Annual contributions to the Genealogical World of Phylogenetic Networks

Johann-Mattis List
mattis.list@shh.mpg.de

Department of Linguistic and Cultural Evolution
Max Planck Institute for the Science of Human History, Jena

2018
Contents

Networks of pronunciation glosses in Traditional Chinese phonology ............ 1
Tossing coins: linguistic phylogenies and extensive synonymy .................. 7
It’s the system, stupid! More thoughts on sound change in language history .... 14
Stratification: how linguists traditionally identify borrowings .................... 19
Comparing reconstruction systems in historical linguistics ...................... 23
Horizontal and vertical language comparison ........................................... 28
Networks of polysemous and homophonous words .................................... 34
Regular cognates: A new term for homology relations in linguistics ............ 37
Structural data in historical linguistics .................................................... 41
Controversies about structural data in historical linguistics ...................... 45
How languages lose body parts: once more about structural data in historical linguistics (with G. Grimm) .................................................. 48
Patterns, processes, abduction, and consilience ....................................... 63
Every person who has learned how to read in any language will at some point have to deal with the question of how to pronounce words with unusual spellings. The English writing system offers an abundance of examples; and in my own pronunciation practice of English, I am still frequently corrected by native speakers when mispronouncing words that I know only from books.

For example, what constitutes a big problem for me is the stress on words of Latin origin, which have different stress patterns in German, my native tongue. While we speak of a Theo·rem in German, stressing the final syllable, English pronounces the word as the·orem, stressing the first. But my problems with the English writing system (and probably the problems of many other non-native and native speakers as well) do not usually end with the placement of stress, but may often go much deeper.

**Pronunciation**

Determining how an unknown word is pronounced in English is very easy nowadays. One can just use one of the numerous online dictionaries, where pronunciations are given in form of sound files. Another, more old-fashioned, alternative is to consult a classical dictionary that illustrate pronunciation with help of the International Phonetic Alphabet (IPA 1999). As linguists, we use it on a regular basis in order to compare pronunciations of words across different languages and language families. The original purpose of the IPA, however, was essentially the correct pronunciation for first and second language acquisition; and many teachers were involved in its creation in the late 19th century (compare Kalusky 2017).

Even earlier than the standardization efforts by the International Phonetic Association are ad hoc practices of glossing the pronunciation of difficult words, by comparing them with the pronunciation of more common words in the same language. In order to explain the pronunciation of the English word *digest*, for example, we could say that word is pronounced as *die* in *dead* and as *gest* in *adventure*. In the English context, it seems furthermore to be common to make use of some very basic syllables that most people will read and pronounce unambiguously, like *ah* for the *a* we find in *abacus* as opposed the the *a* we find in *and*, or *toe* for the "normal" *o*-sound we find in *no* as opposed to the sound of the *o* in words like *to*.

That these ad hoc systems, which humans use to gloss pronunciations in writing, are not very reliable can be easily understood when recalling that writing systems have often grown over
centuries, reflecting different layers of pronunciation practices applied to words that were imported into the languages at different stages in history. English is, of course, an extremely messy case, but even writing systems like German or Russian, of which speakers would say that the pronunciation is close to the spelling, are far from reaching the explicitness of the International Phonetic Alphabet.

**Chinese pronunciation**

A particularly interesting case concerns historical glossing practices in the history of Chinese. As I mentioned in an earlier blogpost on networks in Chinese poetry, the Chinese writing system gives only minimal hints regarding the pronunciation of its characters. A character like 手 "hand", which is pronounced as shǒu (or [ʂɔu²¹⁴] in the IPA), does not tell us anything about its pronunciation; and even its meaning is difficult to derive from its modern written form.

Chinese scholars became aware of the problem rather early, around the 1st century AD when they tried to read the ancient texts produced by their intellectual and poetic masters some 500 years before. In order to make sure that the pronunciation of infrequent characters would not be forgotten over time, they developed different ways to gloss character pronunciations in a more or less systematic manner.

The ancient Chinese scholars didn't have an alphabet to simply transcribe their sounds — intensive contact with Indian phoneticians started much later. So, they started from simple equations, according to which one character was pronounced similarly to another character.

For example, the Shuōwén Jiězì (Explaining Simple and Complex Characters) is an early Chinese character dictionary by the famous scholar Xǔ Shèn (58-148 AD), which was published in 121 AD. In it, the author occasionally uses the formula "read [this character] as X" (in Chinese 读若 dúruò X), in addition to his explanations of the meanings and the structure of the characters. The disadvantage of this duruo method, as linguists often call it (Coblin 1983), is that it only allows glossing of characters for which a simple character with an identical pronunciation exists. It is also not clear whether the formula consistently points to strictly identical pronunciations or whether certain deviations are allowed.

In order to overcome these problems, much more precise ways of glossing character pronunciations were developed from about the 2nd century AD. One of the most interesting glossing systems in this context is the so-called fǎnqiè spellings (Coblin 1983, Branner 2000). This spelling method, which seems to go back to at least the third century, is based on breaking the character pronunciation into two parts, the *initial* and the *final*, and selecting one character for glossing each of the two parts — one with an identical initial sound and one with the identical final. If we applied this method to English, we could think of explaining the pronunciation of *rice* as *rye-nice*, with *rye* pointing to the initial sound *r* and *nice* pointing to the final of the word.
In the following figure, I have tried to illustrate how both methods (the dúruò and the fǎnqiè method) are applied in concrete examples of text.

森木多貌。從林從木。讀若曾參之參。

The character 森 (sēn «forest») denotes many trees. It consists of the characters 林 (lín «forest») and 木 (mù «wood»). To read as the character 參 in 曾參 zēngcēn.

(a) dúruò glossing method

东 德 红 反/切
dōng (tuwng) dé (tok) hóng (huwng) fān / qiè

The character 东 «east» is read as t[ok]-[h]uwng.

(b) fǎnqiè glossing method

Given their straightforwardness and simplicity, fǎnqiè pronunciation glosses became quite popular among Chinese scholars. Even today, people may occasionally use them in order to explain pronunciations without having to rely on foreign writing systems, like the Latin alphabet. As a result, there is an abundance of sources that use this pronunciation device throughout the history of the Chinese language. Although the pronunciation is only given indirectly, with respect to the pronunciation traditions that were active during a given epoch, the fǎnqiè spellings offer great help to explore how the pronunciation of the Chinese language changed over time.

Pronunciation networks

Most of this research on the usage of fǎnqiè spellings has been carried out manually. The first work on fǎnqiè spelling goes back to the early 19th century, when scholars like Chén Lǐ (1818-1882) began to investigate systematically which characters were used to denote certain initial sounds (in Chinese, these are called the upper fǎnqiè characters, fǎnqiè shàngzì 反切上字), and which characters were used to denote the finals (called the lower fǎnqiè characters, fǎnqiè xiàzì 反切下字).

As we might expect, instead of using the same character for the pronunciation of the initial sound all the time, scholars would often alternate the characters, but the alternations were more or less consistent, with some characters being used more frequently and some characters being used less
frequently. Scholars like Chén Lǐ figured out that the characters could be classified in a rather rigorous manner which would allow us to reconstruct direct pronunciations of the fānqiè spellings. For example, based on the spellings reported in the Qièyùn, an early rhyme book published in 601 AD, we can say that the characters ㄍ, ㄕ, ㄖ, etc. were regularly used to indicate initials that would be spelled as [k] in the International Phonetic Alphabet, while ㄢ, ㄣ, and ㄩ were used to pronounce [kʰ] (a k with strong aspiration).

What I find even more interesting and important than these concrete findings, is that Chinese scholars inherently employed rudimentary network thinking to arrive at their clusters (Gēng 2004). The system of glossed character and glossing character can be easily translated into a system of directed networks, in which we draw a link from the glossing character to the glossed character.

For a talk held earlier during the last year (List 2017), I constructed such a network from the Guǎngyùn (ca. 1000 AD), a later edition of the aforementioned rhymebook Qièyùn, which gives fānqiè spellings for more than 20,000 characters. In this network, I concentrated only on the initials, that is, the initial consonants of the language encoded in the source, and constructed a network of all internal relations among the glossing characters. The full network is shown in the following figure.

Eyeballing the network, we can see that the system does look rather systematic. The network is not connected and, apart from a few large connected components, we find a lot of discrete groups that seem to reflect individual initial sounds that were clearly distinguished from other sounds in the fānqiè spellings.

The following figure shows a part of the network, namely the second cluster in the big network (above) when going from left to right and staying at the top. In this figure, we can see that the
network has two highly connected source characters linking to almost all of the other characters.

I have to admit that I am still having trouble interpreting the network satisfactorily, let alone designing more complex methods to analyse it. Nevertheless, I have the hope that the network analysis of Chinese pronunciation glosses can give us new insights into the phonetic history of Chinese. Importantly, the structures reflected by the network are true pronunciation differences, and that we can indeed find concrete sounds in the indirect fǎnqiè spelling system, becomes specifically clear when comparing the reconstructed pronunciations of the characters in the sample with each other.

For example, when you look at the figure below, you can see that our connected component represents two different clusters of initials, namely a simple \( k \) and an aspirated \( k^h \). The node that links the two groups is given the pronunciation \( k^h \) in our example, but its original reading is ambiguous. The character has two readings and two meanings reflecting both ancient \( k \) and ancient \( k^h \) (today pronounced as jiē «Chinese pistachio tree» and kǎi «template», respectively).
Networks of pronunciation glosses in Chinese Traditional Phonology are still under-explored, both with respect to traditional scholarship and with respect to the way they are best handled and analyzed in modern network approaches. If we could develop an approach that would infer the clusters of glosses that point consistently to the same sound, they could give us fascinating insights, not only into the phonological system of Chinese varieties spoken during a given time period, but perhaps also into the dynamics underlying pronunciation changes, when comparing different networks across different times and places.

References

List, J.-M. (2017) Network approaches to the reconstruction of Old Chinese phonology. Talk, held at the "Center for Chinese Linguistics" (2017/03/07, Hong Kong, The Hong Kong University of Science and Technology).

Tossing coins: linguistic phylogenies and extensive synonymy

Johann-Mattis List

Max-Planck Institute for the Science of Human History, Jena

The procedures by which linguists sample data when carrying out phylogenetic analyses of languages are sometimes fundamentally different from the methods applied in biology. This is particularly obvious in the matter of the sampling of data for analysis, which I will discuss in this post.

Sampling data in historical linguistics

The reason for the difference is straightforward: while biologists can now sample whole genomes and search across those genomes for shared word families, linguists cannot sample the whole lexicon of several languages. The problem is not that we could not apply cognate detection methods to whole dictionaries. In fact there are recent attempts that try to do exactly this (Arnaud et al. 2017). The problem is that we simply do not know exactly how many words we can find in any given language.

For example, the Duden, a large lexicon of the German language, for example, recently added 5000 more words, mostly due to recent technological innovations, which then lead to new words which we frequently use in German, such as twittern "to tweet", Tablet "tablet computer", or Drohnenangriff "drone attack". In total, it now lists 145,000 words, and the majority of these words has been coined in complex processes involving language-internal derivation of new word forms, but also by a large amount of borrowing, as one can see from the three examples.

One could argue that we should only sample those words which most of the speakers in a given language know, but even there we are far from being able to provide reliable statistics, not to speak of the fact that it is also possible that these numbers vary greatly across different language families and cultural and sociolinguistic backgrounds. Brysbaert et al. (2016), for example, 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.

But in order to count as "near-native" in a certain language, including the ability to pursue studies at a university, the Common European Framework of Reference for Languages, requires only between 4000 and 5000 words (Milton 2010, see also List et al. 2016). How many word families this includes is not clear, and may, again, depend directly on the target language.

Lexicostatistics

When Morris Swadesh (1909-1967) established the discipline of lexicostatistics, which represented the first attempt to approach the problems we face in historical linguistics with the help of quantitative methods. He started from a sample of 215 concepts (Swadesh 1950), which he later reduced to only 100 (Swadesh 1955), because he was afraid that some concepts would often be denoted by words that are borrowed, or that would simply not be expressed by single words in certain language families. Since then, linguists have been trying to refine this list further, either by modifying it (Starostin 1991...
added 10 more concepts to Swadesh's list of 100 concepts), or by reducing it even further (Holman et al. 2008 reduced the list to 40 concepts).

While it is not essential how many concepts we use in the end, it is important to understand that we do not start simply by comparing words in our current phylogenetic approaches, but instead we sample parts of the lexicon of our languages with the help of a list of comparative concepts (Haspelmath 2010), which we then consecutively translate into the target languages. This sampling procedure was not necessarily invented by Morris Swadesh, but he was first to establish its broader use, and we have directly inherited this procedure of sampling when applying our phylogenetic methods (see this earlier post for details on lexicostatistics).

**Synonymy in linguistic datasets**

Having inherited the procedure, we have also inherited its problems, and, unfortunately, there are many problems involved with this sampling procedure. Not only do we have difficulties determining a universal diagnostic test list that could be applied to all languages, we also have considerable problems in standardizing the procedure of translating a comparative concept into the target languages, especially when the concepts are only loosely defined. The concept "to kill", for example, seems to be a rather straightforward example at first sight. In German, however, we have two words that could express this meaning equally well: töten (cognate with English dead) and umbringen (partially cognate with English to bring). In fact, as with all languages in the world, there are many more words for "to kill" in German, but these can easily be filtered out, as they usually are euphemisms, such as eliminieren "to eliminate", or neutralisieren "to neutralize". The words töten and umbringen, however, are extremely difficult to distinguish with respect to their meaning, and speakers often use them interchangeably, depending, perhaps, on register (töten being more formal). But even for me as a native speaker of German, it is incredibly difficult to tell when I use which word.

One solution to making a decision as to which of the words is more basic could be corpus studies. By counting how often and in which situations one term is used in a large corpus of German speech, we might be able to determine which of the two words comes closer to the concept "to kill" (see Starostin 2013 for a very elegant example for the problem of words for "dog" in Chinese). But in most cases where we compile lists of languages, we do not have the necessary corpora.

Furthermore, since corpus studies on competing forms for a given concept are extremely rare in linguistics, we cannot exclude the possibility that the frequency of two words expressing the same concept is in the end the same, and the words just represent a state of equilibrium in which speakers use them interchangeably. Whether we like it or not, we have to accept that there is no general principle to avoid these cases of synonymy when compiling our datasets for phylogenetic analyses.

**Tossing coins**

What should linguists do in such a situation, when they are about to compile the dataset that they want to analyze with the modern phylogenetic methods, in order to reconstruct some eye-catching phylogenetic trees? In the early days of lexicostatistics, scholars recommended being very strict, demanding that only one word in a given language should represent one comparative concept. In cases like German töten and umbringen, they recommended to toss a coin (Gudschnisky 1956), in order to
guarantee that the procedure was as objective as possible.

Later on, scholars relaxed the criteria, and just accepted that in a few — hopefully very few — cases there would be more than one word representing a comparative concept in a given language. This principle has not changed with the quantitative turn in historical linguistics. In fact, thanks to the procedure by which cognate sets across concept slots are dichotomized in a second step, scholars who only care for the phylogenetic analyses and not for the real data may easily overlook that the Nexus file from which they try to infer the ancestry of a given language family may list a large amount of synonyms, where the classical scholars simply did not know how to translate one of their diagnostic concepts into the target languages.

**Testing the impact of synonymy on phylogenetic reconstruction**

The obvious question to ask at this stage is: does this actually matter? Can't we just ignore it and trust that our phylogenetic approaches are sophisticated enough to find the major signals in the data, so that we can just ignore the problem of synonymy in linguistic datasets? In an early study, almost 10 years ago, when I was still a greenhorn in computing, I made an initial study of the problem of extensive synonymy, but it never made it into a publication, since we had to shorten our more general study, of which the synonymy test was only a small part. This study has been online since 2010 (Geisler and List 2010), but is still awaiting publication; and instead of including my quantitative test on the impact of extensive synonymy on phylogenetic reconstruction, we just mentioned the problem briefly.

Given that the problem of extensive synonymy turned up frequently in recent discussions with colleagues working on phylogenetic reconstruction in linguistics, I decided that I should finally close this chapter of my life, and resume the analyses that had been sleeping in my computer for the last 10 years.

The approach is very straightforward. If we want to test whether the choice of translations leaves traces in phylogenetic analyses, we can just take the pioneers of lexicostatistics literally, and conduct a series of coin-tossing experiments. We start from a "normal" dataset that people use in phylogenetic studies. These datasets usually contain a certain amount of synonymy (not extremely many, but it is not surprising to find two, three, or even four translations in the datasets that have been analysed in the recent years). If we now have the computer toss a coin in each situation where only one word should be chosen, we can easily create a large sample of datasets each of which is synonym free. Analysing these datasets and comparing the resulting trees is again straightforward.

I wrote some Python code, based on our LingPy library for computational tasks in historical linguistics (List et al. 2017), and selected four datasets, which are publicly available, for my studies, namely: one Indo-European dataset (Dunn 2012), one Pama-Nyungan dataset (Australian languages, Bowern and Atkinson 2012), one Austronesian dataset (Greenhill et al. 2008), and one Austro-Asiatic dataset (Sidwell 2015). The following table lists some basic information about the number of concepts, languages, and the average synonymy, i.e., the average number of words that a concept expresses in the data.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Concepts</th>
<th>Languages</th>
<th>Average Synonymy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indo-European</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pama-Nyungan</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Austronesian</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Austro-Asiatic</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Concepts</th>
<th>Languages</th>
<th>Average Synonymy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indo-European</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pama-Nyungan</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Austronesian</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Austro-Asiatic</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
For each dataset, I made 1000 coin-tossing trials, in which I randomly picked only one word where more than one word would have been given as the translation of a given concept in a given language. I then computed a phylogeny of each newly created dataset with the help of the Neighbor-joining algorithm on the distance matrix of shared cognates (Saitou and Nei 1987). In order to compare the trees, I employed the general Robinson-Foulds distance, as implemented in LingPy by Taraka Rama. Since I did not have time to wait to compare all 1000 trees against each other (as this takes a long time when computing the analyses for four datasets), I randomly sampled 1000 tree pairs. It is, however, easy to repeat the results and compute the distances for all tree pairs exhaustively. The code and the data that I used can be found online at GitHub (github.com/lingpy/toss-a-coin).

Some results

As shown in the following table, where I added the averaged generalized Robinson-Foulds distances for the pairwise tree comparisons, it becomes obvious that — at least for distance-based phylogenetic calculations — the problem of extensive synonymy and choice of translational equivalents has an immediate impact on phylogenetic reconstruction. In fact, the average differences reported here are higher than the ones we find when comparing phylogenetic reconstruction based on automatic pipelines with phylogenetic reconstruction based on manual annotation (Jäger 2013).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Concepts</th>
<th>Languages</th>
<th>Synonymy</th>
<th>Average GRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austro-Asiatic</td>
<td>200</td>
<td>58</td>
<td>1.08</td>
<td>0.20</td>
</tr>
<tr>
<td>Austronesian</td>
<td>210</td>
<td>45</td>
<td>1.12</td>
<td>0.19</td>
</tr>
<tr>
<td>Indo-European</td>
<td>208</td>
<td>58</td>
<td>1.16</td>
<td>0.59</td>
</tr>
<tr>
<td>Pama-Nyungan</td>
<td>183</td>
<td>67</td>
<td>1.1</td>
<td>0.22</td>
</tr>
</tbody>
</table>

The most impressive example is for the Indo-European dataset, where we have an incredible average distance of 0.59. This result almost seems surreal, and at first I thought that it was my lazy sampling procedure that introduced the bias. But a second trial confirmed the distance (0.62), and when comparing each of the 1000 trial trees with the tree we receive when not excluding the synonyms, the distance is even slightly higher (0.64).

When looking at the consensus network of the 1000 trees (created with SplitsTree4, Huson et al. 2006), using no threshold (to make sure that the full variation could be traced), and the mean for the calculation of the branch lengths, which is shown below, we can see that the variation introduced by the synonyms is indeed real.
Notably, the Germanic languages are highly incompatible, followed by Slavic and Romance. In addition, we find quite a lot of variation in the root. Furthermore, when looking at the table below, which shows the ten languages that have the largest number of synonyms in the Indo-European data, we can see that most of them belong to the highly incompatible Germanic branch.

<table>
<thead>
<tr>
<th>Language</th>
<th>Subgroup</th>
<th>Synonymous Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLD_NORSE</td>
<td>Germanic</td>
<td>83</td>
</tr>
<tr>
<td>FAROESE</td>
<td>Germanic</td>
<td>77</td>
</tr>
<tr>
<td>SWEDISH</td>
<td>Germanic</td>
<td>68</td>
</tr>
<tr>
<td>OLD_SWEDISH</td>
<td>Germanic</td>
<td>65</td>
</tr>
<tr>
<td>ICELANDIC</td>
<td>Germanic</td>
<td>64</td>
</tr>
<tr>
<td>OLD_IRISH</td>
<td>Celtic</td>
<td>61</td>
</tr>
<tr>
<td>NORWEGIAN_RIKSMAL</td>
<td>Germanic</td>
<td>54</td>
</tr>
<tr>
<td>GUTNISH_LAU</td>
<td>Germanic</td>
<td>50</td>
</tr>
<tr>
<td>ORIYA</td>
<td>Indo-Aryan</td>
<td>50</td>
</tr>
<tr>
<td>ANCIENT_GREEK</td>
<td>Greek</td>
<td>46</td>
</tr>
</tbody>
</table>

**Conclusion**

This study should be taken with some due care, as it is a preliminary experiment, and I have only tested it on four datasets, using a rather rough procedure of sampling the distances. It is perfectly possible that Bayesian methods (as they are "traditionally" used for phylogenetic analyses in historical...
linguistics now) can deal with this problem much better than distance-based approaches. It is also clear that by sampling the trees in a more rigorous manner (eg. by setting a threshold to include only those splits which occur frequently enough), the network will look much more tree like.

However, even if it turns out that the results are exaggerating the situation due to some theoretical or practical errors in my experiment, I think that we can no longer ignore the impact that our data decisions have on the phylogenies we produce. I hope that this preliminary study can eventually lead to some fruitful discussions in our field that may help us to improve our standards of data annotation.

I should also make it clear that this is in part already happening. Our colleagues from Moscow State University (lead by George Starostin in the form of the Global Lexicostatistical Database project) try very hard to improve the procedure by which translational equivalents are selected for the languages they investigate. The same applies to colleagues from our department in Jena who are working on an ambitious database for the Indo-European languages.

In addition to linguists trying to improve the way they sample their data, however, I hope that our computational experts could also begin to take the problem of data sampling in historical linguistics more seriously. A phylogenetic analysis does not start with a Nexus file. Especially in historical linguistics, where we often have very detailed accounts of individual word histories (derived from our qualitative methods), we need to work harder to integrate software solutions and qualitative studies.

References


---

It's the system, stupid! More thoughts on sound change in language history

Johann-Mattis List

Max-Planck Institute for the Science of Human History, Jena

In various blog posts in the past I have tried to emphasize that sound change in linguistics is fundamentally different from the kind of change in phenotype / genotype that we encounter in biology. The most crucial difference is that sound sequences, i.e., our words or parts of the words we use when communicating, do not manifest as a physical substance but — as linguists say — "ephemerically'', i.e. by the air flow that comes out of the mouth of a speaker and is perceived as an acoustic signal by the listener. This is in strong contrast to DNA sequences, for example, which are undeniably somewhere "out there''. They can be sliced, investigated, and they preserve information for centuries if not millenia, as the recent boom in archaeogenetics illustrates.

Here, I explore the consequences of this difference in a bit more detail.

Language as an activity

Language, as Wilhelm von Humboldt (1767-1835) — the boring linguist who investigated languages from his armchair while his brother Alexander was traveling the world — put it, is an activity (energeia). If we utter sentences, we pursue this activity and produce sample output of the system hidden in our heads. Since the sound signal is only determined by the capacity of our mouth to produce certain sounds, and the capacity of our brain to parse the signals we hear, we find a much stronger variation in the different sounds available in the languages of the world than we find when comparing the alphabets underlying DNA or protein sequences.

Despite the large variation in the sound systems of the world's languages, it is clear that there are striking common tendencies. A language without vowels does not make much sense, as we would have problems pronouncing the words or perceiving them at longer distances. A language without consonants would also be problematic; and even artificial communication systems developed for long-distance communication, like the different kinds of yodeling practiced in different parts of the world, make use of consonants to allow for a clearer distinction between vowels (see the page about Yodeling on Wikipedia). But, between both extremes we find great variation in the languages of the world, and this does not seem to follow any specific pattern that could point to any kind of selective pressure, although scholars have repeatedly tried to demonstrate it (see Everett et al. 2015 and the follow-up by Roberts 2018).

What is also important here is that, not only is the number of the sounds we find in the sound system of a given language highly variable, but there is also variation in the rules by which sounds can be concatenated to form words (called the phonotactics of a language), along with the frequency of the sounds in the words of different languages. Some languages tolerate clusters of multiple consonants (compare Russian vzroslye or German Herbst), others refuse them (compare the Chinese name for Frankfurt: fàlánkèfù), yet others allow words to end in voiced stops (compare English job in standard pronunciation), and some turn voiced stops into voiceless ones (compare the standard pronunciation of Job in German as jop).
Language as a system

Language is a system which essentially concatenates a fixed number of sounds to sequences, being only restricted by the encoding and decoding capacities of its users. This is the core reason why sound change is so different from change in biological characters. If we say that German d goes back to Proto-Germanic *θ (pronounced as th in path), this does not mean that there were a couple of mutations in a couple of words of the German language. Instead it means that the system which produced the words for Proto-Germanic changed the way in which the sound *θ was produced in the original system.

In some sense, we can think metaphorically of a typewriter, in which we replace a letter by another one. As a result, whenever we want to type a given word in the way we know it, we will type it with the new letter instead. But this analogy would be too restricted, as we can also add new letters to the typewriter, or remove existing ones. We can also split one letter key into two, as happens in the case of palatalization, which is a very common type of sound change during which sounds like [k] or [g] turn into sounds like [ʧ] and [ʤ] when being followed by front vowels (compare Italian cento "hundred", which was pronounced [kɛntum] in Latin and is now pronounced as [ʧɛnto]).

Sound change is not the same as mutation in biology

Since it is the sound system that changes during the process we call sound change, and not the words (which are just a reflection of the output of the system), we cannot equate sound change with mutations in biological sequences, since mutations do not recur across all sequences in a genome, replacing one DNA segment by another one, which may not even have existed before. The change in the system, as opposed to the sequences that the system produces, is the reason for the apparent regularity of sound change.

This culminates in Leonard Bloomfield's (1887-1949) famous (at least among old-school linguists) expression that 'phonemes [i.e., the minimal distinctive units of language] change' (Bloomfield 1933: 351). From the perspective of formal approaches to sequence comparison, we could restate this as: 'alphabets change'. Hruschka et al. (2015) have compared sound change with concerted evolution in biology. We can state the analogy in simpler terms: sound change reflects systemics in language history, and concerted evolution results from systemic changes in biological evolution. It's the system, stupid!

Given that sound systems change in language history, this means that the problem of character alignments (i.e. determining homology/cognacy) in linguistics cannot be directly solved with the same techniques that are used in biology, where the alphabets are assumed to be constant, and alignments are supposed to identify mutations alone. If we want to compare sequences in linguistics, where we have to compare sequences that were basically drawn from different alphabets, this means that we need to find out which sounds correspond to which sounds across different languages while at the same time trying to align them.
An artificial example for the systemic grounding of sound change

Let me provide a concrete artificial example, to illustrate the peculiarities of sound change. Imagine two people who originally spoke the same language, but then suffered from diseases or accidents that inhibited them from producing their speech in the way they did before. Let the first person suffer from a cold, which blocks the nose, and therefore turns all nasal sounds into their corresponding voiced stops, i.e., \( n \) becomes a \( d \), \( ng \) becomes a \( g \), and \( m \) becomes a \( b \). Let the other person suffer from the loss of the front teeth, which makes it difficult to pronounce the sounds \( s \) and \( z \) correctly, so that they sound like a \( th \) (in its voiced and voiceless form, like in \textit{thing} vs. \textit{that}).

\[
\begin{array}{lll}
m & \rightarrow & b \\
n & \rightarrow & d \\
ng & \rightarrow & g \\
ss & \rightarrow & f \\
s & \rightarrow & w \\
\end{array}
\]

Artificial sound change resulting from a cold or the loss of the front teeth.

If we now let both persons pronounce the same words in their original language, they won’t sound very similar anymore, as I have tried to depict in the following table (\( dh \) points to the \( th \) in words like \textit{father}, as opposed to the voiceless \( th \) in words like \textit{thatch}).

<table>
<thead>
<tr>
<th>No.</th>
<th>Speaker Cold</th>
<th>Speaker Tooth</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>bass</td>
<td>math</td>
</tr>
<tr>
<td>2</td>
<td>buzic</td>
<td>mudhic</td>
</tr>
<tr>
<td>3</td>
<td>dose</td>
<td>nothe</td>
</tr>
<tr>
<td>4</td>
<td>boizy</td>
<td>moidehy</td>
</tr>
<tr>
<td>5</td>
<td>sig</td>
<td>thing</td>
</tr>
<tr>
<td>6</td>
<td>rizig</td>
<td>ridhing</td>
</tr>
</tbody>
</table>

By comparing the words systematically, however, bearing in mind that we need to find the best alignment and the mapping between the alphabets, we can retrieve a set of what linguists call \textit{sound correspondences}. We can see that the \( s \) of speaker \textit{Cold} corresponds to the \( th \) of speaker \textit{Tooth}, \( z \) corresponds to \( dh \), \( b \) to \( m \), \( d \) to \( n \), and \( g \) to \( ng \). Having probably figured out by now that my words were taken from the English language (spelling voiced \( s \) consequently as \( z \)), it is easy even to come up with a reconstruction of the original words (\textit{mass}, \textit{music}=[\textit{muzik}], \textit{nose}, \textit{noisy}=[\textit{noizy}], etc.).
Reconstructing ancestral sounds in our artificial example with help of regular sound correspondences.

Summary

Systemic changes are difficult to handle in phylogenetic analyses. They leave specific traces in the evolving objects we investigate that are often difficult to interpret. While it has been long since known to linguists that sound change is an inherently systemic phenomenon, it is still very difficult to communicate to non-linguistics what this means, and why it is so difficult for us to compare languages by comparing their words. Although it may seem tempting to compare languages with simple sequence-alignment algorithms with differences in biological sequences resulting from mutations (see for example Wheeler and Whiteley 2015), it is basically an oversimplifying approach.

Simple models undeniably have their merits, especially when dealing with big datasets that are difficult to inspect manually — there is nothing to say against their use. But we should always keep in mind that we can, and should, do much better than this. Handling systemic changes remains a major challenge for phylogenetic approaches, no matter whether they use trees, networks, bushes, or forests.

Given the peculiarity of sound change in linguistic evolution, and how well the phenomena are understood in our discipline, it seems worthwhile to invest time in exploring ways to formalize and model the process. During the past two decades, linguists have taken a lot of inspiration from biology. The time will come when we need to pay something back. Providing models and analyses to deal with systemic processes like sound change might be a good start.

References


changes in linguistics as events of concerted evolution. *Curr. Biol.* 25.1: 1-9

In my previous blog post, I illustrated how important it is to take the systemic aspects of sound change into account when comparing languages. What surfaces as a surprisingly regular process is in fact a process during which the sound system of a language changes. Since the words in a given language are derived from the sound system, a change in the system will necessarily change all words in which the respective sound occurs.

On one hand, this makes it much more difficult for linguists to identify homologous words across languages. On the other hand, however, it enables us to identify borrowings, by searching for exceptions to regular sound correspondences. I will be discussing the latter here.

Sound changes and borrowing

In order to illustrate how this can be done in practice, consider the examples of 15 cognates between German and English in the following table:

<table>
<thead>
<tr>
<th>No.</th>
<th>German</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Dach</td>
<td>thatch</td>
</tr>
<tr>
<td>2</td>
<td>Daumen</td>
<td>thumb</td>
</tr>
<tr>
<td>3</td>
<td>Degen</td>
<td>thane</td>
</tr>
<tr>
<td>4</td>
<td>Ding</td>
<td>thing</td>
</tr>
<tr>
<td>5</td>
<td>drei</td>
<td>three</td>
</tr>
<tr>
<td>6</td>
<td>Durst</td>
<td>thirst</td>
</tr>
<tr>
<td>7</td>
<td>denken</td>
<td>think</td>
</tr>
<tr>
<td>8</td>
<td>Dieb</td>
<td>thief</td>
</tr>
<tr>
<td>9</td>
<td>dreschen</td>
<td>thresh</td>
</tr>
<tr>
<td>10</td>
<td>Drossel</td>
<td>throat</td>
</tr>
</tbody>
</table>

When comparing these words quickly, it is easy to see that in all cases where German has a *d* as the initial sound, English has a *th*. This *sound correspondence*, as we call it in historical linguistics, reflects a very typical systematic similarity between English and German, which we can identify for all related words in English and German which go back to Proto-Germanic *θ*-, a very regular sound change which is well accounted for in Indo-European linguistics.

Not all homologous words between English and German, however, show this correspondences, as we can easily see from the five examples provided in the next table:
It is easy to see that these words don't fit our expected pattern ($d$ matching $th$ as the first consonant). It is also clear from the overall similarity of the words that it is rather unlikely that they trace back to different words, and thus turn out to be not cognate at all. One of the simplest possible explanations for the divergence from our initial $d$ in German corresponding to $\theta$ in English, which now surfaces as $d = d$, is borrowing, be it from German to English, from English to German, or from some third language.

Among the five examples, the final one, *Dollar* is the easiest to explain, as we are dealing with a recent borrowing of the name of the U.S. currency. English *dollar* itself has another cognate with German, namely German *Taler*, the name of a currency from ancient times (see [here](#) for the full etymology, based on Pfeifer 1993).

The other four terms in the table may seem less straightforward to explain as borrowings, as they are by no means of recent origin; but we can confirm their exceptional status by contrasting them with older Middle High German readings (11-14th century), which are listed in the following table for all 15 of our examples:

<table>
<thead>
<tr>
<th>No.</th>
<th>German</th>
<th>English</th>
<th>Middle High German</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Dach</td>
<td>thatch</td>
<td>dah</td>
</tr>
<tr>
<td>2</td>
<td>Daumen</td>
<td>thumb</td>
<td>düm</td>
</tr>
<tr>
<td>3</td>
<td>Degen</td>
<td>thane</td>
<td>degan</td>
</tr>
<tr>
<td>4</td>
<td>Ding</td>
<td>thing</td>
<td>ding</td>
</tr>
<tr>
<td>5</td>
<td>drei</td>
<td>three</td>
<td>drī</td>
</tr>
<tr>
<td>6</td>
<td>Durst</td>
<td>thirst</td>
<td>durst</td>
</tr>
<tr>
<td>7</td>
<td>denken</td>
<td>think</td>
<td>denken</td>
</tr>
<tr>
<td>8</td>
<td>Dieb</td>
<td>thief</td>
<td>diob</td>
</tr>
<tr>
<td>9</td>
<td>dreschen</td>
<td>thresh</td>
<td>dreskan</td>
</tr>
<tr>
<td>10</td>
<td>Drossel</td>
<td>throat</td>
<td>drozze</td>
</tr>
<tr>
<td>11</td>
<td>Dill</td>
<td>dill</td>
<td>tilli</td>
</tr>
<tr>
<td>12</td>
<td>dumm</td>
<td>dumb</td>
<td>tumb</td>
</tr>
<tr>
<td>13</td>
<td>Damm</td>
<td>dam</td>
<td>tam</td>
</tr>
<tr>
<td>14</td>
<td>Dunst</td>
<td>dunst</td>
<td>tunst</td>
</tr>
<tr>
<td>15</td>
<td>Dollar</td>
<td>dollar</td>
<td>—</td>
</tr>
</tbody>
</table>

As can be easily seen from this table, examples 11-14 all have a $t$ as the initial consonant in Middle High German, and not $d$, as in the other cases. The change from original Proto-Germanic $d$ to $t$ in German is a well-attested sound change, for which we have many examples in the form of sound correspondences (cf. *day* vs. *Tag*, *do* vs. *tun*, etc.). We can therefore conclude that the Middle High German readings like *tilli* vs. English *dill* reflect the readings we would expect if all words had
changed according to the rules. Since no regular change from $t$ in Middle High German to $d$ in Standard High German can be attested, it is furthermore safe to assume that the words have been modified under the influence of contact with other Germanic language varieties.

Here, English is not the most obvious candidate for contact; and the influence is rather due to contact with neighboring language varieties in the North-West of Germany, such as Frisian or Dutch. Similar to English, they have retained the original $d$ (cf. Dutch *dille* vs. English *dill*). If speakers of High German varieties borrowed the term from speakers of Low German varieties, they would re-introduce the original $d$ into their language, as we can see in our examples 11-14.

Why some of these borrowings took place and some did not is hard to say. That people in the North-West, living on the coast, know more about the building of dams, for example, is probably a good explanation why High German borrowed the term: obviously, the High German speakers did not use the word *tam* all that frequently, but instead heard the word *dam* often in conversations with neighboring varieties closer to the coast. For the other words, however, it is difficult to tell what was the reason for the success of the alternative forms.

Conclusions

Despite its important role for historical language comparison, the kind of analysis described here, by which linguists infer exceptional patterns in order to identify borrowings, is not well documented, either in handbooks of historical linguistics or in the journal literature. Following Lee and Sagart (2008), it is probably best called stratification analysis, since linguists try to identify the layers of contact and inheritance which surface in the form of sound correspondences. If these layers are correctly identified, linguists can often not only determine the direction in which a borrowing occurred, but also the relative time window in which this borrowing must have happened. This is the reason why linguists can often give very detailed word histories, which show where a word was first borrowed and how it then traveled through linguistic landscapes.

As for so many methods in historical language comparison, it is difficult to identify a straightforward counterpart of this technique in biology. What probably comes closest is the usage of $GC$ content as a proxy for the inference of directed networks of lateral gene transfer (as described in, for example, Popa et al. 2011). In contrast to lateral gene transfer in biology, however, our linguistic word histories are often much more detailed, especially in those cases where we have well-documented languages.

For the future, I hope that increased efforts to formalize the process of cognate identification, cognate annotation, and phonetic alignments in computer-assisted frameworks to historical language comparison may help to improve the way we infer borrowings in linguistics. There are so many open questions about lateral word transfer in historical linguistics that we cannot answer by sifting manually through datasets. We will need all the support we can get from automatic and semi-automatic approaches, if we want to shed some light on the many mysterious non-vertical aspects of language evolution.

References


Comparing reconstruction systems in historical linguistics

Johann-Mattis List

Max-Planck Institute for the Science of Human History, Jena

The term *linguistic reconstruction* has a very specific meaning in historical linguistics, pointing usually to the techniques that are used in order to infer how a given language was originally pronounced, even though it has not been attested in written sources. In previous posts, I have occasionally pointed to reconstructed forms, the so-called *proto-forms*, which linguists usually mark as such by putting an asterisk in from of them. For example, the word Indo-European *ph₂tér-* is a reconstructed proto-form for the supposed Indo-European word "father".

While the reconstruction techniques are usually limited to languages for which we have no written record, they can in principle also be applied in order to find out how ancient languages like, for example, Latin and Greek, were pronounced in detail (Sturtevant 1920). For languages like Chinese, whose writing system leaves almost no clues about pronunciation, linguistic reconstruction is the only way to investigate the pronunciation of the oldest stages of the language.

When dealing with different reconstruction systems for Old Chinese phonology, it is quite difficult, even for experienced scholars, to spot the actual differences between the systems. That these differences exist, and that they can be quite substantial, is beyond question — and easy to understand, if one takes into account that Old Chinese is reconstructed with the help of a philological (as opposed to a mainly comparative) approach, by which data from different sources is sifted and individually weighed (cf. Jarceva 1990: 409 and List 2008).

When comparing different reconstruction systems, it is not enough simply to look at the inventories of proto-phonemes proposed by different scholars. Even if two proto-inventories (the sets of the reconstructed sounds) are exactly the same, it is possible that scholars will provide different reconstructions for individual characters. The only way to compare two or more reconstruction systems is therefore to compare the concrete reconstructions for a certain number of characters.

In addition to the sample of words, however, we also need a clear account of which segments (which proto-sounds) should be compared with each other. When comparing proto-forms for Chinese *yī* ‘one’ in different Old Chinese reconstruction systems, such as Karlgren (1950) *ʔi̯ît*, Li (1971) *ʔjit*, Wáng (1980) *i̯et*, and Baxter and Sagart (2014) *ʔi[t], we would obviously not compare the medial *ʔi̯* of Karlgren with the initial *ʔ* of Baxter and Sagart.

When adding more reconstructions, such as the one for *qī* ‘seven’ across the four systems, for which the authors give *tsʰi̯ît*, *tsʰjit*, *tshiet*, and *[tsh]i[t], respectively, we can further see that there are not only differences for the different segments in the same positions, but also for the interpretation of the words. Although all authors give different medials, main vowels, and finals in the two words, they are structurally consistent in giving both words the same sound segments for medial, nucleus, and coda.
What we can see from this example is that any difference in the sound segments, like the choice of initials, or the concrete solution proposed for a problem, does not immediately reflect important differences in the reconstruction systems. If two scholars just choose another symbol for a distinction that they both recognize and acknowledge, this does not render the reconstructions incompatible. It should therefore not be used as a criterion for dismissing a given reconstruction system, at least not in a first step. If two systems are structurally equivalent, then they have equivalent predictive power for the descendant language(s) they are supposed to reconstruct.

This abstractionist notion of proto-forms, which can be found in the early work of Saussure (1916) and Meillet (1903), is problematic for the endeavour of linguistic reconstruction, and usually not strictly followed (Lass 2017). Nevertheless, the potentially abstract notion of proto-forms is important to be kept in mind when comparing different reconstruction systems. When distinguishing the structural differences (which result from the direct interpretation of the data and the identification of regular sound correspondences) from the substantial differences (resulting from a phonetic and phonological interpretation of the identified correspondences), we have a much clearer account of the core of the differences, and whether they are worth our consideration or not.

But how can we compare reconstruction systems structurally? Firstly, we need to have the data assembled in aligned form, in order to make sure that we only compare like with like (e.g., medial with medial, and vowel with vowel). A sample illustration in which alignments of the proto-forms for ‘seven’ and ‘one’, produced with the help of the EDICTOR tool (List 2017), is given in the figure below.

![Figure 1: Comparing reconstruction proposals with the help of alignments.](image)

Alternatively, we can also select a single aspect, such as, for example, the vowel system proposed in different reconstruction systems. Having assembled a substantial amount of different proto-forms in this way, the structural comparison between different reconstruction systems can be modeled as a comparison of different cluster analyses, or, more accurately, partitioning analyses. A partitioning analysis assigns a given number of objects to a certain number of different groups. When dealing only with the vowels proposed by different reconstruction systems, we can say that a given reconstruction, like the one by Karlgren, for example, assigns each Chinese character, for which a proto-form is given, to a particular group depending on the main vowel selected for the reconstruction.

If, for a given number of reconstructions, we model each reconstruction system as a partitioning analysis, based on the main vowel proposed by the system, we can use standard metrics from graph theory and Natural Language Processing to compare different reconstruction systems with each other. Very straight-forward measures for the comparison of two partitioning analyses are the so-called B-Cubed scores (Amigó et al. 2009), which have proven specifically useful for the evaluation of automatic cognate detection methods in historical linguistics, compared to a gold standard (Hauer and
Being an evaluation measure, B-Cubed scores come in the typical three flavors of precision, recall, and F-Score. Precision is similar to the notion of true positives, and recall is similar to true negatives. For the purpose of comparing reconstruction systems, only the F-score is needed, as it is a symmetric measure, and the notion of true positives and true negatives is meaningless, unless we decide that we blindly trust one of the given systems. As also for the scores for precision and recall, the F-score ranges between 0 and 1, with 1 indicating that the two partitioning analyses are identical.

In order to compare more than one reconstruction system, we can make use of techniques for exploratory data analysis (Morrison 2014); and the most straightforward way to do this, is, of course, to use the NeighborNet algorithm (Bryant and Moulton 2004), as provided by the SplitsTree package (Huson 1998).

In order to illustrate how data-display networks can be used to study differences among Old Chinese reconstruction systems, I designed a little experiment, based on data taken from (List et al. 2017b), who provide Old Chinese reconstructions for all rhyme words in the Shijing based on eight different reconstruction systems (Baxter and Sagart 2014, Karlgren 1950, Li 1971, Pān 2000, Schuessler 2007, Starostin 1989, Wáng 1980, Zhèngzhāng 2003).

In order to keep the analysis simple, I extracted only the different reconstructions of the main vowel for each character in each system, and carried out a pairwise comparison of all eight systems, computing the B-Cubed F-scores for each pair, omitting characters for which no reconstruction could be found in the data. These scores were then converted to a distance matrix, and fed to the NeighborNet algorithm (the source code can be downloaded here). The resulting network is provided in the figure below.

![Figure 2: NeighborNet reflecting the closeness of the different reconstruction systems](image)

As one can see, the data roughly clusters into three subgroups, namely Schuessler, Baxter and Sagart, and Starostin vs. Pān and Zhèngzhāng vs. Karlgren, Li, and Wáng. On a larger scale, we can divide the
data into all six-vowel systems versus the non-six-vowel systems (Karlgren, Wáng, Li). Given that Pān is a direct student of Zhèngzhāng, the closeness between their reconstruction systems is not surprising.

What may be surprising is the closeness of the Schuessler, Starostin, and Baxter and Sagart systems, given their notable differences with respect to the criterion of vowel purity tested by List et al. (2017b). Even if the network analysis cannot directly explain all of these differences in detail, it seems like a worthwhile enterprise, which should be further expanded by comparing not only the vowels, but fully aligned proto-forms.

Given the straightforwardness of the application, it seems also useful to test it on other language families where there is similar disagreement, as in the reconstruction of Old Chinese phonology.

References


Horizontal and vertical language comparison

Johann-Mattis List

Max-Planck Institute for the Science of Human History, Jena

In the traditional handbooks on historical language comparison, one can often find the claim that there are two fundamentally different, but equally important, means of linguistic reconstruction. One is usually called "external reconstruction" (or alternatively the "comparative method"), and one is called "internal reconstruction". If we think of sequence comparison in historical linguistics in the form of a table, in which concepts are arranged on the vertical axis, and different languages on the horizontal axis, we can look at the two different modes of language comparison (external vs. internal) as the horizontal and the vertical axes of the table. Horizontal language comparison refers to external reconstruction — scholars compare forms (not necessarily of the same meaning) across the horizontal axis, that is, across different languages. Internal language comparison is vertical — scholars search inside one and the same language for structures that allow to infer its older stages.

In past blog posts I have been talking a lot about horizontal / external language comparison, for which especially the notion of sound correspondences is crucial. But in the same way in which we use the evidence across languages to infer the past states of a given language family, we can make use of language-internal evidence to learn more about the history — not only of a given language, but also of a group of languages.

Vertical Language Comparison

A classical example of vertical or internal language comparison is the investigation of paradigms, that is, the inflection systems of the verbs or nouns in a given language. This, of course, makes sense only if the respective languages have verbal or nominal morphology, ie. if we find differences in the verb forms for the first, second, or third person singular or plural, or for the case system. The principle would not work in Chinese, although we have different means to compare languages without inflection vertically, as I'll illustrate below.

As a simplified example of internal reconstruction, consider the verbal paradigm of the verb esse "to be" in Latin:

<table>
<thead>
<tr>
<th>Person</th>
<th>Singular</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>first</td>
<td>sum</td>
<td>sumus</td>
</tr>
<tr>
<td>second</td>
<td>es</td>
<td>estis</td>
</tr>
<tr>
<td>third</td>
<td>est</td>
<td>sunt</td>
</tr>
</tbody>
</table>
If you try to memorize this pattern, you will quickly realize that it is not regular, and you will have difficulties to identify patterns that assist in memorizing the forms. A much more regular pattern would be the following:

<table>
<thead>
<tr>
<th>Person</th>
<th>Singular</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>first</td>
<td>es-um</td>
<td>es-umus</td>
</tr>
<tr>
<td>second</td>
<td>es-Ø</td>
<td>es-tis</td>
</tr>
<tr>
<td>third</td>
<td>es-t</td>
<td>es-unt</td>
</tr>
</tbody>
</table>

This pattern would still require us to memorize six different endings, but we could safely remember that the beginning of all forms is the same, and that there are six different endings, accounting for person and number at the same time (which is anyway typical for inflecting languages).

An alternative pattern that would be easier to remember is the following one:

<table>
<thead>
<tr>
<th>Person</th>
<th>Singular</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>first</td>
<td>es-um</td>
<td>s-umus</td>
</tr>
<tr>
<td>second</td>
<td>es-Ø</td>
<td>s-tis</td>
</tr>
<tr>
<td>third</td>
<td>es-t</td>
<td>s-unt</td>
</tr>
</tbody>
</table>

While it may seem that this pattern is slightly more complicated at first glance, it would still be more regular than the pattern we actually observe, and we would now have two different aspects expressing the meaning of the different forms: the alternation of the root *es-* vs. *s-* accounts for the singular-plural distinction, while the endings express again both number and person.

If we look at older stages of Latin, we can, indeed, find evidence for the first person singular, which was written *esom* in ancient documents (see Meier-Brügger 2002 for details on the reconstruction of this paradigm in Indo-European). If we look at other languages, like Sanskrit and Ancient Greek, we can further see that our alternation between *es-* and *s-* in the root (thus our last example) comes also much closer to the supposed ancient state, even if we don’t find complete evidence for this in Latin alone.

What we can see, however, is that the inspection of *alternating* forms of the same root can reveal ancient states of a language. The key assumption is that observed irregularities usually go back to formerly regular patterns.
Horizontal language comparison

The classical example for horizontal or external language comparison is the typical wordlists in which words with similar meanings across different languages are arranged in tabular form. I have mentioned before that it was in great part Morris Swadesh (1909-1967) who popularized the simple tabular perspective that puts a concept and its various translations in the center of historical language comparison. Before the development of this concept-based approach to historical linguistics, scholars would pick examples based on their similarity in form, allowing for great differences in the semantics of the words being assigned to the same slot of cognate words; and this exclusively form-based approach to external language comparison is still the prevalent one in most branches of historical linguistics.

No matter what approach we employ in this context — be it the concept- or the form-based — as long as we compare forms across different languages, we carry out external language comparison, and our main concern is then the identification of regular sound correspondences across the languages in our sample, which enable us to propose ancestral sounds for the ancestral language.

Problems of vertical language comparison

As can be seen from my above example of the inflection of esse in Latin, it is not obvious how the task of internal language comparison could be formalized and automated. There are two main reasons for this. First, inflection paradigms vary greatly among the languages of the world, which makes it difficult to come up with a common way to investigate them.

Second, since we are usually looking for irregular cases that we try to explain as having evolved from former regularities, it is clear that our data will be extremely sparse. Often, it is only the paradigm of one word that we seek to explain, as we have seen for Latin esse, and patterns of irregularities across many verbs are rather rare (although we can also find examples for this). As a result, internal reconstruction is dealing with even fewer data than external reconstruction, where data are also not necessarily big.

Formalizing the language-internal analysis of word families

Despite the obvious problems of exploiting the language-internal perspective in historical language comparison, there are certain types of linguistic analysis that are amenable to a more formal treatment in this area. One example that we are currently testing is the inference and annotation of word families within a given language. It is well known that large number of words in human languages are not unrelated atomic units, but have themselves been created from smaller parts. Linguists distinguish derivation and compounding as the major techniques here, by which new words are created from existing ones.
Derivation refers to those cases where a word is being modified by a form unit that could not form a word of its own, usually a suffix or a prefix. As an example, consider the suffix -er in English which can be attached to verbs in order to form a noun that usually describes the person that regularly carries out the action denoted by the original verb (e.g., examine → examiner, teach → teacher, etc.). While the original verb form exists without the suffix in the English language, the form -er only occurs as part of verbs. In contrast to derivation, compounding refers to the process by which two word forms that can be used in isolation are merged to form a new expression (compare foot and ball with football).

Searching for suffixes and compounds in unannotated language data is a very difficult task. Although scholars have been working on automatic methods that split a given monolingual dictionary into its smallest meaning-bearing form units (morphemes), these methods usually only work on very large datasets (Creutz and Laugs 2005). Trained linguists, on the other hand, can easily detect patterns, even when working on smaller datasets of a few hundred words.

The reason why linguists are successful in analysing the morphology of languages, in contrast to machine-learning approaches, is that they make active use of their external knowledge about the potential semantics underlying the patterns, while current methods for automatic morpheme detection usually only consider the forms, and disregard the semantics. Semantics, however, are important to distinguish words that form a true family (in that they share cognate material) from words that are similar only due to chance.

It is clear that languages may have words that sound alike but convey different meanings. As an extreme example, consider French paix [pɛ] "peace" vs. pet [pɛ] "fart". Although both words are pronounced the same, we know that they are not cognate, going back to different ancestral forms, as is also reflected in the French writing system. But even if we lacked the evidence of the French orthography, we could easily justify that the words do not form a family, since (a) their meaning is quite different, and (b) their genus is different as well (la paix vs. le pet). An automatic method that disregards semantics and external evidence (like the orthography or the gender of nouns in our case) cannot distinguish words that are similar due to chance from words that are similar due to their history.

As a further example illustrating the importance of semantics, consider the data for Achang, a Burmish language, spoken in Myanmar (data from Huáng 1992), which is shown in the following graphic (derived from the EDICTOR tool and analyzed by Nathan W. Hill).
In this figure, we can see six words which all share $tëh^{55}$ (high numbers represent tones) as their first part. As we can see from the detailed analysis of these compounds in Achang, which is given in the column "MORPHEMES" in the figure, our analysis claims that the form $tëh^{55}$, which expresses the concepts "foot" or "leg" in isolation, recurs in the words for "hoof", "claw", "knee", and "thigh", but not in the word for "ant". While the semantic commonalities among the former are plausible, as they all denote body parts which are closely related to "feet" or "legs", we do not find any transparent motivation for why the speakers should have used a compound containing the word for "foot" to denote an ant. Although we cannot demonstrate this at this point, we are hesitant to add the Achang word for "ant" to the word family based on compounds containing the word for "foot".

### Bipartite networks of word families

For the time being, we cannot automate this analysis, since we lack data for the testing and training of potential algorithms. We can, however, formalize it in a very straightforward way: with help of a bipartite network (see Hill and List 2017). Bipartite networks are networks with two kinds of nodes, which are usually thought of as representing different types. While we can easily assign different types to all nodes in any network we are dealing with, bipartite networks only allow us to link nodes of different types. In our bipartite network of word families, the first type of nodes represent the forms of the words, while the second type represent the meanings attributed to the sub-parts of the words. In the figure above, the former can be found in the column "tokens", where the symbol "+" marks the boundaries, and the latter can be found in the column "MORPHEMES".

The following figure shows the bipartite network underlying the word family relations following from our analysis of words built with the morpheme "foot" in Achang.
**Figure 2:** Bipartite network of word families: nodes in red text represent the (reconstructed) meaning of the morphemes, and blue nodes the words in which those occur as parts.

**Conclusion**

The bipartite network above shows only a small part of the word family structure of one language, and the analysis and formalization of word families with help of bipartite networks thus remains exemplary and anecdotal. I hope, however, that the example illustrates how important it is to keep in mind that language change is not only about sound shifts that can be analyzed with help of language-external, horizontal comparison. Investigating the vertical (the language-internal) perspective of language evolution is not only fascinating, offering many so far unresolved methodological problems, it is at least as important as the horizontal perspective for a proper understanding of the dynamics underlying language change.

**References**


Networks of polysemous and homophonous words

Johann-Mattis List

Max Planck Institute for the Science of Human History, Jena

When I was very young, maybe even before I went to school, we often played a game with my parents and grandparents, during which we had to select two homophonous words (that is, one word form that expresses two rather different meanings), and the other people had to guess which word we had selected. This game is slightly different from its Anglo-Saxon counterpart, the homophone game.

In Germany, this game is called Teekesselchen: "little teapot". Therefore, people now also use the word Teekesselchen to denote cases of homophonoy or very advanced polysemy. In this sense, the word Teekesselchen itself becomes polysemous, since it denotes both a little teacup, and the phenomenon that word forms in a given language may often denote multiple meanings.

Homophony and polysemy

In linguistics, we learn very early that we should rigorously distinguish the phenomenon of homophony from the phenomenon of polysemy. The former refers to originally different word forms that have become similar (and even identical) due to the effects of sound change — compare French paix "peace" and per "fart", which are now both pronounced as [pe]. The latter refers to cases where a word form has accumulated multiple meanings over time, which are shifted from the original meaning — compare head as in head of department vs. head as in headache.

Given the difference of the processes leading to homophony on the one hand and polysemy on the other, it may seem justified to opt for a strict usage of the terms, at least when discussing linguistic problems. However, the distinction between homophony and polysemy is not always that easy to make.

In German, for example, we have the same word Decke for "ceiling" and "blanket" (Geyken 2010). This may seem to reflect a homophony at first sight, given that the meanings are so different, so that it seems simpler to assume a coincidence. However, it is in fact a polysemy (cf. Pfeiffer 1993, s. v. «Decke»). This can be easily seen from the verb (be)decken "to cover", from which Decke was derived. While the ceiling covers the room, the blanket covers the body.

Given that we usually do not know much about the history of the words in our languages, we often have difficulties deciding whether we are dealing with homophones or with polysemy when encountering ambiguous terms in the languages of the world. The problem of the two terms is that they are not descriptive, but explanatory (or ontological): they do not only describe a phenomenon ("one word form is ambiguous, having multiple meanings"), but also the origin of this phenomenon (sound change or semantic change).

In this context, the recently coined term colexification (François 2008) has proven to be very helpful, as it is purely descriptive, referring to those cases where a given language has the same word form to express two or more different meanings. The advantage of descriptive terminology is that it allows us to identify a certain phenomenon but analyze it in a separate step — that is, we can already talk about...
the phenomenon before we have found out its specific explanation.

**A new contribution**

Having worked hard during recent years writing computer code for data curation and analysis (cf. List et al 2018a), my colleagues and I have finally managed to present the fascinating phenomena of colexifications (homophonies and polysemies) in the languages of the world in an interactive web application. This shows which colexifications occur frequently in which languages of the world.

In order to display how often the languages in the world express different concepts using the same word, we make use of a network model, in which the concepts (or meanings) are represented by the nodes in the networks, and links between concepts are drawn whenever we find that any of the languages in the sample *colexifies* the concepts. The following figure illustrates this idea.

![Figure 1: Colexification network for concepts centering around "FOOD" and "MEAL".](image)

This database and web application is called **CLICS**, which stands for the *Database of Cross-Linguistic Colexifications* (List et al. 2018b), and was published officially during the past week (http://clics.clld.org) — it can now be freely accessed by all who are interested. In addition, we describe the database in some more detail in a forthcoming article (List et al. 2018c), which is already available in form of a draft.

The data give us fascinating insights into the way in which the languages of the world describe the world. At times, it is surprising how similar the languages are, even if they do not share any recent ancestry. My favorite example is the network around the concept *FUR*, shown below. When inspecting this network, one can find direct links of *FUR* to *HAIR, BODY HAIR, and WOOL* on one hand, as well as *LEATHER, SKIN, BARK*, and *PEEL* on the other. In some sense, the many different languages of the world, whose data was used in this analysis, reflect a general principle of nature, namely that the bodies of living things are often covered by some protective substance.
Although we have been working with these networks for a long time, we are still far from understanding their true potential. Unfortunately, nobody in our team is a true specialist in complex networks. As a result, our approaches are always limited to what we may have read by chance about all of those fascinating ways in which complex networks can be analyzed.

For the future, we hope to convince more colleagues of the interesting character of the data. At the moment, our networks are simple tools for exploration, and it is hard to extract any evolutionary processes from them. With more refined methods, however, it may even be possible to use them to infer general tendencies of semantic change in language evolution.

References


Regular cognates: A new term for homology relations in linguistics

Johann-Mattis List

Max Planck Institute for the Science of Human History, Jena

The identification of homologous words between genealogically related languages is one of the crucial tasks in historical linguistics. In contrast to biology where, especially at the level of genetic sequences, we find a rather rich terminology contrasting different types of homology among genes and gene sequences, linguistic terminology is still not very precise. Most scholars seem to be content if they can claim that they have identified words that are cognate, which means that they are homologous but have not been borrowed throughout their history.

On various occasions in the past, I have tried to work on a more precise terminology for linguistic frameworks (see for example List 2014 and List 2016, or this earlier blogpost on homology in linguistics). In this context, I have often tried to emphasize that we need to be specifically more careful with the problem of partial cognacy in linguistics, since many words across related languages are not fully homologous, but show homology only in specific parts (List et al. 2016).

Thanks to an increase in accurately annotated linguistic data, resulting specifically from my very productive collaboration with Nathan W. Hill (SOAS, London) on the Burmish languages (see Hill and List 2017), my view has now again changed a bit, and I thought it would be useful to share it here.

Cognacy and homology

The starting point for my earlier proposals to refine the notion of cognacy in linguistics was the rather refined distinction between orthologs, paralogs, and xenologs in molecular biology (Fitch 2000). To account for the distinction between directly inherited (orthologs), duplicated (paralogs), and laterally transferred genes (xenologs), I proposed the terms direct cognates, indirect cognates (inspired by the term oblique cognates by Trask 2000), and indirectly etymologically related words or morphemes (word parts).

While the first and last term are more or less straightforward with respect to linguistic processes, the notion of indirect cognates, however, turned out to be insufficient, given that it is not clear which processes lead to indirect cognacy. Originally, I thought of morphological processes, that is, processes of word formation, by which a word is slightly modified to account for a slightly derived meaning (usually involving processes like suffixation or compounding). My idea was that words that have “experienced” these processes would behave similarly to genes that have been duplicated in biological evolution, and that it would be sufficient to just assign them to a common sub-class of cognates.

However, the research with Nathan W. Hill recently revealed that these terms are insufficient to capture the processes underlying lexical change in historical linguistics. In order to understand this idea, it is useful to get back to the biological terms and have a closer look at how they distinguish the underlying processes. As far as I understand it, a directly inherited gene sequence may differ from its ancestral sequence due to processes of random mutation, by which the original gene sequence becomes modified throughout its history. In cases of paralogy, the original gene sequence is duplicated...
and both copies are subsequently inherited. The copies may, during this process, become more different from each other than would be expected when assuming direct inheritance and random mutation. Similarly, in cases of lateral transfer of genetic material, the changes may again be different from the ones introduced by "normal" random mutation.

If we adopt the view of "normal change", as it is employed in the biological processes, we find a counterpart in the process of sound change in linguistics. As I have mentioned earlier, sound change is a systemic process by which certain sounds in certain environments change regularly across all words in the lexicon of a given language. This process is definitely not comparable with random mutation in sequence evolution, since the process involves a class of "letters" in the sound system of a language that are systematically turned into another sound. However, regarding the crucial role that sound change plays in language evolution, it seems that it is in some sense comparable with random mutation resulting in orthologous genes. Sound change is somewhat the baseline of what happens if languages change, and we have the means to identify its traces by searching for regular sound correspondence patterns across related languages (see my earlier blogpost on this matter).

That sound change is the default which can be handled with some confidence, while other processes, like word formation, semantic change, or the notorious process of analogical leveling, by which not only complex paradigms are transformed to reduce complexity, but other complexities can emerge (compare the German irregular plural of *Morgen-de* "mornings", which is built on the template of "evenings" *Abend-e*), is also the reason why Gévaudan (2007) does not include it into the major processes of lexical change. If we take sound change as the default process of language change and as our key evidence for homologous word relations, however, this means that we can no longer make the distinction between *direct* and *indirect* cognates following my earlier proposal, since indirect cognates do not necessarily reflect instances of irregular sound change.

This is in fact easy to illustrate. If we follow the former definition of indirect cognacy, the comparison of German *Handschuh* "glove" (lit. *hand-shoe*) with English *hand* would reflect indirect cognacy, since the German word is a compound of *Hand* "hand" and *Schuh"shoe", and thus a derived word form. The morpheme *Hand* in this example, however, is phonetically identical with German *Hand*, and the sound correspondences between the English word and the first element of the German compound are still regular by all means. In fact, only a small amount of word formation processes in language evolution also impact on the pronunciation of the base forms.

This means, in turn, that any distinction of cognate word forms (and word parts, i.e., morphemes) into direct and indirect ones that is based on the absence or presence of morphological (= word formation) processes, does not tell us much about the degree to which the sound change affecting these word forms was regular. We could state that direct cognates should always reflect regular sound change, since any irregularity would have to be accounted for by alternative explanations (e.g. shortening of a given word due to frequent use, assimilation of sounds serving the ease of pronunciation, etc.).

I wonder whether this would be useful for the initial idea behind the concept of direct cognacy. If we find direct cognates, that is, words that we assume were used by a couple of languages without further modification, apart from regular sound change and potentially sporadic sound changes, it seems still useful to assume that these reflect vertical language history better than cognate sets with residues that were exposed to various morphological processes. Thus, when coding direct cognacy in linguistic
datasets, sporadic sound change (if it can be illustrated properly) should not serve as an argument against direct cognacy.

The only way around this problem seems to be to establish a further shade of cognacy, which describes the relations among words and morphemes that have been only affected by sound change, in contrast to words whose history reflects various morphological derivations that impact directly on pronunciation, or processes of irregular sound change due to analogical leveling or assimilation. While I first thought that the biological term ortholog would be useful to describe these specific word relations in linguistics, I realized later that, judging from the Ancient Greek meaning of ortholog (ortho "straight, direct" + logos "relation"), the fact that differences are due to regular sound change is not that neatly reflected.

For now, I think that it should be sufficient to use the term regular cognates for those words or word parts for which we can demonstrate that their change was following the regular "laws" of sound change. Regular cognates are thus defined as words or word parts that have been affected only by sound change during their history. This notion deliberately excludes differences in meaning, frequency of use, or whether the word forms are only reflected in compounds or derived word forms. In fact, for some cases, we could even propose that only parts of a word form that no longer bear any meaning of their own (eg. the first two sounds of a word form) are regular cognates, as long as we can propose good arguments for the regularity of the correspondences.

Note that our tools for alignment analyses in historical linguistics already account for this property. The EDICTOR (http://edictor.digling.org, List 2017), a web-based tool for editing, analyzing, and publishing etymological dictionaries, allows users to exclude those parts from an alignment that are assumed to be irregular, as can be seen in the following illustrative alignment of Proto-Germanic *bakanan "to bake". Scholars who want to be explicit about what parts of an alignment they consider to be regular can use this annotation framework to provide more refined analyses.

<table>
<thead>
<tr>
<th>ID</th>
<th>CONCEPT</th>
<th>ALIGNMENT</th>
<th>IPA</th>
<th>COGID</th>
<th>DOCULECT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1431</td>
<td>*bakanan</td>
<td>b a k e</td>
<td>bake</td>
<td>391</td>
<td>Dutch</td>
</tr>
<tr>
<td>1430</td>
<td>*bakanan</td>
<td>b e t k</td>
<td>betk</td>
<td>391</td>
<td>English</td>
</tr>
<tr>
<td>1429</td>
<td>*bakanan</td>
<td>b a k e n</td>
<td>baken</td>
<td>391</td>
<td>German</td>
</tr>
<tr>
<td>1428</td>
<td>*bakanan</td>
<td>b a k k a n a n a n</td>
<td>bakanan</td>
<td>391</td>
<td>Proto-Germanic</td>
</tr>
</tbody>
</table>

**Figure 1:** EDICTOR alignment of regular cognates for Proto-Germanic *bakanan "to bake"

A crucial consequence of using only regularity in the sound correspondences as the criterion to distinguish regular from irregular cognates is that regular cognacy may also be found to hold for borrowings, since borrowings can, as well, be shown to be regular, especially when the language contact between languages was intensive. Identifying regular cognates is furthermore the first and most important step of the classical comparative method (Weiss 2015) for historical language comparison, since (unless we have written evidence for the true relations between languages) regular cognates (as proven by readily aligned cognate sets) are the fundament upon which we build all our hypotheses regarding the external history of languages.
References


Structural data in historical linguistics

Johann-Mattis List

Max-Planck Institute for the Science of Human History, Jena

The majority of historical linguists compare words to reconstruct the history of different languages. However, in phylogenetic studies focusing on cognate sets reflecting shared homologs across the languages under investigation, there exists another data type that people have been trying to explore in the past. The nature of this data type is difficult to understand for non-linguists, given that it has a very abstract nature. In the past, it has led to a considerable amount of confusion both among linguists and among non-linguists who tried to use this data for quick (and often also dirty) phylogenetic approaches. For this reason, I figured it would be useful to introduce this type of data in more detail.

This data type can be called "structural". To enable interested readers to experiment with the data themselves, this blogpost comes along with two example datasets that we converted into a computer-readable format (with much help from David), since the original papers only offered the data as PDF files. In future blogposts, we will try to illustrate how the data can, and should, be explored with network methods. In this first blogpost, I will try to explain the basic structure of the data.

Structural data in historical linguistics and language typology

In order to illustrate the type of data we are dealing with here, let's have a look at a typical dataset, compiled by the famous linguist Jerry Norman to illustrate differences between Chinese dialects (Norman 2003). The table below shows a part of the data provided by Norman.

<table>
<thead>
<tr>
<th>No.</th>
<th>Feature</th>
<th>Beijing</th>
<th>Suzhou</th>
<th>Meixian</th>
<th>Guangzhou</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The third person+ pronoun is tā, or cognate to it</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Velars palatalize+ before high-front vowels</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>The qu-tone+ lacks a register distinction</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>The word for+ &quot;stand&quot; is zhàn or cognate to it</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

In this example, the data is based on a questionnaire that provides specific questions; and for each of the languages in the sample, the dataset answers the question with either + or -. Many of these datasets are binary in their nature, but this is not a necessary condition, and questionnaires can also query categorical variables, such as, for example, the major type of word order might have three categories (subject-object-verb, subject-verb-object or other).
We can also see is that the questions can be very diverse. While we often use more or less standardized concept lists for lexical research (such as fixed lists of basic concepts, List et al. 2016), this kind of dataset is much less standardized, due to the nature of the questionnaire: asking for the translation of a concept is more or less straightforward, and the number of possible concepts that are useful for historical research is quite constrained. Asking a question about the structure of a language, however, be it phonological, lexical, based on attested sound changes, or on syntax, provides an incredible number of different possibilities. As a result, it seems that it is close to impossible to standardize these questions across different datasets.

Although scholars often call the data based on these questionnaires "grammatical" (since many questions are directed towards grammatical features, such as word order, presence or absence of articles, etc.), most datasets show a structure in which questions of phonology, lexicon, and grammar are mixed. For this reason, it is misleading to talk of "grammatical datasets", but instead the term "structural data" seems more adequate, since this is what the datasets were originally designed for: to investigate differences in the structure of different languages, as reflected in the most famous World Atlas of Language Structures (Dryer and Haspelmath 2013, https://wals.info).

**Too much freedom is a restriction**

In addition to mixed features that can be observed without knowing the history of the languages under investigation, many datasets (including the one by Norman we saw above) also use explicit "historical" (diachronic in linguistic terminology) questions in their questionnaires. In his paper describing the dataset, Norman defends this practice, as he argues that the goal of his study is to establish an historical classification of the Chinese dialects. With this goal in mind, it seems defensible to make use of historical knowledge and to include observed phenomena of language change in general, and sound change in specific, when compiling a structural dataset for group of related language varieties.

The problem of the extremely diverse nature of questionnaire items in structural datasets, however, makes their interpretation extremely difficult. This becomes especially evident when using the data in combination with computational methods for phylogenetic reconstruction. This is problematic for two major reasons:

1. Since questions are by nature less restricted regarding their content, scholars can easily pick and choose the features in such a way that they confirm the theory they want them to confirm rather than testing it objectively. Since scholars can select suitable features from a virtually unlimited array of possibilities, it is extremely difficult to guarantee the objectivity of a given feature collection.

2. If features are mixed, phylogenetic methods that work on explicit statistical models (like gain and loss of character states, etc.) may often be inadequate to model the evolution of the characters, especially if the characters are historical. While a feature like "the language has an article" may be interpreted as a gain-loss process (at some point, the language has no article, then it gains the article, then it looses it, etc.), features showing the results of processes, like "the words that originally started in [k] followed by a front vowel are now pronounced as [ʨ]", cannot be interpreted as a process, since the feature itself describes a process.

For these reasons, all phylogenetic studies that make use of structural data, in contrast to purely lexical datasets, should be taken with great care, not only because they tend to yield unreliable results, but
more importantly because they are extremely difficult to compare across different language families, given that they have way too much freedom when compiling them. Feature collections provided in structural datasets are an interesting resource for diversity linguistics, but they should not be used to make primary claims about external language history or subgrouping.

**Two structural datasets for Chinese dialects**

Before I start to bore the already small circle of readers interested in these topics, it seems better to stop discussing the usefulness of structural data at this point, and to introduce the two datasets that were promised at the beginning of the post.

Both datasets target Chinese dialect classification, the former being proposed by Norman (2003), and the latter reflecting a new data collection that was recently used by Szeto et al. (2018) to propose a North-South-split of dialects of Mandarin Chinese with help of a Neighbor-Net analysis (Bryant and Moulton 2004). Both datasets have been uploaded to Zenodo, and can be found in the newly established community collection cldf-datasets. The main idea of this collection is to collect various structural datasets that have been published in the literature in the past, and allow those people interested in the data, be it for replication studies or to test alternative approaches, easy access to the data in various formats.

The basic format is based on the format specifications laid out by the CLDF initiative (Forkel et al. 2018), which provides a software API, format specifications, and examples for best practice for both structural and lexical datasets in historical linguistics and language typology. The collection is curated on GitHub (cldf-datasets), and datasets are converted to CLDF (with all languages being linked to the Glottolog database, glottolog.org, Hammarström et al. 2018) and also to Nexus format. The dataset is versionized, it may be updated in the future, and interested readers can study the code used to generate the specific data format from the raw files, as well as the Nexus files, to learn how to submit their own datasets to our initiative.

**Final remarks on publishing structural datasets online**

By providing only two initial datasets for an enterprise whose general usefulness is highly questionable, readers might ask themselves why we are going through the pain of making data created by other people accessible through the web.

The truth is that the situation in historical linguistics and language typology has for a very long time been very unsatisfactory. Most of the research based on data did not supply the data with the paper, and often authors directly refuse to share the data when asked after publication (see also the post on Sharing supplementary data). In other cases, access to the data is exacerbated by providing data only in PDF format in tables inside the paper (or even worse: long tables in the supplement of a paper), which force scholars wishing to check a given analysis themselves to reverse-engineer the data from the PDF. That data is provided in a form difficult to access is not even necessarily the fault of the authors, since some journals even restrict the form of supplementary data to PDF only, giving authors wishing to share their data in an appropriate form a difficult time.

Many colleagues think that it is time to change this, and we can only change it by offering standard
ways to share our data. The CLDF along with the Nexus file, in which the two Chinese datasets are now published in this open repository collection, may hopefully serve as a starting point for larger collaboration among typologists and historical linguistics. Ideally, all people who publish papers that make use of structural datasets, would — similar to the practice in biology where scholars submit data to GenBank (Benson et al. 2013) — submit their data in CLDF format and Nexus, so that their colleagues can easily build on their results, and test them for potential errors.

References


Controversies about structural data in historical linguistics

Johann-Mattis List

Max Planck Institute for the Science of Human History, Jena

In the past, there have been many controversies about structural data, — that is, the kind of that data I introduce in the post written last month. Given the misinterpretation of structural data as being "grammatical", along with the unproven and misleading claim by Nichols (2003) that certain grammatical features are more stable than lexical ones, one can often read about a controversy in linguistics: which aspects are more stable, and therefore more useful to study deep linguistic relationships, the lexicon or the grammar?

In this context, it is often ignored that we are not talking chiefly about the grammar when applying phylogenetic studies to structural datasets. It is also ignored that the original idea of the importance of "grammar" was pointing to homologies in complex and concrete morphological paradigms, as has been most prominently discussed by Meillet (1925), later popularized by Nichols (1996) (i.e., individual word forms, that is: predominantly lexical traits). "Grammar" never pointed to abstract similarities as they are captured in most structural datasets (see the excellent discussion by Dybo and Starostin).

"Grammar" as evidence for deep language relations

Leading scholars in historical linguistics have provided convincing arguments that genetic relationships among languages can only be demonstrated by illustrating regular sound correspondences in concrete form-meaning pairs across the languages under investigation (see especially the very good analysis by Campbell and Poser 2008). In spite of this, the rumor that "grammar" (i.e., structural datasets) might provide a shortcut to detect deep, so far unnoticed, relationships among the languages of the world is very persistent, as reflected in many different studies.

Among the examples, Dunn et al. (2008) claimed that language relationships for Papuan languages of Island Melanesia could be uncovered by means of phonological and grammatical (abstract) structural features; and Longobardi et al. (2015) used syntactic features to compare the development of European languages with the development of European populations. Zhang et al. (2018) used phonological inventories of more than 100 different Chinese dialects, coding the data for simple presence and absence of each of the more than 200 different sounds in the database, and analyzing the data with the STRUCTURE software (Pritchard et al. 2000), whose results tend to be notoriously misinterpreted.

What is important about these studies is that none of them (maybe with exception of the study by Dunn et al. 2008, but I am in no position to actually judge the findings) could make a convincing claim why the structural datasets would provide evidence of deeper relationships than could the lexicon. Even the study by Dunn et al., which tests the suitability of their small questionnaire of only 115 structural traits on Oceanic languages, has since then not led to any new insights into so far undetected language relationships, contrary to the hope expressed by the authors, "that structural phylogeny is an important new tool for exploring historical relationships between languages" (ibid. 45).
Structural data as a shortcut?

Some scholars who work on structural datasets may find my claims harsh and unjustified. In fact, there are studies that seem to provide evidence that structural datasets perform similarly or equally well compared to phylogenetic methods based on lexical data.

For example, Longobardi et al. (2016) carry out experiments on structural data of phoneme inventories, syntactic features, and "traditional" cognate sets for very small Indo-European datasets, concluding that all of the datasets yield similar results, and that syntactic or phonological features in structural datasets could be used instead of lexical phylogenies.

Contrary to this, Grennhill et al. (2017) also experiment on lexical datasets in comparison with structural data for 81 Austronesian languages, but they find that, in general, lexical data is much more stable than structural data, although some structural features seem to be similar to lexical items regarding their stability.

A wish list for future tests

I see two major problems in the debate about the usefulness of structural data in historical linguistics.

First, the studies that confirm that structure might work equally well compared with lexical data, are all based on small samples of one specific language family that was analyzed based on very diverse features that were specifically designed to study the languages under question. For me, a true test that some features carry deep historical signal would need to be illustrated for a large set of related and unrelated languages, not only just for selected datasets.

Furthermore, to allow for an honest comparison with the lexicon, the selection of features should not contain any lexical characters or characters that could only be extracted with the help of lexical characters. Thus, asking whether the words for "fish", "I", and "five" are pronounced similarly in a language would not be allowed in such a feature collection, because this would follow lexical criteria, and we know very well that this property is a very good proxy for identifying Sino-Tibetan languages (Handel 2008).

Second, and more problematic, is the fact that structural datasets do not provide information on the relatedness of the traits under comparison. While this is no problem for typologists who study shared structural features out of interest in universal tendencies in the languages of the world, it is a problem for the application of phylogenetic software, since the typical approaches in biology treat homoplasy as an exception, while it may often be rather the norm than an exception in structural datasets.

Conclusion

In order to make structural data suitable for historical analyses, much more research needs to be carried out, including specifically a much thorougher study of parallel evolution and geographic convergence (due to language contact) in different language families of the world — a nice illustration
for the Indo-European languages is provided by Cathcard et al. (2018).
I would be happy for our field if such research could reveal markers of deep genetic ancestry in the languages of the world, and help us to push the boundaries of linguistic reconstruction. For the time being, however, I remain highly skeptical, especially when scholars try to demonstrate the suitability of "grammatical" comparison with small datasets and idiosyncratically selected feature sets that are not comparable across datasets.

References


How languages lose body parts: once more about structural data in historical linguistics

Guido Grimm and Johann-Mattis List

Max Planck Institute for the Science of Human History, Jena

Mattis’ last two blog posts dealt with problems of what linguists call "structural data". Here we discuss what this means for the inference of relationships between languages.

A closer look at structural data: the questionnaire issue

As pointed out before, what is called structural data in comparative linguistics is a very diverse mix of data solely unified by the idea of having some kind of questionnaire that a linguist may use when going into the field and trying to describe a certain language. These questionnaires are a bit different from the traditional concept lists usually used for the purpose of historical language comparison (see the collection of different lists in the Concepticon project by List et al. 2016). The main difference is that they are based on an imaginative question that a field worker asks an informant (which could as well be a written grammar of the language under question). Since questions can be asked in many different ways, while concepts in historical language comparison are usually restricted to the so-called "basic vocabulary", the diversity of structural datasets is much greater than the diversity we encounter when comparing questionnaires based on concept lists.

When analyzing these data, we deal with characters of very different nature, and likely different evolutionary pathways or histories. A biological analogy would probably be (true) total evidence data sets that combine genetic data from: genes/genomes with different inheritance pathways (paternally, maternally, biparentally; basic information level), morphological-anatomical data (visible form, phenotypic), palaeontological data (historical evidence), ontogenetic (life-history stages, developmental features), and biochemical data (expression level). The only difference is probably that the linguistic characters’ histories may be more complex. [Side-remark: ‘total evidence’ datasets found in the biological literature are typically just combination of genetic and morphological data, allowing for the inclusion of extinct/fossil taxa.]

To give a specific example, let's have a look at a the Chinese dataset by Szeto et al. (2018), mentioned in Mattis' blogpost from September. This dataset is now accessible as a GitHub repository (https://github.com/cldf-datasets/szetosinitic). Mattis added some information regarding the different features of the questionnaire. We list these features in slightly abbreviated form in the table below, adding rough categorizations by Mattis in the Comment column.
<table>
<thead>
<tr>
<th>ID</th>
<th>Description</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-1</td>
<td>5 or more tone categories</td>
<td>phonological / diachronic</td>
</tr>
<tr>
<td>p-2</td>
<td>Retroflex fricative initials</td>
<td>phonological / diachronic</td>
</tr>
<tr>
<td>p-3</td>
<td>Bilabial nasal coda</td>
<td>phonological / diachronic</td>
</tr>
<tr>
<td>p-4</td>
<td>Stop codas</td>
<td>phonological / diachronic</td>
</tr>
<tr>
<td>p-5</td>
<td>Monosyllabic word for 'snake'</td>
<td>lexical</td>
</tr>
<tr>
<td>p-6</td>
<td>Differentiation between 'hand' lexical / semantic and 'arm'</td>
<td></td>
</tr>
<tr>
<td>p-7</td>
<td>Differentiation between 'defecate' lexical / semantic and 'urinate'</td>
<td></td>
</tr>
<tr>
<td>p-8</td>
<td>Differentiation between 'eat' and lexical / semantic 'drink'</td>
<td></td>
</tr>
<tr>
<td>p-9</td>
<td>Semantically void suffix in 'table'</td>
<td>lexical</td>
</tr>
<tr>
<td>p-10</td>
<td>Different classifiers for humans lexical / semantic and pigs</td>
<td></td>
</tr>
<tr>
<td>p-11</td>
<td>[CLF N] constructions in subject syntactic position with definite reference</td>
<td></td>
</tr>
<tr>
<td>p-12</td>
<td>Reduplicated monosyllabic nouns</td>
<td>morphological</td>
</tr>
<tr>
<td>p-13</td>
<td>Post-verbal modal auxiliarsyntactic / diachronic developed from 'ge/acquire'</td>
<td></td>
</tr>
<tr>
<td>p-14</td>
<td>Modified-modifier order inmorphological / syntactic animal gender marking</td>
<td></td>
</tr>
<tr>
<td>p-15</td>
<td>Post-verbal adverb meaning 'first'</td>
<td>lexical / syntactic</td>
</tr>
<tr>
<td>p-16</td>
<td>[V DO IO] order in double objectsyntactic dative constructions</td>
<td></td>
</tr>
<tr>
<td>p-17</td>
<td>'Give' as a disposal marker</td>
<td>syntactic / diachronic</td>
</tr>
<tr>
<td>p-18</td>
<td>'Give' as a passive marker</td>
<td>syntactic / diachronic</td>
</tr>
<tr>
<td>p-19</td>
<td>'Go' as a post-VP associatedsyntactic / diachronic motion marker</td>
<td></td>
</tr>
<tr>
<td>p-20</td>
<td>Marker-Standard-Adjective ordersyntactic in comparatives</td>
<td>morphological / syntactic</td>
</tr>
<tr>
<td>p-21</td>
<td>case system</td>
<td></td>
</tr>
</tbody>
</table>

Mattis has tried to characterize the features, i.e. matrix’ characters, by generalizing linguistic categories: "phonological", pointing roughly to questions about pronunciation (the biological equivalent would be phenotypic traits in morphology or anatomy); "lexical", pointing to the words in the lexicon (this would be the DNA of a language); "morphological", pointing to the ways in which words are constructed; and "syntactic", pointing to the ways in which words are combined to form sentences. In combination, “morphological” and “syntactic” are equal to ‘meta-level’ biological traits, such as development-related features, ontogenetic evidence, and biochemical composition — the ways in which the genetic code is expressed or used in a living organism in adaption to the environment.

Mattis also flagged some characters as "diachronic", to mark whether the respective feature was selected by the authors due to their independent knowledge about the history of the Chinese dialects.
This is something rarely possible in biology, but imagine that we could go back in time to literally observe the evolution of a lineage over a given time-period, and code this observed evolution as traits. Note that this is not entirely science-fiction — there are two examples where we can observe directly pathways of biological evolution: mutation patterns in viruses, and horizontal modification of marine morphs in high-resolution sediment cores.

While one can discuss to what degree a certain feature should belong to this category, it is rather obvious that all phonological features are diachronic, because they name distinctions that reflect well-known processes of sound change, which happened in a couple of Chinese dialects and have been proposed in the past by dialectologists in order to classify the Chinese dialects historically.

For example, consider feature p-3 of the questionnaire: Does a given dialect have a syllable that ends in [-m]? From the history of the Chinese dialects we know that the [-m] was present in Middle Chinese, but later merged with [-n] and [-ŋ] in many varieties. Given that we know that this happened, and that we know that people have used this to mark a split, especially between the "innovative" dialects in the North and the South, it is clear that this feature bears explicit historical information. The same holds for all phonological features that we find in the data: p-1, the number of different tones in the dialects is again roughly reflecting the differences between languages in the North and in the South (the North having lost many tones); p-2 reflects the retention or specific development of retroflex sounds (similar to sh in English as opposed to s) mostly in the North; and p-4 reflects if a variety has syllables that can end in [-p, -t, -k], again a feature characteristic for the more "conservative" varieties in the South of China.

Figure 1: Overlap of features in Szeto et al.'s (2018) structural feature collection of Chinese dialects
Four lexical features have further been flagged as "semantic"; we query here existing or missing distinctions of concepts. People who learned, for example, Russian or certain German dialects know that it is rather common to have a single word for what other languages call "arm" and "hand" (see the respective entry in the CLICS database) or "foot" and "leg".

This diverse feature collection is coded as binary characters, reflected by presence/absence, or a yes/no answer to the question in the questionnaire. The choice of features is very selective. A biological analogy would be a matrix collecting incompatible splits of paternal (molecular) genealogies, along with a few prominent phenotypical traits (reflecting major evolutionary steps), and some traits that we expect to be primarily triggered not by genetics (inheritance) but by expression or adaptation to the environment. Biologists would not phylogenetically analyze such diverse and complex, potentially selection-biased data (although it could be very interesting), but linguists do.

In this context, it is remarkable, but also typical for these kind of data, that the 21-character feature collection by Szeto et al. (2018) has no feature in common with the collection by Norman (2003), a 15-character-matrix, which we also converted to our Cross-Linguistic Data Formats (see Forkel et al. 2018) in order to increase the data comparability.

![Figure 2: A Neighbor-net splits graph of the structural data by Szeto et al. (2018).](image)

The typification, coded as binary matrix to infer the Neighbor-net splits graph in Figure 2, demonstrates some basic characteristics of such 2-dimensional graphs. Note four of the 'characters' (typification categories) correlate with an edge(-bundle) in the network, separating the 'taxa' (the queried features). All "semantic" taxa are also "lexical", but "lexical" is more comprehensive, hence, "semantic" is placed as 'descendant' of "lexical" (Neighbor-nets can visualize ancestor-descendant relationships to some degree). "Morphological" taxa are either just "morphological" or also "syntactic", hence the pronounced box.
For "diachronic" and "syntactic", we have no corresponding edge(-bundle), because one taxon is also "lexical", but the others are "diachronic" and "syntactic" — this is a conflict that cannot be resolved with two dimensions. To visualize all the resultant 'taxon' splits, called also taxon bipartitions, we would need a third dimension. Lacking a third dimension, the Neighbor-net prioritizes keeping most "syntactic" together, because the "diachronic-syntactic" are closer to "syntactic" (max. 1 'character' difference) than to "diachronic-phonological" (2 character difference). The "syntactic-lexical" has to be placed apart because it is equally close to "lexical" and "syntactic" 'taxa', but differs much from "morphological-syntactic" or "diachronic-syntactic", the closest two relatives of "syntactic"-only 'taxa'. It is resolved closer to the centre of the graph, because it is more closely related to the other "syntactic" taxa than to the rest of the "lexical" taxa. This is also the reason why the "syntactic"-only taxa have to be placed farther out: "Diachronic-phonological" and "syntactic-lexical" are closer to the other endpoints, and the distance of "syntactic"-only to "diachronic-phonological", "lexical" and "morphological" should be as large as possible.

**Losing body parts: How data coding masks underlying processes**

Most typologists collecting structural data are not per se interested in phylogenies. Yet, given that scholars deliberately collect historical (diachronic) features, this shows that even if they would not necessarily admit it, they have a genuine interest in uncovering the history of the languages under question; or at least, how closely related languages (or here: dialects) are. But this requires understanding the characters we analyze, the collected "structural data".

In evolutionary biology, the key question people (should) ask when trying to select characters is how their change can be modeled on a tree or a network. What processes could be expected that shaped the data? What is behind the diversity? Is similarity or dissimilarity instigated by:

- [A] inheritance, i.e. passed from an ancestor to all / some of its descendants,
- [B] random mutation and/or sorting, i.e. the product of a stochastic, evolutionary neutral process,
- [C] non-random mutation, i.e. processes that recur frequently, may be beneficial and positively (gain, or negatively: loss) selected for, or
- [D] secondary contact, mixing of lineages by hybridization (symmetric mixing) and introgression (asymmetric mixing)?

[A]–[C] are vertical processes following a tree, even if the tree does not necessarily need to be the same; [D] is (mostly) horizontal and can only be modeled using a network. For each of the above, we can find an analogy in the evolution of languages.

In addition, process [3], and to a lesser extent [4], can lead to what biologists call 'homoplasy', meaning that the same feature is observed in two unrelated or distantly related taxa. In the context of phylogenetic inferences, homoplacies inflict tree-incompatible signals, seemingly reticulate patterns originating from a tree-like evolution. Structural (or other) linguistic data and phenotypical biological data have a lot in common — complex processes are boiled down to mere absence or presence of features (or traits, as they are called in biology).
Figure 3: Basic evolutionary processes, we need to consider when looking at linguistic data. Or biological traits, when we replace simplification by adaptive evolution, positively selected traits.

If we check the features in our table above, and ask: to which degree can they be used to model these processes (see also David's last post on illogic in phylogenetics), e.g. simply distinguish between similarity by chance, relatedness, or secondary contact (mixing), we can easily see that they are by no means optimal for evolutionary investigations. This is not necessarily because of the processes they involve, but more importantly because of the data sampling, which makes modeling almost impossible, with each character needing its own model.

As an example, take the feature p-6 in our table. Whether or not a language makes a distinction between "arm" and "hand" does not seem to follow specific geographic or genealogical patterns. The following figure shows a plot from the CLICS database (List et al. 2018), visualizing the most frequently recurring polylemies (or colexifications) centering around the concept "arm". The full visualization in CLICS can be found here, and when hovering with the mouse over the link between "arm" and "hand" (marked in green below).
From eye-balling the data, it is hard to find a consistent geographic / language-family pattern, which suggests that the feature p-6 is likely to show a high degree of homoplasy in the languages of the world. Obviously, different people decided not to distinguish between "hand" or "arm". But, the example of the Sami languages in northern Scandinavia also demonstrate that some people using related, long-isolated languages, consistently don't make the distinction. Here, the homoplasy is inherited (lineage-conserved). A biological analogy would be the rarely applied difference between a 'convergence' (a trait is independently evolved in different lineages) and a 'parallelism' (a trait is expressed by different but not all members of the same lineage).

A specific analogy to the "hand-arm" colexification / differentiation pattern is leaf shedding in oaks and their relatives (Fagaceae, the beech family). Some oak lineages (section Cerris of oaks, beech
trees, chestnuts) are essentially or strictly deciduous, others (sections *Cylcobalanopsis*, *Ilex*, the sister sections of *Cerris*; *Castanopsis*, the sister genus of chestnuts) are always evergreen, and the biggest group (number of species) of all Fagaceae, subgenus *Quercus* includes evergreen (1 section), mixed (the two by far largest sections), and deciduous (1 nearly extinct section) sublineages. To some extent this is linked to the climate in which the species thrive (high latitudes and/or per-humid = deciduous, low latitude and/or seasonally dry = evergreen), but consistently evergreen and deciduous lineages do co-exist.

Looking at the Chinese dialects, we see that p-6 represents a trivial split in the network.

Figure 6: A Neighbor-net inferred from the Szeto et al. matrix. Dialects that distinguish "arm" and "hand" with filled dots ("1" for character 6 in the matrix), those that don't ("0") with empty dots. We can put a single line separating all don't- from do-taxa (dialects), i.e. a bipartition of the taxon set fitting the character partition seen in (p-)6.

But, given the general patterning of the feature on a global scale, does this really mean that it is inherited — that is, a good feature to reflect relatedness?

Whether a feature is likely to be homoplastic is just one part of the story. Linguists typically have more information about how things change than do biologists, putting a double-edged sword in their hands (that they hardly ever use). Asking whether "hand" and "arm" are expressed by distinctive concepts does not consider the underlying processes. Here, we can assume at least three different character states, namely:

1. "arm" and "hand" are expressed by the same word, which is the original word for "arm",
2. "arm" and "hand" are expressed by the same word, which is the original word for "hand", and
3. "arm" and "hand" are expressed by different word.

We could even have a forth state, in which "arm" and "hand", in the whole long history of the ancestral languages, was always used to express "arm or hand" (i.e., both body parts). No differentiation and no later generalization from either arm nor hand took place.
From Ancient Chinese, we know that "1" (Yes, I do differ between "arm" and "hand") was most likely the original state. We can further assume that once the distinction is dropped, it is less likely to come back again (although this can, of course, also happen). That is, our model involves two possible mutations (vertical process): we lose the word for "arm" due to its replacement by "hand", or we lose the word for "hand" due to its replacement by "arm", each with its own probability.

The probability, mutation or not, and which mutation, relates to four principal driving factors:

1. probability of random loss (mutation)
2. probability of random gain (mutation)
3. global linguistic tendencies
4. regional socially-enforced preference

Establishing $p_{\text{arm}}$ (loss "arm") and $p_{\text{hand}}$ (loss "hand") is not trivial, because they may be affected by what is the word for "arm" and "hand" (for simplicity we will assume that $P_{\text{arm}}$ and $P_{\text{hand}}$ are close to 0). We could expect a higher tendency to keep the word that is easier to pronounce or less easy to confuse with other words and, hence, is easier to understand. If two dialects with different states come into contact, this may also influence the decision to take over a state or not. In everyday language, a distinction between "arm" and "and" may be useless because of the clear context in which both words are used, so $p_{1\text{-word}} > p_{2\text{-words}}$. However, closeness to administration centers or areas with a higher percentage of educated people could decrease $p_{1\text{-word}}$, because it may be considered a sign of poor social standard to not make the difference between "arm" and "hand".

![Diagram of vertical and horizontal processes involving transitions of "hand" and "arm".](image)

**Figure 9:** Vertical and horizontal processes involving transitions of "hand" and "arm".

Estimating $p$ can only be left to phylogenetic algorithms (unless more detailed information is available). But we can (and should) design the questionnaire to capture as many of the processes as possible. In this case, to not only ask whether there is a distinction between "arm" and "hand", but also to find out whether the word "arm" or "hand" is used, e.g. by using two questions/binary characters:

- Do we use "hand"?
- Do we use "arm"?

Note that this question requires quite a deal of knowledge about the languages under investigation, since it may not be trivial to find out what was the "original" word for "arm" or "hand".

Therefore, a further step would be to replace the binary characters by a value measuring the similarity between the words used for "hand" and those used for "arm". One could again argue that adding this information would add historical information to the feature, but it is clear that the abstract nature of the question is hiding important phylogenetic (and also typological) information from us.
It seems therefore, that, instead of asking whether or not there is a distinction between "arm" and "hand", it would make much more sense to trace the cognacy (or homology) of the expressions for "arm" and "hand" across all taxa (languages, dialects), and think of ways how this could be scored and modeled by phylogenetic analyses. The structural data framework with its features based on simple yes-no questions therefore inevitably leads to a misinterpretation of processes when analyzing the data with phylogenetic software.

The need for exploratory data analysis

In reality, structural (or other) data sets in linguistics face problems similar to the ones palaeontologists face when trying to establish phylogenetic relationships between fossils (extinct organisms) — the probability for a mutation (visible change) is largely unknown, and differs not only from character to character but also within the same characters. A state 0, 1, 2 etc. may have a higher probability to manifest (or get lost) in one lineage than in another.

In addition, the linguistic problems recur in a similar way to that of biologists working close to and below the species level (see also Guido's post on population dynamics and individual-based fossil phylogenies) — reticulation is rather the rule than the exception, as similarity is triggered by contact, so that horizontal processes, not inheritance, may dominate evolutionary dynamics. Thus, the diversity pattern cannot be modeled by a tree alone. Establishing explicit probabilistic frameworks to deal with this may not only be difficult but even impossible (given the available data). Meanwhile, however, one can embrace exploratory data analysis as a heuristic tool.

So, let's look at the example. As in the original paper, we used the binary matrix of the 21 characters to infer a planar, 2-dimensional (meta-)phylogenetic network, a Neighbor-net splits graph. The resulting graph is a longitudinally inflated spider-web, with its endpoints defined by the southern Chinese dialects (e.g. Guangzhou, Nanning, Taishan) and the north-central (e.g. Linxia and Xining) dialects. The latter are significantly closer (geographically and data-wise) to the Beijing version of Chinese.

![Figure 10: The Neighbor-net based on simple mean (Hamming) pairwise binary character distances](image)

The first thing to note is that the matrix includes dialects that are indistinct (green stars) for all 21 characters, and some that are geographically and data-wise very similar to each other, while being...
distinct from all others (green ovals). In biology, we call this (taxic, lineage-)coherence. In addition to Linxia and Xining, we have Nanchang and Lichuan characterized by elongated ('tree-like') terminal edge-bundles. These obviously represent closely related dialects sharing a long(er) common history. Others have more than one possible closest relative. For instance, Liuzhou may share quite a few features with Guangzhou, but it is equally close to the Nanchang-Lichuan pair (yellow fields). Dongtai (orange star) is unique, but its 'neighborhood' (orange-ish brackets) as defined by shared edge-bundles that include Changsha (which again is most related to Jiujang) and Taiyuan plus Baotou, the latter two substantially closer to the Beijing (red star) group.

Similar to Dongtai, and also connected to the central part of the graph, are dialects with long-terminal branches (edges). Hefeng (blue star) is substantially different from Dongtai, and only has one further dialect in its neighborhood (blue bracket), Wangrong, a close relative of the Beijing group. The Wuhan, Chengdu, and Guiyang (gray field) dialects appear, on the other hand, to be completely isolated.

As explained above, there are different processes, vertical and horizontal ones, that may trigger similarity, and we want to get an idea as to which character may be influenced by which process. From the graph, several aspects are obvious:

- geographic closeness plays a major role,
- the signal provided by the data is not tree-like,
- the data is highly homoplastic, and includes internal conflict.

Not so obvious is whether this situation is due to random or evolutionary directed similarity, or reticulation. Since the graph is planar, and puts the Chinese dialects in a circular order, we can order the character matrix accordingly to see how the traits form groups (which could be called cliques in this context). In the next step, we can then map each character onto this network, to see how well they fit with the overall similarity pattern. We showed this above for p-6 (hand-arm-distinction, one split), and here we add a character with quite a poor fit, p-17 (syntactic-diachronic), "give" as a disposal marker.

Figure 11: Character mapping for p-17 (filled dots, "give" used as disposal marker; empty, not used), with the p-6 split indicated as well. Red, splits (taxon bipartitions defined by character cliques) that
have no corresponding edge-bundle (neighborhood); blue, splits with neighborhood; green, unique, isolated change (deviation from the rule) within the neighborhood.

The number of inferred mutations in the map uses Ockham’s Razor, upon which parsimony (tree and network) inference relies as well. Using such a map, we can even provide an estimate for how likely (qualitatively spoken) a change is under the assumption that neighborhoods in the graph represent either exchange (homogenization) between closely related dialects or are inherited, reflecting both horizontal and vertical relatedness. Mapping characters on a 2-dimensional network allows finding a scenario beyond a single tree hypothesis.

For p-6, we need just one change (i.e. loss in all more south-bound dialects), but we don't find an edge bundle corresponding to this unique change. Given what we discussed above about p-6, we have more independent losses than the simple reconstructed one. Social preference or general contact for retaining the primitive state of having two words could explain why dialects closer to the Beijing dialect area have a "0", although not all are closely related in general.

For p-17, we need at least four (independent) changes from "0" → "1", two of which have a corresponding edge bundle (blue, Nanchang plus Lichuan, Changsha plus Dongtai), one isolated (green, Luoyang), and one without a corresponding edge bundle (Wuhan and Hefeng dialects). The (equally parsimonious) alternative for p-17 would be a series of gains and losses, with the same number of steps:

![Figure 12: Alternative scenario for p-17.](image)

This is where one needs to consider additional knowledge about the probability of getting or retaining a certain feature. The state shared by most dialects across the entire net is "0", irrespective of overall similarity, which would make it a natural pick for the primitive state. Thus, assuming four (or more) changes from 0 → 1 (acquisition of the queried feature), rather than two independent acquisitions (starting with the Beijing group; note, the position of the root will not change the number of needed changes), then a loss (1 → 0) in many southbound dialects and a re-gain (0 → 1) in the Nanchang + Lichuan dialects.

The same assessment can be made for all of the characters, and we end up with something like this:
Figure 13: Fully annotated split network of the data. Changes relating to edge-bundles accordingly colored, arc indicate changes without a corresponding edge-bundle. Note, the prominent yellow split that defines a neighborhood of dialects most similar to the Beijing dialect, albeit there is no character supporting this edge. The rather poor fit of many character splits (cliques) with edge-bundles relate to the fact that we visualize a highly complex diversification (multi-dimensional processes) using a planar, 2-dimensional graph.

While this figure may be confusing at first sight, it comprehensively shows what the characters contribute to the overall graph. We can discriminate more-likely from less-likely mutations (how many changes are needed at least), but also the character assemblies shared by putatively closely related dialects.

- p-3 and p-11 are a typical feature of Guangzhou and allied dialects within the southern Chinese complex. p-3 is also present in Lichuan, and p-11 in Jixi (thus in not so distant dialects).
- Features p-6 to p-9, p-16, and p-19 form a diagnostic suite for the Guangzhou dialects and other dialects related to them in the one or other fashion and distinguish them from, e.g., the Beijing group.
- The latter, the Beijing group, has fewer diagnostic character assemblies. One characteristic sequence could be p-1, p-2, p-12, p-14, but this includes three features with a minimum of 3+ changes. Similarity here is mostly the result of a lack of (potentially) derived features (hence, the character-unsupported yellow edge-bundle defining a Beijing-including neighborhood).

Outlook and summary

In this re-investigation, we have, once more, commented on the problems we see with the use of structural features for the purpose of historical language comparison and phylogenetic reconstruction. We see the major problems in the (often) unfortunate choice of question, resulting in elicitations of features that cannot be easily modeled with current software for phylogenetic analyses. It is important to keep in mind, in linguistics and phylogenetics, that we can infer trees or networks based on data of no matter what quality and information content. But before we present the result, we should have taken a look at the primary data.
Does it fit with the resulting graph, or not?
Where does it fit, and where not?

In the context of our critique of linguistic questionnaires, the mapping strategy discussed above opens a potential avenue to identify:

- stable / unstable features (geographically or evolution-wise) and
- coherent / incoherent features.

Based on this, we can then inquire as to which degree language (or dialect) groups influenced, stabilized or modified each other by geographic proximity.

Inference-wise, the natural next step would be to use the information about the minimum number of necessary changes to counter-weight characters. This would eventually allow to use median networks (and related) approaches on the data, which is currently the only way to explicitly identify ancestors using phylogeographic reconstructions. With the current matrices, the extreme homoplasy makes an unweighted application of median networks and related methods impossible.

References


Supplementary data

The data we used to create the analyses and figures provided in this post are available at https://github.com/cldf-datasets/szetosinitic/tree/master/examples

Patterns, processes, abduction, and consilience

Johann-Mattis List

Max Planck Institute for the Science of Human History, Jena

In a recent blog post, David emphasized how important it is to distinguish patterns from processes in evolutionary biology, with phylogenetic analysis concentrating on the description of patterns (and not on the direct investigation of processes. David's major point is that we need to be careful to not forget about the logical limitations of our approaches:

In the world of logic, propositions cannot be converted; and yet converting propositions is exactly what is done by all descriptive data analyses.

As David correctly points out, in phylogenetic analysis, we tend to observe a pattern (some similarity between different species or languages, for example), and use this pattern to conclude that a specific process has happened (eg. the languages are so similar that we think they are identical).

Given that this problem is also important in historical linguistics, I want to share some thoughts from a linguistic perspective. Most of these were elaborated much earlier, in my PhD dissertation; and if you have read the original chapter (List 2014: 51-57), what I write below may seem repetitive. I have also alluded to these ideas in a couple of previous posts: What we know, what we know we can know, and what we know we cannot know; and Killer arguments and the nature of proof in historical sciences.

However, it is worthwhile to elaborate on these thoughts here, as David's comments are extremely interesting for historical sciences in general, and I think they deserve a more proper discussion across different disciplines.

Ontological fact and epistemological reality

The basic pattern/process problem may be even more complex than it is in evolutionary biology. In quite a few branches of science, most prominently in the historical and social sciences, even the object of investigation is not directly accessible to the researcher. All researchers can do is to try to infer the research object with the help of tests. In historiography we infer the res gestae by comparing direct and indirect (usually written) sources (Schmitter 1982: 55f). In psychology, attributes of people, such as "intelligence" cannot (yet) be directly measured but have to be inferred by measuring how they are "reflected in test performance" (Cronbach and Meehl 1955: 178).

The same holds for ancestors in historical linguistics and evolutionary biology. All we can do in order to examine whether some languages or species share a specific kind of ancestry is comparing them systematically, trying to identify patterns that provide evidence for close relationship. Given that we lack direct evidence of its existence, the ancestral languages or species we infer through comparison cannot be
treated like an *ontological fact* but only as an *epistemological reality* (Kormišin 1988: 92). We address what psychologists call the *construct*, that is, the "fiction or story put forward by a theorist to make sense of a phenomenon" (Statt 1981/1998), not the "real" object.

**Abduction as our sole mode of logical reasoning**

In historical linguistics, we can address our research objects only via constructs, and so we have to rely on abduction as our sole mode of logical reasoning (Anttila 1972: 196f). The term abduction was originally coined by Charles Sanders Peirce (1839-1914) and refers, as opposed to *induction* and *deduction*, to a "mode of reasoning [...] in which rather than progressing 'logically' [...], one infers an antecedent condition by heuristic guessing from a present case" (Lass 1997: 334). In Peirce's word:

> Accepting the conclusion that an explanation is needed when facts contrary to what we should expect emerge, it follows that the explanation must be such a proposition as would lead to the prediction of the observed facts, either as necessary consequences or at least as very probable under the circumstances. A hypothesis then, has to be adopted, which is likely in itself, and renders the facts likely. This step of adopting a hypothesis as being suggested by the facts, is what I call abduction. I reckon it as a form of inference, however problematical the hypothesis may be held. (Peirce 1931/1958: 7.202).

Due to the specific aspects of knowledge we are given in the historical sciences, abduction is the only mode of reasoning that we can employ. According to Peirce (ibid.: 2.623), all three modes of reasoning, induction, deduction, and abduction, "involve the triad of 'rule', 'case' and 'result', but inference moves in different directions" (Lass 1997: 334). While induction infers a *rule* from a situation in which one is given *case* (initial situation) and *result*, deduction infers a *result* from a situation in which one is given *case* and a *rule*. Abduction, however, starts from a *result* (or a *pattern* in David's words) and a *rule* from which we try to infer a *case*.

As an example, consider the problem of language evolution. Given two languages with no written records of their previous history, we may observe as a pattern (or *result*) that they show striking systematic regularities in terms of sound correspondences. Given that we know, that — as a *rule* — languages change their sound systems slowly over time, we can conclude that the initial situation, the *case*, was that the two languages were once a single language. There is no way we employ any other mode of reasoning here, as long as we start from individual languages (or species) whose past we want to understand and describe.

We can think of situations in which we try to induce rules in historical linguistics, for example, when dealing with the development from Latin into its descendant languages, where we could ask about the individual processes of sound change (or sound change rules) by which the former was transformed into the latter. We can also think of situations in which we try to decide results from rules and initial situations, for example when trying to predict unobserved cognate words in languages that have not yet been completely documented by fieldwork (Bodt et al. 2018), by applying rules of sound change (or sound correspondences) to aligned cognate sets (List, forthcoming). But the big bulk of our work in historical
linguistics (and also in evolutionary biology) works only via abduction: given a result (a pattern / observation in the present), we use our knowledge of rules and processes to infer an ancestral state.

**Problems of reasoning based on abduction**

According to Schurz (2008), different patterns of abduction can be distinguished, depending on: (1) "the kind of hypothesis which is abduced", (2) "the kind of evidence which the abduction intends to explain", and (3) "the beliefs or cognitive mechanisms which drive the abduction" (ibid.: 205). The kind of abduction that is commonly used in historical linguistics and evolutionary biology belongs to the family of **factual abductions**, that is, abductions in which "both the evidence to be explained and the abduced hypothesis are singular facts" (ibid.: 206). Since we mainly deal with unobservable facts (i.e. **constructs**), we can further characterize it as **historical-fact abduction** (ibid.: 209).

The problem of historical-fact abduction is not necessarily that what we are try to "observe" lies in the past, but more importantly, that — due to the logic underlying abduction as a mode of reasoning — we usually have to infer both the rules and the initial situation from the patterns we observe. Given (as David emphasized) that a pattern can result from different processes, our inference of a specific, individual historical fact requires that we decide on a specific, individual process at the same time. Given that we have to infer both the process and initial state at the same time, it is not surprising that our inferences about the past are often so vague, and may easily change so quickly, specifically in a situation where we can't just travel back in time to see whether we were right.

In contrast to David, who suggested that we cannot directly investigate processes in the evolutionary sciences, however, I think that we still can indirectly, be it with help of experiments, with simulations, or in those cases where we are lucky enough to find history documented in sources. These cases where we can study processes, however, are — and here I agree completely with David — not what we normally do in our research. What we usually do is investigating patterns and trying to infer both the process and the original state by which the patterns can be explained.

**Cumulative evidence**

The problem of abduction, in general (or historical-fact abduction, in specific), is to make sure that we protect ourselves from giving in to wild speculations. That we are not necessarily good at doing so is reflected in the numerous debates in historical linguistics, and evolutionary biology, where scholars at times invoke completely contrary scenarios explaining the past based on identical patterns. In addition, in historical linguistics, people often do not even agree regarding the patterns they believe can be observed in the data.

Earlier, in my dissertation (List 2014), I identified two aspects that I deem important in order to minimize the speculative aspect of our research, claiming that historical-fact abduction should be based on: (1) **unique hypotheses**, and (2) **cumulative evidence**. That we need unique hypotheses may seem self-evident at first sight, since it seems to be silly to claim that a certain pattern could be explained by a range of
processes. Looking back at this point now, however, I tend to see this less strictly. In fact, I think that I
would even prefer it if scholars would list all potential (individual) processes that may seem likely to have
yielded a pattern, instead of focusing only on one possibility (and disregarding alternative solutions).
Since we are not doctors who need to heal our patients as quickly as possible, we can afford a certain
amount of doubt in our research.

Regarding the second point, what I had in mind earlier was that it is best if we have multiple results or
different patterns (observed for the same species or languages under investigation) that can all be
explained by the same hypothesis. In order to justify the claim that one specific hypothesis explains the
evidence better than any alternative hypotheses, we can profit from combining multiple pieces of evidence
that might "[fall] short of proof [when taking] each item separately" but become convincing when "all the
items [are] combined" (Sturtevant 1920: 11).

Being forced to rely on multiple pieces of evidence (that only when taken together allow one to draw a
rather convincing picture of the past) is not a unique problem of historical linguistics and evolutionary
biology, but also of historiography – and even crime investigations, as was pointed out by Georg von der
Gabelentz (1840-1893, cf. Gabelentz 1891: 154), and in later work on semiotics (cf. the papers in Eco
and Sebeok 1983). The fact that historical linguistics theories are built about cases (events, unique
objects), as opposed to theories about general laws, may also be the reason for the philological "style"
prevalent in historical linguistic studies. I also believe that it is due to the complex nature of the inference
process that a systematization of our methods has never been carried out efficiently.

While, for example, we can claim (at least to some degree) that the identification of cognate words in
historical linguistics can be systematized (and even to some extent automatized, List et al. 2017), we are
at a loss when it comes to systematizing the methods that we use to determine whether words have been
borrowed or not. Instead of using one single method, we use a whole range of indicators, and only take
borrowings for granted if at least a few of them point into the same direction (List 2018).

**Consilience and conclusion**

In a talk by James McInerney, held in 2015 in Paris (presenting an overview of his research as reflected
in part in McInerney et al. 2014), I realized that the question of "cumulative evidence", which I had
thought would have been discussed only in linguistic circles, belongs to a larger complex of discussions
about consilience, as opposed to the Popperian tradition that claims that knowledge in science can only
advance via falsification and the identification of general laws, as opposed to singular facts (Popper 1945:
Chapter 25:II). We find this view, that we need to employ cumulative evidence when trying to infer
individual facts, clearly stated in the work of William Whewell (1794-1866), who originally introduced
the term consilience:

*The Consilience of Inductions [ie. abductions] takes place when an Induction obtained from one
class of facts, coincides with an Induction, obtained from another different class. This Consilience is
a test of the truth of the Theory in which it occurs.* (Whewell 1840: 469)
As far as I understand from James McInerney’s talk, the idea of consilience has long been disregarded in the historical sciences but is now gaining popularity (also thanks to the influential book by Wilson 1998). Although at first I felt delighted when I realized that I was not alone with the problem that I had called "cumulative evidence", based on the old book by Sturtevant (1920), I have to admit that I still do not really know what to do with this information, as it is extremely hard to operationalize the concept of consilience. When confronted with numerous different pieces of evidence, how can we identify the hypothesis that explains them all? How can we compare two opposing hypotheses that each convincingly explain some but not all the data? How can we arrive at an objective weighting of our evidence, based on its importance?

What is clear to me is that a "probabilistic evaluation of causes and elimination of implausible causes plays a central role in factual abductions" (Schurz 2008: 207), since it reduces the search space when seeking an explanation for a given phenomenon (ibid.: 210f). But it is not clear how to arrive at such an evaluation when dealing with patterns in practice. For the time being, thinking and discussing about consilience seems interesting; but until we find ways to operationalize it, it will just remain a nice idea without any concrete value for our scientific endeavors. I dearly hope that this won’t be the case.

References

List

Patterns, processes, abduction, and consilience
