List, Johann-Mattis (2019): Open problems in computational historical linguistics. Invited talk presented at the 24th International Conference of Historical Linguistics (2019-07-01/05, Canberra, Australian National University).

Open Problems in Computational Historical Linguistics

Johann-Mattis List (DLCE, MPI-SHH, Jena)

Contents

Abstract

Despite a period of almost two decades in which quantitative approaches in historical linguistics have been increasingly used, gaining constantly more popularity even among predominantly qualitatively oriented linguists, we find many problems in the field of computational historical linguistics, which have only sporadically been addressed. In the talk, I will present a previously published list of 10 problems I personally deem important for historical linguistics, quickly explaining why I think that these problems are not solved yet, why they are hard to solve, but why I have confidence that they might be solved in the nearer or farer future. I will then present a computer-assisted framework for the development of problem-solving strategies in computational historical linguistics. The core of the framework is its interdisciplinary perspective, which helps to adapt existing solutions to similar problems in other disciplines to the problems in historical linguistics, while at the same time relying on a rigorous formalization and inspection of existing strategies in the ”classical”, qualitative approaches to historical language comparison. In contrast to the now very popular machine-learning approaches, which are widely applied to many problems in general and comparative linguistics, the framework is furthermore open with respect to the strategies which are used to solve a given problem, and it generally prefers explicit solutions (if they are available) over black-box strategies, as they are preferred in standard machine-learning frameworks. I will then pick four specific problems (automated morpheme detection, automated borrowing detection, automated induction of sound laws, and automated phonological reconstruction), and present initial strategies for their computer-assisted solution.

1 Introduction (2-8)

1.1 Problems

There are problems that we never discuss in our discipline, such as the problem of language origin, which was banned from the linguistic menu already considerably early, as we can see, for example, from the *statuts* published by the *Société de Linguistique de Paris*.

> La Société n'adment aucune communication concernant, soit l'origine du langage, soit la création d'une langue universelle. (“Statuts” 1871: III)

That there are good reasons to avoid these questions becomes obvious when having a look at the large amount of speculative accounts on the origin of language, ranging from Herder's 1778 onomatopoetic speculation of early human beings running through the woods and imitating the sounds of the things surrounding them, up to recent mystic accounts, published in largely unrecognized journals.

> The Proto-Sapiens grammar was so simple that the sporadic references in previous paragraphs have essentially described it. The prime importance of sound symbolism for the people of nature should be noted again before we further detail that the vowel “E” was felt as indicating the “yin” element, passivity, femininity etc., while “O” indicated the “yang” element, activeness, masculinity etc.; “A” was neutral or spiritual, indicating things conceived by the mind and emotions rather than with the physical senses. (Papakitsos and Kenanidis 2018: 8)

But at times, we may forget that there are valid problems in our field which we do not address, because we focus too much on the hard problems of the mainstream, or on tiny problems for which we know we might never find a sufficient answer. These problems may become evident when talking with laypeople, who may at times simply ask a question that would appear silly for a trained linguist. An example for such a question is the number of words that a language disposes of. While this sounds silly for linguists at the first sight, the question is in fact important for our science in multiple ways.[1] In a paper on similarities

[1] It is important for the field of didactics, where it could help us to provide more efficient lessons on the most important words, it is important for historical linguistics, as it would allow us to measure how many of the words we can actually trace back in

between linguistic and biological evolution, we circumvented the question by giving a simple assessment on the words one needs in order to reach a level of proficiency according to different didactic studies (List 2016). But in the same year, Brysbaert et al. (2016) proposed a way to measure the amount of words that an English speaking person knows:

> Based on an analysis of the literature and a large scale crowdsourcing experiment, we estimate that an average 20-year-old native speaker of American English knows 42,000 lemmas and 4,200 non-transparent multiword expressions, derived from 11,100 word families. (ibid.: 1)

As a historical linguist, I would personally be interested to which degree the estimate of Starostin (1989), which says that every language has about 1000 roots which reflect its ancestry, holds cross-linguistically, and how much variation we could expect when comparing the languages of the world.

1.2 Hilbert and Hilpert Problems

At the and of the last year, inspired by a discussion I had with students who asked me about the biggest challenges for computational historical linguistics, I decided to sit down and make a short list of tasks that I consider challenging, but of which I think that they could still be solved some time in the nearer or further future.

The idea to make such a list of questions is not new to mathematicians, who have their well-known Hilbert Problems, proposed by David Hilbert in 1900 (published in Hilbeert 1902). In linguistics, I first heard about them from Russell Gray, who himself was introduced to this by a talk of the linguist Martin Hilpert, who gave a talk on challenging questions for linguistics in 2014, called “Challenges for 21st century linguistics”. Russell Gray since then has emphasized the importance to propose “Hilbert” questions for the fields of linguistic and cultural evolution, and has also presented his own challenges in the past.

Due to my methodological background, the problems I identified and assembled are by no means big and in some sense also not necessarily extremely challenging (at least on first sight). Instead, the problems I decided for, when being asked, are problems I would like to see tackled, since I think they could help us to further advance our knowledge indirectly, by giving us the possibility to use the solutions of the problems to then answer deeper question on problems in historical linguistics in specific and diversity linguistic in general. One further aspect of the problems that I selected is that these challenges can all be solved by algorithms or workflows. Even when being “small” in some sense, this does not mean, of course, that these problems are not challenging in the big sense. It also does not automatically mean that they can be solved in the near future. But given that the work in the field of computational and computer-assisted language comparison, progresses steadily, at times even at an impressive paste, I have some trust that these problems will indeed be solvable within the next 5-10 years.

2 Open problems in computational historical linguistics (9-19)

When writing down my ten open problems for computational diversity linguistics, I announced this in a blog post with the blog *The genealogical world of phylogenetic networks*, edited by David Morrison (`http://phylonetworks.blogspot.com/`), in January, with the plan of discussing each of the problems in detail in monthly blog posts throughout the year. So far, five problems have already been published (*Automatic borrowing detection*, List 2019a, *Automatic morpheme detection*, List 2019c, *Automatic sound law induction*, List 2019e, and *Automatic phonological reconstruction*, List 2019d).

history, and it would be important for cognitive research, as it would allow us to assess the amount of information individuals can make use of when speaking.

The 10 problems, which are listed in Table 1 can be further classified into three different groups, which roughly correspond to three different categories important for research in general, namely *modeling*, *inference*, and *analysis*. This triad, inspired by Dehmer et al. (2011: XVII), follows the general idea that scientific research in the historical disciplines usually starts from some kind of idea we have about our research object (the *model* stage), and based on which we then apply methods to infer the phenomena in our data (the *inference* stage). Having inferred enough examples for the phenomenon, we can then *analyze* it qualitatively or quantitatively (the *analysis* stage) and use this information to update our model.

The first group in my list of problems deals with questions of *inference*, including the *detection of morpheme boundaries* (# 1), the *induction of sound laws* (# 2), the *detection of borrowings* (# 3), and *phonological reconstruction* (# 4). What all these problems have in common is that they deal with inference in the sense described above, in so far as they start from linguistic data in some specific form, and the task is to find specific patterns in the data, which have not been annotated in the data beforehand.

Number	Problem	Class
1	automated morpheme segmentation	inference
2	automated sound law induction	inference
3	automated borrowing detection	inference
4	automated phonological reconstruction	inference
5	simulating lexical change	modeling
6	simulating sound change	modeling
7	statistical proof of language relatedness	modeling
8	typology of semantic change	analysis
9	typology of semantic promiscuity	analysis
10	typology of sound change	analysis

Table 1: 10 problems of computational diversity linguistics. Entries shaded in gray have already been published.

The second group of problems deals with questions of *modeling*, including the *simulation of lexical change*, i.e., the design of consistent models that describe how the lexemes of a language change over time, the *simulation of sound change*, i.e., the simulation of the sound-change process by which sounds in a language change in dependence of the context in which they occur, and *the statistical proof of language relatedness*. While the simulation problems are clear problems of *modeling*, given that a simulation requires a model to be then applied to some artificial or existing datasets, the statistical proof or language relationship is a specific case, since it requires a model of language relatedness in order to test this model against a random model in which languages are thought to be unrelated.[2]

The last group of problems all have *typology* in their title, and belong to the class of *analysis* problems, dealing with the analysis of *semantic change*, *semantic promiscuity*, and *sound change*. What is meant by *typology* in this context is a data-driven estimate of the overall cross-linguistic frequency of these phenomena. Since we lack consistent accounts on the general tendencies of these processes and phenomena when excluding areal and genetic factors, the task is simply to come up with a consistent estimate on each of them. While semantic change and sound change are probably self-explaining in this context, the question of semantic promiscuity deserves some more attention. What is essentially meant

[2] While there are numerous attempts in the literature to come up with a convincing statistical model to prove genetic relationship (Baxter and Manaster Ramer 2000, Kassian et al. 2015, Kessler 2001, Mortarino 2009, Ringe 1992), none of the attempts which have been proposed so far deals with lexical comparisons in all their complexity. Either, scholars only compare initial consonants with each other (Kessler 2001, Ringe 1992), or they resort to sound classes (Baxter and Manaster Ramer 2000, Kassian et al. 2015), and even if scholars compute random models for whole alignments of potentially related words (List 2014), they have the problem of not accounting for the factor of closeness due to borrowing.

by this term is the degree to which certain words, due to their original meanings, are re-used or re-cycled in the human lexicon. While the term *promiscuity* has been used before in other contexts in linguistics, the specific usage of promiscuity to denote what one could also call *semantic productivity* or *concept productivity* was first proposed in List et al. (2016), where biological and linguistic processes were consistently compared with each other, and semantic promiscuity was identified as a phenomenon similar to *domain promiscuity* in protein evolution in biology, with an explicit analogy being identified between the processes of *word formation* in linguistics and *protein assembly* in biology (ibid.: 5). For further elaborations of the concept of *semantic promiscuity*, compare List (2018) and Schweikhard (2018).

3 Problem solving strategies (20-25)

3.1 Computer-assisted language comparison

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

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

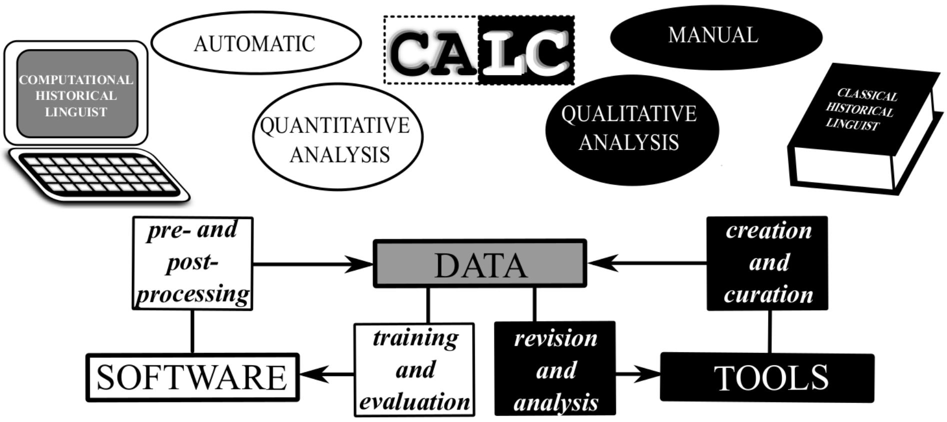


Figure 1: Basic idea of data managment within the CALC framework.

The basic idea behind computer-*assisted* as opposed to computer-*based* language comparison is to allow scholars to do qualitative and quantitative research are done at the same time. In order to allow scholars to do this, **data must always be available in *machine-* and *human-readable* form.** Figure 1 shows a tentative workflow for the CALC framework, in which data is constantly passed back and forth between computational and classical linguists.

Three different aspects are essential for this workflow:

(a) New software allows for the application of transparent methods which increase the accuracy and the application range of current methods and also treat the peculiarities of specific language families (like, e.g., Sino-Tibetan).

(b) Interactive tools provide an interface between human and machine, allowing experts to correct errors and to inspect the automatically produced results in detail.

(c) Specific data is used to test and train the software algorithms.

3.2 Mind the Machines

An alternative to computer-assisted approaches in historical linguistics would be pure computational approaches. As I have tried to argue before, these purely automatic approaches tend to lag behind human analyses, as they lack the accuracy and the flexibility of human judgments. Proponents of machine learning techniques often argue that human judgment is similarly problematic, pointing to problems in inter-annotator agreement, or situations in which machines are now better than humans, such as the Go game (Silver et al. 2016). Interestingly, in computational historical linguistics, the superiority of purely computer-based approaches has still not been convincingly proven, although it has been claimed in the past[3] Thus, in tests on the task of automatic cognate detection, methods based on machine-learning paradigms, such as support vector machines (Jäger 2018), neural networks (Rama 2016), or generally *unsupervised* approaches (Rama et al. 2017) have so far not shown to clearly outperform the more "hand-crafted" algorithms, which make use of intuitive linguistic knowledge (Rama et al. 2018).

In my opinion, the major reason why pure machine-learning approaches have problems in detecting the signal that trained linguists detect easily in linguistic datasets is that these methods follow the *big data paradigm*, which is essentially useless when working with data in historical linguistics. The idea that one can solve all problems, without spending too much time to thinking about their proper solutions, if one only has enough data, is very wide-spread among computer scientists who work with neural networks or Bayesian inference. Unfortunately, the big data promise stands and falls with the availability of big data, and big data is essentially not existent in historical linguistics.

Another problem resulting from the big data paradigm is that it does not seek to search for scientific solutions to a given problem, i.e., by telling us the major processes involved in language change, but instead normally uses annotated data (for example, images of horses) to train a certain model (so that it recognizes a horse if it appears on an image) and then apply them to real data. This framework is useful in specific tasks that are too tedious to be carried out by humans, such as face recognition, the automatic detection of certain kinds of content when people upload pictures on social networks, or to spy on people in general. The framework, however, does not work when it comes to solving real questions that drive scientific research. In our research we do not only want to know, whether a given image is a cat or not, but we also want to know what makes cats different from dogs, that is, we also want to determine the *gestalt* of a phenomenon

3.3 Basic aspects of computer-assisted problem solving

The framework for computer-assisted problem solving which I try to pursue in my own research and which I try to propagate does not neglect the possibility of using machine-learning techniques to tackle

[3] As can be seen from the famous quote attributed to Fred Jelinek: "[...] it was at the 1985 workshop [...] that Fred Jelinek uttered the now immortal phrase «Every time we fire a phonetician/linguist, the performance of our system goes up»" (Moore 2005: 1).

specific problems, but it does also not necessarily require that they be used exclusively. We do not naively accept machine learning solutions, but start instead from a careful inspection of the problems we actually want to solve. In many cases, a complex solution involving neural networks or Bayesian inference techniques may actually not be needed, since there are smart heuristics, or even complete solutions that do not require any stochastic component. In the same way in which we would not use a machine learning method to tackle the problem of multiplication, it is futile to have an algorithm searching for sound correspondences without any underlying model of sequence comparison or alignments.

That does not mean that machine learning solutions should be excluded per se, and in fact, many of the algorithms for cognate detection, which scholars call *supervised* or based on *linguistic knowledge*, make use of classical techniques, like random works, in specific stages of their workflow. But the decision when to use a specific technique is usually always based on some explicit reasoning that takes the phenomenon to be investigated into account, as well as the existing qualitative solutions that were developed within the field itself, and actual solutions in computer science or similar disciplines, such as bioinformatics, which are consulted to provide inspiration for possible solutions.

The current strategy, which has been applied to propose automatic solutions for various aspects of historical linguistics (List 2014, List 2019) starts from a detailed investigation (also in collaboration with experts on the topic) of the existing qualitative solutions to a given problem in historical linguistics. As a second step, we try to describe the task in a clear way, by naming explicitly the input data and the output data we expect from the automatic method. We then try to model the process, while at the same time being prepared to further modify the requirements regarding the input data. The solution for the problem is then sought by looking at neighboring disciplines and topics, specifically graph theory, sequence comparison techniques in computer science and bioinformatics, in order to come up with a solution to the problem.[4]

4 Possible solutions for the inference problems

4.1 Morpheme segmentation (27-33)

The first task on my list of 10 open problems in computational diversity linguistics deals with morphemes, that is, the minimal meaning-bearing parts in a language. A morpheme can be a word, but it does not have to be a word, since words may consist of more than one morpheme, and —depending on the language in question —may do so almost by default. The task of automatic morpheme segmentation is thus a pretty straightforward one: given a list of words, potentially along with additional information, such as their meaning, or their frequency in the given language, try to identify all morpheme boundaries, and mark this by adding dash symbols where a boundary has been identified.

4.1.1 Background on proposed methods

One may ask why automatic identification of morphemes should be a problem —and some people commenting on my presentation of the 10 open problems last month did ask this. The problem is not unrecognized in the field of Natural Language Processing, and solutions have been discussed from the 1950s onwards (Benden 2005, Bordag 2008, Hammarström 2006, Harris 1955, see also the overview by Goldsmith et al. 2017).

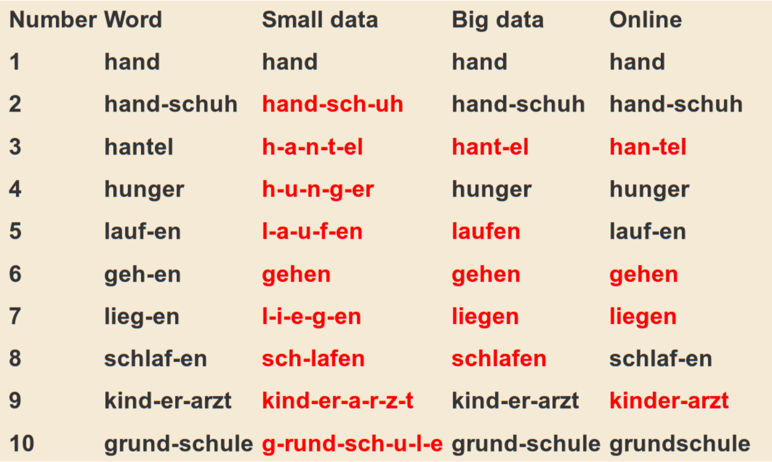
[4] An example for a problem solved in this way was the handling of sound correspondence patterns across multiple languages. After a careful data modeling involving multiple sequence alignments as input data, we realized that the problem could be modeled as the well-known clique cover problem (Bhasker and Samad 1991), for which an approximate solution exists (Welsh and Powell 1967). Based on this solution, we then designed an algorithm that essentially searches for sound correspondence patterns across multiple languages and presents the results in an interactive framework (List 2019).

Roughly speaking, all approaches build on statistics about n-grams, i.e., recurring symbol sequences of arbitrary length. Assuming that n-grams representing meaning-building units should be distributed more frequently across the lexicon of a language, they assemble these statistics from the data, trying to infer the ones which ”matter”. With Morfessor (Creutz and Lagus 2005), there is also a popular family of algorithms available in form of a very stable and easy-to-use Python library (Virpioja et al. 2013). Applying and testing methods for automatic morpheme segmentation is thus very straightforward nowadays.

4.1.2 Problems of current solutions

The issue with all of these approaches and ideas is that they require a very large amount of data for training, while our actual datasets are small and sparse, by nature. As a result, all currently available algorithms fail graciously when it comes to determining the morphemes in datasets of less of 1,000 words. Interestingly, even when having been trained on large datasets, the algorithms still commit surprising errors, as can be easily seen when testing the online demo of the Morfessor software for German (`https://asr.aalto.fi/morfessordemo/`). When testing words like auftürmen “pile up”, for example, the algorithm yields the segmentation auf-türme-n, which is probably understandable from the fact that the word Türme “towers” is quite frequent in the German lexicon, thus confusing the algorithm; but for a German speaker, who knows that verbs end in -en in their infinitive, it is clear that the auftürmen can only be segmented as auf-türm-en. If I understand the information on the website correctly, the Morfessor algorithm offered online was trained with more than 1 million different word forms in German. Given that in our linguistic approaches we can usually dispose of 1,000 words, if not less, per language, it is clear that the algorithms won’t provide help in finding the morphemes in our data.

To illustrate this, I ran a small test on the Morfessor software, using two datasets for training, one big dataset with about 50000 words from Baayen et al. (CELEX), and one smaller dataset of about 600 words which I used as a cognate detection benchmark when writing my dissertation List (2014). I then used these two datasets to train the Morfessor software and then applied the trained models to segment a list of 10 German words.[5] The results for the two models (small data and big data) as well as the segmentations proposed by the online application (online) are given in the table below (with my own judgments on morphemes given in the column word).



Number	Word	Small data	Big data	Online
1	hand	hand	hand	hand
2	hand-schuh	hand-sch-uh	hand-schuh	hand-schuh
3	hantel	h-a-n-t-el	hant-el	han-tel
4	hunger	h-u-n-g-er	hunger	hunger
5	lauf-en	l-a-u-f-en	laufen	lauf-en
6	geh-en	gehen	gehen	gehen
7	lieg-en	l-i-e-g-en	liegen	liegen
8	schlaf-en	sch-lafen	schlafen	schlaf-en
9	kind-er-arzt	kind-er-a-r-z-t	kind-er-arzt	kinder-arzt
10	grund-schule	g-rund-sch-u-l-e	grund-schule	grundschule

Figure 2: Results of the small test of the Morfessor software.

What can be seen clearly from the table, where all forms deviating from my analysis are marked in red font, is that none of the models makes a convincing job in segmenting my ten test words. More

[5] See `https://gist.github.com/LinguList/04bba6d97595d7e474ab109a639fecce` for data and code

importantly, however, we can clearly see that the algorithm's problems increase drastically when dealing with small training data. Since the segmentations proposed in the Small data column are clearly the worst, splitting words in a seemingly random fashion into letters. What is interesting in this context is that trained linguists would rarely fail at this task, even when all they were given is the small data list for training. That they do not fail is shown by the numerous studies where linguistic fieldworkers have investigated so far under-investigated languages, and quickly figured out how the morphology works.

4.1.3 Why is morpheme segmentation difficult?

What makes the detection of morpheme boundaries so difficult, also for humans, is that they are inherently ambiguous. A final *-s* can mark the plural in German, especially on borrowings, as in *Job-s*, but it can likewise mark a short variant of *es* "it", where the vowel is deleted, as in *ist's* "it's", and in many other cases, it can just mark nothing, but instead be part of a larger morpheme, like Haus "house". Whether or not a certain substring of sounds in a language can function as a morpheme depends on the meaning of the word, not on the substring itself. We can —once more —see one of the great differences between sequences in biology and sequences in linguistics here: linguistic sequences derive their "function" (i.e. their meaning) from the context in which they are used, not from their structure alone.

If speakers are no longer able to clearly understand the morphological structure of a given word, they may even start to change it, in order to make it more "transparent" in its denotation. Examples for this are the numerous cases of *folk etymology*, where speakers re-interpret the morphemes in a word, with English *ham-burger* as a prominent example, since the word originally seems to derive from the city *Hamburg*, which has nothing to do with *ham*.

4.1.4 How humans find morphemes

The reasons why human linguists can relatively easy find morphemes in sparse data, while machines cannot, is still not entirely clear to me (i.e. humans are good at pattern recognition and machines are not). However, I do have some basic ideas about why humans largely outperform machines when it comes to morpheme segmentation; and I think that future approaches that try to take these ideas into account might drastically improve the performance of automatic morpheme segmentation methods.

As a first point, given the importance of meaning in order to determine morphemic structure, it seems almost absurd to me to try to identify morphemes in a given language corpus based on a pure analysis of the sequences, without taking their meaning into account. If we are confronted with two words like Spanish *hermano* "brother" and *hermana* "sister", it is clear —if we know what they mean —that the -o vs. -a most likely denotes a distinction of gender. While the machines compare potential similarities inside the words independent of semantics, humans will always start from those pairs where they think that they could expect to find interesting alternations. As long as the meanings are supplied, a human linguist —even when not familiar with a given language —can easily propose a more or less convincing segmentation of a list of only 500 words.

A second point that is disregarded in current automatic approaches is the fact that morphological structures vary drastically among languages. In Chinese and many South-East Asian languages, for example, it is almost a rule that every syllable represents one morpheme (with minimal exceptions being attested and discussed in the literature). Since syllables are again easy to find in these languages, since words can often only end in a specific number of sounds, an algorithm to detect words in those languages would not need any n-gram statistics, but just a theory on syllable structures. Instead of global strategies, we may rather have to use for local strategies of morpheme segmentation, in which we identify different types of languages for which a given algorithm seems suitable.

This brings us to a third point. A peculiarity of linguistic sequences in spoken languages is that they are built by specific phonotactic rules that govern their overall structure. Whether or not a language tolerates more than three consonants in the beginning of a word depends on its phonotactics, its set of rules by which the inventory of sounds is combined to form morphemes and words. Phonotactics itself can also give hints on morpheme boundaries, since they may prohibit combinations of sounds within morphemes which can occur when morphemes are joined to form words. German *Ur-instinkt* “basic instinct”, for example, is pronounced with a glottal stop after the *Ur-*, which can only occur in the beginning of German words and morphemes, thus marking the word clearly as a compound (otherwise the word could be parsed as *Urin-stinkt* “urine smells”).

A fourth point that is also generally disregarded in current approaches to automatic morpheme segmentation is that of cross-linguistic evidence. In many cases, the speakers of a given language may themselves no longer be aware of the original morphological segmentation of some of their words, while the comparison with closely related languages can still reveal it. If we have a potentially multi-morphemic word in one language, for example, and only one of the two potential morphemes reflected as a normal word in the other language, this is clear evidence that the potentially multi-morphemic word does, indeed, consist of multiple morphemes.

4.1.5 Suggestions for solutions

Linguists regularly use multiple types of evidence when trying to understand the morphological composition of the words in a given language. If we want to advance the field of automatic morpheme segmentation, it seems to me indispensable that we give up the idea of detecting the morphology of a language just by looking at the distribution of letters across word forms. Instead, we should make use of semantic, phonotactic, and comparative information.

Although one may think that it is difficult to gather sufficient information on all these aspects, we are currently building resources that could help in this task, including the Concepticon (List et al. 2016), a catalogue of meanings for cross-linguistic approaches, CLICS^2 (List et al. 2018), a database providing information on cross-linguistic lexical associations, LingPy (List et al. 2018), a Python library that offers ways to model prosody in phonetic data, as well as offering basic algorithms to compare words across different languages.

We should further give up the idea of designing universal morpheme segmentation algorithms, but rather study which approach works best on which morphological type. How these aspects can be combined in a unified framework, however, is still not entirely clear to me; and this is also the reason why I list automatic morpheme segmentation as the first of my ten open problems in computational diversity linguistics.

Even more important than the strategies for the solutions of the problem, however, is that we start to work on extensive datasets for testing and training of new algorithms that seek to identify morpheme boundaries on sparse data. As of now, no such datasets exist. Approaches like Morfessor were designed to identify morpheme boundaries in written languages, they barely work with phonetic transcriptions. But if we had the datasets for testing and training available, be it only some 20 or 40 languages from different language families, manually annotated by experts, segmented both with respect to the phonetics and to the morphemes, this would allow us to investigate both existing and new approaches much more profoundly, and I expect it could give a real boost to our discipline and greatly help us to develop advanced solutions for the problem.

4.1.6 Current work

Our initial goal for work in the near future in the CALC project is to create a first *Morpheme-Annotated Lexical Database* (working title *MOALD*), using aggregation strategies developed for the CLICS database (List et al. 2018), building on the standardization efforts of the CLDF initiative (Forkel et al. 2018). While we have already assembled morpheme-segmented data for Sino-Tibetan and other language families of the SEA area, we are currently directly collaborating with experts on language families from South America (Tukanoan, with T. C. Chacon) and Africa (Dogon and Mande with A. Hantgan) to further expand the availability of larger morpheme-annotated wordlists.

4.2 Borrowing detection (34-40)

The second task on my list of 10 open problems in computational diversity linguistics deals with detecting borrowings or language contact. The prototypical case of language contact would be lexical borrowing. More complex cases involve semantic borrowing (*calques*). Even less well understood are cases where specific aspects of grammar have been transferred. German has, for example, a certain number of neuter nouns, all borrowed from Ancient Greek or Latin, in which the plural is built according to (or inspired by) the Greek model: *Lexikon* has *Lexika* as plural, *Komma* has *Kommata* as plural, and *Kompositum* has *Komposita* as plural. While these cases are spurious in German and thus rather harmless (as are the similar examples in English), there are other cases of language contact where scholars not only suspect that plural forms have been borrowed along with the words (as in German), but that entire paradigms and strategies of grammatical marking have been adopted by one language from a neighboring variety as a result of close language contact.

4.2.1 Background on proposed methods

In principle, all algorithms for contact inference proposed so far make use of the strategies used in the classical approaches. Thus, they infer or determine shared traits among two or more languages, and then determine conflicts in these traits, taking geographical closeness and borrowability into account. In contrast to classical approaches, which combine different types of evidence, computational approaches are usually restricted to one type.

The automatic methods proposed so far can be divided into three classes. The first class employs phylogeny-related conflicts to identify those traits whose evolution cannot be explained with a given phylogenetic tree, explaining the conflicts as resulting from contact. Examples include work where I was involved myself (List et al. 2014, Nelson-Sathi et al. 2011), some early and interesting approaches which did not receive too much attention (Minett and Wang 2003), or have been mostly forgotten by now (Nakhleh et al. 2005), along with a recent study on grammatical features (Cathcart et al. 2018).

The second class uses techniques for automatic sequence comparison to search for similar words, but not cognate words, across different languages. Here, the most prominent examples include the work by Ark et al. (2007), and later Mennecier et al. (2016), who searched for similar words among languages known to be not related. Further examples include the work by Boc et al. (2010) and Willems et al. (2016), who experimented with tree reconciliation approaches, based on word trees derived from sequence-alignment techniques. There is also an experimental study where I was again involved myself (Hantgan and List forthcoming), in which we tried to identify borrowings by comparing two automatically inferred similarities among words from related and unrelated languages: surface similarities, as reflected by naive alignment algorithms, and deep similarities, reflected by advanced methods that take sound correspondences into account (List 2014).

The third class searches for distribution-related conflicts by comparing the amount of shared words within sublists of differing degrees of borrowability. This class is best represented by Sergey Yakhontov's (1926-2018) work on stable and unstable concept lists (Starostin 1991), which assumed that deep historical relations should surface in those parts of the lexicon that are stable and resistant to borrowing, while recent contact-induced relations would surface rather in those parts of the lexicon that are more prone to borrowing. Yakhontov's work was independently re-invented by Chén (1996), and McMahon et al. (2005); but given how difficult it turned out to distinguish concepts prone to borrowing from those resistant to borrowing, it has been largely disregarded for some time now.

4.2.2 Problems of current solutions

All three classes of approaches discussed so far have certain shortcomings. Phylogeny-based inference of borrowing, for example, tends to drastically overestimate the number of borrowed traits, simply because conflicts in a phylogeny can result from undetected borrowings in the data but they never need to (see Appendix 1 of Morrison 2011 on causes of reticulation in biology, which has many parallels to linguistics). Saying that all instances in which a dataset conflicts with a given phylogeny are borrowings is therefore generally a bad idea. It can be used as a very rough heuristics to come up with potentially wrongly annotated homologies in a dataset, which could then be checked again by experts, but deriving stronger claims from it seems problematic.

While sequence comparison techniques applied to unrelated languages are basically safe in my opinion, and the results are very reliable, unless one compares words that occur in all languages, such as *mama* and *papa* (Blasi et al. 2016, Jakobson 1960).

Using methods for tree reconciliation on individual word trees, calculated from word distances based on phonetic alignment techniques or similar, yields the same problems of over-counting conflicts as we get for phylogeny-based approaches to borrowing. The problem here is a general misunderstanding of the concept differences between gene trees in biology, where surface similarity of gene sequences is thought to reflect evolutionary history, and word trees in linguistics. While we can use qualitative methods to draw a word tree for a given set of homologous words, the surface similarity among the words says little, if anything, about their evolutionary history.

Attempts to distinguish borrowed from inherited traits with sublists have lost their popularity in most recent studies. When properly applied, they might, indeed, provide some evidence in the search for borrowings or deep homologies. So far, however, all stability rankings of concepts that have been proposed have been based on too small an amount of either concepts (we would need rankings for some 1,000 concepts at least), or languages from which the information was derived. If we could manage to get reliable counts on some 1,000 concepts for a larger sample of the world's languages, this might greatly help our field, as it would provide us with a starting point from which people could search (even qualitatively) for borrowings in their data.

4.2.3 Why is borrowing detection difficult?

Unless we witness them happening directly, most cases of borrowing are difficult to demonstrate consistently. By comparison with lexical borrowing, however, the borrowing of grammar is probably the hardest to show, especially when dealing with abstract categories that could have actually emerged independently. The reason why borrowing is generally hard to deal with, not only in computational approaches, is that detecting borrowing and demonstrating language contact presupposes that alternative explanations are all excluded, such as universal tendencies of language change (i.e., "convergent evolution" in the biological sense), common inheritance, or simple chance.

While we need to exclude alternative possibilities to prove any of the four major types of similarities (coincidental, natural, genealogical, or contact-induced, see List 2014: 55-57), we have a much harder time in doing so when dealing with borrowings, because linguistics does not know even one procedure for the identification of borrowings. Instead, we resort to a mix of different types of evidence, which are qualitatively weighted and discussed by the experts. While historical linguistics has developed sophisticated techniques to show that language similarities are genealogical, it has not succeeded to reach the same level of sophistication for the identification of borrowings.

In this regard, techniques for contact detection are not much different from other, more specific, types of linguistic reconstruction, such as the "philological reconstruction" of ancient pronunciations (Jarceva 1990, Sturtevant 1920), the reconstruction of detailed etymologies (Malkiel 1954), or the reconstruction of syntax (Willis 2011).

4.2.4 How humans detect borrowings

It is not easy to give an exhaustive and clear-cut overview of all of the qualitative methods that scholars make use of in order to detect borrowings among languages. This is at least partially due to the nature of "cumulative-evidence arguments" (Berg 1998: 66) —or arguments based on consilience (Whewell 1847, Wilson 1998) —which are always more difficult to formalize than clear-cut procedures that yield simple, binary results. Despite the difficulty in determining exact workflows, we can identify a couple of proxies that scholars use to assess whether a given trait has been borrowed or not (List forthcoming).

One important class of hints are conflicts with possible genealogical explanations. A first type of conflict is represented by similarities shared among unrelated or distantly related languages. Since English mountain is reflected only in English, with similar words only in Romance, we could take this as evidence that the English word was borrowed. Since these conflicts arise from the supposed phylogeny of the languages under consideration, we can speak of *phylogeny-related arguments for interference*.

A second conflict involves the traits themselves, most prominently observed in the case of irregular sound correspondence patterns. German *Damm*, for example, is related to English *dam*, but since the expected correspondence for cognates between English and German would yield a German reflex *Tamm* (as it is still reflected in Old High German, see Kluge 2002), we can take this as evidence that the modern German term was borrowed Pfeifer (1993). We can call these cases *trait-related arguments for contact*.

In addition to observations of conflicts, two further types of evidence are of great importance for inferring contact. The first one is *areal proximity*, and the second one is the assumed *borrowability* of traits. Given that language contact requires the direct contact of speakers of different languages, it is self-evident that geographical proximity, including proximity by means of travel routes, is a necessary argument when proposing contact relations between different varieties.

Furthermore, since direct evidence confirms that linguistic interference does not act to the same degree on all levels of linguistic organisation, the notion of borrowability also plays an important role. Although scholars tend to have different opinions about the concept, most would probably agree with the borrowability scale proposed by Aikhenvald (2007: 5), which ranges from "inflectional morphology" and "core vocabulary", representing aspects resistant to borrowing, up to "discourse structure" and the "structure of idioms", representing aspects that are easy to borrow. How core vocabulary can be defined, and how the borrowability of individual concepts can be determined and ranked, however, has been subject to controversial discussions (Lee and Sagart 2008, Starostin 1995, Tadmor 2009, Zenner et al. 2014).

4.2.5 Suggestions for solutions

Assuming that currently we have no realistic way to operationalize arguments based on consilience, there is no direct hope to have a fully automatic method for detecting borrowings any time soon. By developing promising existing methods further, however, there is a hope that we can learn a lot more about borrowing processes in the world's languages. What is needed here are, of course, the data that we need in order to apply the methods.

In addition to the above-mentioned automatic approaches for borrowing detection, so far, nobody has tried to use trait-related conflicts to infer borrowings. Since these are usually considered to be quite reliable by experts in historical linguistics, it seems inevitable to work in this direction as well, if we want to tackle the problem of consistent automatic detection of borrowing. Here, my recently proposed framework for a consistent handling and identification of patterns of sound correspondences across multiple languages (List 2019), could definitely be useful, although it will again be challenging to find the right balance of parameters and interpretation, since not all conflicts in sound correspondences necessarily result from borrowings.

Whether it will be possible to identify even the direction of borrowings, when developing these methods further, is an open question. Borrowability accounts might help here, but again, since no clear-cut strategies are being used by scholars, it is difficult to formalize any of the existing qualitative approaches. The greatest challenge will perhaps consist in the creation of a database of known borrowings that could assist digital linguists in testing and training new approaches.

4.2.6 Current work

In the CALC project, we currently develop cross-linguistic datasets to test borrowing relations in contact areas, which are based on high-quality data in phonetic transcriptions for well-known contact areas with well-annotated and carefully selected concept lists. We also try to develop new feature-based metrics of phonetic word similarity, based on the features developed for the CLTS project on Cross-Linguistic Transcription Systems (Anderson et al. 2019). We are also currently testing methods for contact-zone detection based on a new method for cognate set partitioning (first presented in List 2019).

4.3 Automatic sound law induction (41-46)

The third problem I want to discuss in this context is a problem that my not even be considered as a true problem in computational historical linguistics, as it has usually been overlooked greatly, and only indirectly been discussed by colleagues. This problem, which I call the *automatic induction of sound laws*, can be summarized as follows: starting from a list of words in a proto-language and their reflexes in a descendant language, try to find the rules by which the ancestral language is converted into the descendant language. Note that under *rules*, in this context, I understand the classical notation that phonologists and historical linguists use in order to convert a source sound in a target sound in a specific environment. They are similar to a regular expression in computer science, but they differ in the scope and the rules for annotation. Normally, they pick one sound in the proto-language (or what phonologists call the *underlying sound* in the synchronic description of a language) and show how this sound is converted to another sound (including that the sound may be lost) by applying some kind of *conditioning context*. The notation in historical linguistics can thus be summarized as follows:

$$s_P > s_D / e_p\_e_f e_a \tag{1}$$

Here, s_P represents the sound in the proto-language, s_D the sound in the descendant language, and _ represents the position of s_P in the description of the environment, which can be divided into the preceding environment e_p, the following environment e_f, and what I call the *abstract* environment e_a, which refers to suprasegmental properties, like stress or tone. Note that linguists do not necessarily follow this schema completely. If one type of context cannot be observed, they will consequently drop it, but they may also use additional ways to encode conditioning context, making use of annotations on syllables, using abstract symbols that are supposed to represent classes of sounds instead of concrete sounds, and they may even provide annotations by which a class of sounds in the ancestor language changes into a class of sounds in the descendant language.

It is not the right place to have a complete discussion about the consistency of the annotation practice for sound change. All that needs to be emphasized here is that we expect a certain amount of inconsistency and also incomparability in the linguistic literature, due to the ad-hoc nature in which these formulas are normally used, and the fact that classes of sounds rather than concrete sounds are often presented.[6]

[6] Note also that the use of the term *induction* in this context was deliberately done, as it reflects the prototype of induction, when following the original framework of Peirce (Peirce 1931/1958), since the original state of a phenomenon is given, as well as the later state, and the task is to find the *rules* by which the original state was converted to the later state (see also List 2014). While most enterprises in historical linguistics can be seen as reflecting *abduction*, the mode of reasoning, by which one starts from a given result state, and the knowledge of processes, to abduce the initial state as well as the rules which which the initial state was turned into the result state (ibid.), the task of finding the sound laws that turned a proto-language into a descendant language, is a clear case of *induction* in historical linguistics.

4.3.1 Background on computational approaches to sound laws

To my knowledge, the question of how to *induce* sound laws from data on proto- and descendant language has barely been addressed so far by scholars in concrete. What comes closest to the problem are attempts to *model* sound change from known ancestral languages, such as Latin, to daughter languages, such as Spanish, as reflected, for example, in the PHONO program (Hartmann 2003), where one can insert data for a proto-language along with a set of sound change rules, which need to be ordered, and then check if they correctly predict the descendant forms. Another class of approaches are word prediction experiments, such as the one by Ciobanu and Dinu (2018) (but see also Bodt and List 2019), in which training data consisting of the source and the target language are used to create a model which is then successively applied to more data, in order to test how well this model predicts target words from the source words. Since the model itself is not reported in these experiments, but only used, in form of a black box, to predict new words, the task cannot be considered as the same as the task for sound law induction that I propose as one of my ten challenges for computational historical linguistics, given that we are interested in a method that explicitly returns the model in order to allow linguists to inspect it.

4.3.2 Problems of current solutions

Given that – to my knowledge – no current solutions exist, it seems useless to point to problems of current solutions. What I want to mention in this context, however, are the problems of the solutions presented for word prediction experiments, be they fed by manual data on sound changes (Hartmann 2003), or based on inference procedures (Ciobanu and Dinu 2018, Dekker 2018). While manual solutions like PHONO suffer from the fact that they are tedious to apply, given that linguists have to present all sound changes in their data in an ordered fashion, with the program converting them step by step (not allowing for simultaneous changes to take place), the word prediction approaches suffer from poor feature design. The method by Ciobanu and Dinu (2018), for example, is based on orthographic data alone, using the Needleman-Wunsch algorithm for sequence alignment (Needleman and Wunsch 1970), and the approach by Dekker (2018) only allows for the use for the limited alphabet of 40 symbols proposed by the ASJP project (Holman et al. 2008). In addition to a poor representation of linguistic sound sequences, be it by resorting to abstract orthography or to abstract reduced phonetic alphabets, none of the methods can handle those kinds of contexts which I labelled as *abstract* in Equation 1. But we know well that abstract contexts are vital for certain aspects of sound change, with Verner's law being one of the most prominent examples (Verner 1877).

4.3.3 Why is automatic sound law induction difficult?

The handling of the *abstract* context types mentioned in the paragraph before is in my opinion also the reason why sound law induction is so difficult, not only for machines, but also for humans. In addition, the context by *preceding* or *following* environment is also tricky, since it not necessarily points to the first preceding or the first following segment, but may well relate to contexts of longer distance (such as we notice for phenomena like vowel harmony). As an additional problem of handling linguistic context, there is the problem of the *systemic aspect* of sound change, which often reflects in situations where not only one sound in a language changes in a certain environment, but instead full classes of sounds. Thus, in Spanish, for example, all voiced stops are subjected to fricativization when occurring intervocalically, while in German *Auslautverhärtung*, all voiced stops are devoiced and aspirated. Terms like *fricativization* and *devoicing* point to the change of a feature rather than the direct change of one sound symbol being replaced by another one. Resorting to feature explanations has the advantage of reducing the number of rules needed to explain sound change phenomena (thus increasing their parsimony), while at the same time allowing to back up the list of observed phenomena by more concrete evidence. If, for example, there was only one instance of the sound [ɣ] occurring intervocalically in a language, with no instances of intervocalic [g], but plenty examples of [v] (with lack of intervocalic [b]) and [ð] (with lack of intervocalic [d]), inducing a rule like *fricativization* would not suffer from lack of evidence for the case of g > ɣ. On the contrary, the argument would be completed by the single case of velar fricativization. When taken in isolation, however, one would have to reject the argument for the fricativization of the voiced velar, since the evidence would look only spurious at best.

To summarize these points, the difficulties in handling conditioning context in sound law induction lie in the nature of conditioning context in linguistics, going beyond a simple notion of *preceding* or *following* sound as type of context, along with the existence of non-linear, "abstract" context, reflected in suprasegmental phenomena, and the problem of data sparseness in all cases where systemic processes are at work, which apply to classes of sounds, rather than to single sound units.

4.3.4 How humans detect sound laws

Given that there are only a few examples in the literature, where scholars have tried to provide detailed lists of sound changes from proto- to descendant language (Baxter 1992, Chén 1996), with most of these contributions not even checking whether their sound changes would be successfully applied, it is difficult to assess what humans usually do in order to detect sound laws. What is clear is that historical linguists who have been working a lot on linguistic reconstruction tend to acquire a very good intuition that helps them to quickly check sound laws applied to word forms in their head and convert the output forms. This ability is developed in a *learning-by-doing* fashion, with no specific techniques ever being discussed in the classroom, reflecting the general tendency in historical linguistics to trust that students will learn how to become a good linguist from examples, sooner or later (Schwink 1994: 29). For this reason, it is difficult to take inspiration from current practice in historical linguistics in order to develop computer-assisted approaches to solve this task.

4.3.5 Proposed solutions

In contrast to sequences as we meet them in mathematics and informatics, *words* in spoken languages do not consist solely of letters drawn from an alphabet that is lined up in some unique order. They are instead often composed of *multiple layers*, which are in part hierarchically ordered. Words, morphemes, and phrases in linguistics are thus *multi-layered constructs*, which cannot be represented by one sequence alone, but could be more fruitfully thought of as the same as a *partitura* in music – the score of a piece of orchestra music, in which every voice of the orchestra is given its own sequence of sounds, and all different sequences are aligned with each other to form a whole.

Based on this insight into the multi-layered character of the form part of the linguistic sign, the solution that I propose is a radically new model of sound sequences in historical linguistics. The idea for this enhanced sequence modeling was originally developed to handle prosodic context in phonetic alignments (List 2014: 130-133), by representing one sequence (a word or morpheme) not only by its sounds in form of transcriptions, but by additional sequences that would encode for the prosodic environment. That this idea could be further expanded was then shown in a toy application that would show how conditioning context that would change Proto-Germanic *p to German [p], [pf], and [f] could be inferred (List and Chacon 2015). The solution proposed here is called *multi-tiered sequence representation*, and the basic idea is to represent phonetic context by representing a sound sequence in a matrix in which each type of context can be represented in different degrees of abstraction in a row aligned to the original sequence, such as shown in Figure 3.

In Figure 4, I have furthermore tried to display how we can use the contexts across multiple occurrences of the same sound in a given proto-language, aligned to their reflex sound in a given descendant language, to search actively for conditioning context.

Orthography	k	i	n	d	e	r	g	a	r	t	e	n
IPA	k	ɪ	n	d	ɐ		g	a	ʁ	t	ə	n
Stress	2	2	2	0	0	0	1	1	1	0	0	0

(a) orthography, transcription, and stress

IPA	k	ɪ	n	d	ɐ	g	a	ʁ	t	ə	n
Preceding	#	k	ɪ	n	d	ɐ	g	a	ʁ	t	ə
Following	ɪ	n	d	ɐ	g	a	ʁ	t	ə	n	$
Prec. CV	#	C	V	c	C	V	C	V	c	C	V
Foll. CV	V	c	C	V	C	V	c	C	V	C	$

(b) context

Figure 3: German *Kindergarten* in multi-tiered sequence representation, with (a) representing tiers for transcription and stress, and (b) representing tiers for context.

Proto	p	p	p	p	p	p	p	p	p	p	p
Prec. CV	#	#	C	C	V	V	V	#	C	#	C
Foll. CV	C	V	c	c	V	V	V	C	V	c	V
Stress	2	2	2	2	2	1	0	1	0	1	0
Descendant	p	p	p	p	f	f	f	h	h	h	h

(a) proto-sounds, contexts, and reflexes

Proto	p	p	p	p	p	p	p	p	p	p	p
Prec. CV	#	#	C	C	V	V	V	#	C	#	C
Foll. CV	C	V	c	c	V	V	V	C	V	c	V
Stress	2	2	2	2	2	1	0	1	0	1	0
Descendant	p	p	p	p	f	f	f	h	h	h	h

(b) manually identified conditioning contexts

Figure 4: Searching for sound laws with help of multi-tiers.

Once context is handled in such a way, one could start to systematically search for those contexts which allow for a unique conversion of a proto-sound in a descendant sound, and would thus yield unambiguous results. With the small datasets we usually observe in historical linguistics, this can be easily done in an exhaustive fashion, by testing systematically all possible context combinations with respect to the accuracy by which they predict the reflex sounds. Since sound change should in theory proceed without exceptions, a sound law can be said to hold, if in all cases (or the majority of cases) the same context in the proto-language triggers the same reflex in the descendant.

4.3.6 Current work

In the CALC project, we currently develop a Python library to work with multi-tiered sequence representations (together with T. Tresoldi), and try to create first datasets for testing and training (with T. Tresoldi). In addition, we work on metrics to test to which degree a given reconstruction system and a given set of tiers predict the outcome in the target language (for initial ideas in this regard, see List forthcoming). While it may still seem problematic to handle the problem of *systemic changes* with the segment-oriented tier-model, a first step towards its solution is provided by the database of Cross-Linguistic Transcription Systems (CLTS, `https://clts.cldf.org`, Anderson et al. 2019), which provides distinct feature sets for more than 8000 distinct speech sounds. By creating multi-tiered sequences from features instead of raw sound segments, it may be possible to account for systematic changes as well in this framework (although we still do not know how to do this in concrete).

4.4 Automated phonological reconstruction (47-53)

The fourth problem in my list of open problems in computational diversity linguistics is devoted to the problem of linguistic reconstruction, or, more specifically, to the problem of phonological reconstruction, which can be characterized as follows: Given a set of cognate morphemes across a set of related languages, try to infer the hypothetical pronunciation of each morpheme in the proto-language.

This task needs to be distinguished from the broader task of linguistic reconstruction, which would usually include also the reconstruction of full lexemes, i.e. lexical reconstruction – as opposed to single morphemes or "roots" in an unknown ancestral language. In some cases, linguistic reconstruction is even used as a cover term for all reconstruction methods in historical linguistics, including such diverse approaches as phylogenetic reconstruction (finding the phylogeny of a language family), semantic reconstruction (finding the meaning of a reconstructed morpheme or root), or the task of demonstrating that languages are genetically related (see, e.g., the chapters in Fox 1995).

When listing phonological reconstruction as one of my ten problems, I am deliberately distinguishing this task from the tasks of lexical reconstruction or semantic reconstruction, since they can (and probably should) be carried out independently. Furthermore, by describing pronunciation of the morphemes as "hypothetical pronunciations" in the ancestral language, I want not only to emphasize that all reconstruction is hypothetical, but also to point to the fact that it is very possible that some of the morphemes for which one proposes a proto-form may not even have existed in the proto-language. They could have evolved only later as innovations on certain branches in the history of the languages. For the task of phonological reconstruction, however, this would not matter, since the question of whether a morpheme existed in the most recent common ancestor becomes relevant only if one tries to reconstruct the lexicon of a given proto-language. But phonological reconstruction seeks to reconstruct its phonology, i.e. the sound inventory of the proto-language, and the rules by which these sounds could be combined to form morphemes (phonotactics).

4.4.1 Background on computational approaches to phonological reconstruction

Not many attempts have been made so far to automate the task of reconstruction. The most prominent proposal in this direction has been made by (Bouchard-Côté et al. 2013). Their strategy radically differs from the strategies used by classical linguists (outlined below), since they do not make use of correspondence patterns, but instead use a stochastic transducer and known cognate words in the descendant languages, along with a known phylogenetic tree that they traverse, inferring the most likely changes that could explain the observed distribution of cognate sets.[7]

In a forthcoming paper, Gerhard Jäger forthcoming illustrates how classical methods for ancestral state reconstruction applied to aligned cognate sets could be used for the same task. While Jäger's method is more in line with "linguistic thinking", in so far as he uses alignments, and applies ancestral state reconstructions to each column of the alignments, it does not make use of correspondence patterns, which would be the general way by which linguists would proceed. This may also explain the performance, which shows an error rate of 0.48 (also using edit distance for evaluation) – although this is also due to the fact that the method was tested on Romance languages and compared with Latin, which is believed to be older than the ancestor of all Romance languages.

4.4.2 Problems with current solutions

Both the method of Bouchard-Côté et al. and the approach of Jäger suffer from the problem of not being able to detect unobserved sounds in the data. Jäger side-steps this problem in theory, by using a shortened alphabet of only 40 characters, proposed by the ASJP project, which encoded more than half of the world's languages in this form. Bouchard-Côté's test data, Proto-Austronesian (and its subgroups), are fairly simple in this regard. It would therefore be interesting to see what would happen if the methods are tested with full phonetic (or phonological) representations of more challenging language families (for example, the Chinese dialects). While Jäger's approach assumes the independence of all alignment sites, Bouchard-Côté's stochastic transducers handle context on the level of bigrams (if I read their description properly). However, while bigrams can be seen as an improvement over ignoring conditioning context, they are not the way in which context is typically handled by linguists.

[7] So far, this method has been tested only on Austronesian languages and their subgroups, where it performed particularly well (with error rates between 0.25 and 0.12, using edit distance as the evaluation measure). Since it is not available as a software package that can be conveniently used and tested on other language families, it is difficult to tell how well it would perform when being presented with more challenging test cases.

Apart from the handling of context and unobserved characters, the evaluation measure used in both approaches seems also problematic. Both approaches used the edit distance ("Dvoičnye kody s ispravleniem vypadenij, vstavok i zameščenij simvolov"), which is equivalent to the Hamming distance (Hamming 1950) applied to aligned sequences. Given the problem of *unobserved characters* and the abstract nature of linguistic reconstruction systems, however, any measure that evaluates the surface similarity of sequences is essentially wrong: even if two reconstruction systems do not share a single letter when aligned to the reflexes they reconstruct, they may still be structurally identical (List forthcoming).[8]

Currently, there is (to my knowledge) no accepted solution for the comparison of structural differences among aligned sequences. Finding an adequate evaluation measure to compare reconstruction systems can therefore be seen as a sub-problem of the bigger problem of phonological reconstruction. To illustrate why it is so important to compare the structural information and not the pure substance, consider the three cases in which Jäger's reconstruction gives a *v* as opposed to a *w* in Latin (data in List forthcoming): while evaluating by the edit distance yields a score of 0.48, this score will drop to 0.47 when replacing the *v* instances with a *w*. Jäger's system is doing something right, but the edit distance cannot capture the fact that the system is deviating systematically from Latin, not randomly.

4.4.3 Why is phonological reconstruction hard?

That phonological reconstruction is hard should not be surprising. What the task entails is to find the most probable pronunciation for a bunch of morphemes in a language for which no written records exist. Imagine you want to find the DNA of LUCA as a biologist, not even in its folded form, with all of the pieces in place, but just a couple of chunks, in order to get a better picture of how this LUCA might have looked like. But while we can employ some weak version of uniformitarianism when trying to reconstruct at least some genes of our LUCA (we would still assume that it was using some kind of DNA, drawn from the typical alphabet of DNA letters), we face the specific problem in linguistics that we cannot even be sure about the letters.[9]

But the very idea, that we may have good reasons to reconstruct something in our ancestral language that has been lost in all descendant languages, is something completely normal for linguists. In 1879, for example Ferdinand de Saussure used internal and comparative evidence to propose the existence of what he called coefficients sonantiques in Proto-Indo-European. His proposal included the prediction that – if ever a languages was found that retained these elements – these new sounds would surface as segmental elements, as distinctive sounds, in certain cognate sets, where all known Indo-European languages had already lost the contrast.

These sounds are nowadays known as laryngeals (**h1*, **h2*, **h3*, see Meier-Brügger 2002), and when Hittite was identified as an Indo-European language (Hrozný 1915), one of the two sounds predicted by Saussure could indeed be identified. I have discussed before on this blog the problem of unattested character states in historical linguistics, so there is no need to go into further detail. What I want to emphasize is that this aspect of linguistic reconstruction in general, and phonological reconstruction specifically, is one of the many points that makes the task really hard, since any algorithm to reconstruct

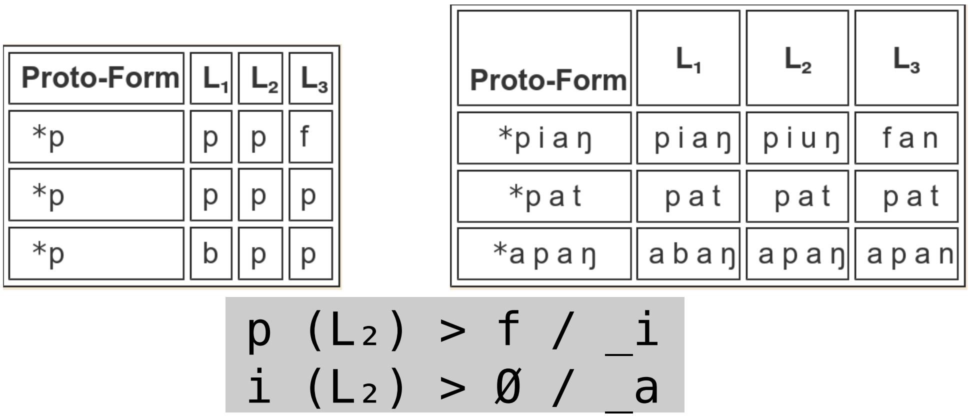
[8] Since scholars do not necessarily select phonetic values in their reconstructions that derive directly from the descendant languages, and moreover they may differ often regarding the details of the phonetic values they propose, a valid evaluation of different reconstruction systems (including automatically derived ones) needs to compare the structure of the systems, not their substance (see List 2014: 48-50 for a discussion of structural and substantial differences between sequences).

[9] Only recently, Blasi et al. (2019) argued that sounds like f and v may have evolved later than the other sounds we can find in the languages of the world, driven by post-Neolithic changes in the bite configuration, which seem to depend on what we eat. As a rule, and independent of these findings, linguists do not tend to reconstruct an *f* for the proto-language in those cases where they find it corresponding to a *p*, since we know that in almost all known cases a p can evolve into an *f*, but an *f* almost never becomes a *p* again. This can lead to the strange situation where some linguists reconstruct a *p* for a given proto-language even though all descendants show an *f*, which is, of course, an exaggeration of the principle (see the discussion in Jacques 2019).

the phonological system of some proto-language would have to find a way to formalize the complicated arguments by which linguists infer that there are traces of something that is no longer there.[10]

4.4.4 How humans reconstruct phonemes

Given the lack of methodological literature on phonological reconstruction, it is not easy to describe how it should be done in an ideal scenario. What seems to me to be the most promising approach is to start from correspondence patterns. A correspondence pattern is an abstraction from individual alignment sites distributed over cognate sets drawn from related languages. As I have tried to show in a paper published earlier this year (List 2019), a correspondence pattern summarizes individual alignment sites in an abstract form, where missing data are imputed. I will avoid going into the details here but, as a short-cut, we can say that each correspondence pattern should, in theory, only correspond to one proto-sound in the language, although the same proto-sound may correspond to more than one correspondence pattern. As an example, consider the left table in Figure 5, showing three (fictive) patterns that would all be reconstructed by a **p*. To justify that the same proto-sound is reconstructed by a **p* in all three patterns, linguists invoke the rule of context, by looking at the real words from which the pattern was derived. An example is shown in the next table.



Proto-Form	L$_1$	L$_2$	L$_3$
*p	p	p	f
*p	p	p	p
*p	b	p	p

Proto-Form	L$_1$	L$_2$	L$_3$
*piaŋ	piaŋ	piuŋ	fan
*pat	pat	pat	pat
*apaŋ	abaŋ	apaŋ	apan

p (L$_2$) > f / _i
i (L$_2$) > Ø / _a

Figure 5: Example for correspondence-pattern-based phonological reconstruction.

What one should be able to see from the table is that we can find in all three patterns a conditioning factor that allows us to assume that the deviation from the original **p* is secondary. In language L$_3$, the factor can be found in the palatal environment (followed by the front vowel **i*) that we find in the ancestral language. We would assume that this environment triggered the change from **p* to *f* in this language. In the case of the change from **p* to *b* in L$_1$, the triggering environment is that the *p* occurs inter-vocalically.

To summarize: what linguists usually do in order to reconstruct proto-forms for ancestral languages that are not attested in written sources, is to investigate the correspondence patterns, and to try to find some neat explanation of how they could have evolved, given a set of proto-forms along with triggering contexts that explain individual changes in individual descendant languages.

[10] There are many more things that I could mention, if I wanted to identify the difficulty of phonological reconstruction in its entirety. What I find most difficult to deal with is that the methodology is insufficiently formalized. Linguists have their success stories, which helped them to predict certain aspects of a given proto-language that could later be confirmed, and it is due to these success stories that we are confident that it can, in principle, be done. But the methodological literature is sparse, and the rare cases where scholars have tried to formalize it are rarely discussed when it comes to evaluating concrete proposals (as an example for an attempt of formalizing, see Hoenigswald 1960).

4.4.5 Suggestions for improvement

There are many things that we can easily improve when working on automated methods for phonological reconstruction. As a first point, we should work on enhanced measures of evaluation, going beyond the edit distance as our main evaluation measure. In fact, this can be easily done. With B-Cubed scores (Amigó et al. 2009), we already have a straightforward measure to compare whether two reconstruction systems are structurally identical or similar. In order to apply these scores, the automatic reconstructions have to be aligned with the gold standard. If they are identical, although the symbols may differ, then the scores will indicate this, as shown in (List forthcoming).[11]

Furthermore, given that we lack test cases, we might want to work on semi-automatic instead of fully automatic methods, in the meantime. Given that we have a first method to infer sound correspondence patterns from aligned data (List 2019), we can infer all patterns and have linguists annotate each pattern by providing the proto-sound they think would fit best – we are testing this at the moment in studies on Burmish (with N. W. Hill) and Kho-Bwa (with T. A. Bodt). Having created enough datasets in this form, we could then think of discussing concrete algorithms that would derive proto-forms from correspondence patterns, and use the semi-automatically created and manually corrected data as gold standard.

Last but not least, one straightforward way by which it is possible to formally create unknown sounds from known data, is to represent sound as vectors of phonological features instead of bare symbols (e.g. representing *p* as *voiceless bilabial plosive consonant* and *b* as *voiced labial plosive consonant*). If we then compare alignment sites or correspondence patterns for the feature vectors, we could check to what degree standard algorithms for ancestral state reconstructions propose unattested sounds similar to the ones proposed by experts. In order to do this, we would need to encode our data in transparent transcription systems. This is not the case for most current datasets, but with the Cross-Linguistic Transcription Systems initiative (Anderson et al. 2019) we already have a first attempt to provide features for the majority of sounds that we find in the languages of the world.

4.4.6 Current work

In our work in the CALC project, we are currently establishing linguistic reconstructions by collaborating with different researchers on specific subgroups. We are also testing semi-automatic methods for reconstruction, based on the algorithm for sound correspondence pattern detection by List (2019), evaluating metrics for the comparison of reconstruction systems (List forthcoming), and testing multi-tier-based methods to test the predictive strength of different reconstruction systems.

4.5 General ideas (54-56)

Given the aforementioned problems, there are some general ideas on problem solving, which I would like to share in the end of this review. The first idea relates to the obvious problem that we lack good benchmark datasets and evaluation standards in computational historical linguistics. These problems can be addressed in two ways. First, we could try to develop models for simulation, as they would guarantee us "cheap" data for evaluation and allow us directly to test the power of a given method to identify the processes underlying the simulation study. In order to develop high-level benchmark datasets for the specific tasks at hand, we need to develop *interfaces for data annotation* that help linguists to produce the relevant data in a machine-readable, efficient, and consistent way.

To allow for machine- and human-readable data, we need to invest in *standards* for data annotation. First efforts in this direction are currently being undertaken, as reflected by the *Cross-Linguistic Data*

[11] The problem of comparing reconstruction systems is, of course, more difficult, as we can face cases where systems are not structurally identical (i.e. you can directly replace any symbol *a* in system `A` by any symbol *a'* in system `B` to produce `B` from `A` and vice versa), but they would be a start.

Formats initiative (Forkel et al. 2018). The CLDF initiative is based on a straightforward tabular format that can be edited with help of a spreadsheet editor, and offers code to check if a given datasets conforms to the standards. In addition, CLDF encourages scholars to use *reference catalogs*, such as Concepticon (List et al. 2016), Glottolog (Hammarström et al. 2018), or CLTS (Anderson et al. 2019), to represent elicitation glosses for concepts, language varieties, and speech sounds.

As a first interface that enhances standard tasks of manual annotation and analysis in computational historical linguistics, the EDICTOR tool (List 2017) is freely available, web-based (and therefore platform-independent) application, that can be used to annotated which words in a wordlist are cognate, to align the cognate words, to check the phonological transcriptions in a dataset, and to inspect automatically inferred sound correspondence patterns. The tool is still under heavy development, but it has already proven useful in a couple of publications (Chacon 2017, Chen 2019, Kaiping and Klamer 2018, Sagart et al. 2019).

5 Outlook

Often (but wrongly, see Kleinert 2009) attributed to Galileo Galilei, the quote to “Measure what is measurable, and make measurable what is not so.” can be seen as one of the major goals of the CALC initiative. In the past, I have met quite a few linguists who emphasized the impossibility of some of the tasks I essentially wanted to achieve, and often, they argued that the task was impossible, since it could not be represented in a computer-readable (i.e., “measurable”) way. But if somethings seems impossible to measure, this should not prevent scientists from trying to do so.[12] Formulating open problems for our field, by mentioning things we still cannot measure successfully, is a first in this direction. Searching open problems in our field that may have been overlooked so far is a first step to a deeper understanding of our research and our research object.

References

Aikhenvald, A. Y. (2007). “Semantics and pragmatics of grammatical relations in the Vaups linguistic area” . In: *Grammars in contact: A cross-linguistic typology*. Ed. by A. Y. Aikhenvald and R. M. W. Dixon. Vol. 4. Explorations in linguistic typology. Oxford: Oxford University Press, 237–266.

Amigó, E., J. Gonzalo, J. Artiles, and F. Verdejo (2009). “A comparison of extrinsic clustering evaluation metrics based on formal constraints” . *Information Retrieval* 12.4, 461–486.

Anderson, C., T. Tresoldi, T. C. Chacon, A.-M. Fehn, M. Walworth, R. Forkel, and J.-M. List (2019). “A Cross-Linguistic Database of Phonetic Transcription Systems” . *Yearbook of the Poznań Linguistic Meeting*, 1–27.

Ark, R. van der, P. Mennecier, J. Nerbonne, and F. Manni (2007). “Preliminary identification of language groups and loan words in Central Asia” . In: *Proceedings of the RANLP Workshop on Acquisition and Management of Multilingual Lexicons*. (Borovets, 09/03/2007), 13–20.

Baayen, R. H., R. Piepenbrock, and L. Gulikers, comp. (1995). *The CELEX Lexical Database*. Version 2. Linguistic Data Consortium.

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

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

Baxter, W. H. and A. Manaster Ramer (2000). “Beyond lumping and splitting: Probabilistic issues in historical linguistics” . In: *Time depth in historical linguistics*. Ed. by C. Renfrew, A. McMahon, and L. Trask. Cambridge: McDonald Institute for Archaeological Research, 167–188.

[12] While I expect the tasks which I discussed above to remain very challenging in the nearer and farther future, we can see drastic achievements in some other problems that were also treated as rather difficult if not impossible to be measured. One example is the problem of semantic change. While we are still not able to provide preference pathways of semantic change, the introduction of *colexification networks* (Cysouw 2010, List et al. 2013) has drastically changed the potential of computational studies dealing with semantics. This is probably best illustrated with the recently published update of the CLICS^2 database (List et al. 2018), in which we made consequent use of the standardization efforts of the CLDF initiative to aggregate data from several different sources into a common framework that almost reaches “big data status”.

Benden, C. (2005). “Automated detection of morphemes using distributional measurements” . In: *Classification – the ubiquitous challenge*. Ed. by C. Weihs and W. Gaul. Berlin and Heidelberg: Springer, 490–497.

Berg, T. (1998). *Linguistic structure and change: An explanation from language processing*. Gloucestershire: Clarendon Press.

Bhasker, J. and T. Samad (1991). “The clique-partitioning problem” . *Computers & Mathematics with Applications* 22.6, 1 –11.

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.

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.

Boc, A., A. M. Di Sciullo, and V. Makarenkov (2010). “Classification of the Indo-European languages using a phylogenetic network approach” . In: *Classification as a tool fo research*. Ed. by H. Locarek-Junge and C. Weihs. Berlin and Heidelberg: Springer, 647–655.

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 langauges” . *Papers in Historical Phonology* 4.1, 22–44.

Bordag, S. (2008). “Unsupervised and knowledge-free morpheme segmentation and analysis” . In: *Advances in multilingual and multimodal information retrieval*. Ed. by C. Peters, V. Jijkoun, T. Mandl, H. Müller, D. W. Oard, A. Peñas, V. Petras, and D. Santos. Lecture Notes in Computer Science 5152. Berlin and Heidelberg: Springer, 881–891.

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.

Brysbaert, M., M. Stevens, P. Mandera, and E. Keuleers (2016). “How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant ’ s Age” . *Frontiers in Psychology* 7, 1116.

Cathcart, C., G. Carling, F. Larson, R. Johansson, and E. Round (2018). “Areal pressure in grammatical evolution. An Indo-European case study” . *Diachronica* 35.1, 1–34.

Chacon, T. C. (2017). “Arawakan and Tukanoan contacts in Northwest Amazonia prehistory” . *PAPIA* 27.2, 237–265.

陈保亚, C. B. (1996). *Lùn yǔyán jiēchù yǔ yǔyán liánméng* 论语言接触与语言联盟 [Language contact and language unions]. Běijīng 北京: Yǔwén 语文.

Chen, E. (2019). “Phonological reconstruction of Proto-Kampa consonants” . *Berkeley Papers in Formal Linguistics* 2.1, 1–56.

Ciobanu, A. M. and L. P. Dinu (2018). “Simulating language evolution: A tool for historical linguistics” . In: *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*. (Santa Fe). Association of Computational Linguistics, 68–72.

Creutz, M. and K. Lagus (2005). *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0*. Publications in Computer and Information Sciences 81. Helsinki: Helsinki University of Technology.

Cysouw, M. (2010). “Semantic maps as metrics on meaning” . *Linguistic Discovery* 8.1, 70–95.

Dehmer, M., F. Emmert-Streib, A. Graber, and A. Salvador, eds. and introd. (2011). Weinheim: Wiley-Blackwell.

Dekker, P. (2018). “Reconstructing language ancestry by performing word prediction with neural networks” . Master. Amsterdam: University of Amsterdam.

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

Fox, A. (1995). *Linguistic reconstruction. An introduction to theory and method*. Oxford: Oxford University Press.

Goldsmith, J. A., J. L. Lee, and A. Xanthos (2017). “Computational learning of morphology” . *Annual Review of Linguistics* 3.1, 85–106.

Hammarström, H. (2006). “A Naive Theory of Affixation and an Algorithm for Extraction” . In: *Proceedings of the Eighth Meeting of the ACL Special Interest Group on Computational Phonology and Morphology at HLT-NAACL 2006*. New York City, USA: Association for Computational Linguistics, 79–88.

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

Hamming, R. W. (1950). “Error detection and error detection codes” . *Bell System Technical Journal* 29.2, 147–160.

Hantgan, A. and J.-M. List (forthcoming). “Bangime: Secret language, language isolate, or language island?” *Journal of Language Contact* 0.0.

Harris, Z. S. (1955). “From phoneme to morpheme” . *Language* 31.2, 190–222.

Hartmann, L. (2003). “Phono. Software for modeling regular historical sound change” . In: *Actas VIII Simposio Internacional de Comunicación Social*. (Santiago de Cuba). Southern Illinois University, 606–609.

Herder, J. G. (1778). *Abhandlung über den Ursprung der Sprache, welche den von der königl. Academie der Wissenschaften für das Jahr 1770 gesetzten Preis erhalten hat. Welche den von der Königl. Academie der Wissenschaften für das Jahr 1770 gesetzten Preis erhalten hat*. Berlin: Christian Friedrich Voß. Google Books: `QP4TAAAAQAAJ`.

Hilbeert, D. (1902). “Mathematical problems” . *Bulletin of the New York Mathematical Society* 8.1, 437–479.

Hoenigswald, H. M. (1960). “Phonetic similarity in internal reconstruction” . *Language* 36.2, 191–192. JSTOR: `410982`.

Holman, E. W., S. Wichmann, C. H. Brown, V. Velupillai, A. Müller, and D. Bakker (2008). "Explorations in automated lexicostatistics". *Folia Linguistica* 20.3, 116–121.

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

Jacques, G. (2019). "L' antiquité des fricatives labiodentales [The antiquity of labiodental fricatives]". *Panchronica*.

Jakobson, R. (1960). "Why 'Mama´and 'Papa?". In: *Perspectives in psychological theory: Essays in honor of Heinz Werner*. Ed. by B. Kaplan and S. Wapner. New York: International Universities Press, 124–134.

Jarceva, V. N., ed. (1990). *Lingvističeskij ėnciklopedičeskij slovar (Linguistical encyclopedical dictionary)*. Moscow: Sovetskaja Enciklopedija.

Jäger, G. (2018). "Global-scale phylogenetic linguistic inference from lexical resources". *Scientific Data* 5.180189, 1–16.

– (forthcoming). "Computational historical linguistics". *Theoretical Linguistics*.

Kaiping, G. A. and M. Klamer (10/2018). "LexiRumah: An online lexical database of the Lesser Sunda Islands". *PLOS ONE* 13.10, 1–29.

Kassian, A., M. Zhivlov, and G. S. Starostin (2015). "Proto-Indo-European-Uralic comparison from the probabilistic point of view". *The Journal of Indo-European Studies* 43.3-4, 301–347.

Kessler, B. (2001). *The significance of word lists. Statistical tests for investigating historical connections between languages*. Stanford: CSLI Publications.

Kleinert, A. (2009). "Der messende Luchs. Zwei verbreitete Fehler in der Galilei-Literatur [The Measuring Lynx. Two widespread mistakes in the Galileo literature]". *NTM Zeitschrift für Geschichte der Wissenschaften, Technik und Medizin* 17.2, 199–206.

Kluge, F., found. (2002). *Etymologisches Wörterbuch der deutschen Sprache*. Cont. by E. Seebold. 24th ed. Berlin: de Gruyter.

Lee, Y.-J. and L. Sagart (2008). "No limits to borrowing: The case of Bai and Chinese". *Diachronica* 25.3, 357–385.

Levenshtein, V. I. (1965). "Dvoičnye kody s ispravleniem vypadenij, vstavok i zameščenij simvolov [Binary codes with correction of deletions, insertions and replacements]". *Doklady Akademij Nauk SSSR* 163.4, 845–848; English translation: – (1966). "Binary codes capable of correcting deletions, insertions, and reversals". *Soviet Physics Doklady* 10.8 (1966), 707–710.

List, J.-M. (2014a). "Investigating the impact of sample size on cognate detection". *Journal of Language Relationship* 11, 91–101.

– (2014b). *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.

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

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

– (2018). "Von Wortfamilien und promiskuitiven Wörtern [Of word families and promiscuous words]". *Von Wörtern und Bäumen* 2.10.

– (2019a). "Automatic detection of borrowing (Open problems in computational diversity linguistics 3)". *The Genealogical World of Phylogenetic Networks* 6.3.

– (2019b). "Automatic inference of sound correspondence patterns across multiple languages". *Computational Linguistics* 1.45, 137–161.

– (2019c). "Automatic morpheme segmentation (Open problems in computational diversity linguistics 1)". *The Genealogical World of Phylogenetic Networks* 6.2.

– (2019d). "Automatic phonological reconstruction (Open problems in computational diversity linguistics 4)". *The Genealogical World of Phylogenetic Networks* 6.5.

– (2019e). "Automatic sound law induction (Open problems in computational diversity linguistics 3)". *The Genealogical World of Phylogenetic Networks* 6.4.

– (2019f). *Studying language contact within a computer-assisted framework*. Talk, held at the "64th Annual Conference of the International Linguistic Association" (Buenos Aires, 05/30/2019–06/01/2019).

– (forthcoming[a]). "Automated methods for the investigation of language contact situations". *Language and Linguistics Compass* ????, 1–19.

– (forthcoming[b]). "Beyond Edit Distances: Comparing linguistic reconstruction systems". *Theoretical Linguistics* ????, 1–10.

List, J.-M. and T. Chacon (2015). *Towards a cross-linguistic database for historical phonology? A proposal for a machine-readable modeling of phonetic context*. Paper, presented at the workshop "Historical Phonology and Phonological Theory [organized as part of the 48th annual meeting of the SLE]" (Leiden, 09/04/2015).

List, J.-M., A. Terhalle, and M. Urban (2013). "Using network approaches to enhance the analysis of cross-linguistic polysemies". In: *Proceedings of the 10th International Conference on Computational Semantics – Short Papers*. "IWCS 2013" (Potsdam, 03/19/2013–03/22/2013). Association for Computational Linguistics. Stroudsburg, 347–353. PDF: `http://aclweb.org/anthology-new/W/W13/W13-0208.pdf`.

List, J.-M., S. Nelson-Sathi, H. Geisler, and W. Martin (2014). "Networks of lexical borrowing and lateral gene transfer in language and genome evolution". *Bioessays* 36.2, 141–150.

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, 05/23/2016–05/28/2016). Ed. by N. C. C. Chair), K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis. European Language Resources Association (ELRA), 2393–2400.

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

List, J.-M., S. J. Greenhill, C. Anderson, T. Mayer, T. Tresoldi, and R. Forkel (2018a). "CLICS^2. 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. Greenhill, T. Tresoldi, and R. Forkel (2018b). *LingPy. A Python library for quantitative tasks in historical linguistics*. URL: http://lingpy.org.

Malkiel, Y. (1954). "Etymology and the Structure of Word Families" . *Word* 10.2-3, 265–274.

McMahon, A., P. Heggarty, R. McMahon, and N. Slaska (2005). "Swadesh sublists and the benefits of borrowing: An Andean case study" . *Transactions of the Philological Society* 103, 147–170.

Meier-Brügger, M. (2002). *Indogermanische Sprachwissenschaft*. In collab. with M. Fritz and M. Mayrhofer. 8th ed. Berlin and New York: de Gruyter.

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

Minett, J. W. and W. S.-Y. Wang (2003). "On detecting borrowing" . *Diachronica* 20.2, 289–330.

Moore, R. K. (2005). "Results from a survey of attendees at ASRU 1997 and 2003" . In: *Proceedings of INTERSPEECH 2005*. (Lisbon, 09/04/2005–09/08/2005), 117–120.

Morrison, D. A. (2011). *An introduction to phylogenetic networks*. Uppsala: RJR Productions.

Mortarino, C. (2009). "An improved statistical test for historical linguistics" . *Statistical Methods and Applications* 18.2, 193–204.

Nakhleh, L., D. Ringe, and T. Warnow (2005). "Perfect Phylogenetic Networks: A new methodology for reconstructing the evolutionary history of natural languages" . *Language* 81.2, 382–420. JSTOR: 4489897.

Needleman, S. B. and C. D. Wunsch (1970). "A gene method applicable to the search for similarities in the amino acid sequence of two proteins" . *Journal of Molecular Biology* 48, 443–453.

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.

Papakitsos, E. C. and I. K. Kenanidis (2018). "Going to the root: Paving the way to reconstruct the language of homo-sapiens" . *International Linguistics Research* 1.2, 1–16.

Peirce, C. S. (1931/1958). *Collected papers of Charles Sanders Peirce*. Ed. by C. Hartshorne and P. Weiss. Cont. by A. W. Burke. 8 vols. Cambridge, Mass.: Harvard University Press.

Pfeifer, W., comp. (1993). *Etymologisches Wörterbuch des Deutschen*. 2nd ed. 2 vols. Berlin: Akademie. URL: http://www.dwds.de/.

Rama, T. (2016). "Siamese convolutional networks for cognate identification" . In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. "COLING 2016" (Osaka, 12/11/2016–12/17/2016), 1018–1027.

Rama, T., J. Wahle, P. Sofroniev, and G. Jäger (2017). "Fast and unsupervised methods for multilingual cognate clustering" . *CoRR* abs/1702.04938. arXiv: 1702.04938.

Rama, T., J.-M. List, J. Wahle, and G. Jäger (2018). "Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics?" In: *Proceedings of the North American Chapter of the Association of Computational Linguistics*. "NAACL 18" (New Orleans, 06/01/2018–06/06/2018), 393–400.

Renfrew, C., A. McMahon, and L. Trask, eds. (2000). *Time depth in historical linguistics*. Cambridge: McDonald Institute for Archaeological Research.

Ringe, D. A. (1992). "On calculating the factor of chance in language comparison" . *Transactions of the American Philosophical Society*. New Series 82.1, 1–110. JSTOR: 1006563.

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

Schweikhard, N. E. (11/07/2018). "Semantic promiscuity as a factor of productivity in word formation" . *Computer-Assisted Language Comparison in Practice* 1.11.

Schwink, F. (1994). *Linguistic typology, universality and the realism of reconstruction*. Washington: Institute for the Study of Man.

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

Starostin, S. A. (1991). *Altajskaja problema i proischoždenije japonskogo jazyka [The Altaic problem and the origin of the Japanese language]*. Moscow: Nauka.

Starostin, S. A. (1995). “Old Chinese vocabulary: A historical perspective” . In: *The ancestry of the Chinese language*. Ed. by W. S.-Y. Wang. Berkeley: University of California Press, 225–251.

Starostin, S. A. (1989). “Sravnitel’no-istoričeskoe jazykoznanie i leksikostatistika [Comparative-historical linguistics and lexicostatistics]” . In: *Lingvističeskaja rekonstrukcija i drevnejšaja istorija Vostoka* [Linguistic reconstruction and the oldest history of the East]. Vol. 1: *Materialy k diskussijam na konferencii* [Materials for the discussion on the conference]. Ed. by S. V. Kullanda, J. D. Longinov, A. J. Militarev, E. J. Nosenko, and V. A. Shnirelḿan. Moscow: Institut Vostokovedenija, 3–39; English translation: – (2000). “Comparative-historical linguistics and lexicostatistics” . In: *Time depth in historical linguistics*. Vol. 1: *Time depth in historical linguistics*. Ed. by C. Renfrew, A. McMahon, and L. Trask. Trans. from the Russian by I. Peiros. Papers in the prehistory of languages. Cambridge: McDonald Institute for Archaeological Research, 2000, 223–265.

“Statuts” (1871). “Statuts. Approuvés par décision ministérielle du 8 Mars 1866” . *Bulletin de la Société de Linguistique de Paris* 1, III–IV.

Sturtevant, E. H. (1920). *The pronunciation of Greek and Latin*. Chicago: University of Chicago Press. Internet Archive: `http://www.archive.org/details/pronunciationgr00unkngoog`.

Tadmor, U. (2009). “Loanwords in the world’ s languages. Findings and results” . In: *Loanwords in the worldś languages. A comparative handbook*. Ed. by M. Haspelmath and U. Tadmor. Berlin and New York: de Gruyter, 55–75.

Verner, K. A. (1877). “Eine Ausnahme der ersten Lautverschiebung [An exception to the first sound shift]” . *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete der Indogermanischen Sprachen* 23.2, 97–130.

Virpioja, S., P. Smit, S.-A. Grönroos, and M. Kurimo (2013). *Morfessor 2.0: Python Implementation and Extensions for Morfessor Baseline*. Helsinki: Aalto University.

Welsh, D. J. A. and M. B. Powell (1967). “An upper bound for the chromatic number of a graph and its application to timetabling problems” . *The Computer Journal* 10.1, 85–86.

Whewell, W. D. D. (1847). *The philosophy of the inductive sciences, fofound upon their history*. 2nd ed. Vol. 2. London: John W. Parker.

Willems, M., E. Lord, L. Laforest, G. Labelle, F.-J. Lapointe, A. M. Di Sciullo, and V. Makarenkov (2016). “Using hybridization networks to retrace the evolution of Indo-European languages” . *BMC Evolutionary Biology* 16.1, 1–18.

Willis, D. (2011). “Reconstructing last weekś weather: Syntactic reconstruction and Brythonic free relatives” . *Journal of Linguistics* 47.2, 407–446.

Wilson, E. O. (1998). *Consilience. The unity of knowledge*. New York: Vintage Books.

Zenner, E., D. Speelman, and D. Geeraerts (2014). “Core vocabulary, borrowability and entrenchment” . *Diachronica* 31.1, 74–105.

Acknowledgments

This research is funded by the ERC Starting Grant 715618 “Computer-Assisted Language Comparison” (`http://calc.digling.org`). I am very thankful to the organizing committee of the ICHL-24 at ANU, Bethwyn Evans, Nick Evans, Simon Greenhill, and Jennifer Hendriks, for inviting me to the conference. I want to express my special thanks to current and past CALC associates, as well as to the many colleagues who helped me with advise and critics in the past: Cormac Anderson, Wolfgang Behr, Timotheus A. Bodt, Thiago Chacon, Michael Cysouw, Robert Forkel, Hans Geisler, Guido Grimm, Simon Greenhill, Russell Gray, Guillaume Jacques, Gerhard Jäger, Harald Hammarström, Martin Haspelmath, Gereon Kaiping, Yunfan Lai, Nathan W. Hill, David Morrison, Justin Power, Taraka Rama, Christoph Rzymski, Laurent Sagart, Nathanael Schweikhard, George Starostin, Tiago Tresoldi, Mary Walworth, Søren Wichmann, Mei-Shin Wu.