Computational Historical Linguistics

Lecture, given at the LOT Winter School 2023 (Netherlands National Graduate School of Linguistics, 16-20 January, Amsterdam)

Johann-Mattis List
johann-mattis.list@uni-passau.de

Chair of Multilingual Computational Linguistics
University of Passau

January 2023

Contents

Lecture 1: Getting Started

Abstract

The goal of this session is to discuss briefly the basic aspects of the research that has been carried out in the field of computational historical linguistics since the quantitative turn.

1 The Quantitative Turn in Historical Linguistics

1.1 Background

In the early 1950s, Morris Swadesh (1909–1967) presented a method to measure the genetic closeness between languages on the basis of a statistical formula that was ultimately based on counting the amount of shared cognates across standardized wordlists of different languages (Swadesh 1950). Although it seemed at first that the methods could revive the discipline of historical linguistics, which had past its prime after the structuralist turn in the begin of the 1920s , and had not seen any major methodological or analytical improvement since the begin of the 20th century.[1] Unfortunately, the original interest in the new ideas did not last long, and soon after it was first published, the new method was heavily criticized (Bergsland and Vogt 1962), and went out of vogue some 10 years later.

In the begin of the second millennium, Gray and Atkinson (2003) used similar data but different statistical methods to date the age of the Indo-European language family. They caused a similar stir as Swadesh had done almost half a century ago. But while Swadesh's method was filed away soon after it had been proposed, the method of Gray and Atkinson was part of a general *quantitative turn in historical linguistics*, which started at the begin of the second millennium. This quantitative turn is reflected in a large bunch of literature on such different topics as phonetic alignment (Kondrak 2000, Prokić et al. 2009), automated cognate detection (List 2014), and phylogenetic reconstruction (Atkinson and Gray 2006).

What may have been the reasons why Swadesh's approach was abandoned so quickly by historical linguists?

1.2 New Studies on Language Evolution

We can distinguish four different aspects of research approaches in the course of the quantitative turn. As a first and most prominent aspect, we have research dealing with questions of *phylogenetic reconstruction* which usually involved *dating* as well. Language data are not only analyzed to yield a topology of the branching structure of the language family in question, but in addition, absolute branch lengths are often also inferred, which allow to estimate when a given language family has originated. The software and methods used for these studies are usually taken or inspired from approaches developed first in evolutionary biology. As of now, quite a few different language families have been analyzed in this way, including Indo-European (Chang et al. 2015, Gray and Atkinson 2003), Austronesian (Gray et al. 2009), Dravidian (Kolipakam et al. 2018), Bantu (Grollemund et al. 2015), Pama-Nyungan (Bowern et al. 2011), Japonic (Lee and Hasegawa 2011), and Sino-Tibetan (Sagart et al. 2019). In addition, scholars have also attempted to provide unified methods that could be applied in a completely automated fashion to all languages of the world (Holman et al. 2011).

[1] The last major improvement, the decipherment of Hittite, which also helped to proof that it was an Indo-European language dated back to Hrozný (1915).

Another strand of research deals with the computation of inference procedures which were traditionally only carried out manually. Most prominently, we find here various attempts to automate different aspects of the general workflow of the traditional *comparative method* for historical language comparison (Weiss 2015). Breaking down the workflow into some of its major parts, we thus find (1) automated methods for the comparison of words, as reflected in methods for phonetic alignment (Kondrak 2000, Prokić et al. 2009) and automated cognate detection (Hauer and Kondrak 2011, List et al. 2016b, Turchin et al. 2010), (2) automated approaches for the detection of borrowings (List 2015, Mennecier et al. 2016, Nelson-Sathi et al. 2011),[2] (3) automated approaches for linguistic reconstruction (Bouchard-Côté et al. 2013, Jäger 2019), and (4) automated approaches for the detection of sound correspondences (List 2019b).

While the second strand deals mostly with questions of inference, a third strand organizes inferred data in form of large-scale online databases that aggregate different kinds of information on the world's languages. The most prominent of these databases is beyond doubt the *World Atlas of Language Structures* (Dryer and Haspelmath 2013), but in addition we also find attempts to aggregate cross-linguistic information on phoneme inventories (Maddieson et al. 2013, Moran and McCloy 2019), polysemies (List et al. 2018), phonotactics (Donohue et al. 2013), borrowings (Haspelmath and Tadmor 2009), as well as datasets like D-Place, that compare cultural, environmental, and linguistic diversity (Kirby et al. 2016).

While the popular phylogenetic approaches deal with concrete languages in concrete times, trying to answer very specific (or *particular*) questions about their past, a fourth strand of research makes use of the new cross-linguistic databases along with results drawn from the phylogenetic approaches to investigate general aspects of language change, including questions like the rate of linguistic change and its correlates (Calude and Pagel 2011, Greenhill et al. 2017), the question to which degree environmental factors might have an impact on language evolution (Everett et al. 2015), or how language structures converge independent of contact or inheritance (Blasi et al. 2016).

Why is the aspect of dating, i.e., the inference of absolute phylogenies, so important for the new methods in historical linguistics?

1.3 Benefits of computational historical linguistics

Apart from the obvious benefit that the new quantitative methods have drastically revived the interest of scholars in historical linguistics, which also resulted in an increased amount of funding and a new generation of young scholars who are highly collaborative in their research and well trained in computational methods, the quantitative turn has also led to a considerable amount of rethinking in the field of historical linguistics, which offers new perspectives on the subject which have been ignored so far. First, we can see that the new methods shift the focus from *internal* to *external language* history, while at the same time turning away from the traditional focus on Indo-European alone.[3] We can also see that the new methods lead to the raise of new questions, specifically addressing *general* questions of language history.

This is also reflected in new research approaches, which are more explicitly *data-centered* nowadays and often based on statistical or stochastic modeling. While research in historical linguistics has always been data-centered, the new methods have shown that the classical approaches to deal with data – namely the individual collection of extensive personal notes from the literature, and the publication of new insights from these personal collections in form of extensive prose – are reaching their limits

[2] See List (2019a) for an overview on these approaches.

[3] Compare classical handbooks such as the *Einführung in die vergleichende Sprachwissenschaft* by Szemerényi (1970), where the term *comparative linguistics* (which should be a general discipline) is seen as a synonym for *Indo-European linguistics*.

in times where the amount of data is constantly increasing. Although the attempts to automate the classical methods have so far not yet led to a situation where computers could beat the experts,[4] we have won many important and new insights into the methods and the practice of historical language comparison, specifically also because the new methods challenged classical (traditional) linguists to revise the methods they use and to increase the degree of explicitness by which they apply them.

That languages interact with different factors is evident. What are the aspects that make it so difficult to study language change with help of computational frameworks?

1.4 Problems and Criticisms

Not all linguists have enthusiastically welcomed the new methods. While the various critics range from justified criticism, via exaggerations, up to complete ignorance for the initial goals of the computational approaches, and at times rather reflect the insulted ego of those who consider themselves as indisputable experts, the new field faces a couple of serious problems that are worth being criticized and rigorously analyzed. Among the most important of these are (1) problems with the data that is used in quantitative analyses, (2) problems of applicability of the computational approaches, and (3) problems of transparency and (4) comparability with respect to the results and methods which scholars report, and (5) problems of the general accuracy of the computational methods in comparison with experts.

The data problems related to the way in which data are compiled and curated, and what judgments they are based upon. The general problem here is that most of the phylogenetic approaches still make use of human-annotated data, trusting the expertise of only a small amount of experts to be enough to annotated data for at times more than 100 different languages. The danger of this procedure (which is to some degree difficult to avoid) are potential problems of inter-annotator-agreement, which may themselves, of course, impact the results (Geisler and List 2010). The problem of applicability and transparency is reflected in large amounts of software solutions and datasets that are only discussed in the literature, but have not been openly shared (List et al. 2017). As a result, there are quite a few methods out there that could provide valid solutions, but which have only been tested on one dataset and never officially been published, which comes close to a crisis of irreproducibility as it has been noted in many branches of science since the beginning of this millennium (Nature 2013).[5]

The problem of comparability results from missing standards in our field, which make it difficult to compare results across datasets, since it is often very tedious to lift the data used by different scholars to a level where they could be easily compared. The problem of accuracy, finally, is probably the hardest problem to address, since the problems of historical linguistics are often quite hard to solve automatically, specifically also because – as a rule – data is sparse, while most computational methods have been built based on the assumption that data to test and train algorithms would be abundantly available.

What solutions can you think of to overcome the problems of transparency and comparability, which were mentioned above?

[4] This is also not to be expected shortly, given that the only areas in which machines outperform humans so far are restricted fields, such as chess, or the go-game (Silver et al. 2016), and not in problems that need to be solved in open worlds.

[5] Luckily, this picture is slowly changing, thanks to extensive efforts to propagate free data and free code. A our department, for example, we have now decided to refuse to review papers where we are not given code and data, if they are needed for replication, following the idea of referee's rights as expressed by the editorial board of the journal Nature in 2018.

2 Towards a Qualitative Turn in Computational Historical Linguistics

2.1 Reconciling Classical and Computational Research

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

Do you have experience with computer-assisted translation? If not, what role do computers and computer tools play for your research?

2.2 Computer-Assisted Language Comparison

Following the idea of computer-assisted frameworks in translation and biology, a framework for computer assisted language comparison (CALC) is the key to reconcile classical and computational approaches in historical linguistics. Computational approaches may still not be able to compete with human experts, but when used to pre-process the data with human experts systematically correcting the results, they can drastically increase the efficiency of the classical comparative method and make up for the insufficiencies of of current computational solutions. At the same time, bringing experts closer to computational and formal approaches will also help to increase the consistency or classical research, forcing experts to annotated their specific findings and corrections in due detail, without resorting to texts in prose and ad-hoc explanations.

Classical linguists working on etymological research often emphasize the importance of looking into all details of language history, invoking the slogan "chaque mot a son histoire", which is, according to Campbell (1999: 189) traditionally attributed to Jules Gilliéron (1854-1926). Even if this was completely true, how can we still defend the recent attempts of computer-assisted and computer-based strategies in historical linguistics to work on a more formal and more quantitative handling of linguistic data?

2.3 Data, Software, and Interfaces

In the framework of computer-assisted language comparison, data are constantly passed back and forth between computational and classical linguists. Three different aspects are essential for this workflow: Specific *software* allows for the application of transparent methods which increase the accuracy and the application range of current methods in historical linguistics and linguistic typology. Interactive *interfaces* serve as a bridge between human and machine, allowing experts to correct errors and to inspect the automatically produced results in detail. To guarantee that software and interfaces can interact directly, *data* need to be available in human- and machine-readable form.

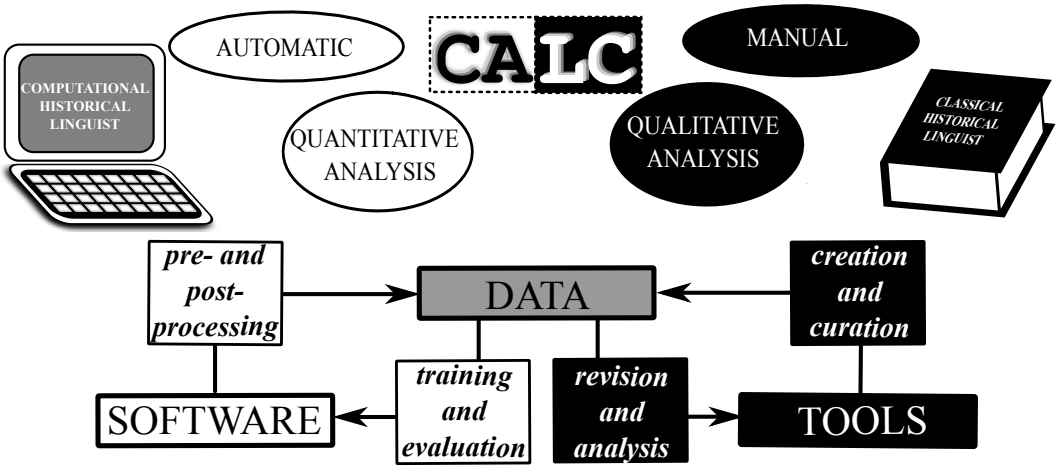


Fig. 1: Interplay of data, software, and interfaces in computer-assisted language comparison.

How exactly should one imagine data that are human- and machine-readable at the same time?

2.4 ERC Project on Computer-Assisted Language Comparison

In the ERC-funded research project CALC (Computer-Assisted Language Comparison, List 2016, 2017-2022), we tried to establish a computer-assisted framework for historical linguistics. Even with the end of the project, this task has not finished, specifically also since the project was extended by the Max Planck Society (2022-2024) and will play a crucial role in our new ERC project on Productive Signs, in which we look at word formation from a cross-linguistic perspective. In our work on CALC, we pursue an interdisciplinary approach that adapts methods from computer science and bioinformatics for the use in historical linguistics. While purely computational approaches are common today, Computer-Assisted Language Comparison focuses on the communication between classical and computational linguists, developing interfaces that allow historical linguists to produce their data in machine readable formats while at the same time presenting the results of computational analyses in a transparent and human-readable way. All technologies which we will discuss in the remaining lectures of this course can be seen as a direct output of our work on Computer-Assisted Language Comparison.

What may be the reason for choosing an interdisciplinary approach, and what are the most likely disciplines from which the project could take inspiration?

3 Important Aspects of Computational Historical Linguistics

To get a better understanding of the state of the art, the potential, and the limitations of computational approaches in historical linguistics after the quantitative turn, it is important to have a closer look at the problems, as they were outlined before, and how scholars try to address(which can relate to both p- and g-linguistic questions) them today. Even more important, however, is to understand the basic ideas that underlie the new methods, and the topics that the methods deal with. To provide a short overview on these different aspects, we will follow the triad of *modeling*, *inference*, and *analysis*, as outlined in the session before. In this context, however, it is important to note explicitly that the division into the three aspects has its limits in practice, since what counts as inference in a given research framework may at times count as analysis in another one and vice versa.

3.1 Modeling

The models that are used so far in computational historical linguistics are all rather simple. While this may at times be surprising for classical linguists, who have a very complex idea of change process and also very detailed knowledge of the complex range of what is possible in language change, reducing the complexity of models is a necessary step in all scientific research. Rather then trying to establish the most complex models before we start to infer something, we should investigate how far we can go with a simplifying model and where its specific limits lie.

Crucial aspects for the models in diversity linguistics are the concept of *language*, *word* (or linguistic sign), *word form*, and *word meaning*. Higher dimensions relevant for questions of language use, such as the speaker-listener interaction, are usually disregarded in the initial stages of investigation. The most common model for a language ist to treat a given language as a bag of words (or a bag of linguistic signs). Depending on the perspective, one can invoke a set of grammatical rules by which these signs are combined to form sentences. The linguistic sign itself follows the basic idea of Saussure (*Cours de linguistique générale*) with the modification that the sign is not seen as a duplet of *form* and *meaning*, but a triplet of form, meaning, and the *language* to which the sign belongs (List 2014).

The sign form is usually modeled as a *sequence of sounds*, which implies that we can *segment* each word into a certain number of sounds. The sequences are constructed or constrained by *phonotactic rules*. If needed, one can add an additional layer of segmentation, dependent on the research question (e.g., one could look at a word consisting of morphemes consisting of sound segments, or a word consisting of syllables consisting of sound segments). These *secondary sequence structures* are of a certain importance in modern approaches for sequence comparison (List 2014, List et al. 2016b), but they are often also deliberately disregarded. While the sign form is best treated as a sequence of sounds, the sign meaning is usually handled as a *network of senses*.

While this model of language as a bag of words may seem very simply, it is effectively the model that was underlying most of the phylogenetic analyses that have been published so far. Additionally one should say, that even classical historical linguists tend to use this model in their analyses. When needed, throughout this course, we will discuss more complex models in due time.

To address the problem that we face a drastic lack of comparability with respect to the data that has been produced in diversity linguistics, the Cross-Linguistic Data Initiative (`https://cldf.clld.org`, Forkel et al. 2018) has published a set of recommendations for unified data standards in diversity linguistics, which are now gaining more and more popularity among scholars. These recommendations build more or less directly on the above-mentioned language model, and the current plan is to expand these further, based on the need and the availability of more complex models. As a very important aspect of standardization, CLDF comes along with *reference catalogs*, which are basically meta-datasets, that offer standards for the handling of languages (Glottolog, `https://glottolog.org`, Hammarström et al. 2018), concepts (Concepticon, `https://concepticon.clld.org`, List et al. 2016a), and sounds in transcription (CLTS, `https://clts.clld.org`, Anderson et al. 2018).

> In addition to the modeling of the data, the modeling of the processes, which has been not mentioned here, is of great importance. What models can you think of that would explain, for example, the process of sound change, or the process of lexical change?

3.2 Inference

As mentioned before, the inference of dated language phylogenies is by far the most popular of the computational methods proposed so far in the field of computational historical linguistics. Discussing the details of these approaches would, unfortunately, go beyond the scope of this session, but good review literature that provides some basic insights is now readily available (Greenhill 2015). What

seems important to mention in this context is that the bag-of-words model mentioned before can be seen as the standard model that is essentially used to search for a language phylogeny. When discussing the simulation of language change in a later session, we will discuss more complex ways to simulate language change, which in theory also allow to handle the interaction between speaker and listener.

Second in popularity are methods for automated sequence comparison, which are very popular in dialectology, where methods for phonetic alignment are used to compute aggregate distances between dialect varieties, based on pronunciation distances derived from pre-selected lists of words (Nerbonne et al. 2011). In addition, methods for phonetic alignments are also used for the task of automated cognate detection, which tries to infer which words in a multi-lingual wordlist go back to the same ancestor. Techniques for automated cognate detection are quite well-developed by now, and have been shown to work surprisingly well, with accuracy scores of up to 90% on shallower language families (List et al. 2017), while the accuracy usually drops to around 60%-70% when dealing with larger datasets (Jäger et al. 2017). Further aspects of inference include automated borrowing detection (Mennecier et al. 2016), the detection of sound correspondences and sound correspondence patterns (List 2019b), and also the automated prediction of so far unobserved words (Bodt and List 2019), which is specifically useful to support fieldworkers working on small groups of related languages.

How can automated word prediction be useful for linguistic field work?

3.3 Analysis

As it was mentioned briefly before, the distinction between what counts as inference and what counts as analysis are not always easy to draw. Intuitively, analysis should involve g-linguistic questions in the sense discussed in the first session, but it is clear that there is no formal justification for it, and it seems to depend more on the workflow, whether a certain step (such as – for example – phylogenetic inference) is labeled as part of the inference or the analysis step. An example for such a borderline case is the *Database of Cross-Linguistic Colexifications* (CLICS, `https://clics.clld.org`, List et al. 2018), which offers cross-linguistic accounts on polysemies, which are displayed in form of a network analysis that provides information on the relative cross-linguistic closeness of more than 1500 different concepts, reflected in more than 1000 of the world's languages. While CLICS is offering an analysis that shows – similar to Youn et al. (2016) – that lexical structure is surprisingly similar across languages, the analysis itself could be treated as some kind of inference, and analysed to answer bigger questions related to human cognition. The more classical analyses which are usually presented, however, try to test certain theories by analysing the data which has been inferred previously. In these cases, the large-scale cross-linguistic databases, which are increasingly produced, play an important role, as they allow scholars to test their hypotheses on a global scale, allowing them, for example, to test hypotheses regarding the transmission of Creole languages (Blasi et al. 2017), the evolution of syntax (Widmer et al. 2017), or the impact of our diet on evolution of our speech sounds (Blasi et al. 2019).

What hypotheses can be derived from historical linguistics that could be tested with the help of cross-linguistic approaches?

References

Anderson, C., T. Tresoldi, T. C. Chacon, A.-M. Fehn, M. Walworth, R. Forkel, and J.-M. List (2018). "A Cross-Linguistic Database of Phonetic Transcription Systems." *Yearbook of the Poznań Linguistic Meeting* 4.1, 21–53.

Atkinson, Q. D. and R. D. Gray (2006). "How old is the Indo-European language family? Illumination or more moths to the flame?" In: *Phylogenetic methods and the prehistory of languages*. Ed. by P. Forster and C. Renfrew. Cambridge, Oxford, and Oakville: McDonald Institute for Archaeological Research, 91–109.

Barrachina, S. et al. (2008). "Statistical approaches to computer-assisted translation." *Computational Linguistics* 35.1, 3–28.

Bergsland, K. and H. Vogt (1962). "On the validity of glottochronology." *Current Anthropology* 3.2, 115–153. JSTOR: 2739527.

Blasi, D. E., S. M. Michaelis, and M. Haspelmath (2017). "Grammars are robustly transmitted even during the emergence of creole languages." *Nature Human Behaviour* 1, 723–729.

Blasi, D. E., S. Moran, S. R. Moisik, P. Widmer, D. Dediu, and B. Bickel (2019). "Human sound systems are shaped by post-Neolithic changes in bite configuration." *Science* 363.1192, 1–10.

Blasi, D. E., S. Wichmann, H. Hammarström, P. Stadler, and M. H. Christiansen (2016). "Sound–meaning association biases evidenced across thousands of languages." *Proceedings of the National Academy of Science of the United States of America* 113.39, 10818–10823.

Bodt, T. A. and J.-M. 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, A., D. Hall, T. L. Griffiths, and D. 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, 4224–4229.

Bowern, C., P. Epps, R. Gray, J. Hill, K. Hunley, P. McConvell, and J. Zentz (2011). "Does Lateral Transmission Obscure Inheritance in Hunter-Gatherer Languages?" *PLoS ONE* 6.9, e25195.

Calude, A. S. and M. Pagel (2011). "How do we use language? Shared patterns in the frequency of word use across 17 world languages." *Philosophical Transactions of the Royal Society B* 366, 1101–1107.

Campbell, L. (1999). *Historical linguistics. An introduction*. 2nd ed. Edinburgh: Edinburgh Univ. Press.

Chang, W., C. Cathcart, D. Hall, and A. Garret (2015). "Ancestry-constrained phylogenetic analysis ssupport the Indo-European steppe hypothesis." *Language* 91.1, 194–244.

Donohue, M., R. Hetherington, J. McElvenny, and V. Dawson (2013). *World phonotactics database*. Canberra: Department of Linguistics. The Australian National University.

Dryer, M. S. and M. Haspelmath, eds. (2013). *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Everett, C., D. E. Blasi, and S. G. Roberts (2015). "Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots." *Proceedings of the National Academy of Sciences of the United States of America* 112.5, 1322–1327.

Forkel, R., J.-M. List, S. J. Greenhill, C. Rzymski, S. Bank, M. Cysouw, H. Hammarström, M. Haspelmath, G. A. Kaiping, and R. D. Gray (2018). "Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics." *Scientific Data* 5.180205, 1–10.

Geisler, H. and J.-M. List (2010). "Beautiful trees on unstable ground. Notes on the data problem in lexicostatistics." In: *Die Ausbreitung des Indogermanischen. Thesen aus Sprachwissenschaft, Archäologie und Genetik*. Ed. by H. Hettrich. Document has been submitted in 2010 and is still waiting for publication. Wiesbaden: Reichert.

Gray, R. D. and Q. D. Atkinson (2003). "Language-tree divergence times support the Anatolian theory of Indo-European origin." *Nature* 426.6965, 435–439.

Gray, R. D., A. J. Drummond, and S. J. Greenhill (2009). "Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement." *Science* 323.5913, 479–483.

Greenhill, S. J., C. H. Wu, X. Hua, M. Dunn, S. C. Levinson, and R. D. Gray (2017). "Evolutionary dynamics of language systems." *Proceedings of the National Academy of Sciences of the United States of America* 114.42, E8822–E8829.

Greenhill, S. (2015). "Evolution and Language: Phylogenetic Analyses." In: *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*. Ed. by J. D. Wright. Second Edition. Oxford: Elsevier, 370 –377.

Grollemund, R., S. Branford, K. Bostoen, A. Meade, C. Venditti, and M. Pagel (2015). "Bantu expansion shows that habitat alters the route and pace of human dispersals." *Proceedings of the National Academy of Sciences of the United States of America* 112.43, 13296–13301.

Hammarström, H., R. Forkel, and M. Haspelmath (2018). *Glottolog*. Version 3.3. URL: http://glottolog.org.

Haspelmath, M. and U. Tadmor, eds. (2009). Berlin and New York: de Gruyter.

Hauer, B. and G. Kondrak (2011). "Clustering semantically equivalent words into cognate sets in multilingual lists." In: *Proceedings of the 5th International Joint Conference on Natural Language Processing*. (Chiang Mai, Thailand, 11/08–11/13/2011). AFNLP, 865–873.

Holman, E. W. et al. (2011). "Automated dating of the world's language families based on lexical similarity." *Current Anthropology* 52.6, 841–875. JSTOR: 10.1086/662127.

Hrozný, B. (1915). "Die Lösung des hethitischen Problems [The solution of the Hittite problem]." *Mitteilungen der Deutschen Orient-Gesellschaft* 56, 17–50.

Jäger, G. (2019). "Computational historical linguistics." *Theoretical Linguistics* 45.3-4, 151–182.

Jäger, G., J.-M. List, and P. Sofroniev (2017). "Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists." In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Long Papers*. "EACL 2017". Valencia: Association for Computational Linguistics, 1204–1215.

Kirby, K. R. et al. (2016). "D-PLACE: A Global Database of Cultural, Linguistic and Environmental Diversity." *PLOS ONE* 11.7, 1–14.

Kolipakam, V., F. M. Jordan, M. Dunn, S. J. Greenhill, R. Bouckaert, R. D. Gray, and A. Verkerk (2018). "A Bayesian phylogenetic study of the Dravidian language family." *Royal Society Open Science* 5.171504, 1–17.

Kondrak, G. (2000). "A new algorithm for the alignment of phonetic sequences." In: *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*. (Seattle, 04/29–05/03/2000), 288–295.

Lee, S. and T. Hasegawa (2011). "Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages." *Proc. Biol. Sci.* 278.1725, 3662–3669.

List, J.-M. (2014). *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.

— (2015). "Network perspectives on Chinese dialect history." *Bulletin of Chinese Linguistics* 8, 42–67.

— (2016). *Computer-Assisted Language Comparison: Reconciling Computational and Classical Approaches in Historical Linguistics*. Jena: Max Planck Institute for the Science of Human History.

— (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, 137–161.

List, J.-M., M. Cysouw, and R. Forkel (2016a). "Concepticon. A resource for the linking of concept lists." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. "LREC 2016" (Portorož, 05/23–05/28/2016). Ed. by N. C. C. Chair), K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis. Luxembourg: European Language Resources Association (ELRA), 2393–2400.

List, J.-M., S. J. Greenhill, C. Anderson, T. Mayer, T. Tresoldi, and R. Forkel (2018). "CLICS². An improved database of cross-linguistic colexifications assembling lexical data with help of cross-linguistic data formats." *Linguistic Typology* 22.2, 277–306.

List, J.-M., S. J. Greenhill, and R. D. Gray (2017). "The potential of automatic word comparison for historical linguistics." *PLOS ONE* 12.1, 1–18.

List, J.-M., P. Lopez, and E. Bapteste (2016b). "Using sequence similarity networks to identify partial cognates in multilingual wordlists." In: *Proceedings of the Association of Computational Linguistics 2016 (Volume 2: Short Papers)*. Association of Computational Linguistics. Berlin, 599–605.

Maddieson, I., S. Flavier, E. Marsico, C. Coupé, and F. Pellegrino. (2013). "LAPSyD: Lyon-Albuquerque Phonological Systems Database." In: *Proceedings of Interspeech*. (Lyon, 08/25–08/29/2013).

Mennecier, P., J. Nerbonne, E. Heyer, and F. Manni (2016). "A Central Asian language survey." *Language Dynamics and Change* 6.1, 57–98.

Moran, S. and D. McCloy, eds. (2019). *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History.

Nature, E. B. (2013). "Reducing our irreproducibility." *Nature* 496.4, 398.

— (2018). "Referees´rights." *Nature* 560, 409.

Nelson-Sathi, S., J.-M. List, H. Geisler, H. Fangerau, R. D. Gray, W. Martin, and T. Dagan (2011). "Networks uncover hidden lexical borrowing in Indo-European language evolution." *Proceedings of the Royal Society of London B: Biological Sciences* 278.1713, 1794–1803.

Nerbonne, J., R. Colen, C. Gooskens, P. Kleiweg, and T. Leinonen (2011). "Gabmap – A web application for dialectology." *Dialectologia* Special Issue II, 65–89.

Prokić, J., M. Wieling, and J. Nerbonne (2009). "Multiple sequence alignments in linguistics." In: *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*. "LaTeCH-SHELT&R 2009" (Athens, 03/30/2009), 18–25. acm: 1642052.

Sagart, L., G. Jacques, Y. Lai, R. Ryder, V. Thouzeau, S. J. Greenhill, and J.-M. List (2019). "Dated language phylogenies shed light on the ancestry of Sino-Tibetan." *Proceedings of the National Academy of Science of the United States of America* 116 (21), 10317–10322.

Saussure, F. de. *Cours de linguistique générale*. Ed. by C. Bally. Lausanne: Payot, 1916; German translation: — . *Grundfragen der allgemeinen Sprachwissenschaft*. Trans. from the French by H. Lommel. 2nd ed. Berlin: Walter de Gruyter & Co., 1967.

Silver, D. et al. (2016). "Mastering the game of Go with deep neural networks and tree search." *Nature* 529.7587, 484–489.

Swadesh, M. (1950). "Salish internal relationships." *International Journal of American Linguistics* 16.4, 157–167. JSTOR: 1262898.

Szemerényi, O. (1970). *Einführung in die vergleichende Sprachwissenschaft*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Turchin, P., I. Peiros, and M. Gell-Mann (2010). "Analyzing genetic connections between languages by matching consonant classes." *Journal of Language Relationship* 3, 117–126.

Weiss, M. (2015). "The comparative method." In: *The Routledge handbook of historical linguistics*. Ed. by C. Bowern and N. Evans. New York: Routledge, 127–145.

Widmer, M., S. Auderset, J. Nichols, P. Widmer, and B. Bickel (2017). "NP recursion over time: Evidence from Indo-European." *Language* 93.4, 799–826.

Youn, H., L. Sutton, E. Smith, C. Moore, J. F. Wilkins, I. Maddieson, W. Croft, and T. Bhattacharya (2016). "On the universal structure of human lexical semantics." *Proceedings of the National Academy of Sciences of the United States of America*.

Lecture 2: From Cognates to Correspondences

Abstract

Our goal for the second session in our little course on computational historical linguistics is to discuss some major developments which were made in the field of historical linguistics in the past years, and which – in my opinion – may also have some importance for the practitioners of language comparison, especially those who are not independently interested in quantitative methods.

1 Introduction

1.1 Modeling and Representation

In the sciences, scholars often talk about *modeling*. Scholars *model* sound change, they *model* language change, and they try to *model* lexical borrowing. It is not always clear what is meant with the term *modeling*, and it seems that scholars use it with varying ideas in mind. If we talk about *modeling* in the context of quantitative and formal approaches to historical language comparison, I use the term *model* in the sense of what Bröker and Ramscar (2021) call an *implemented model*. While a general model can also exist of a prose explanation of the mechanisms underlying a phenomenon, an *implemented model* is a model which can be shown to work in some piece of software and applied to some data.

> To explain why the contributions of representations, algorithms, and computations will only rarely manifest themselves in fully independent ways [...], it is important to recognise that in practice, models in the brain and cognitive sciences are typically presented in one of two distinct ways: either as abstract model descriptions, or as implemented models. Abstract model descriptions typically comprise symbolic (i.e. verbal or algebraic) descriptions of the relationships between what are typically quite loosely defined quantities or entities. Accordingly, while abstract models can appear to be more or less "formal", they typically fail to fully specify representations (what exactly will be counted and in which format) and typically fail to fully specify the algorithms that will transform these representations into predictions (Figure 6). It is in fact only when these latter steps are made, and an abstract model is actually implemented, that it can be considered formal in any meaningful sense. (ibid.: 17/25)

Of crucial importance for implemented models is the way in which data are *represented*, since this determines how the implementation works. In the work I will present, for example, we may conveniently represent language data (words in the lexicon of a language, etc.) in the form of tables. These can be printed to paper, but they can also be typed into spreadsheets on the computer. The representation of data is thus the basis upon which we build our models and implement them in computer code.

> To recapitulate: Representational choices can significantly alter the performance of a model, the predictions it makes and thus the way it is interpreted. (ibid.: 20/25)

The distinction between models, implemented models, and representations, does not define the term "model" itself. Atkinson and Gray (2006: 94) write about models, that they are "lies that lead us to the truth". Is this a useful characterization of models?

1.2 Integrated Data Representations

When working with data, scholars often use very different representations of their data. They may have one file for their syntactic properties they collect, one word document, where they collect their favorite quotes, another spreadsheet where they started to collect sound changes, and some old FileMaker database, which they still use for convenience, to enlarge their personal etymological dictionary of their favorite language. When working with data, scholars also often commit certain common errors in data collection. The most common errors are to extract information from sources without storing a reference to the original sources, or to copy text from some resource into a cell in a spreadsheet and later modify this content manually without keeping the original raw data.

As of now, there are many good guidelines for working transparently with data out in the internet (Perkel 2022), and I recommend that all who feel a bit insecure about how to collect data properly to inform themselves about these resources and generally take much more time in planning or experimenting with different formats of data representation than starting to collect data and eventually destroying information without having intended to do so. I also recommend to think about *integrated data representation*, that is, to think about ways to work on different questions with the same data, and to extract certain important aspects of the annotation of a dataset rather than paste it into a separate data sheet. As an example, scholars may store a dictionary of a given language written in orthography, and additionally they may type off the phoneme inventory of that language from another resource and collect these separately. It would be much better to work on a dictionary in phonetic transcription from which the same information could be derived (the inventory should be extractable from the dictionary). Examples for integrated data handling have been recently published by our group in the form of the Lexibank repository (List et al. 2022a), where we compute several lexical and phonological features of various languages from the wordlists, which we have collected and standardized.

Why is it so important to keep the raw data when collecting data for one's studies?

2 Cognates

The starting point for many analyses in the field of historical language comparison is the identification of cognate words across related languages. It is well known that this is typically done in an iterative fashion when applying the comparative method (Ross and Durie 1996). This means that scholars usually first identify some potential cognates, then search for sound correspondences among those, then identify more cognates, kick out some that turn out to be wrong, etc. In order to formalize this procedure, we do not only need to be clear about the way in which we want to represent words in our data (we will discuss this in detail when discussing phonetic alignments), but also how we want to define *cognacy* in this context. It turns out that this is far more difficult than one might think at first sight.

2.1 Terminology

In order to approach the question of how to model, represent, infer, and annotate relations of cognacy, it is important to get a clearer picture of the terminology we use in the field of historical linguistics and beyond. Although most historical linguistics would probably confirm that they have a very clear idea of the term *cognacy* and its meaning, we can find quite divergent applications in the practice. Historical linguists usually agree that cognacy describes the relation between words that share a common ancestor form and have descended from this common form only via vertical inheritance – as opposed to lateral transfer, which would point to borrowing events. When it comes to identifying cognates in multilingual datasets, however, we can quickly see that there may be various disagreements in practice.

I attribute these problems to the insufficiency of the term *cognacy* to capture fundamental relations we want to handle in our field. As I have shown before, our terminology would profit a lot from being modified and made more precise and clear, reflecting fundamental relations of descent. The following table (taken from List 2016, which is an extension of List 2014), we find terms for cognacy in linguistics (along with suggestions for new terms) contrasted with typical terms for *homology* in evolutionary biology.

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

It seems that the linguistic view on relatedness, when comparing it with the view reflected in the terminology of evolutionary biology, is very biased towards borrowing. What reasons may contribute to this view?

2.2 Relations

While the specification of the terminology on cognacy was helpful for some time, it turned out that it was still not capturing all important aspects which we need for a formal representation of cognate sets for the purpose of historical language comparison. What was specifically missing was a dimension that would allow us to determine if a cognate relation can be considered *regular* or not, since this in turn is crucial for the identification of regular sound correspondences. The following table (taken from Schweikhard and List 2020) revises the previous table on cognate relations by adding regularity as a new dimension and asking more clearly how each of the dimension must be specified.

Relation	**Biological term**	**Regularity**	**Morphological continuity**	**Semantic continuity**	**Stratic continuity**
traditional notion of cognacy	-	+/−	+/−	+/−	+
cognacy à la Swadesh	-	+/−	+/−	+	+
direct cognacy	~~orthology~~	+/−	+	+/−	+
oblique cognacy	-	+/−	−	+/−	+
etymological relation	homology	+/−	+/−	+/−	+/−
oblique etymological relation	xenology	+/−	+/−	+/−	−
strict cognacy	-	+	+/−	+/−	+/−

This table also introduces a new term, following List (2018), *strict cognacy*, which is unfortunately misrepresented in the table, in so far as it must require morphological and stratic continuity, since strict cognates are defined as word forms which differ only with respect to regular sound change and

semantics. We will see that these cognates are crucial for alignment analyses and for the identification of regular sound correspondences, which depend on alignments.

How can we deal with cases in which parts of a word are strictly cognate (the root) while other parts aren't?

3 Alignments

Phonetic alignments are a central concept of the formal and quantitative work on historical language comparison which our group pursues. The crucial idea of phonetic alignments is to represent words as a sequence of sounds, which in turn allows us to handle the comparison of words as the problem of comparing two or more sequences, which is most typically done with the help of alignment analyses (List 2014). Implicitly, alignment analyses have been used for a long time in historical linguistics, but explicitly, it has only been recently that they started to trigger the interest of scholars, as reflected in a growing amount of studies devoted to the topic (Kondrak 2000, List 2014, Prokić et al. 2009).

3.1 Representing Words as Sequences of Sounds

Before we can align words, we need to discuss how we want to represent them as sequences of sounds. Here, we can see a lot of variation in scholarly practice, mostly resulting from the fact that scholars insist to stick to traditional orthography or traditional phonological representations when comparing cognate words in their specific language family. Traditions are very strong here, and scholars are often unwilling to give them up. For a proper representation of words as sequences of sounds, however, it is indispensable to have some clear notion of how sounds should be represented and how they should be chained up to form a sequence.

The solution we pursue by now is to use a specific – and much stricter – version of the International Phonetic Alphabet, named B(road)IPA, published as part of the Cross-Linguistic Transcription Systems Reference catalog (`https://clts.clld.org`, Anderson et al. 2018, List et al. 2021), which defines more than 8000 different sound segments and is based on a generative component which allows to define more possible sounds on the fly. We represent words in this transcription, but we do not have any strong opinion on whether the representation should be phonetic or phonological. We tend to recommend phonological representations which do not exaggerate phonological theory (e.g., by proposing that Chinese has only one vowel).

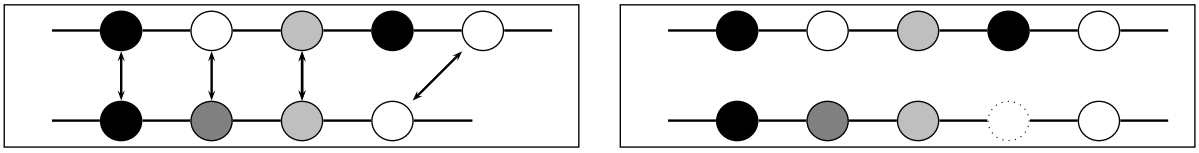
In order to represent a word, distinct *sounds* need to be written with a space as separator, so that we can avoid ambiguities which may result from pre- vs. post-aspirated plosives, or various affricates, such as [ts] or [tʃ]. Morpheme boundaries are written with the help of a plus symbol, thus, German *Hausmeister* "housekeeper" could be represented as [h au s + m ai s t ə r].

In addition, we are currently experimenting with the possibility to *cluster* sounds (which are still listed as separate sounds but marked as belonging together, using a dot instead of a space), if it is known that they tend to evolve together (Hill and List in preparation). Thus one could represent all instances of German [s t] as [s.t], since – as we know – the two sounds tend to behave in a specific manner when it comes to their sound change patterns.

Are there certain sounds which would universally tend to evolve together and thus qualify as candidate for a grouped representation?

3.2 Aligning Sound Sequences

An alignment between a certain number of sequences is a technique by which the sequences are arranged in a matrix in such a way that all corresponding segments in the sequences appear in the same columns, while empty slots, resulting from segments which correspond with not other segment in another sequence are filled with *gap* symbols. We follow the tradition of evolutionary biology in representing gaps with a dash. This is also the reason why we use the plus character for morpheme boundaries. In the following figure, a pairwise alignment of two chains is shown, which illustrates the major idea of alignment analyses, which introduce gaps in order to adjust the lengths of two different sequences.



For phonetic alignments, we do nothing else than that, we introduce gap symbols when we find that a sound in a word from one language has no counterpart in the other language (such as the [n] in German *anderer* "other" not having a counterpart in English *other*).

How can we deal with cases of metathesis when carrying out pairwise or multiple word alignments?

3.3 "Alignability"

In Schweikhard and List (2020) we introduced the term *alignability*, in order to emphasize that not all alignments which we can think of in historical linguistics are also meaningful.

> This distinction accounts for the relation of strict as opposed to loose cognates and embraces the fact that only word forms which are strictly cognate can be aligned in a meaningful way. (ibid.: 9/25)

As an example for problems of alignability, compare the two alignments given in the following table (originally outlined in List 2014).

Language	Alignment						
Russian	s	-	ɔ	n	ts	ə	-
Polish	s	w	ɔ	n^j	ts	ɛ	-
French	s	-	ɔ	l	-	ɛ	j
Italian	s	-	o	l	-	e	-
German	s	-	ɔ	n	-	ə	-
Swedish	s	-	uː	l	-	-	-

(a) Global Alignment

Language	Alignment							
Russian	s	ɔ	-	-	n	ts	ə	
Polish	s	-	w	ɔ	n^j	ts	ɛ	
French	s	ɔ	l	-	-	-	-	ɛj
Italian	s	o	l	-	-	-		e
German	s	ɔ	-	-	-	-		nə
Swedish	s	uː	l	-	-	-	-	

(b) Local Alignment

Here, we have different words for *sun* in various Indo-European languages, which we could align in a naive way as shown on the right, or in a historically more informed way, as shown on the right. It turns out, that in the proposed informed alignment, we are only left with the initial, since we have not (yet) figured out a proper way of handling metathesis (as observed in the Polish word form).

However, a closer look at the word forms also reveals that they may not be *alignable* after all, given that they reflect two distinct Indo-European roots of *sun*, which are thought to reflect a unique (!) root alternation, only observed for the term *sun* in Proto-Indo-European, involving **-l-* and **-n-*. Thus, instead of aligning the Polish form with the German form, which both have inherited different reflexes of the alternating paradigm in Indo-European, it may be better to avoid aligning both words after all, and instead try to align only those word forms which go back to the same original root form.

Can the dot-representation for the grouping of sounds help to resolve some potential problems of multiple alignments in historical language comparison?

4 Correspondences

One of the biggest breakthroughs which we have made in our group during the past years was to find ways to handle correspondence patterns formally. The major ideas was presented in a study by List (2019) which also introduces an algorithm by which correspondence patterns can be inferred from phonetic alignments across multiple languages. Since then, we are exploring correspondence patterns in different language families and try hard to make their handling easier. However, until now, the handling and our understanding of correspondence patterns is still not perfect, and much more work will be needed in the future. In order to understand the major idea behind the notion of correspondence patterns, it is important to go back to the work of Anttila (1972).

4.1 Alignment Sites

The following table shows regular sound correspondences across four Indo-European languages, illustrated with help of alignments along the lines of Anttila (ibid.: 246). In contrast to the original illustration, lost sounds are displayed with help of the dash '-' as a gap symbol, while missing words (where no reflex in Gothic or Latin could be found) are represented by the '∅' symbol.

		A		B			C		D			E					F		
Sanskrit	y	u	g	a	m	dh	u	h	i	(tar)	s	n	u	ṣ	(ā)	-	r	u	dh (iras)
Greek	z	u	g	o	n	th	u	g	a	(ter-)	-	n	u	-	(os)	e	r	u	th (rós)
Latin	i	u	g	u	m	Ø	Ø	Ø	Ø	(Ø)	-	n	u	r	(us)	-	r	u	b (er)
Gothic	j	u	k	-	-	d	au	h	-	(tar)	Ø	Ø	Ø	Ø	(Ø)	Ø	Ø	Ø	Ø (Ø)
Gloss		'yoke'					'daughter'					'daughter-in-law'						'red'	

This illustration highlights six *alignment sites*, that is, columns of an alignment. These sites are chosen, because they are useful to illustrate the concept of *sound correspondence patterns* or simply *correspondence* patterns, which we have been trying to formalize during the past years. In contrast to an alignment site in isolation, a correspondence pattern is a patterns of sounds across different related languages (one sound per language) which is reflected in *many* alignment sites. When looking at the illustration above, we can identify two major patterns in the six sites, namely AEF on the one hand and CEF on the other hand. A and C cannot form a pattern, since they are in conflict with respect to the sound in Gothic. E and F are compatible with both A and C, because they both have *missing data*, which masks the true structure of the pattern to which they belong. The major task of historical language comparison consists in the identification of correspondence patterns in a language family.

Correspondence patterns have been used to predict word forms for which no cognate reflex had been identified before. How could this work, when considering the example above?

4.2 Correspondence Patterns in the Literature

In the literature, correspondence patterns are – unfortunately – typically only shown in a condensed form, which does not allow for concrete predictions or applications. As an example, consider the

following table illustrating correspondence patterns identified by Clackson (2007: 37), taken from List (2019: 142).

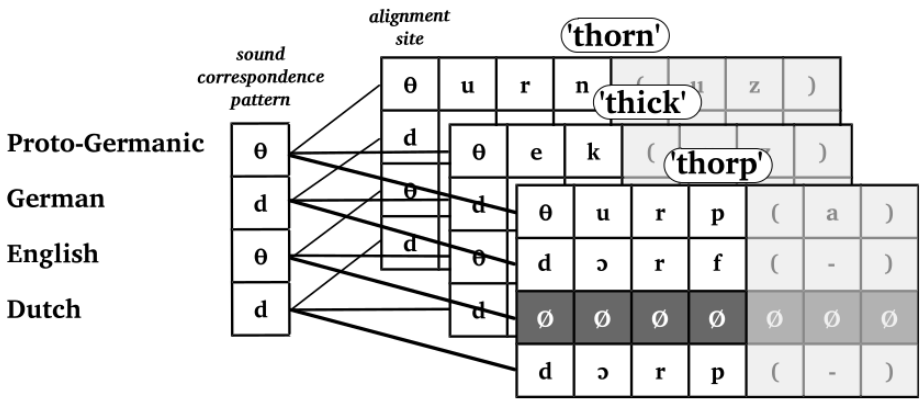
PIE	Hittite	Sanskrit	Greek	Latin	Gothic	...
*p	p	p	p	p	f b	...
*b	b p	b	b	b	p	...
*b^h	b p	$b^ɦ$/bh	p^h/ph	f b	b	...
*t	t	t	t	t	θ/þ d	...
*d	d t	d	d	d	t	...
*d^h	d t	d^h/dh h	t^h/th	f d b	d	...
...	...	...	...	...	...	...
*k^w	k^w/ku	k c	k p t	k^w/qu	h^w/hw g	...
*g^w	k^w/u	g j	g b d	g^w/gu u	q	...
*g^{wh}	k^w/ku g^w/gu	$g^ɦ$/gh h	p^h/ph t^h/th k^h/kh	f g^w/gu u	g b	...

The problem of the table is that it does not show how individual patterns look (e.g., if we have two sounds in Hittite as reflexes of Proto-Indo-European *b^h, are their distributions the same as for the two sounds in Gothic? One can surely argue that this depends on the conditioning context, which is missing here, but one could as well argue that the concrete patterns derived from concrete alignments would give us much more information here, since they would also allow us to assess the frequency of the patterns in the data, etc.

The representation distinguishes different sound reflexes separated by a space and by a dash, what is the difference here?

4.3 Correspondence Pattern Identification

The following figure (taken from ibid.) summarizes some of the terminology discussed so far. We consider the identification of correspondence patterns in aligned cognate sets as one of the key tasks of historical language comparison, since it is the basis of linguistic reconstruction and crucial for the evaluation of cognate sets and the justification of regular sound change.



In List (ibid.), I proposed an algorithm that helps to identify regular sound correspondence pattern from aligned cognate sets. This method was then applied to *predict* unobserved reflexes of cognate sets of Western Kho-Bwa languages in a field word experiment which were later verified through additional field work (Bodt and List 2022). It was also expanded as a method for automated supervised

phonological reconstruction (List et al. 2022b) and later used as a baseline for a shared task on word prediction (List et al. 2022c). But apart from the fact that this method works quite well, I consider it much more important now, that we have established a much stricter *representation* of correspondence patterns than what has been used in the literature so far. This representation is specifically useful in combination with proto-forms, as it allows scholars to check language by language how regular their cognate sets are, and to search for conditioning contexts that explain why one and the same proto-form show more than one reflex in the same language. We will illustrate the manual inspection of correspondence patterns in more detail during the fifth lecture. The following table (taken from List 2019) gives a short idea on the potential of this representation in showing correspondence patterns for Chinese dialects derived from Middle Chinese voiced alveolar initials (**d*) in dependence of the Middle Chinese tone of the syllables in question. As can be seen, we find a strict division in diverging patterns resulting from **d* under the even (P) tone, as opposed to the other three tones (SQR).

#	Cogn.	MC	MC Tones	BJ	SZ	CS	NC	MX	TY	GZ	FZ	TB
30	13	*t	PSQR	t	t	t	t	t	t	t	t	t
41	9	*th	PSQR	t^h	t^h	t^h	t^h	t^h	t^h	t^h	t^h	t^h
747	3	*d	P	t^h	d	t	t^h	t^h	t^h	t^h	t^h	t^h
718	2	*d	P	t^h	d	t	t^h	t^h	t^h	t^h	t	t^h
719	1	*d	P	t^h	d	t	t^h	t	t^h	t^h	t	t^h
1,096	1	*d	P	t^h	Ø	t	t^h	t^h	t^h	t^h	t	t
73	5	*d	QR	t	d	t	t^h	t^h	t^h	t	t	t^h
484	1	*d	R	t	d	t	l	t^h	t^h	t	t	t^h
654	1	*d	S	t	d	t	t^h	t	t^h	t	t	t^h

How can we explain the irregularity of certain patterns, and how can represent irregularity in our data?

References

Anderson, C., T. Tresoldi, T. C. Chacon, A.-M. Fehn, M. Walworth, R. Forkel, and J.-M. List (2018). "A Cross-Linguistic Database of Phonetic Transcription Systems." *Yearbook of the Poznań Linguistic Meeting* 4.1, 21–53.

Anttila, R. (1972). *An introduction to historical and comparative linguistics*. New York: Macmillan.

Atkinson, Q. D. and R. D. Gray (2006). "How old is the Indo-European language family? Illumination or more moths to the flame?" In: *Phylogenetic methods and the prehistory of languages*. Ed. by P. Forster and C. Renfrew. Cambridge, Oxford, and Oakville: McDonald Institute for Archaeological Research, 91–109.

Bodt, T. A. and J.-M. List (2022). "Reflex prediction. A case study of Western Kho-Bwa." *Diachronica* 39.1, 1–38.

Bröker, F. and M. Ramscar (2021). "Representing absence of evidence: Why algorithms and representations matter in models of language and cognition." *Language, Cognition and Neuroscience* 37.1, 1–24.

Clackson, J. (2007). *Indo-European linguistics*. Cambridge: Cambridge University Press.

Kondrak, G. (2000). "A new algorithm for the alignment of phonetic sequences." In: *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*. (Seattle, 04/29–05/03/2000), 288–295.

List, J.-M. (2014). *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.

— (2016). "Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction." *Journal of Language Evolution* 1.2, 119–136.

— (2018). *Regular cognates: A new term for homology relations in linguistics*. Vol. 5. 8.

— (2019). "Automatic inference of sound correspondence patterns across multiple languages." *Computational Linguistics* 45.1, 137–161.

List, J.-M., C. Anderson, T. Tresoldi, and R. Forkel (2021). *Cross-Linguistic Transcription Systems. Version 2.1.0*. Jena: Max Planck Institute for the Science of Human History. URL: https://clts.clld.org.

List, J.-M., R. Forkel, S. J. Greenhill, C. Rzymski, J. Englisch, and R. D. Gray (2022a). "Lexibank, A public repository of standardized wordlists with computed phonological and lexical features." *Scientific Data* 9.316, 1–31.

List, J.-M., N. W. Hill, and R. Forkel (2022b). "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*. (Dublin, 05/26–05/27/2022). Association for Computational Linguistics. Dublin, 89–96.

List, J.-M., E. Vylomova, R. Forkel, N. Hill, and R. D. Cotterell (2022c). "The SIGTYP shared task on the prediction of cognate reflexes." In: *Proceedings of the 4th Workshop on Computational Typology and Multilingual NLP*. "SIGTYP 2022" (Seattle, 07/14/2022). Association for Computational Linguistics. Seattle: Max Planck Institute for Evolutionary Anthropology, 52–62.

Perkel, J. M. (2022). "Six tips for better spreadsheets." *Nature* 608, 229–230.

Prokić, J., M. Wieling, and J. Nerbonne (2009). "Multiple sequence alignments in linguistics." In: *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*. "LaTeCH-SHELT&R 2009" (Athens, 03/30/2009), 18–25. acm: 1642052.

Ross, M. and M. Durie (1996). "Introduction." In: *The comparative method reviewed. Regularity and irregularity in language change*. Ed. by M. Durie. New York: Oxford University Press, 3–38.

Schweikhard, N. E. and J.-M. List (2020). "Developing an annotation framework for word formation processes in comparative linguistics." *SKASE Journal of Theoretical Linguistics* 17.1, 2–26.

Lecture 3: From Words to Trees

Abstract

In this session, we will discuss how we can get from the histories of individual words to the inference of full phylogenies that represent how languages evolve over time.

1 Lexical Motivation

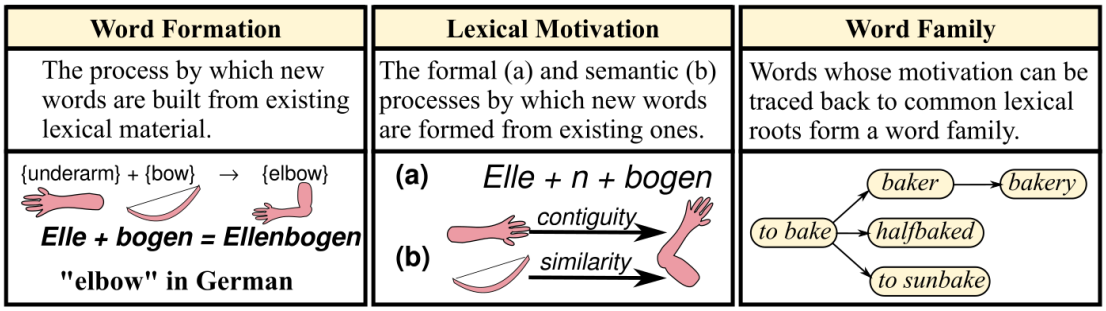
In the second session, we have discussed some major questions regarding cognates, alignments, and sound correspondences. In this session, we will take a rather radical turn and discuss phylogenetic trees and phylogenies in a broader sense. In order to keep the transition smooth, however, we will first look at *lexical motivation* as a major process underlying word formation, which is indispensable in order to understand the major processes of lexical change, which themselves open the doors towards phylogenies and phylogenetic trees.

As we have seen before, in our discussion about alignments and alignability, it can be quite challenging to find examples for regular sound correspondences, not because sound change is not regular, but rather because it has been superseded by numerous morphological processes by which the original shape of word forms has been modified. In order to cope with this problem of drastic information loss when searching for regular sound change processes, scholars have for a long time made use of specific techniques of *internal reconstruction* in order to find the original shape of the word forms which had been changed under the influence of morphological change. These techniques have never been sufficiently formalized, not only, because a formalization might be difficult and language or language-family-specific, but also because the workflows which linguists use often switch back and forth between internal and external reconstructions.

In our group, we have began to work towards a formalization of internal reconstruction by focusing on word formation processes in general and lexical motivation patterns and word families in particular. The results of these efforts have been in part published, but our approaches have evolved with the publications. For this reason, we are still in a stage where we test our methods and try to find a sufficiently large enough number of examples that illustrate how the new techniques can be used. In the following, we will only quickly look at word families, and then show how we try to handle motivation structures with the help of *morpheme glosses*, and how we handle allomorphy with *inline alignments*. I hope that we will be able to publish some more detailed studies presenting our new approaches in the nearer future.

1.1 Word Families

Following Koch (2001), lexical motivation can be defined as summarizing the formal and semantic processes by which new words are formed from existing ones during word formation (Gruaz 2002). Words whose motivation can be traced back to common lexical roots form a *word family* (Hundsnurscher 2002). The following image (taken from a grant application, with the grant to be started in January 2023, which may later be published independently) provides some examples for word formation, lexical motivation, and word families.



Interestingly, scholars do not often talk about word families in historical linguistics. An exception is the field of Sino-Tibetan linguistics, where it enjoys a doubtful reputation, due to the fact that the grouping of words into families is not often done with the help of rigorous principles, see Fellner and Hill 2019). In scholarly practice, however, it is rather the norm than the exception that scholars apply language-specific knowledge about language- or subgroup-internal word formation processes in order to identify those comparanda which substantiate their sound correspondences when comparing words across languages. This internal reconstruction – when carried out rigorously – is often the reason why scholars are convinced of the regularity of sound change, since the "raw" material derived from concrete word forms for concrete terms often does not show regular sound correspondences throughout the whole word without further explanation.

Word families – the result of assembling words in the same language into groups of common descent – are thus central to scholarly practice of historical language comparison, even if there are no common ways to formalize scholars' knowledge on them.

How do compounds fit into the word family schema?

1.2 Morpheme Glosses

When comparing words across languages and inside one and the same language, it is crucial to formally annotate their lexical motivation, that is, the semantic and morphological processes by which they have been derived. As a first attempt to handle these processes – originally tested only on compounds in Burmish languages – Hill and List (2017) introduced *morpheme glosses* as a device to make individual components of words transparent through annotation. Morpheme glosses are similar to *interlinear-glossed text* (Lehmann 2004) but not applied to the sentences in a corpus, but rather to the words in a wordlist or a lexicon. Thus, starting from a morpheme-segmented word form (in our space-segmented representation by which words are modeled as sound sequences), we can annotated individual word parts by providing short glosses which give hints to the original meaning or the grammatical function (inside the complex word). The following table (taken from our grant application) provides some examples of morpheme glosses as represented in the EDICTOR tool (`https://digling.org/edictor`, List 2021a, List et al. 2017).

ID	CONCEPT	TOKENS	MORPHEMES	COGIDS
1794	bow	b o: g + ə n	bow + -:S.m.s/en	11[4] 922[47]
408	elbow	ɛl + ə + n + bo:g + ən	underarm + -S.f.s/e + -l/n + bow + -S.m.s/en	11[4] 91[219] 870 922[47] 922[47]
82	rain (noun)	r e: g ə n	rain	190[2]
71	rainbow	re:gən + bo:g + ən	rain + bow + -:S.m.s/en	190[2] 11[4] 922[47]

After our initial experiments, we successively expanded the idea of using glosses for the annotation of lexical motivation in follow-up studies, trying to show that morpheme glosses are not only apt for the handling of etymological relations between words in the case of compounds but also in the case of derivation (Schweikhard and List 2020). Currently, we are furthermore testing, to which degree these glosses can help us to improve phylogenetic reconstruction and the identification of alignable word parts (Wu and List forthcoming).

Would it be possible to capture universal motivation patterns with the morpheme gloss annotation approach?

1.3 Inline Alignments

So far, we have not fully resolved the problem of handling allomorphy and exceptions in sound correspondence patterns. The problem here is that both allomorphy and (known) exceptions, due to some kinds of irregular sound change, such as, for example, assimilation of frequently used words, analogical processes like contamination, or sandhi phenomena, easily mask the regularity of the correspondences exhibited by the root of a word family. In our formal annotations of linguistic data, we want to capture both the knowledge that certain sounds are unexpected, following our idea of sound laws, but at the same time we want to list them along with the major patterns, showing specifically also what we would *expect* a word to look like if it had been changed by regular processes. A current approach which we are testing on different datasets at the moment are *inline alignments*. An inline alignment is an alignment of a sequence that is represented along with the sequence itself in the same *line* (which is why I call it in-line).

As a very general example, consider German *Eltern* “parents” and its sequential phonetic representation [ɛ l t ɐ n], which we can contrast with its phonological representation [ɛ l t ə ʁ n]. An inline alignment of these two forms, which would take the phonological form as the primary one would look like [ɛ l t Ø/ə ɐ/ʁ n]. So what we do in an inline alignment is we “entangle” both forms in a single sequence in such a way that we list identical segments of the alignment only once, while alternating segments are represented with the help of a dash symbol that separates the “source” of the “target” element. From the perspective of its information, we can generate both the phonological and the phonetic form from this new sequence.

But the principle of representing two sequences in one line does not have to be limited to the representation of phonological and phonetic forms, it can also be used for morphological alternation. Thus [ɛ l t ə ʁ n] can be analyzed as consisting of the comparative of *alt* “old” and the plural ending *-en*, thus referring to “the older (people)”. We could represent the original form as [a l t + ə ʁ + ə n], which we can describe with a morpheme gloss as ○LD :comparative :plural, and in an inline alignment as [ɛ/a l t + ə ʁ + Ø/ə n] (this has been outlined in more detail in List 2021b). The advantage of using the analyzed form as the primary form is that we can now directly compare the inline alignment with

other members of the word family *alt* in German. Furthermore, using the root [a l t] as an anchor point, we can also automatically align all kinds of its derivations to this root form and thus provide a multiple alignment of word forms derived from the same root (the use of multiple alignments for language-internal handling of morphology is uncommon, but there are some examples for recent attempts to make use of them, as shown by Beniamine and Guzmán Naranjo 2021 for inflectional morphology).

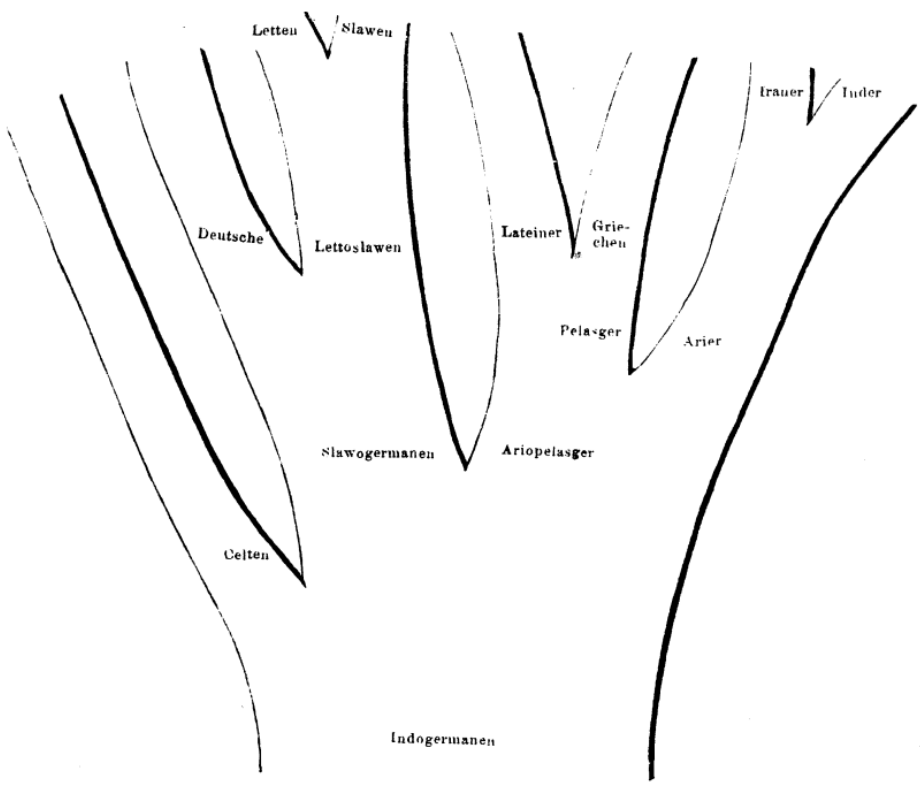
What is the advantage of inline alignments over a listing of all words that belong to one and the same language family, and where are the limits of this approach?

2 Phylogenies

Linguists have known for a long time that languages evolve and that the languages we observe today may stem from common sources which themselves no longer exists. First speculations on the common descent and the tree- or network-like separation of languages can already be found in early studies of the 17th century and thereafter (List et al. 2016). Until the late 18th century, however, the dominant view among scholars in Europe was that all human languages were products of the mythical Confusion of Tongues which prevented the construction of the Tower of Babel (Klein 2004).

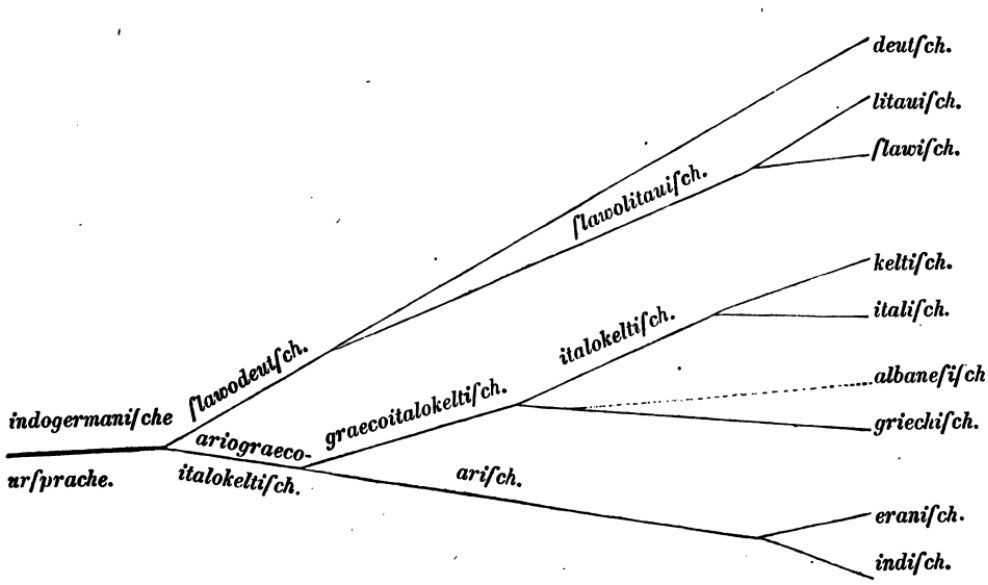
2.1 From Oaks to Tree Diagrams

Unlike modern phylogenetic trees, early linguistic trees were much less formal and systematic, but had the tendency to resemble true trees much more closely. As an example, consider Schleicher's tree from 1853 (Schleicher 1853), which has the appearance of a massive oak with a big trunk. Only later, the family tree visualizations became more schematized, but the interpretation was still far away from being formalized.



As an example for the lack of formalization, consider again a tree by Schleicher, this time from 1861 (Schleicher 1861). While this tree looks much more formalized than the earlier tree from 1853, the description of this tree in the text is interesting, since Schleicher points to branch lengths as representing the supposed time which had elapsed since separation while at the same time emphasizing that the distance between extant languages reflected their synchronic closeness.

> The oldest splits of Indo-European until the development of the funda- mental languages of the language families which constitute the stem of the language [sprachstamm] can be visualized by the following schema. The length of the lines in- dicates the elapsed time, the distance of the lines from each other indicates the degree of relationship. (ibid.: 6f)[1]



While it is difficult to understand the passage by Schleicher completely, it is possible that Schleicher thought of some additional closeness between languages independent of their evolutionary history and tried to mark this in his tree drawing by separating the major subgroups visually from each other in the tree and by placing languages like Albanian and Greek horizontally close to each other while at the same time assigning them a larger divergence time than given for Celtic and Italian. What could such a close placement of languages in a phylogenetic tree reflect from a contemporary perspective on language evolution?

2.2 From Trees to Webs

Not long after Schleicher and some colleagues had propagated their family tree models for the first time, scholars began to contest them. One of the most prominently cited opponents of Schleicher's family trees was Johannes Schmidt (1843-1901), who devoted a complete booklet to contradict Schleicher (Schmidt 1872).

In this study, Schmidt presented concrete data in the form of sets of ho- mologous words ("cognate sets" in linguistic terminology) for the major Indo-European branches. He noted that one could easily find examples for homologs shared exclusively among different possible pairings (Greek vs. Old Indian,

[1] My translation, original text: "Die ältesten teilungen des indogermanischen bis zum entstehen der grundsprachen der den sprachstamm bildenden sprachfamilien laßen sich durch folgendes schema anschaulich machen. Die länge der linien deutet die zeitdauer an, die entfernung derselben von einander den verwantschaftsgrad."

Greek vs. Slavic, Slavic vs. Old Indian) with no residues ("reflexes" in linguistic terminology) in any of the other branches. Based on this finding, Schmidt refuted the family tree hypothesis, arguing that a tree could not explain the observed data.

What Schmidt proposed instead was the rather fuzzy idea of a wave-like expansion of the major branches of the Indo-European languages which contributed to their gradual separation and would explain the specific commonalities between individual pairs which seemed to contradict each other. Unfortunately, Schmidt did not see that the cases he listed could be perfectly explained by the traditional tree model assuming well-known phenomena like differential loss (Geisler and List 2013) or incomplete lineage sorting (Evans et al. 2021, Jacques and List 2019, List et al. 2016). But although his critic was not valid and his alternative model, the "wave theory" (Wellentheorie), as it was called thereafter, did not offer any concrete instructions with respect to the formal modeling of language divergence and spread, many linguists started to present it as a valid alternative to the family tree model.

Nowadays, the wave theory is often presented as some kind of diffusion model in which languages gradually diverge without splitting abruptly. This model of language evolution was already mentioned by Hugo Schuchardt (Schuchardt 1870 [1900]), and family tree models are not capable of modeling the split process in detail. But is it justified to attribute diffusion models to Schmidt's wave theory?

3 Phylogenetic Reconstruction

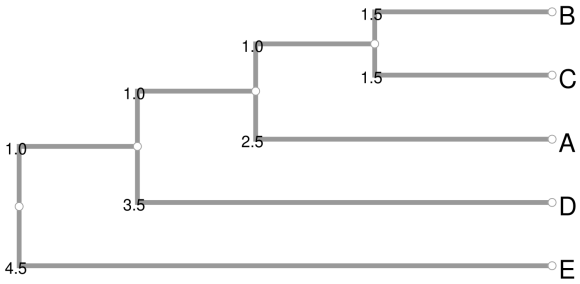
In the following, we will quickly introduce some major concepts of phylogenetic reconstruction as it is practiced in recent approaches to historical language comparison and heavily inspired from evolutionary biology.

3.1 Distance-Based Approaches to Phylogenetic Reconstruction

Distance-based approaches were among the earliest approaches to phylogenetic reconstruction which were proposed by scholars. The basic assumption of distance-based approaches is that one can aggregate certain information about the taxonomic units (languages, species, etc.) in such a way that these aggregated similarities or dissimilarities among the taxonomic units provide enough information to reconstruct a phylogeny of the units in question. While most scholars would contradict this claim nowadays, both in evolutionary biology and in linguistics, distance-based approaches are still useful in order to quickly check a certain dataset, since they are easy to understand and fast and easy to apply.

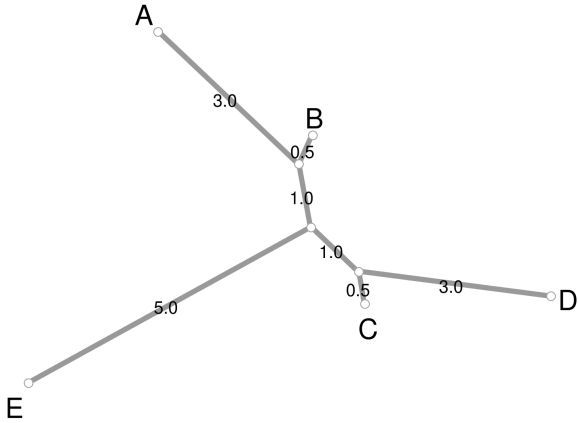
As a first algorithm that can be used to infer phylogenetic trees from the data, there is the algorithm that is nowadays simply called UPGMA by Sokal and Michener (1958). This algorithm is fairly simple, by subsequently merging those languages with each other which show the lowest distance score, and then averaging all distances between the merged languages with the rest, thus creating a new matrix that is then again investigated for the pair with the lowest distance. What is essential about this algorithm is that it yields a rooted tree in which branch lengths are supposed to reflect the true, steady, evolution that occurred. If a distance matrix fulfils the criterion of really reflecting steady evolution (in which change proceeds at the same speed), the distances in the distance matrix are the same as between the branches in the tree. If evolution is not ultra-metric, the UPGMA branch lengths will differ from the distances in the matrix. In such a situation, no rooted tree can display the matrix truthfully. In An unrooted tree, however, can display certain distances, which are called additive. If a matrix shows data which are truly additive, the Neighbor-joining algorithm can be used to find the tree (Saitou and Nei 1987). The difference is illustrated in the following figures.

	A	B	C	D	E
A	0.0	5.0	5.0	7.0	9.0
B	5.0	0.0	3.0	7.0	9.0
C	5.0	3.0	0.0	7.0	9.0
D	7.0	7.0	7.0	0.0	9.0
E	9.0	9.0	9.0	9.0	0.0



(a) *ultrametric distances and corresponding rooted tree*

	A	B	C	D	E
A	0.0	3.5	5.5	8.0	9.0
B	3.5	0.0	3.0	5.5	6.5
C	5.5	3.0	0.0	3.5	6.5
D	8.0	5.5	3.5	0.0	9.0
E	9.0	6.5	6.5	9.0	0.0



(b) *additive distances and corresponding unrooted tree*

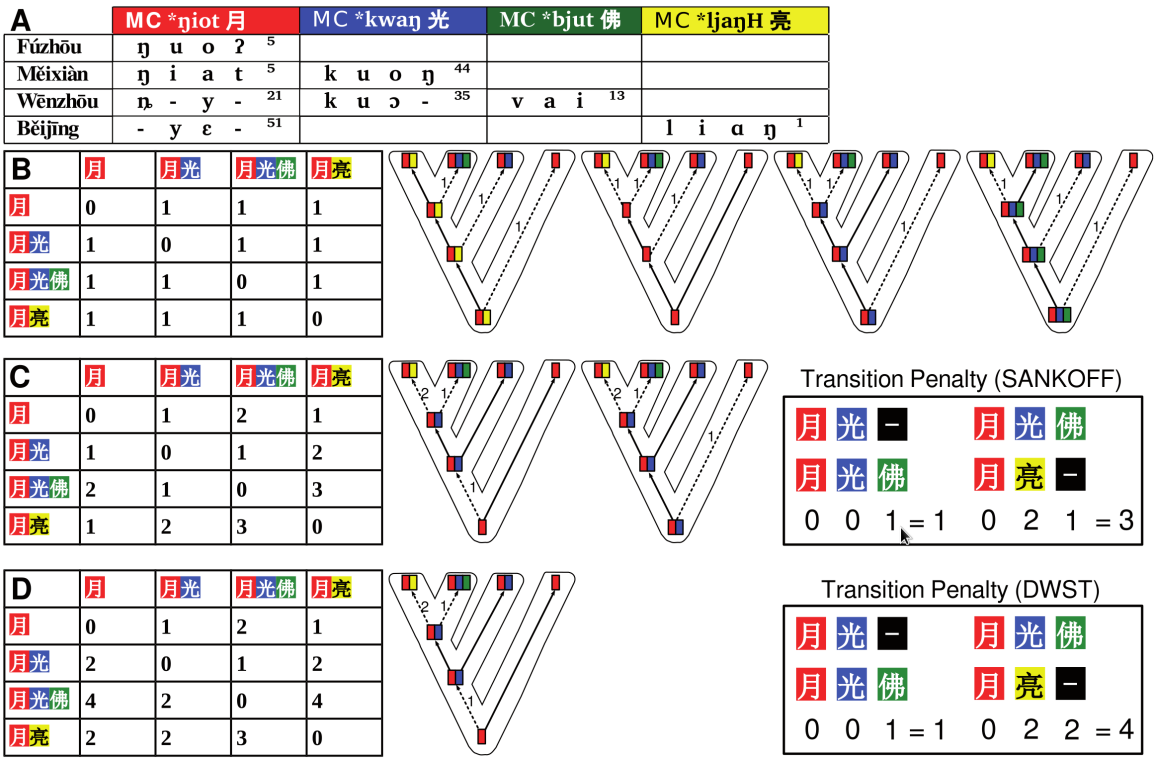
Neighbor-Nets (Bryant and Moulton 2004) show uncertainty in distance-data by representing distance data in the form of a network when data are not additive, but resolving data as a tree, when they are truly additive. Why is it problematic when scholars interpret Neighbor-Nets literally, reflecting concrete evolutionary processes?

3.2 Character-Based Approaches to Phylogenetic Reconstruction

In contrast to distance-based approaches, character-based approaches start from individual features and try to model their evolution along a potential phylogeny. When being given a larger number of features, specific techniques are used in order to infer the phylogeny that explains best how all features evolved under the same phylogeny.

The normal case in which we think about language evolution is by using certain models. Swadesh turned his cognate sets into distance matrices, which means he lost a lot of interesting information on individual processes of lexical replacement, which is a pity, since he had such an interesting model, as we saw last time. Later, in biology, methods were developed to account for the evolution of individual traits, and the most prominent one (at least for a long time) is parsimony or *maximum parsimony*. The idea is that one checks for a certain traits how it evolves along a tree, and while doing so, we penalize certain processes. If the trait does not change, we do not penalize it, but if it turns into something else, we might decide to give it a penalty of 1, 2 or any other value. We could even go so far as to provide specific weights for individual processes. We represent the processes by assuming a finite set of *character states*, and a corresponding *step matrix* in which the transitions by which one character state changes into another state are penalized. The matrix is typically symmetric (in traditional parsimony), but nothing prevents us from using an individually designed matrix. Once this

matrix has been established, we can calculate parsimony in a straightforward way. When working with small trees, one can actually count the cases oneself, without using any complex algorithm. When working with larger trees, using an algorithm is of course a better idea (Sankoff 1975).



The figure above shows different transition matrices (step matrices) and corresponding scenarios for parsimony analyses. What are the major differences between these approaches, and what are the consequences of using a non-symmetric matrix?

3.3 Maximum Likelihood

Maximum parsimony has a great disadvantage which consists in the lack of branch lengths, which are typically not estimated when doing a parsimony analysis. That means, the underlying model assumes that change from an ancestral to a descendant language always occurs in the same *rate*. This is of course not very *realistic* with respect to language change (and biological change), and as a result, alternative approaches were proposed already early, and one of the most important approaches is the so-called *pruning algorithm* by Felsenstein (1973), which provides a new model, based on the likelihood calculation that no longer solves the problem of finding the perfect scenario directly, but instead evaluates all possible scenarios that could happen at once, calculating the *likelihood* of each character state to appear in a certain position in the tree, while the tree is assumed to have branch lengths, which have a direct influence on the likelihood.

What is the advantage of a likelihood model over a parsimony model apart from the possibility to include branch lengths?

References

Beniamine, S. and M. Guzmán Naranjo (2021). "Multiple alignments of inflectional paradigms." *Proceedings of the Society for Computation in Linguistics* 4.21, 1–8.

Bryant, D. and V. Moulton (2004). "Neighbor-Net. An agglomerative method for the construction of phylogenetic networks." *Molecular Biology and Evolution* 21.2, 255–265.

Evans, C. L., S. J. Greenhill, J. Watts, J.-M. List, C. A. Botero, R. D. Gray, and K. R. Kirby (2021). "The uses and abuses of tree thinking in cultural evolution." *Philosophical Transcactions of the Royal Society B* 376.20200056, 1–12.

Fellner, H. A. and N. W. Hill (2019). "Word families, allofams, and the comparative method." *Cahiers de Linguistique Asie Orientale* 48.2, 91 –124.

Felsenstein, J. (1973). "Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters." English. *Systematic Zoology* 22.3, pp. 240–249. JSTOR: 2412304.

Geisler, H. and J.-M. List (2013). "Do languages grow on trees? The tree metaphor in the history of linguistics." In: *Classification and evolution in biology, linguistics and the history of science. Concepts – methods – visualization*. Ed. by H. Fangerau, H. Geisler, T. Halling, and W. Martin. Stuttgart: Franz Steiner Verlag, 111–124.

Gruaz, C. (2002). "The analysis of word families and their motivational relations." Handbooks of linguistics and communication sciences 1.21. Ed. by A. Cruse, F. Hundsnurscher, M. Job, and P. R. Lutzeier, 700–704.

Hill, N. W. and J.-M. List (2017). "Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages." *Yearbook of the Poznań Linguistic Meeting* 3.1, 47–76.

Hundsnurscher, F. (2002). "Das Wortfamilienproblem in der Forschungsdiskussion [Word families in scientific discussion]." Handbooks of linguistics and communication sciences 1.21. Ed. by A. Cruse, F. Hundsnurscher, M. Job, and P. R. Lutzeier, 675–680.

Jacques, G. and J.-M. 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, 128–166.

Klein, W. P. (2004). "Was wurde aus den Wörtern der hebräischen Ursprache? Zur Entstehung der komparativen Linguistik aus dem Geist etymologischer Spekulation." In: *Gottes Sprache in der philologischen Werkstatt. Hebraistik vom 15. bis zum 19. Jahrhundert*. Proceedings of the Symposium "Die Geburt der Philologie aus dem Geist der Hebraistik" (Wittenberg, 10/06–10/06/2002). Ed. by G. Veltri and G. Necker. Studies in European Judaism 11. Leiden: Brill, 3–23.

Koch, P. (2001). "Lexical typology from a cognitive and linguistic point of view." In: *Linguistic typology and language universals*. Handbook of Linguistics and Communication Science 20.2. Berlin and New York: de Gruyter, 1142–1178.

Lehmann, C. (2004). "Interlinear morphemic glossing." In: *Morphology. An international handbook*. Ed. by G. E. Booij, C. Lehmann, J. Mugdan, and S. Skopeteas. Vol. 2. Berlin and New York: De Gruyter, 1834–1857.

List, J.-M. (2021a). *EDICTOR. A web-based tool for creating, editing, and publishing etymological datasets. Version 2.0.0*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: https://digling.org/edictor.

— (2021b). "Using EDICTOR 2.0 to Annotate Language-Internal Cognates in a German Wordlist." *Computer-Assisted Language Comparison in Practice* 4.4. URL: https://calc.hypotheses.org/2735.

List, J.-M., S. J. Greenhill, and R. D. Gray (2017). "The potential of automatic word comparison for historical linguistics." *PLOS ONE* 12.1, 1–18.

List, J.-M., J. S. Pathmanathan, P. Lopez, and E. 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.

Saitou, N. and M. Nei (1987). "The neighbor-joining method: A new method for reconstructing phylogenetic trees." *Molecular Biology and Evolution* 4.4, 406–425.

Sankoff, D. (1975). "Minimal mutation trees of sequences." *SIAM Journal on Applied Mathematics* 28.1, 35–42.

Schleicher, A. (1853). "O jazyku litevském, zvláště ohledem na slovanský. Čteno v posezení sekcí filologické král. České Společnosti Nauk dne 6. června 1853 On the Lithuanian language, with a special focus on Slavic." *Časopis Českého Museum* 27, 320–334. google: cLMDAAAAYAAJ.

— (1861). *Compendium der vergleichenden Grammatik der indogermanischen Sprache*. Vol. 1: *Kurzer Abriss einer Lautlehre der indogermanischen Ursprache*. Weimar: Böhlau.

Schmidt, J. (1872). *Die Verwantschaftsverhältnisse der indogermanischen Sprachen* On the genetic relations among the Indo-European languages. Weimar: Hermann Böhlau.

Schuchardt, H. (1870 [1900]). *Über die Klassifikation der romanischen Mundarten. Probe-Vorlesung, gehalten zu Leipzig am 30. April 1870* On the classification of Romance dialects. Test lecture given in Leipzig on April 30, 1870. Graz. URL: http://schuchardt.uni-graz.at/cgi-bin/print.cgi?action=show\&type=pdf\&id=724.

Schweikhard, N. E. and J.-M. List (2020). "Developing an annotation framework for word formation processes in comparative linguistics." *SKASE Journal of Theoretical Linguistics* 17.1, 2–26.

Sokal, R. R. and C. D. Michener (1958). "A statistical method for evaluating systematic relationships." *University of Kansas Scientific Bulletin* 28, 1409–1438.

Wu, M.-S. and J.-M. List (forthcoming). "Annotating cognates in phylogenetic studies of South-East Asian languages." *Language Dynamics and Change* 0.0, 1–25.

Lecture 4: From Words to Stars

Abstract

This session is mainly devoted to various questions of phonological reconstruction. We will concentrate on recent approaches on supervised phonological reconstruction and end by discussing some so far unpublished ideas on handling uncertainty when dealing with reconstructions.

1 Introduction

In historical linguistics, *linguistic reconstruction* is a rather important task. It can be divided into several subtasks, like *lexical reconstruction*, *phonological reconstruction*, and *syntactic reconstruction*. From the perspective of evolutionary biology, all of these tasks would be handled in a framework of *ancestral state reconstruction*, in which biologists try to identify the ancestral *state* of a given biological feature. In the linguistic literature, we do not often make a clear distinction between lexical and phonological reconstruction. From the perspective of South-East Asian languages, however, this distinction is quite important, since in this field, linguists often only embark on phonological reconstruction without trying to reconstruct full words in ancestral languages.

In my definition of phonological reconstruction, linguists seek to reconstruct the sound system of the ancestral language, the Ursprache that is no longer attested in written sources, by reconstructing individual morphemes back to the ancestral language. The term lexical reconstruction is less frequently used, then points to the reconstruction of whole lexemes in the proto-language, and requires sub-tasks, like semantic reconstruction where one seeks to identify the original meaning of the ancestral word form from which a given set of cognate words in the descendant languages developed, or morphological reconstruction, where one tries to reconstruct the morphology, such as case systems, or frequently recurring suffixes.

While the distinction of phonological and lexical reconstruction may be useful for the investigation of South-East Asian languages, it may be less clear why one would need it for language families like Indo-European. Or can we find arguments that justify the clear distinction here?

2 Reconstruction without Trees

While methods for the automated detection of cognates in multilingual wordlists (List 2017) are now more and more frequently used by scholars to preprocess their data (Gerardi et al. 2022), before manually correcting obvious errors made by the algorithms, methods for phonological reconstruction have so far only been applied to very specific language families like the Austronesian languages (Bouchard-Côté et al. 2013, Hruschka et al. 2015), and despite the success reported by scholars, they have not made their way into the standard workflow of computer-assisted language comparison.

The reason for the lack of application lies, however, not only in the fact that the success stories were only reported for those language families which are considered as “easy” to reconstruct, but also in the fact that phonological reconstruction itself is often misinterpreted by these methods.

The first erroneous assumption of most proposed automatic methods for phonological reconstruction is that the sounds used in a set of attested languages are necessarily the pool of sounds that would also be the best candidates for the Ursprache. Already Saussure (1879) proposed that Proto-Indo-European had at least two sounds that did not survive in any of the descendant languages. The

laryngeals, nowadays commonly represented as h_1, h_2, and h_3, leave complex traits in the vocalism and the consonant systems of some Indo-European languages. Ever since then, it has been a standard assumption that it is always possible that ancestral sounds in a given proto-language are not attested in any of its descendants.

An additional methodological problem of the methods is that they are based on language trees, which are either given to the algorithm or inferred during the process. In contrast to evolutionary biology, where most if not all approaches to ancestral state reconstruction are based on some kind of phylogeny, the classical methods to infer ancestral sounds and ancestral sound systems can often advance well without a phylogeny as a backbone.

The reason for this lies in the highly *directional nature* of sound change, especially in the consonant systems of languages, which often makes it extremely easy to predict the ancestral sound without invoking any phylogeny more complex than a star tree. For example, if a linguist observers a [k] in one set of languages and a [ts] in another languages in the same alignment site of multiple cognate sets, then they will immediately reconstruct a *k for the proto-language, since they know that [k] can easily become [ts] but not vice versa. The same holds for many sound correspondence patterns that can be frequently observed among all languages of the world, including cases like [p] and [f], [k] and [x], and many more. Why should we bother about any phylogeny in the background, if we already know that it is much more likely that these changes occurred independently? Directed character-state assessments make a phylogeny unnecessary.

In which cases may it still be useful to know the phylogeny of a language family when doing a phonological reconstruction?

3 Supervised Phonological Reconstruction

Supervised phonological reconstruction refers to a specific approach to phonological reconstruction in which a part of the data has already been annotated. Thus, we suppose that a researcher has already identified cognates in a dataset and already started to provide proto-forms for at least some part of the data. A supervised method would now *learn* from the existing annotations (the proto-forms) and then use this knowledge to provide proto-forms for so-called *unseen data*. In this way, supervised techniques do not solve our problems for us, but they can help us to speed up the process of data annotation, which may at times be quite tedious. In the following, we will look at some methods in detail and try to understand the basic techniques of a new framework which comes rather close to the way in which historical linguists would carry out reconstruction manually.

3.1 Supervised Phonological Reconstruction on the Rise

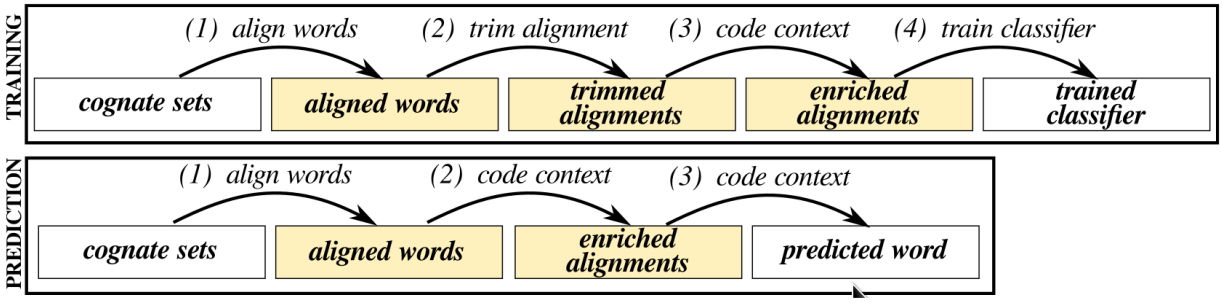
In the past decade, scholars have proposed quite a few new methods for both cognate reflex prediction and supervised phonological reconstruction. List (2019) proposed a new method for the inference of sound correspondence patterns from phonetic alignments, which also allows one to impute sounds in correspondence patterns in which individual reflex sounds are missing, due to data sparsity. Bodt and List (2022) employed this method in order to first predict missing words in Western Kho-Bwa languages and then verify the predictions in subsequent fieldwork. List et al. (2022a) further expanded this workflow by using support vector machines to predict proto-forms from phonetic alignments, improving the performance of the original method for reflex prediction proposed by List (2019). In a recently organized *Shared Task* on cognate reflex prediction List et al. (2022b), the method of List et al. (2022a) proved robust, but was outperformed by two new methods based on neural networks (Kirov et al. 2022), one originally designed for the handling of place name pronunciations in Japan (Jones et al. 2022) and one

designed for the restoration of digital images in which pixels are missing (Liu et al. 2018). All in all, the most successful methods in the shared task all showed good performance: when retaining 90% of the data for training, the methods differed on average by one sound from the attested word forms.

If neural network approaches work so much better than alternative approaches, why should one bother turning to alternative methods for the task of supervised phonological reconstruction?

3.2 Basic Technique for Supervised Phonological Reconstruction

The basic technique we want to present in this context is based on our new framework, which makes use of alignments and correspondence patterns (List et al. 2022a). The framework consists of two major stages, one stage in which the system is trained with existing annotated data, and one stage in which the trained system is applied to new data. The following graphic (taken from ibid.) illustrates the different sub-steps of the major stages.



In the first stage, cognate sets are provided as input to the method (in phonetic transcriptions) and then aligned as a first step (1). The aligned cognate sets are then *trimmed*, that means, that cases in which a sound is only reflected in the proto-language but not in the descendant languages are treated in such a way that the sound in the proto-language is merged with the sound in a preceding column of the alignment (called *trimming* in our study). We then code for conditioning context, by extending the resulting alignments with certain mostly abstract information about the position of the alignment site within the alignment (e.g., if a correspondence pattern is in the beginning of a word) and we finally use the alignments computed in this way to train a classifier, that is, a machine learning system that can learn which proto-sound to yield when seeing only the rest of the data without the proto sound.

When applying this model to then reconstruct proto-forms from alignments, we repeat all steps, apart from the trimming, which relies on the existence of the proto-form in the alignment. The system can thus now be used to predict a proto-form for a given proto-language based on a cognate set.

Why should one code abstract context, why can't we code complex context, like preceding and following sounds, and the like?

3.3 Supervised Phonological Reconstructions in Historical Language Comparison

While the task of unsupervised phonological reconstruction, where algorithms would reconstruct a proto-language from cognate sets from scratch, has not been sufficiently solved so far, we can see that phonological reconstruction in a supervised setting has become a real option and could be integrated into computer-assisted workflows, in which scholars first annotate parts of their data, then compute new reconstructions automatically, and later refine them again. Given that the performance of our systems

was quite satisfying so far, we believe it would be useful to work more with supervised phonological reconstructions, especially on language families with fewer resources.

Are there possibilities to estimate how well a reconstruction works for a given family, in order to make sure that the proposals do not lead one astray?

4 Uncertainty in Phonological Reconstruction

Although the results are inherently preliminary (as witnessed by the changes in Schleicher's "Fabel" over the last decades), linguists typically present their results in the form of discrete phonological units, giving the impression of exactitude and rigor. Thus, although phonological reconstructions change with time, when our knowledge or our assumptions about a language family change, we typically provide the results as if they were final.

4.1 Representation of Uncertainty in Reconstruction in the Literature

With respect to the representation of reconstruction uncertainty, linguists typically come up with ad-hoc solutions for individual language families or individual enterprises. In Indo-European studies, scholars express their uncertainty with respect to the three laryngeals (*h_1, *h_2, or *h_3) by writing a capital *H. In their reconstruction of Old Chinese, Baxter and Sagart (2014) employ a complex notation system that puts uncertain parts of their reconstruction into brackets (with -[n] meaning, for example, that the reconstruction could be either the final *n or to *r). In other cases, scholars mention alternative reconstructions only in comments. While both manual and automated methods are inherently fuzzy with respect to phonological reconstruction, so far, few methods known to us have explicitly embraced fuzziness, trying to present uncertainty in reconstructions explicitly. An exception was the method of List (2019), which offered degrees of uncertainty in the imputation of missing sounds in aligned cognate sets, but the fuzzy reconstructions were not further evaluated or inspected.

What practices of representing uncertainty do you know of and what practices do you use?

4.2 A new Proposal for the Representation of Uncertainty

In Hill and List (2022), we follow Bodt and List (2022) who represent multiple options for the prediction of an individual sound by using the pipe symbol | as a separator for the different options. The symbol is often used in the meaning of "or" in regular expressions, which makes it particularly apt to represent uncertainty, since we can interpret a fictitious proto-form like [p a|i t] as a kind of a regular expression that matches both the form [p a t] and [p i t]. Note that this notation needs to be used with some care when more than one sound is treated as uncertain, since the resulting expression will always match the Cartesian product of the uncertain sounds. Thus, a fictitious proto-form [p a|i t|d] would match four distinct proto-forms, namely the forms [p a t], [p i t], [p a d], and [p i d]. If scholars want to explicitly propose two different proto-forms only, e.g. [p a t] vs. [p i d], our notation cannot be used. We recommend instead to assume two distinct forms, which can both be proposed as possible proto-forms for a given cognate set. Our fuzzy notation is thus only reserved for cases where the uncertainty is independent of contextual information that could be derived from the proto-form.

Why can uncertainty not rather be represented on substrings instead of single characters?

4.3 Computing Fuzzy Reconstructions

Our method for the creation of fuzzy reconstructions is straighforward. We expand the framework for supervised phonological reconstruction proposed by List et al. (2022a), by drawing several samples from the same data, in which different parts of the forms are intentionally ignored. While the framework of List et al. starts from a training set, in which proto-forms are provided and then a model is trained that can be used to predict proto-forms for data that has not been seen before, we draw multiple samples, drop a certain number of words from each sample, and use the method by List et al. to train the "classifier" that can be used to predict proto-forms from aligned cognate sets. Since we drop data in each of the samples, each sample will produce slightly different proto-forms, depending on the data which has been randomly ignored. The different proto-forms offered may point to problems in the original data, or reveal cognate sets that in fact underspecify the proto-form.

In the default settings of our method, we create 10 proto-form predictors from the annotated data and remove 10% of the word forms in each of the samples. When creating an individual reconstruction, we feed our method with a concrete cognates set and then use all 10 predictors to predict proto-forms. The predictions are then summarized, and we count for each position in the original alignment, how often which proto-sound occurs. These fuzzy reconstructions are then represented in the form of a sequence in which column of the alignment is represented by at least one sound, and each possible sound is provided with a frequency in which it occurs in our 10 samples. The table below provides an example from the Burmish data for the fuzzy prediction procedure and the specific output produced by our method.

Reconstruction	Initial	Nucleus	Coda	Tone
Predictor 1	d	u	-	²
Predictor 2	d	u	-	¹
Predictor 3	d	u	-	¹
Predictor 4	d	u	-	⁴
Predictor 5	d	u	-	⁴
Predictor 6	d	u	-	⁴
Predictor 7	d	u	-	⁴
Predictor 8	d	u	-	⁴
Predictor 9	d	u	-	⁴
Predictor 10	ˀt	u	-	⁴
Summary	d:90|ˀt:10	u:100	-:100	⁴:80|¹:20|²:10
Proto-Form	d	u		²

Since certain irregularities in the input data may be filtered from the different samples, irregular patterns which could lead an algorithm to propose erroneous proto-forms, will be filtered out, and in this way the overall robustness of individual reconstruction can be tested.

Robustness can be investigated with respect to individual proto-forms, but what other perspectives on robustness could be invoked with this approach?

4.4 Visualizing Fuzzy Reconstructions

Apart from the technical representation shown above, we have experimented with different ways to represent uncertainty or "fuzziness" in the tools we use to annotate etymological data. Since the manual curation of the cognate sets was carried out with the help of the EDICTOR (List 2021, List et al. 2017,

`https://digling.org/edictor`), a web-based tool for the creation and curation of etymological datasets, we extended the EDICTOR representation of phonetic alignments by adding a representation which we call quintile-representation. In this representation, we represent the frequencies observed in the ten predictions with the help of a table with five rows, in which each row represents the attested symbols (converted from 10 to 5, to keep the table representation neat).

An example of this representation is given in the following figure, where we contrast the original alignment of the cognate set "belly" in the Burmish languages with the quintile representation for the fuzzy reconstruction. As can be seen, the quintile representation does not show all uncertainties in the initial ([ˀt] is missing) and the tone ([ˀ] is missing), since these occur only in 10 percent of all samples.

DOCULECTS	CONCEPTS	ID: 80			
AchangLongchuan	belly	t	au	-	31
Bola	belly	t	au	-	31
Lashi	belly	t	ou	-	33
Maru	belly	t	u	k	31
ProtoBurmish	belly	d	u	-	2

(a) Phonetic alignment of all word forms (including the proto-form)

COGIDS: 80			
d	u	-	2
d	u	-	4
d	u	-	4
d	u	-	4
d	u	-	4
d	u	-	1

(b) Quintile representation

Can you think of other ways in which uncertainty could be further visualized?

4.5 Application of Fuzzy Reconstructions

As a rather simple first approach to study the consequences of uncertainty on reconstructions, we can look at those sounds which are frequently *confused* in our reconstructions. For the Karenic dataset (Luangthongkum 2019) and the Burmish dataset (Gong and Hill 2020) which we investigated in our study (Hill and List 2022), we identified the following sounds to be confused most frequently.

(a) Karenic Data

Sound A	Sound B	Frequency
n	n̥	19
n	ɴ	14
l	l̥	10
55	0	9
ɴ	ŋ	9
ˀ	-	6
m̥	m	6
11	0	5
k	g	5
r	r̥	4

(b) Burmish Data

Sound A	Sound B	Frequency
1	4	18
-	ŋ	9
k	-	8
-	ˀ	8
-	r	8
4	3	7
ˀs	s	6
ˀk	g	6
2	3	6
r	j	6

As can be seen from the individual results for the Karen and Burmish data, the particular problems are quite different across both datasets and cannot be directly compared with each other. A major difficulty in the Karenic data is the reconstruction of voiceless sonorants ([n̥], [l̥], [m̥], etc.), which the author proposes on the basis of the tonal development in some of the descendant languages (Luangthongkum 2019). Since there are quite a few exceptions with respect to the tonal development, we find that the

original reconstruction itself cannot always indicate clearly whether a proto-sound should be voiced or voiceless, which is at times marked by putting the h, which is used to mark a sonorant as voiceless in parentheses (resulting in forms like (h)n-, ibid.). The confusion of the tone marked as [^{0}] with other tones results from our annotation practice of certain weak syllables, in which originally no tone was reconstructed. Since we wanted to indicate a tone nevertheless, to fill the slot in our alignment, the [^{0}] thus marks an underspecified value, which – as the fuzzy reconstructions show – might just as well be given a more concrete reconstruction.

In the Burmish data, on the other hand, we find three major types of confusion. The first relates to the reconstruction of tones. The reconstruction here is often predicted by the nature of the final consonants, which are not actively used in the automated reconstruction method. This may explain the confusion in this case. The second case relates to the reconstruction of gaps (marked by the symbol [-]), which are often confused with sounds occurring in coda position, such as [ŋ], [r], or [ʔ]. The confusion of pre-glottalized initials like [ˀs] and [ˀk] and their non-glottalized counterparts also results from the fact that the reconstruction of pre-glottalized initials depends on the vowels that appear as reflexes in certain Burmish languages. Since this information was not taken into account by our automated method, it is not surprising that results may vary here.

The confusion resulting from information that is not represented in the individual column of an alignment but in other parts shows that additional analyses in which we take the vowel information in the Burmish languages and the tonal information in the Karenic languages into account would be useful in the future. But how could one implement these analyses in concrete?

4.6 Uncertainty due to Problematic Cognate Judgements

As another concrete example for the benefits of checking for uncertainty in reconstruction, the method allows us to identify quite a few cases where individual cognate judgments turned out to be erroneous and should be modified in future versions of our data. As an example, consider cognate set #288 “dung (horse)” in our Burmish data, shown the table below. That erroneous cognate judgments occur in larger etymological projects is inevitable to some degree. Here, our method for the reconstruction of “fuzzy” proto-forms directly helps us to identify and eliminate these problems in future releases of our data.

Achang	m̥	ʐ	a	ŋ	^{31}
Old Burmese	kʰ	j	i	j	^{5}
Fuzzy Proto-Form	ˀm:70|ˀk:30	r:50|j:30|-:20	i:80¦a:20	-:100	^{2}:100
Proto Burmish	ˀk	j	i		^{2}

What other use-cases of the approach for the reconstruction of fuzzy proto-forms could one think of?

4.7 Context-Dependency of Reconstructions

While phonological reconstruction can in the majority of the cases be successfully carried out by considering individual correspondence patterns alone, there are certain cases where it is not enough to look at a pattern in isolation. What needs to be done instead is to evaluate the pattern in combination with other patterns from the same alignment. Although our method for automatic phonological reconstruction was designed in such a way that it can in theory account for this context-dependency of individual reconstructions, we did not take specific and known processes of sound change in the Burmish and the

Karenic data into account, when applying our method to the data. This was done on purpose, since we wanted to see how far we can get with a unified approach. Individual reconstruction errors and cases of uncertainty in the automated reconstruction, however, show that context-dependency should be accounted for in future applications of our approach.

As an example for the problems resulting from ignoring context-dependencies, the following table shows the reconstruction for the cognate set #536 "shy, be / bashful" in the Burmish data. As we can see, Lashi has a retracted vowel (indicated by the bar under the vowel, shaded in gray in the table). Retracted vowels are taken as evidence for the reconstruction of pre-glottalized initials in Proto-Burmish, while the correspondence pattern of the initial itself does not provide concrete evidence for the presence or absence of pre-glottalization. As a result, we can see that the automated method is uncertain, proposing a pre-glottalized initial in 70% of the cases, and a plain initial in 30%.

Achang	s	ɔ	ʔ	55
Atsi	p	a	n	21
Bola	x	a	ʔ	55
Lashi	ʃ	ɔ̱	ʔ	55
Maru	s	o	ʔ	55
Rangoon	tθ	ɑ	-	53
Xiandao	s	ɔ	ʔ	55
Fuzzy Proto-Form	ˀs:70|s:30	a:100	k:100	⁴:100
Proto Burmish	ˀs	a	k	4

Can context-dependency be addressed with the current methods, and what would one have to keep in mind when doing so?

References

Baxter, W. H. and L. Sagart (2014). *Old Chinese. A new reconstruction.* Oxford: Oxford University Press.

Bodt, T. A. and J.-M. List (2022). "Reflex prediction. A case study of Western Kho-Bwa." *Diachronica* 39.1, 1–38.

Bouchard-Côté, A., D. Hall, T. L. Griffiths, and D. 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, 4224–4229.

Gerardi, F., C. Aragon, and S. Reichert (2022). "KAHD: Katukinan-Arawan-Harakmbut Database (Pre-release." *Journal of Open Humanities Data* 8, 18.

Gong, X. and N. W. Hill (2020). *Materials for an Etymological Dictionary of Burmish.* Geneva: Zenodo.

Hill, N. W. and J.-M. List (2022). *Fuzzy reconstructions. A new framework for the representation and computation of uncertainty in phonological reconstruction.* talkconference "International Conference on Historical Linguistics" (Oxford, 08/01–08/05/2022).

Hruschka, D. J., S. Branford, E. D. Smith, J. Wilkins, A. Meade, M. Pagel, and T. Bhattacharya (2015). "Detecting regular sound changes in linguistics as events of concerted evolution." *Curr. Biol.* 25.1, 1–9.

Jones, L., R. Sproat, and H. Ishikawa (2022). *Helpful Neighbors: Leveraging Geographic Neighbors to Aid in Placename Pronunciation.* In preparation.

Kirov, C., R. Sproat, and A. Gutkin (2022). "Mockingbird at the SIGTYP 2022 Shared Task: Two types of models for the prediction of cognate reflexes." In: *The Fourth Workshop on Computational Typology and Multilingual NLP.* Online: Association for Computational Linguistics.

List, J.-M. (2017). "A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets." In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations.* Valencia: Association for Computational Linguistics, 9–12.

— (2019). "Automatic inference of sound correspondence patterns across multiple languages." *Computational Linguistics* 45.1, 137–161.

— (2021). *EDICTOR. A web-based tool for creating, editing, and publishing etymological datasets. Version 2.0.0.* Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: https://digling.org/edictor.

List, J.-M., S. J. Greenhill, and R. D. Gray (2017). "The potential of automatic word comparison for historical linguistics." *PLOS ONE* 12.1, 1–18.

List, J.-M., N. W. Hill, and R. Forkel (2022a). "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.* (Dublin, 05/26–05/27/2022). Association for Computational Linguistics. Dublin, 89–96.

List, J.-M., E. Vylomova, R. Forkel, N. Hill, and R. D. Cotterell (2022b). "The SIGTYP shared task on the prediction of cognate reflexes." In: *Proceedings of the 4th Workshop on Computational Typology and Multilingual NLP.* "SIGTYP 2022" (Seattle, 07/14/2022). Association for Computational Linguistics. Seattle: Max Planck Institute for Evolutionary Anthropology, 52–62.

Liu, G., F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro (2018). "Image Inpainting for Irregular Holes Using Partial Convolutions." In: *Proceedings of the 15th European Conference on Computer Vision (ECCV 2018).* Ed. by V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss. Preprint. Munich, Germany: Springer International Publishing, 89–105.

Luangthongkum, T. (2019). "A view on Proto-Karen phonology and lexicon." *Journal of the Southeast Asian Linguistics Society* 12.1, i–lii.

Saussure, F. d. (1879). *Mémoire sur le système primitif des voyelles dans les langues indo- européennes.* Leipzig: Teubner.

Lecture 5: From Words to Deeds

Abstract

This session focuses mainly on practical applications of the different methods which have been discussed in the previous sessions. We focus on two major tools, one software library (LingPy) and one web-based tool that can be easily used and applied by all who have access to the internet (EDICTOR).

1 Overview

On this last day of our small lecture series, we want to look into the practical consequences of what has been discussed so far. We will focus on two basic tools which I use in my own research on computer-assisted language comparison, LingPy (List and Forkel 2022b) and EDICTOR (List 2021a). Both tools are being actively developed and are freely accessible online. While LingPy is a software library for quantitative tasks in historical linguistics which requires basic programming skills, EDICTOR can in theory be used and applied without any deeper knowledge of programming. However, in order to get started with EDICTOR, it is still important to have a good account of certain data formats that are commonly used, and the major ideas behind the tool, which may at times not be exactly what historical linguists expect, specifically those linguists who have been used to working out their etymologies in very free formats or in prose form. Before we discuss these two tools in detail, however, we will have a look at general principles and recommendations for the handling of data in historical linguistics.

2 Data Preparation

Linguists who prepare their own data often come up with their own, seemingly convenient, formats for data representation which often have huge disadvantages in terms of transparancy and interoperability. It is therefore generally recommended to pay close attention to the basic formats underlying LingPy and EDICTOR on the one hand and the Cross-Linguistic Data Formats (CLDF) initiative on the other hand (Forkel et al. 2018), which are by now largely compatible, with tools that allow for a facilitated conversion from one format to the other (Forkel and List 2020, List 2021b).

2.1 General Rules

Most data annotation in linguistics, even the annotation of texts, can be conveniently represented in a table, and scholars tend to make extensive use of tables – using different kinds of spreadsheet software – when annotating their data. The most obvious failure when preparing data in tables is to include multiple different types of information into one cell. Thus, if a word has a variant, scholars often place it into one cell in their tables and separate the entries by a comma, a colon, a tilde, a dash, or at times even by a back-slash, often even using all of these separators inconsistently for the same dataset. A first and general rule for data creation is therefore (1).

(1) Only one type of information should be put into one cell in a spreadsheet.**

This rule is extremely important and should not be negotiated.In our experience resulting from working with a large number of differently coded datasets, which we retro-standardized as part of the Lexibank project (List et al. 2022a), annotation errors are usually inevitable, even when scholars try to be consistent. Computers are not like humans, and if one wants to profit from computers to ease one's work,

one needs to understand that computers cannot guess whether commas and slashes are used with or without without semantic difference when listing word variants. Even humans often have a hard time to understand the meaning of different separators used in the same dataset, even if they created the data themselves.

A more general rule deriving from this first rule is the rule (2).

(2) All information valid for a given analysis needs to be consistently annotated.**

This means, for example, that, if root alternation is important for one's reconstruction and cognate decisions, one needs to think of ways to represent root alternation in consistent markup. If one's data contains reflexes of an alternating protoform **ka-* vs. **ku-*, for example, it is not sufficient to simply write **ka-* ~ **ku-* and listing the reflexes, assuming that readers will understand which reflex stems from which of the two alternants. Instead, Two proto-forms should be listed, the variants should be assigned to the correct proto-form from which they evolve, and the additional information should be given that **ka-* and **ku-* are variants of the same root. This practice is rarely used in etymological dictionaries, and therefore also often disregarded in databases. It is, however, obvious that it is the only way to transparently list what reflex stems from which proto-variant. In a broader sense, this is not a only matter of a more computational approach to historical linguistics, but rather a matter of improving our common practices in historical linguistics, which have been for too long a time based on very lax guidelines.

What is the problem with lax guidelines in a scientific discipline like historical linguistics?

2.2 Representing Data in Tables

Apart from texts, tables are the most frequently used data structures in linguistics. Specifically in historical linguistics, scholars tend to represent all kinds of data in tables, ranging from wordlists, via cognate sets, up to regular sound correspondence patterns. It is therefore not surprising that computer-assisted tools for historical language comparison also invoke a tabular format. Not all tabular formats used in the field, however, are useful for data analysis.

Concepts	Languages				
	English	German	Dutch	Danish	Swedish
"hand"	hænd	hant	hɑnt	hʌnˀ	han:d
"ashes"	æʃ	aʃə	ɑs	asg	as:ka
"bark"	bɑ:rk	rɪndə	bɑst	bɑ:g	bar:k
...	...	...	...	...	...

A: Concepts/Languages table format.

Concepts	Languages				
	English	German	Dutch	Danish	Swedish
"bark"	bɑ:rk	rɪndə	bɑst	bɑ:g	bar:k
"bark"		bɔrkə			

B: Using additional rows for concepts.

Concepts	Languages				
	English	German	Dutch	Danish	Swedish
"bark"	bɑ:rk	rɪndə, bɔrkə	bɑst	bɑ:g	bar:k

C: Using separators inside cells.

Concepts	Languages					
	English	German	German (b)	Dutch	Danish	Swedish
"bark"	bɑ:rk	rɪndə	bɔrkə	bɑst	bɑ:g	bar:k

D: Additional language column.

As an example, consider the widely used format by which languages are represented in columns and concepts are represented in rows, which is shown in the figure A above. This format lacks flexibility, as there is only one piece of information that we can give for each concept in a given language. When dealing with more languages it becomes more and more impractical, since it is difficult to inspect all languages on a screen, specifically because scrolling horizontally is always more difficult than scrolling vertically.

Despite its shortcomings, the ”language-columns-concepts-rows” format is one of the most widely used formats in historical linguistics, and scholars even have often extended it in order to allow for the display of cognacy among the words listed in the individual cells, or to make it possible to present more than one word form as the translatioon of a given concept in a given language. Here, the problems of the format become even more evident, since cognate information is often added in an ad-hoc manner. Thus, for the handling of synonyms, the STARLING format (Starostin 2000) adds additional rows for

individual concepts, as shown in figure B above. Other studies use commas or other separators to display more than one entry in the same cell (a format commonly used to import data into the ReflLeX database, Segerer and Flavier 2015, as shown in figure C. We even find cases where additional columns for languages with synonyms are added, as illustrated in figure D.

When it comes to the annotation of cognate sets inside these tables, the formats become even more creative, ranging from color-coding to represent cognate words, via multi-sheet formats, up to cases where scholars even binarize their cognate data manually, instead of having this done automatically (compare examples in Forkel et al. 2018).

Software packages like STARLING try to circumvent the problems resulting from the basic tabular format by allowing for additional columns which add additional information for the same language. As a result, STARLING tables currently have three columns for each language, one for the original word form, one for the cognate judgments, and one for comments. LingPy and EDICTOR, the two packages which provide the major methodology discussed in this study, however, employ a different approach which greatly increases the flexibility of the format.

The major principle of this approach is to reserve *one row* in the spreadsheet for exactly *one word form*. Additional information for each word form is provided in additional columns (which can be flexibly added by the user). The content of each column in these EDICTOR-spreadsheets is given in the header of the file, with the first column being reserved for a numeric ID which should be greater than 0. This column should be called ID. Additional columns can be flexibly ordered, but should provide basic information on the name of the language (usually called DOCULECT), on the concept that is expressed by the word (called CONCEPT), and – in order to be able to compare words – a column providing the word form segmented into individual sound units (called ‘TOKENS‘). Depending on the analysis one wants to carry out, additional columns need to be supplied, but they can be empty when starting with an analysis. Thus, in order to store cognate sets, a column for full cognates (often called COGID) or a column for partial cognates (often called COGIDS) should be supplied. In order to annotate morphemes with the help of morpheme glosses, a column storing this information should be added (often called MORPHEMES). In order to store phonetic alignments, an alignment column (often called ALIGNMENT) should be added.

ID	DOCULECT	CONCEPT	VALUE	FORM	TOKENS	BORROWING	COGID
3631	East_Futuna	above	à/luga/	luga	l u g a	0	1382
284	Wallisian	above	'o/luga/	luga	l u g a	1	1382
5391	Futuna_Aniwa	above	weihlunga	weihlunga	w e i + ʰl u ŋ a	0	1382
761	Maori	above	i runga	i runga	i _ r u ŋ a	0	1382
3332	North_Marquesan	above	'una	'una	ʔ u n a	0	1382
4214	Mele-Fila	all	euči	euči	e u tʃ i	0	1115
3917	Pukapuka	all	katoa(toa)	katoa	k a + t o a	0	293
560	Proto-Polynesian	yellow	*reŋareŋa, *felo(-felo)	*reŋareŋa	r e ŋ a + r e ŋ a	0	162
560	Proto-Polynesian	yellow	*reŋareŋa, *felo(-felo)	*felo	f e l o	0	230

The figure above shows a sample table for a dataset on Polynesian languages (List et al. 2018), in which an extra column for borrowings (BORROWING) was added, in which information on borrowing is stored in a binary format (if a word is considered to be borrowed, it is given a 1, otherwise a 0). Additionally, a column for the original value in the original data has been added (VALUE), as well as a column showing the intermediate word format extracted from the value without segmentation (FORM).

If one wants to prepare one’s data in this format, it is recommend to start with a spreadsheet editor (such as Excel, LibreOffice, or GoogleSheets), where one inserts the values as indicated above. In order to convert this spreadsheet into the required tab-separated value format, the easiest way is to

create an empty file with the ending `.tsv`, to open the file with a text editor (a program like Word should not be used), and to copy-paste all columns and rows with values (individual cells can be empty of course, but no rows and no columns should be fully empty and all columns should have a header and all rows should have an ID) into this file. Having done so, the data can be directly accessed with LingPy from within Python scripts, or loaded into EDICTOR, where one can directly manipulate it.

What is so important about tables when discussing data in historical linguistics?

3 EDICTOR

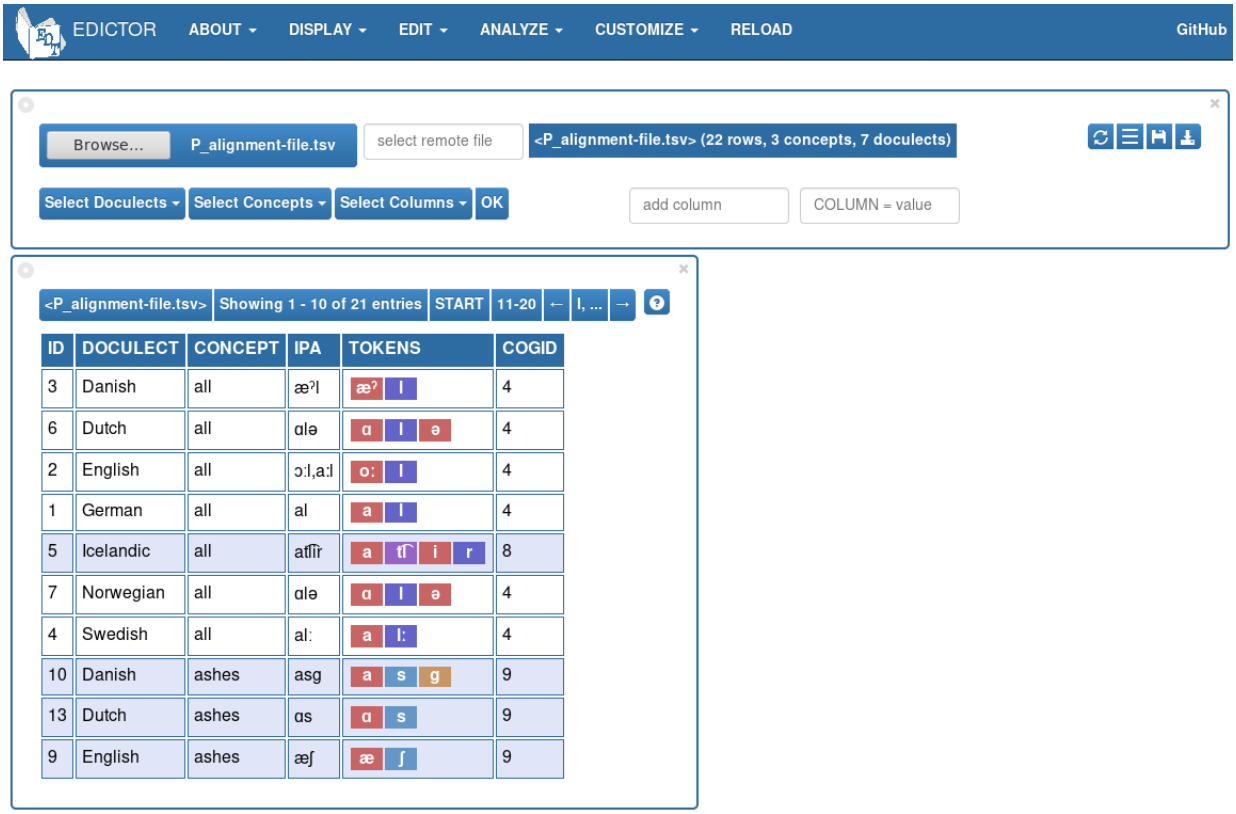
The EDICTOR (`https://digling.org/edictor`, List 2017, List 2021a) is a web-based tool for the curation of etymological data in historical linguistics. The tool has a modular structure which is organized in the form of *panels*. Panels are windows which open once data was loaded into the tool, and users can investigate their data by loading different panels at the same time. The basic panel, the WORDLIST panel, is used to edit data, similar to the way in which this can be done in a spreadsheet editor. Additional panels help to annotate cognates, to align words, or to analyze the data interactively.

3.1 Getting Started with the EDICTOR

What users need in order to use the tool is a text-file encoded in the form in which it was discussed in the previous section, that is, a file in the standard format in which each word is given a row, and a header informs which type of data a certain column contains. In order to use the tool, users need to open the website, located at `https://digling.org/edictor` (for the development version, see `https://lingulist.de/edev` in their browser and drag their file into the BROWSE button which which shows up on the top left of the window.

EDICTOR is written in plain JavaScript. When invoking the tool in this form, the code runs entirely on the system of the user and no data will be send to any servers. As a result, no data can be stolen, users cannot be tracked, and nobody even knows that one is using the tool. Thus, there is no need to be afraid that using the EDICTOR tool will result in data theft in any form. Apart from the fact that it is unlikely that anybody would appropriate somebody else's collected data, it is also not possible, as long as users load a file from their computer into the system and later also export it back from there.

The following figure shows the typical Wordlist panel of the EDICTOR which opens after loading a dataset.



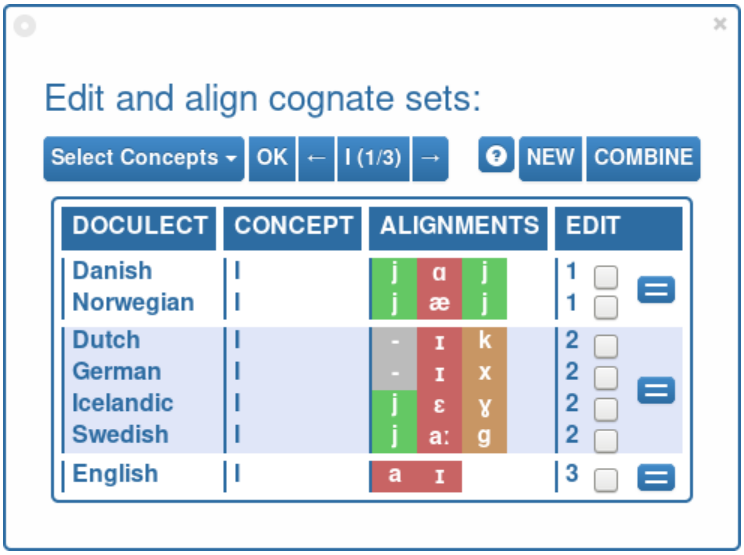
Why is it not possible that the EDICTOR system will save users' data when somebody uses it through the web interface?

3.2 Features of the EDICTOR

The basic structure of the EDICTOR tool is based on *panels*, which allow users to inspect, modify, or analyze their data in specific regards. Panels can also be used to interact together in solving certain tasks. Thus, there is a panel for the annotation of cognate sets, a panel for the annotation of partial cognate sets (which will probably disappear from a future version or frozen and no longer developed actively), and there is a panel to edit the basic data in tabular form and a rather new panel to edit morpheme glosses and partial cognates at the same time.

For the inspection and analysis of the data, there is a panel allowing to check the phonology of a given language variety (along with the possibility to create an IPA chart), there is a panel to inspect various forms of colexifiations (which will be significantly altered in future versions), and there are two important new panels, one showing the cognate set distribution across concepts and languages, allowing also for an export to Nexus format, and one allowing scholars to investigate sound correspondence patterns in their data.

The following figure shows the Cognates panel of the EDICTOR, which can be used to annotate cognate sets in a consistent manner.



What is the difference between annotation and inspection when it comes to the EDICTOR panels?

4 LingPy

LingPy (`https://lingpy.org`, List and Forkel 2022b) is a Python library for quantitative tasks in historical linguistics. With the help of LingPy, several methods which are important for the analysis of wordlists and etymological data can be carried out automatically, including *phonetic alignment analyses* and automated cognate detection (List 2014), basic phylogenetic analyses using the UPGMA (Sokal and Michener 1958) or Neighbor-joining algorithm (Saitou and Nei 1987), and several tasks that help to manipulate word list data in various forms. A couple of years ago, we started to extend LingPy with LingRex (`https://pypi.org/project/lingrex`, List and Forkel 2022c), a Python library dedicated to various tasks related to linguistic reconstruction, which specifically offers access to the new algorithms for correspondence pattern detection (List 2019), and phonological reconstruction (List et al. 2022b). An important feature of LingPy and LingRex is that the basic formats that the libraries read and write are directly compatible with the tab-separated wordlist formats required by the EDICTOR tool. As a result, we now use EDICTOR and LingPy/LingRex in combination, and preprocess a given dataset automatically in order to later refine it manually (Wu et al. 2020).

4.1 Getting Started with LingPy

Installing LingPy and LingRex should by now no longer be a problem for those who have some experience with the installation of Python libraries. Both libraries work without problem on Windows, MacOS, and Linux systems. Numerous tutorials exist, both in the form of articles (List et al. 2018) and in the form of online tutorials (`https://lingpy.org`), and we try to regularly update the basic documentation of our software packages. Additional information can also be found in our blog, where we provide tutorials and howtos on computer-assisted language comparison on a regular basis, with at least one contribution per month (`https://calc.hypotheses.org`). In case of questions it is never wrong to write me an email to inquire, or to file an issue on our GitHub pages (`https://github.com/lingpy/lingpy` and `https://github.com/lingpy/lingrex`).

Where should one start if one does not know anything about programmiing at all?

4.2 Features of LingPy

As mentioned, LingPy and LingRex offer a larger array of methods and implementations for algorithms that were proposed to solve certain problems in historical linguistics, such as the detection of cognates in multilingual wordlists, the alignment of sound sequences, or the reconstruction of phylogenetic trees, as well as initial approaches to phonological reconstruction and borrowing detection (List and Forkel 2022a). It is furthermore important to note that the models for data handling which we have discussed in the course so far are all directly implemented (or represented) in LingPy and thus help us to make sure that our theoretical approaches to data representation and modeling prove useful in practice. In the future we will try to expand LingPy and LingRex further, adding specifically new methods to borrowing detection and developing new methods for the automated segmentation of words into morphemes.

Why would the segmentation of words into morphemes be useful in historical linguistics?

References

Forkel, R. and J.-M. List (2020). "CLDFBench. Give your Cross-Linguistic data a lift." In: *Proceedings of the Twelfth International Conference on Language Resources and Evaluation*. "LREC 2020" (Marseille). Luxembourg: European Language Resources Association (ELRA), 6997–7004.

Forkel, R., J.-M. List, S. J. Greenhill, C. Rzymski, S. Bank, M. Cysouw, H. Hammarström, M. Haspelmath, G. A. Kaiping, and R. D. Gray (2018). "Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics." *Scientific Data* 5.180205, 1–10.

List, J.-M. (2014). *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.

— (2017). "A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets." In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*. Valencia: Association for Computational Linguistics, 9–12.

— (2019). "Automatic inference of sound correspondence patterns across multiple languages." *Computational Linguistics* 45.1, 137–161.

— (2021a). *EDICTOR. A web-based tool for creating, editing, and publishing etymological datasets. Version 2.0.0*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: `https://digling.org/edictor`.

— (2021b). "PyEDICTOR [Python library, Version 0.3.0]."

List, J.-M. and R. Forkel (2022a). "Automated identification of borrowings in multilingual wordlists [version 3; peer review: 4 approved]." *Open Research Europe* 1.79, 1–11.

— (2022b). *LingPy. A Python library for quantitative tasks in historical linguistics [Software Library, Version 2.6.9]*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

— (2022c). *LingRex: Linguistic reconstruction with LingPy*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

List, J.-M., R. Forkel, S. J. Greenhill, C. Rzymski, J. Englisch, and R. D. Gray (2022a). "Lexibank, A public repository of standardized wordlists with computed phonological and lexical features." *Scientific Data* 9.316, 1–31.

List, J.-M., N. W. Hill, and R. Forkel (2022b). "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*. (Dublin, 05/26–05/27/2022). Association for Computational Linguistics. Dublin, 89–96.

List, J.-M., M. Walworth, S. J. Greenhill, T. Tresoldi, and R. Forkel (2018). "Sequence comparison in computational historical linguistics." *Journal of Language Evolution* 3.2, 130–144.

Saitou, N. and M. Nei (1987). "The neighbor-joining method: A new method for reconstructing phylogenetic trees." *Molecular Biology and Evolution* 4.4, 406–425.

Segerer, G. and S. Flavier (2015). *RefLex: Reference Lexicon of Africa*. Version 1.1. URL: `http://reflex.cnrs.fr`.

Sokal, R. R. and C. D. Michener (1958). "A statistical method for evaluating systematic relationships." *University of Kansas Scientific Bulletin* 28, 1409–1438.

Starostin, S. A. (2000). *The STARLING database program*. Moscow: RGGU.

Wu, M.-S., N. E. Schweikhard, T. A. Bodt, N. W. Hill, and J.-M. List (2020). "Computer-Assisted Language Comparison. State of the Art." *Journal of Open Humanities Data* 6.2, 1–14.