.353.

Mennecier, Phillipe, John Nerbonne, Evelyne Heyer, and Franz Manni. 2016. “A Central Asian Language Survey.” *Language Dynamics and Change* 6 (1): 57–98. https://doi.org/10.1163/22105832-00601015.

Miller, John E., Tiago Tresoldi, Roberto Zariquiey, César A. Beltrán Castañón, Natalia Morozova, and Johann-Mattis List. 2020. “Using Lexical Language Models to Detect Borrowings in Monolingual Wordlists.” *PLOS One* 15 (12): e0242709.

Mortarino, Cinzia. 2004. “A Statistical Test Useful in Historical Linguistics.” In *Proceedings of the XLII Scientific Meeting of the Italian Statistical Society*, 107–10.

———. 2009. “An Improved Statistical Test for Historical Linguistics.” *Statistical Methods and Applications* 18 (2): 193–204.

Nakhleh, Luay. 2013. “Computational Approaches to Species Phylogeny Inference and Gene Tree Reconciliation.” *Trends in Ecology and Evolution* 28 (12): 719–28.

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.

Nichols, Johanna. 1996. “The Comparative Method as Heuristic.” In *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*, edited by Mark Durie, 39–71. New York: Oxford University Press.

Prokić, Jelena, and John Nerbonne. 2013. “Analyzing Dialects Biologically.” In *Classification and Evolution in Biology, Linguistics and the History of Science. Concepts – Methods – Visualization*, edited by Heiner Fangerau, Hans Geisler, Thorsten Halling, and William Martin, 147–61. Stuttgart: Franz Steiner Verlag.

Rama, Taraka. 2016. “Chinese Restaurant Process for Cognate Clustering: A Threshold Free Approach.” *ArXiv e-Prints*. https://arxiv.org/abs/1610.06053.

Rama, Taraka, Prasant Kolachina, and Sudheer Kolachina. 2013. “Two Methods for Automatic Identification of Cognates.” In *Proceedings of the 5th Conference on Quantitative Investigations in Theoretical Linguistics*, edited by Thomas Wielfaert, Kris Heylen, and Dirk Speelman, 76–80. University of Leuven.

Rama, Taraka, and Johann-Mattis List. 2019. “An Automated Framework for Fast Cognate Detection and Bayesian Phylogenetic Inference in Computational Historical Linguistics.” In *57th Annual Meeting of the Association for Computational Linguistics*, 6225–35. Association for Computational Linguistics.

Ringe, Donald A. 1992. “On Calculating the Factor of Chance in Language Comparison.” *Transactions of the American Philosophical Society*, New series, 82 (1): 1–110. https://www.jstor.org/stable/1006563.

Ross, Malcom, and Mark Durie. 1996. “Introduction.” In *The Comparative Method Reviewed. Regularity and Irregularity in Language Change*, edited by Mark Durie, 3–38. New York: Oxford University Press.

Rzymski, Christoph, Tiago Tresoldi, Simon Greenhill, Mei-Shin Wu, Nathanael E. Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, et al. 2020. “The Database of Cross-Linguistic Colexifications, Reproducible Analysis of Cross- Linguistic Polysemies.” *Scientific Data* 7 (13): 1–12.

Schweikhard, Nathanael E., and Johann-Mattis List. forthcoming. “Modeling Word Trees in Historical Linguistics. Preliminary Ideas for the Reconciliation of Word Trees and Language Trees.” In *Sprach(en)forschung: Disziplinen Und Interdisziplinarität. Akten Der 27. Fachtagung Der Gesellschaft Für Sprache Und Sprachen*, edited by Bela Brogyanyi and Reiner Lipp. Hamburg: Dr. Kovač.

———. 2020. “Developing an Annotation Framework for Word Formation Processes in Comparative Linguistics.” *SKASE Journal of Theoretical Linguistics* 17 (1): 2–26.

Sokal, Robert. R., and Charles. D. Michener. 1958. “A Statistical Method for Evaluating Systematic Relationships.” *University of Kansas Scientific Bulletin* 28: 1409–38.

“Statuts: Approuvés Par Décision Ministérielle Du 8 Mars 1866.” 1871. *Bulletin de La Société de Linguistique de Paris* 1: III–IV.

Steiner, Lydia, Peter F. Stadler, and Michael Cysouw. 2011. “A Pipeline for Computational Historical Linguistics.” *Language Dynamics and Change* 1 (1): 89–127.

Swadesh, Morris. 1955. “Towards Greater Accuracy in Lexicostatistic Dating.” *International Journal of American Linguistics* 21 (2): 121–37. https://www.jstor.org/stable/1263939.

Szemerényi, Oswald. 1970. *Einführung in Die Vergleichende Sprachwissenschaft*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Turchin, Peter, Ilja Peiros, and Murray Gell-Mann. 2010. “Analyzing Genetic Connections Between Languages by Matching Consonant Classes.” *Journal of Language Relationship* 3: 117–26.

Weiss, Michael. 2015. “The Comparative Method.” In *The Routledge Handbook of Historical Linguistics*, edited by Claire Bowern and Nicholas Evans, 127–45. New York: Routledge.

Wichmann, Søren, André Müller, Annkathrin Wett, Viveka Velupillai, Julia Bischoffberger, Cecil H. Brown, Eric W. Holman, et al. 2013. *The ASJP Database* (version 16). Leipzig: Max Planck Institute for Evolutionary Anthropology. http://email.eva.mpg.de/~wichmann/languages.htm.

Willems, Matthieu, Etienne Lord, Louise Laforest, Gilbert Labelle, François-Joseph Lapointe, Anna Maria Di Sciullo, and Vladimir Makarenkov. 2016. “Using Hybridization Networks to Retrace the Evolution of Indo-European Languages.” *BMC Evolutionary Biology* 16 (1): 1–18. https://doi.org/10.1186/s12862-016-0745-6.

Wu, Mei-Shin, Nathanael E. Schweikhard, Timotheus A. Bodt, Nathan W. Hill, and Johann-Mattis List. 2020. “Computer-Assisted Language Comparison. State of the Art.” *Journal of Open Humanities Data* 6 (2): 1–14.

Zhang, Liqin, Ray Fabri, John Nerbonne, and John Nerbonne. 2021. “Detecting Loan Words Computationally.” In *Contact Language Library*, 269–88. John Benjamins Publishing Company. https://doi.org/10.1075/coll.59.11zha.