Three Open Problems in Computational Historical Linguistics

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 diversity 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 pick three specific problems (automatic morpheme detection, automatic borrowing detection, and automatic induction of sound laws), and present initial strategies for their solution using interdisciplinary and algorithmic approaches which rely closely on existing strategies in classical approaches to historical language comparison.

1 Introduction

1.1 Problems

When working every day on very detailed scientific problems, one always runs danger of loosing track of the broader challenges of one's field. That these challenges exist, and that we often still lack sufficient answers to certain problems becomes specifically clear when listening to the questions which laypeople or scientists from other fields ask with respect to one's area of expertise. In linguistics, for example, people are usually very surprised that the question of how language evolved the first time, the question regarding the origin of language, has been officially banned from the agenda of linguistics already in the 19th century, in the often-quoted *statuts* of 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 in fact 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, or to recent mystic accounts, which have so far been ignored by a larger public:

> 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. 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 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.

In a paper on similarities 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 (“Sravnitel'no-istoričeskoe jazykoznanie i leksikostatistika”), 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 big 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 Ten open problems for historical linguistics

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, three problems have already been

prepared, with two being officially published (*Automatic borrowing detection*, List 2019, *Automatic morpheme detection*, List 2019).

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 trias, 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.

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. 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.

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 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).

Number	Problem	Class
1	automatic morpheme segmentation	inference
2	automatic sound law induction	inference
3	automatic borrowing detection	inference
4	automatic 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

3 Computer-assisted strategies for problem solving

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.

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.

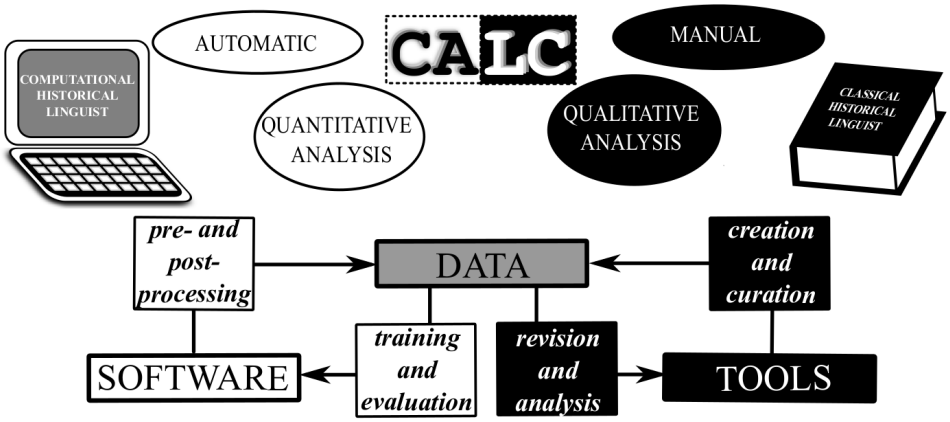

Figure 1: Basic idea of data managment within the CALC framework.

3.2 Machine learning and black boxes

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. 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

One could argue that a machine-learning-based method that could detect borrowed words, for example, would already be good enough: it could help us to find more borrowings, and then follow the outline of *modeling*, *inference*, and *analysis* to increase our knowledge. But the problem is that most of the sophisticated frameworks still royally fail when it comes to addressing tasks in historical linguistics,

and the reason lies in the fact, that the scholars usually believe in machine learning as an easy shortcut to achieve inference without spending too much time to think about a given model. What I call a model in this context is usually called *feature design* or *feature engineering* in the context of machine learning (see Round 2017, who introduces a linguistic notion on the concept), and although it is not often stated in the literature, it is clear that even with the most sophisticated approaches, there is still a definite need in our research to spend some time on thinking how to model our data, in order to come up with a convincing solution.

> One of the promises of deep learning is that it vastly simplifies the feature-engineering process by allowing the model designer to specify a small set of core, basic, or "natural" features, and letting the trainable neural network architecture combine them into more meaningful higher-level features, or representations. However, one still needs to specify a suitable set of core features, and tie them to a suitable architecture. (Goldberg 2017: 18)

Unfortunately, the aspect of feature design, i.e., the careful modeling of the processes we want to investigate, is ignored in most machine learning approaches that have been proposed so far (or it is in fact not ignored, but its role is played down in the description). This reflects the tendency of scholars to dream of a shortcut that would allow them to infer something spectacular without doing the actual work in trying to figure out what the process at hand actually does. In addition, there is a strong misunderstanding among non-linguists who work in the area of linguistics, as they often criticize that the use of *strong models* would bias the analyses and render them less objective, but they misunderstand here, that being objective does not require to be naive.

The problem of the engineering attitude behind most machine learning approaches ("get the job done, don't care how") is reflected in the success of the AlphaGo system (Silver et al. 2016). Since we do not know why trained models make the decisions they make, we have no way to learn from them, and for this reason, humans are now considering to investigate how the AlphaGo system plays the Go game in order to learn from it. Furthermore, since the models in many areas, for example in automatic translation, but also in cognate detection, are trained on human annotations, the "objective" extraction of the most successful feature weights may as well reflect common human bias rather than scientific truth. Scientifically, a machine that tells us if two words are related or not is of no direct value if we're interested in the deeper question of how we can prove word relatedness (it may be useful for other studies, but it does not solve overarching scientific questions).

The black box problem is most prominently reflected in the detection of Lapuschkin et al. (2016), who could show that algorithms which are trained on the wrong data, may only seemingly yield good results, while they base their classifications on wrong aspects of the data, such as, for example, a copyright sign on the bottom of training images, which helped to recognize a horse, when all training data came from the same provider. Trusting black box machine learning approaches blindly is always dangerous, and some even say it might lead a future crisis in science (Ghosh 2019). When using untransparent automatic methods for inference – be it in historical linguistics or any other science – we need to be aware of the fact that we always run the risk of erroneous estimates on a large scale.

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 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

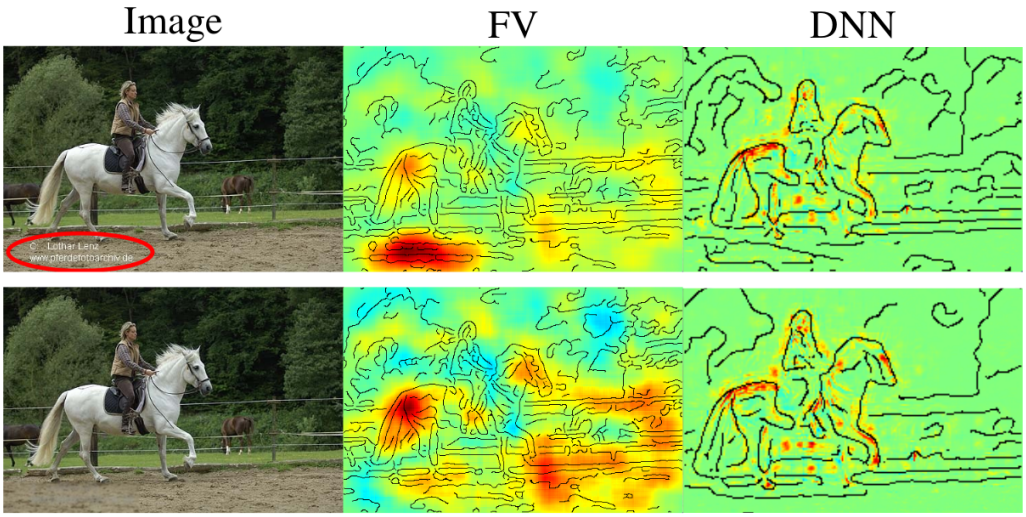

Figure 2: Mistaking copyright marks as being characteristic for horses (from Lapuschkin et al. 2016).

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.

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).

4 Three problems in computational historical linguistics

4.1 Automatic morpheme segmentation

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).

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 (see `https://gist.github.com/LinguList/04bba6d97595d7e474ab109a639fecce` for data and code).

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).

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 3: 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 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.

4.2 Automatic contact inference

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* (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.

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.

4.3 Automatic sound law induction

The last 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.

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 virtually no current solutions virtually 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 experiment, 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

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 Table 2.

Tier	Alignment				
SOURCE	s	w	e	r	d
CV / x_	#	C	C	V	C
CV / _x	C	V	C	C	$
SOUND CLASSES / x_	#	S	W	V	R
SOUND CLASSES / _x	W	V	R	T	$
PROSODIC STRENGTH	7	5	4	3	2
WORD LENGTH	1	1	1	1	1
FEATURES	F	G	V	L	P
ACCENT	1	1	1	1	1
TARGET	ʃ	v	eː	r	t

Table 2: Proto-Germanic **swerd-* in multi-tiered sequence representation.

In Figure 4, I have furthermore tried to display how different contexts can be derived for one single segment in a given sound sequence. The chief idea of this procedure is to avoid any unrealistic modeling of a sound sequence with help of, for example, bi-, tri-, or *n*-grams, by allowing for a representation of phonetic context in a consistent form on each single segment of sound sequence.

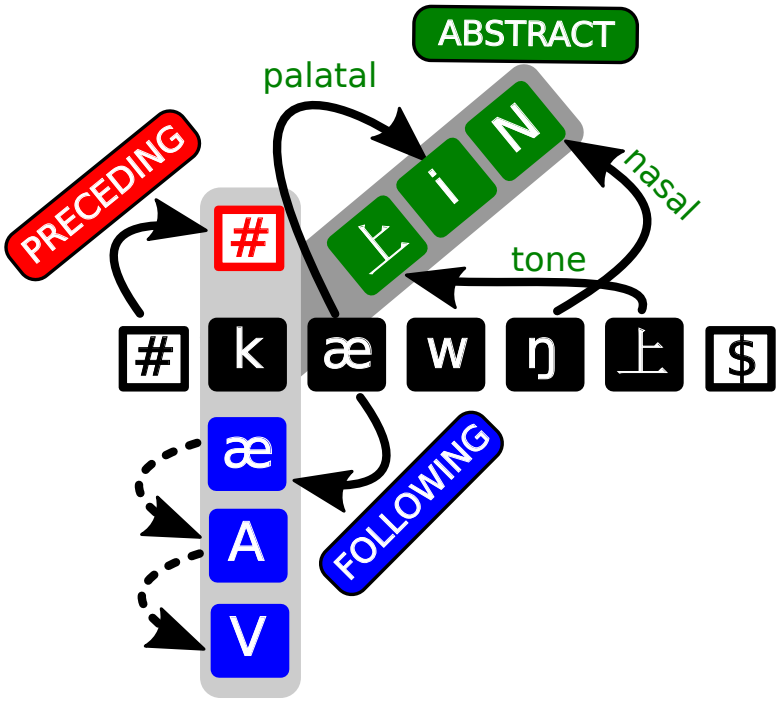

Figure 4: Sources for context displayed in multi-tiered sequence representations.

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. How this search is being carried out in concrete is another question, but an exhaustive listing and checking may be sufficient for initial experiments.

One problem, however, remains in this context: the handling of systematic aspects of sound change. While it may be straightforward to expand the multi-tiered sequence annotation along these lines, we are stuck in this respect, as long as there is not full-fledged feature system that we can trust in historical

linguistics. Given that attempts to provide a cross-linguistic phonetic transcription system that is more explicit than IPA and represented by some rough feature model have shown how easily the number of possible sounds attested in different datasets increases, with more than 8000 distinct sounds being reflected in the bigger datasets (Anderson et al. forthcoming), it is clear that we are only beginning to understand how to handle linguistic complexity in a sufficient, computer-assisted manner.

5 Outlook

This paper is an attempt to raise problems in historical linguistics that have so far not been sufficiently addressed in computational approaches. By providing a list of 10 problems I consider challenging but not impossible to resolve, I have tried to instigate a discussion in our field that would hopefully leave to future solutions of these and many other problems that I have not discussed in this context. Furthermore, by illustrating how initial solutions could be found for three problems, I have tried to show how we could profit from a computer-assisted framework for problem solving, by starting from a close inspection of available solutions for problems in the classical, qualitative literature in order to then use existing solutions from other parts of science to come up with computer-assisted ways to address the problems in our own field.

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.

Anderson, C., T. Tresoldi, T. C. Chacon, A.-M. Fehn, M. Walworth, R. Forkel, and J.-M. List (forthcoming). "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.

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.

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*. Preprint, not peer-reviewed.

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.

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.

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.

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.

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.

Ghosh, P. (02/16/2019). “AAAS: Machine learning ‘causing science crisis’”. *BBC News*.

Goldberg, Y. (2017). *Neural network methods for natural language processing*. San Rafael: Morgan & Claypool.

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.

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.

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.

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.

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.

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

Lapuschkin, S., A. Binder, G. Montavon, K.-R. Müller, and W. Samek (2016). "Analyzing classifiers: Fisher vectors and deep neural networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2912–2920.

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

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.

– (2018). *Von Wörtern und Bäumen* 2.10.

– (2019a). "Automatic detection of borrowing (Open problems in computational diversity linguistics 2)". *The Genealogical World of Phylogenetic Networks* 6.2.

– (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.

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., 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.

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.

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.

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.

Round, E. R. (2017). *The AusPhon-Lexicon project: 2 million normalized segments across 300 Australian languages*. Paper, presented at the conference “Poznań Linguistic Meeting” (Poznań).

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.

– (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. Shnirelʹman. Moscow: Institut Vostokovedenija, 3–39.

“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: 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.

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. eprint: `/oup/backfile/content_public/journal/comjnl/10/1/10.1093/comjnl/10.1.85/2/100085.pdf`.

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.