Computer-Assisted Language Comparison in Practice

Tutorials on Computational Approaches to the History and Diversity of Languages

Foundations of Formal Etymological Analysis

Johann-Mattis List
Chair for Multilingual Computational Linguistics
University of Passau

This study gives a brief overview on formal aspects of etymological analysis, by providing a modified workflow for the classical comparative method in historical language comparison. This workflow is contrasted with the current state-of-the-art in computational historical linguistics, pointing out where computational methods and interactive tools for annotation are lacking, and where they are available already.

1 Introduction

It has been more than 10 years now that my dissertation was published, in which I presented a computational method that was supposed to detect cognates in comparative wordlists, based on an automated analysis of potential sound correspondences (List 2014). Now, ten years later, I know much more about the problem of cognate detection in specific and etymological analysis in general, but I still do not feel that the problem can be considered as solved. However, despite the slow progress in my personal and my colleagues' attempts to automatize the comparative method, I see the general workflow that one should follow in order to carry out a formal etymological analysis, much clearer now. For this reason, I thought it is time to share my views on this workflow, in order to also provide a detailed account on those parts, where we still lack full-fledged automated approaches, while at the same time pointing to those aspects that might become relevant in the future.

Some of these problems have already been mentioned as part of an essay published two years ago (List 2024a), in which I discussed several open problems in the field of comparative linguistics and historical language comparison. The current attempt to summarize the major workflows involved in formal etymological analysis can be seen as an attempt to follow up on this work, by contrasting the solutions we have with the solutions we need.

Since the workflow for etymological analysis involves several steps that consist themselves of at times complex sub-tasks, I will restrict myself here to presenting the

Computer-Assisted Language Comparison in Practice
Volume 9, Number 1
URL: https://calc.hypotheses.org/9208 DOI: 10.15475/calcip.2026.1.3

Published on 25/03/2026

general workflow for etymological analysis. In later studies, I may zoom in to address particular problems in more detail.

2 Background

The central workflow of the comparative method has been discussed in many studies. A very influential workflow can be found in the study of Ross (1996) (pages 6f). Inspired by this workflow, I have proposed a five-step approach (see Figure 1), starting from the (1) proof of relationship, followed by the (2) identification of cognates, and the (3) identification of sound correspondences, and cumulating in the (4) reconstruction of proto-forms and (5) an internal classification of the language family in question (List 2014, 58).

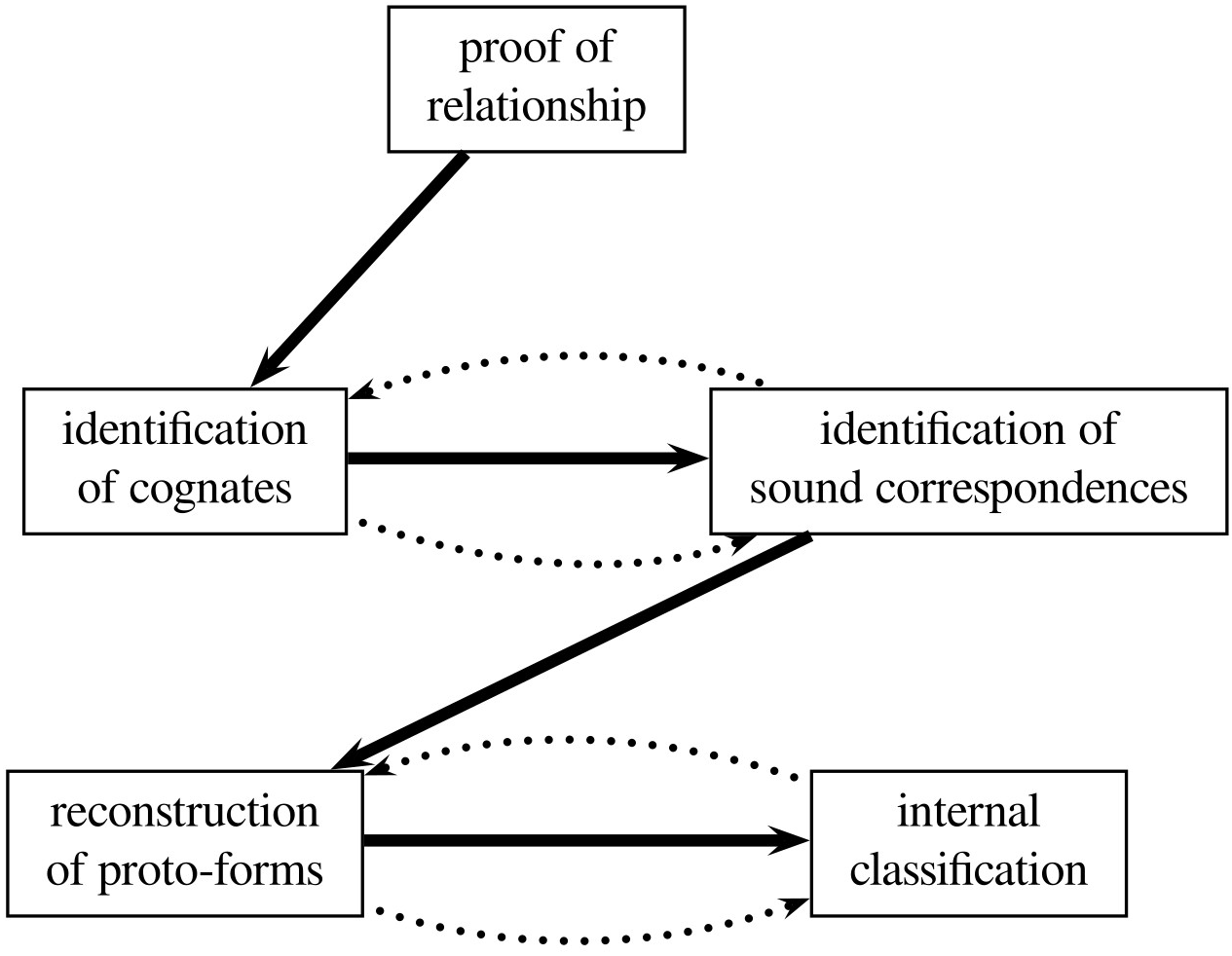

Figure 1: Workflow of the traditional comparative method based on Ross and Durie (1996) taken from List (2014).

It was based on this workflow that I later tried to automatize central parts of the comparative methods. A first step in this direction was my work on phonetic alignment and cognate detection

(List 2014), which can be assigned to step (2). My work on correspondence patterns (List 2019) can be seen as central for step (3). We have used this work to build initial methods for the supervised reconstruction of proto-forms and for the prediction of reflex forms (Bodt and List 2022; List, Hill, and Forkel 2022).

While these methods are all fully automated, I have also tried to establish formal tools that allow to conduct these stages of the workflow manually, but in a way that would account for a formal – computer-readable – treatment of the individual steps when applying them to a standardized dataset. Here, most ideas went into the EDICTOR tool (https://edictor.org, List 2017), that has since then been constantly enhanced (List and Dam 2024, List et al. 202).

3 Workflow

When working on fully automated workflows for different aspects of the comparative method, I soon realized that none of the methods would actually yield the result that I had originally hoped for. Instead of solving a problem once and for all time, each solution would point to additional problems. This has not changed even today, even if I think that we have made great progress in the retrospective.

Right now, however, I would revise the workflow that I had proposed back in 2014 and expect much smaller steps to be carried out – often also in an iterative fashion.

1. initial data preparation -> convert data to standard phonetic transcriptions
2. internal cognate analysis -> segment words into morphemes, carry out internal reconstruction and allomorphic and allophonic analysis, identify language-internal cognates, gloss individual morphemes for their particular meaning
3. external cognate analysis -> identify cognate morphemes, align them, identify potential exceptions
4. external correspondence analysis -> identify correspondence patterns, refine them, mark exceptions
5. linguistic reconstruction -> assign proto-forms to correspondence patterns
6. sound law induction -> determine the sound laws that lead from a proto-form to the individual language varieties
7. phylogenetic reconstruction -> determine major subgroupings of the languages

Most of these steps have been discussed already by me in past studies, but I figured I have so far never found the time to clarify in a much more explicit manner, how I think that these different approaches could or should be best combined to form a full-fledged workflow for historical language comparison. We can arrange the seven steps in the manner shown in Figure 2.

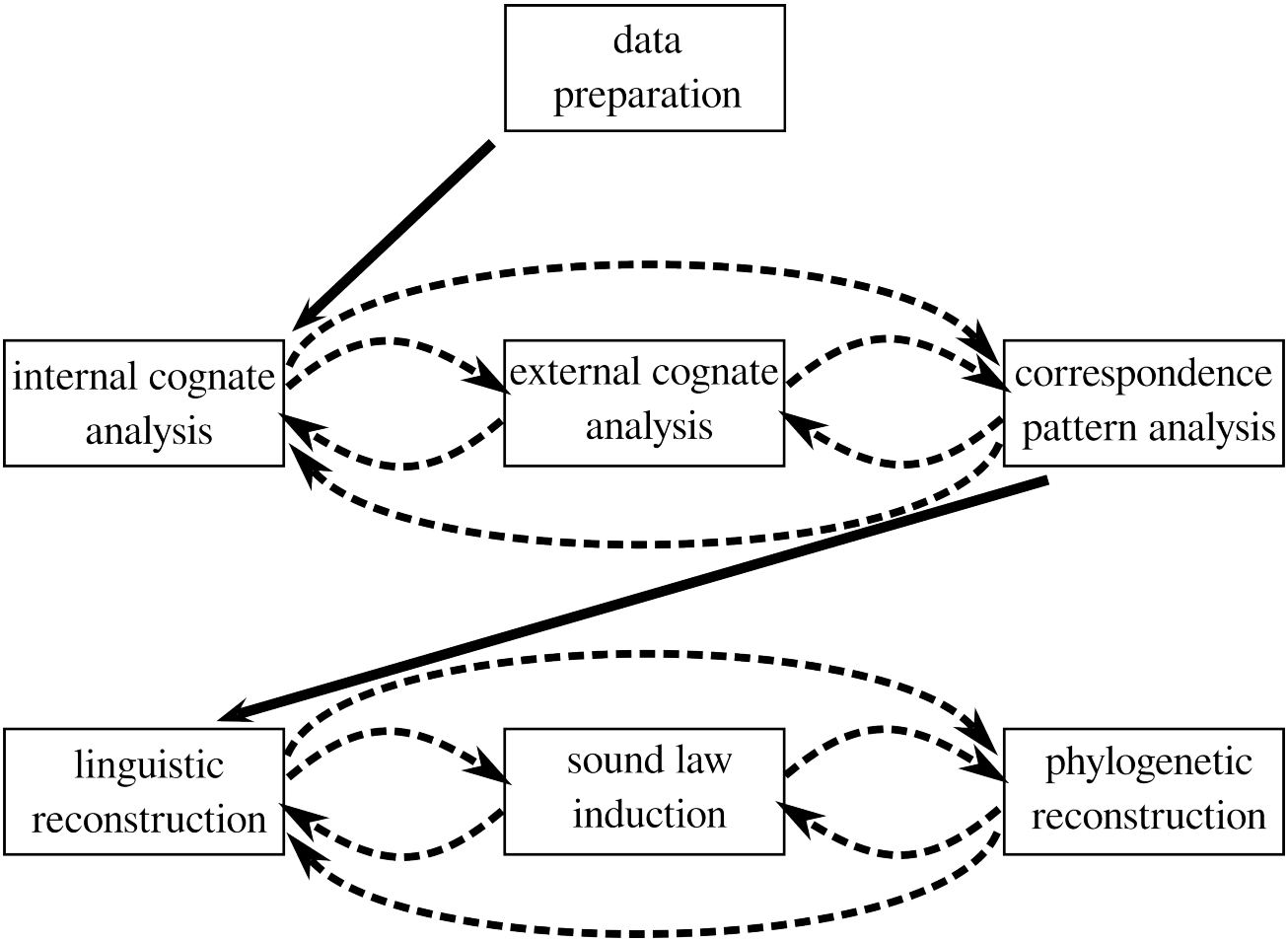

Figure 2: Revised workflow for historical language comparison.

This workflow keeps the major division into two phases of the workflow that I proposed earlier, with one phase being devoted to the identification of cognates and correspondences and one being devoted to the identification of proto-forms and subgroupings. However, both steps are now extended by another step. The identification of cognate sets and correspondence patterns is now extended by what one could best call internal reconstruction. The phonological and phylogenetic reconstruction stage is no accompanied by a detailed search for sound laws. All three subtasks must be solved – as before – in an iterative fashion during which researchers or algorithms jump back and forth through different representations of the original data.

4 Workflow Examples

In the following, I will try to provide more detailed examples on this revised workflow, pointing both to software solutions and to interface and modeling solutions that can be used to address the individual tasks constituting the two major phases of historical language comparison. Before we discuss the phases of cognate and correspondence detection and phonological and phylogenetic reconstruction, however, I will briefly try point to the first stage of data preparation, which is crucial for any kind of linguistic

analysis, but all to often, unfortunately, treated with much less care, both by researchers and by computational solutions.

4.1 Data Preparation

Data preparation constitutes a crucial but also frequently overlooked aspect of formal etymological analysis. Thus, in order to allow for an automated handling of word forms as sound sequences we must insist that sounds are represented in standardized phonetic transcriptions in which each sound is clearly defined. Transcription systems, however, are rarely clearly defined and rely on ad-hoc extensions by scholars who feel that the transcription lacks the sounds they want to transcribe or who simply ignore the rules. Since rules are never explicitly imposed upon transcriptions and also rarely checked, this has led to a considerable diversity of transcription practices and also implementations, as can be seen from the variation underlying cross-linguistic databases and data collections (Anderson et al. 2018, 2023).

With the Cross-Linguistic Data Formats initiative (CLDF, https://cldf.clld.org), we have tried to lay the foundation of standardized datasets in cross-linguistic applications (Forkel et al. 2018). When developing the recommendations that were later adapted in individual CLDF versions, we relied on practical experience resulting from concrete attempts to handle cross-linguistic data both manually – trying to compile new datasets – and automatically – trying to analyze existing datasets. These recommendations, however, were not necessarily developed with CLDF as endpoint in mind, but rather reflect practical requirements for cross-linguistic data that we observed in practice.

Based on these observations, we would now require that a common cross-linguistic dataset, compiled for etymological analysis, has the form of a comparative wordlist provided in long table format. This means, that we work with at least one table that stores the data, with each row reflecting an individual word form (Form) in a individual language variety (Language) that represents the translation of an individual concept (Concept).

The word form itself can be represented in three different ways. In the form of the original value (Value) that one may observe in a dictionary (containing additional non-standard information, like brackets, but also mixed entries in which several forms are provided, separated by slashes or decorated with additional ad-hoc annotations), in the form of the actual form (Form) that one wants to use as the basis of an analysis, and in segmented form (Segments), fully transcribed, with individual sounds being segmented by spaces and morpheme boundaries (including word boundaries) inside the form being marked by a plus symbol (+). While the form may contain idiosyncratic transcriptions, the segmented form in CLDF must account for standard transcription systems, including the B(road) IPA defined by the catalog of Cross-Linguistic Transcription Systems (https://clts.clld.org; Anderson et al. 2018).

This final step of analysis, the actual segmentation thus not only segments a given form of idiosyncratic transcriptions into individual speech sounds, it also standardizes speech sounds. This step of converting forms as they are usually provided in the literature into standardized segmented representations is usually done in a semi-automated fashion with the help of Orthography Profiles (Moran and Cysouw 2018, see also Forkel and List 2024), but could in theory also be done manually.

Depending on the language family in question, the segmentation of word forms into speech sounds is not enough, and additional segmentations of sound sequences into sequences representing individual meaning-bearing units – morphemes – are required. While this step might fall inside the stage of data preparation, practical experience shows that it is often much easier to combine this step with the internal cognate analysis that will be discussed in more detail in the following section.

4.2 Cognate and Correspondence Detection

Starting with Swadesh's work on lexicostatistics (Swadesh 1952, 1955), the compilation of cognate-annotated comparative wordlists has become a common practice in historical linguistics. In a comparative wordlist, words from different languages are arranged according to the concepts they express and later compared for cognacy.

An alternative to this annotation practice consists in the collection of etymologically related words regardless of the meanings they express. The most common format for such collections is the etymological dictionary, which adds hypothetical proto-forms to all previously identified cognate sets.

Both representations have advantages and disadvantages. Etymological dictionaries suffer from the fact that they only pick positive evidence for genetic language relations. Meanings of individual word forms – reflexes of proto-forms in the classical terminology – are usually insufficiently reported and barely standardized, often, scholars even strip off suffixes and prefixes from the reflex forms, making it very difficult to verify the origins of the entries that made it into the etymological dictionary.

Comparative wordlists have the disadvantage of missing out on details. Semantic shift is a common phenomenon, but rarely reported in comparative wordlists. The annotation of cognate sets that goes beyond the concept slots that constitute their organisational basis can be tedious in practice. Furthermore, due to the original selection of a limited number of basic concepts, serving as the main comparanda, many potentially interesting cognate sets are deliberately ignored.

While I understand the arguments of both sides, I see much better chances in enhancing the annotation of comparative wordlists in order to make up for their shortcomings than in trying to address the shortcomings of etymological dictionaries. As a result, the basic format that I consider as relevant for etymological analysis is always

the comparative wordlist that consists of a list of concepts that are in turn translated into a list of languages.

In order to identify cognate sets and sound correspondences in such wordlists, three distinct analysis steps are important, which are deeply intertwined with each other.

The first step consists in the detailed annotation of individual word forms in individual language varieties, with the goal of segmenting the word forms into individual morphemes, thereby identifying allophones and allomorphs and carrying out what is known as an internal reconstruction of the individual language varieties.

The second step consists in the identification of cognate morphemes, which have to be aligned with each other after initial identification. This step is pretty well explored, but quite a few problems remain, especially when carrying out the second step without paying attention to the first step of language-internal analysis.

The third step consists in the identification of correspondence patterns. Here, one must cluster all identical alignment sites that have been identified in step 1 and 2 of the cognate and correspondence detection workflow into partitions that show no conflicts with each other. Conflicts are those cases where one language shows two reflexes for the same correspondence pattern.

Step 1 is typically ignored, but it is of crucial importance, since internal reconstruction decreases language-internal variation (Fox 1995, pp. 211-213), and can thus be seen as an important pre-requisite for any external comparison of word forms. Thus, when comparing word forms in Germanic languages, it would not make much sense to use two sounds [ç] and [x] to represent the grapheme <ch> in German, since both pronunciations can be seen als allophones of the same phoneme. Retaining their phonetic pronunciations without prior language-internal analysis would only increase the variation when trying to identify sound correspondences across Germanic languages. The same holds for allomorphic variation that can be traced to language-internal processes. Language-internal variation must be reduced before comparing one language with another, in order to reduce the complexity of the external comparison.

Another important aspect of language-internal comparison, that is often disregarded, consists in the identification of language-internal cognates or word families. Since our basic workflow does not start from previously identified cognate sets, but rather from a comparative wordlist, in which word forms are selected due to the meanings they express, we may encounter many cases in which a morpheme recurs across many word forms. For this reason, a proper internal analysis should not only identify allophones, allomorphs, and morpheme boundaries, but also tag those cases where morphemes recur across word forms, making sure that morphemes are only compared once across languages.

The second step of the workflow, the actual identification of cognate morphemes through external language comparison is pretty well explored, with many attempts of

automation and formalization. For me, the essential part of this stage consists in three individual tasks, the actual partitioning of morphemes into a first set of partial cognate sets (see List et al. 2016 for details on this notion), the identification of cognates recurring across meaning slots in the wordlist (in Wu et al. 2020, we have called this cross-semantic cognate detection at times, but I am hesitant of using the term nowadays), and the consecutive multiple phonetic alignment of these morphemes (see List 2014 on multiple phonetic alignment). In order to avoid stretching the size of this study too much, I won't comment too much on both steps, but emphasize that I consider them as actually solved for those cases where high-quality comparative wordlists with morpheme-segmented word forms resulting from a thorough internal comparison are provided. Tools like EDICTOR (List et al. 2025; List and van Dam 2024) facilitate this step greatly.

Once morphemes have been assigned to individual cognate sets and consecutively aligned morphologically, we arrive at the third step of the workflow, which requires us to partition the alignment sites, that is, the columns of all alignments reflecting all cognate sets, into correspondence patterns. While individual alignments reflect true correspondences, provided one trusts that the alignments have been carried out properly, thee alignments do not reflect patterns, given that they are but one instance of a correspondence. In order to count as a pattern, a larger number of alignment sites is needed, that must confirm the recurrence of the correspondences observed.

Although they play a crucial role in the traditional literature, correspondence patterns have been ignored for a long time in computational approaches. In 2019, I presented a first workflow that helps to derive correspondence patterns from aligned cognates automatically (List 2019). While this workflow was an eye-opener for myself, helping us to address several additional tasks, such as supervised phonological reconstruction (List et al. 2022) and reflex prediction (Bodt and List 2022), it has not received much attention beyond the work of my own group. The reason may have been that the workflow involves larger amounts of data that are difficult to be handled manually, thus making the actual verification of findings in the interaction with classical linguistics rather difficult.

With EDICTOR 3, new functionalities have been made available that help scholars to carry out the annotation of correspondence patterns in a computer-readable but manual manner (List and van Dam 2024). Unfortunately, we lack clear-cut examples in which actual data have been analyzed from scratch. In practice, it has also turned out to very difficult to account for the different layers of annotation that a full-fledged etymological analysis requires.

While for me it is pretty clear that a very detailed annotation is needed that accounts for the three major aspects and multiple smaller problems involved in all individual steps, we currently lack the full-fledged examples where this annotation has been applied to a larger dataset. I hope to be able to provide such examples in the future. Without such

examples, I fear, it will be very difficult to enhance both the interactive annotation tools and the algorithms that could help to automatize particular tasks.

4.3 Phonological and Phylogenetic Reconstruction

The third and in the present workflow final phase consists again of three steps that are deeply intertwined with each other. While it may be clear that the first step of linguistic (or phonological) reconstruction and the second step of sound law induction should go hand in hand, I see a similar need to integrate the final step of phylogenetic reconstruction with the previous steps, given that phylogenies play a crucial role in informing the reconstruction of proto-sounds and the induction of sound laws.

The first step, linguistic (or more precisely phonological) reconstruction can be seen as an additional clustering or partitioning operation applied this time to the correspondence patterns. Correspondence patterns show – in the definition of the term that I have used in my own work – for each language only one reflex sound. This means that conflicts in inferred correspondence patterns must be flagged as exceptions. Exceptions however, could in theory be resolved as a later stage, especially when they show a high degree of systematicity, as we know from Grimm's studies (Grimm 1822) that were later resolved – among others – by Verner (1877).

While exceptions in correspondence patterns can be considered a specific case, we can also easily think of correspondence patterns that differ strong enough so that we would not merge them (and accepting multiple reflex sounds for the same language variety). This does not, however, mean, that these patterns must reflect different proto-sounds. We know very well that conditioning context can easily trigger different reflex sounds in sound change. As a result, two or more distinct correspondence patterns can often be expected to correspond to the same sound in the proto-language.

Carrying out a phonological reconstruction analysis thus has two consequences. First, we assign each correspondence pattern in our data to a certain proto-sound of which we assume that it was present in the proto-language. Second, by assigning identical proto-sounds to different correspondence patterns, we set the stage for the identification of conditioning context that explains the differences in the reflexes of individual language varieties. Formally, we can thus say that we cluster or partition correspondence patterns into proto-sounds.

Once this step has been carried out, we are left with distinct proto-forms for each aligned cognate sets in our data. We must now identify the conditioning contexts that explain the split of a proto-sound into two or more reflex sounds in the same language variety. I call this step sound law induction (see also List 2024b), since sound laws provide a concrete explanation for the change of a sound based on conditioning context. The basic task of the second step in phonological and phylogenetic reconstruction thus

consists in the identifification of those contexts (or more broadly speaking those sound laws) that explain why the same proto-form splits into different reflexes.

Since splits and mergers (mergers are less relevant for us in this context, since they have no consequences on the reconstruction practice) can occur at any stage during language evolution, it would be wrong to assign their occurrence only to the individual development of a language variaty after having split off its subgroup. Instead, we must assume that splits and mergers can happen throughout the whole evolution of a language family, on each internal node that a phylogenetic reconstruction might propose. As a result, it is crucial to substantiate any findings regarding the proto-phonology and the sound laws that constitute a language family by providing a detailed phylogenetic reconstruction that explains how the proto-language split into branches and individual language varieties over time.

While I have initial ideas of how phonological reconstruction can be carried out formally, based on previously identified corresponcence patterns (see List et al. 2022), as well as an initial idea of modeling sound laws computationally (List 2024b), I have neither concrete ideas for the induction of sound laws from comparative data, nor for the consistent integration of phylogenetic reconstruction with the other two steps of the workflow. I am, however, convinced, that a successfully implemented model of etymological analysis needs to account for all three steps in this stage of the workflows. Future research will show how well these ideas aspects can be integrated with each other.

5 Outlook

This little study has not provided any answers to currently open questions in historical language comparison. What I wanted to provide instead was an overview on the detailed workflow of the comparative method that I consider as fruitful to pursue in the future. I am sure that this overview is by no means the last word on the matter. On the contrary, I expect that the workflow will be enriched by more details with time. It is also quite possible that different language families will require different treatments of certain aspects. In any case, I have the hope that future studies will not only help to improve the current state of the art with respect to the conceptualization of the workflow, but also with respect to the implementation of interactive annotation tools and automated workflows that help to make human annotation more efficient.

References

Anderson, Cormac, Tiago Tresoldi, Thiago Costa Chacon, Anne-Maria Fehn, Mary Walworth, Robert Forkel, and Johann-Mattis List. 2018. “A Cross-Linguistic Database of Phonetic Transcription Systems.” Yearbook of the Poznań Linguistic Meeting 4 (1): 21–53. https://doi.org/10.2478/yplm-2018-0002.

Anderson, Cormac, Tiago Tresoldi, Simon J. Greenhill, Robert Forkel, Russell D. Gray, and Johann-Mattis List. 2023. “Variation in Phoneme Inventories: Quantifying the Problem and Improving Comparability.” Journal of Language Evolution 8 (2): 149–68. https://doi.org/10.1093/jole/lzad011.

Bodt, Timotheus Adrianus, and Johann-Mattis List. 2022. “Reflex Prediction. A Case Study of Western Kho-Bwa.” Diachronica 39 (1): 1–38. https://doi.org/10.1075/dia.20009.bod.

Forkel, Robert, and Johann-Mattis List. 2024. “A New Python Library for the Manipulation and Annotation of Linguistic Sequences.” Computer-Assisted Language Comparison in Practice 7 (1): 17–23. https://doi.org/10.15475/calcip.2024.1.3.

Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. “Cross-Linguistic Data Formats, Advancing Data Sharing and Re-Use in Comparative Linguistics.” Scientific Data 5 (180205): 1–10. https://doi.org/10.1038/sdata.2018.205.

Grimm, Jacob. 1822. Deutsche Grammatik. 2nd ed. Vol. 1. Göttingen: Dieterichsche Buchhandlung.

List, Johann-Mattis. 2014. Sequence Comparison in Historical Linguistics. Düsseldorf: Düsseldorf University Press. https://doi.org/10.1515/9783110720082.

———. 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. Valencia: Association for Computational Linguistics. https://edictor.org.

———. 2019. “Automatic Inference of Sound Correspondence Patterns Across Multiple Languages.” Computational Linguistics 45 (1): 137–61. https://doi.org/10.1162/coli_a_00344.

———. 2024a. “Modeling Sound Change with Ordered Layers of Simultaneous Sound Laws.” [Preprint, not peer-reviewed]. Humanities Commons. https://doi.org/10.17613/4n5z-9y52.

———. 2024b. “Open Problems in Computational Historical Linguistics [Version 2; Peer Review: 5 Approved].” Open Research Europe 3 (201): 1–27. https://doi.org/10.12688/openreseurope.16804.2.

List, Johann-Mattis, and Kellen Parker van Dam. 2024. “Computer-Assisted Language Comparison with EDICTOR 3 [Invited Paper].” In Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change, edited by Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, David Alfter, Francesco Periti, Pierluigi Cassotti, and Netta Huebscher, 1–11. Bangkok, Thailand: Association for Computational Linguistics. https://aclanthology.org/2024.lchange-1.1.

List, Johann-Mattis, Nathan W. Hill, and Robert 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, 89–96. Dublin: Association for Computational Linguistics. https://aclanthology.org/2022.lchange-1.9.

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. https://anthology.aclweb.org/P16-2097.

List, Johann-Mattis, Kellen Parker van Dam, and Frederic Blum. 2025. EDICTOR 3. An Interactive Tool for Computer-Assisted Language Comparison [Software Tool, Version 3.1]. Passau: MCL Chair at the University of Passau. https://edictor.org.

Moran, Steven, and Michael Cysouw. 2018. The Unicode Cookbook for Linguists: Managing Writing Systems Using Orthography Profiles. Berlin: Language Science Press. https://langsci-press.org/catalog/book/176.

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

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): 452–63.

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

Verner, Karl A. 1877. “Eine Ausnahme Der Ersten Lautverschiebung.” Zeitschrift Für Vergleichende Sprachforschung Auf Dem Gebiete Der Indogermanischen Sprachen 23 (2): 97–130.

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. https://doi.org/10.5334/johd.12.

Funding Information

This project has received funding from the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (Grant agreement No. 101044282). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.