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

2016
## Contents

Directional processes in language change ........................................... 1  
Through a glass darkly ................................................................. 4  
Another early tree in linguistics ....................................................... 9  
Monogenesis, polygenesis, and militant agnosticism ............................. 12  
Machine learning, the Go-game, and language evolution ....................... 15  
Alignments and phylogenetic reconstruction in linguistics and biology ...... 18  
Can biologists learn from linguists? .................................................. 22  
More on analogies between biological and linguistic evolution ............... 27  
Inheritance in cultural evolution ..................................................... 29  
Sound change as systemic evolution .................................................. 32  
Once more on artificial intelligence and machine learning ...................... 37  
Isogloss maps are hypergraphs are bipartite networks ........................... 39
Directional processes in language change

Johann-Mattis List

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

Given that we are still in the process of beginning the new year, it seems to be in order to talk about directions — not in general, but rather in specific, namely, about directions in language change. This is important in so far as many processes in language evolution are directional. This means that they follow a specific direction from a state $X$ to a state $Y$, and this is frequently attested across a large number of the languages of the world, while the opposite process, that state $Y$ changes to state $X$, is extremely rare or even unattested.

In language evolution there are a lot of well-known and well-investigated processes with a strong directional tendency. In sound change, for example, a [p] can easily become an [f], whether it is in the Indo-European, the Austronesian, or the Sino-Tibetan languages. Yet the opposite process, that an [f] becomes a [p] is extremely rare. Similar tendencies hold for a [k] becoming a [ʧ], as in Italian [ʧɛnto] *cento "hundred", going back to Latin [kɛntum] centum "hundred", or a [g] becoming a [h], as in Czech [ɦora] hora "mountain", going back to Proto-Slavic *gora "mountain" (Derksen 2008).

In semantic change, unidirectional tendencies can also be observed, although it is often more difficult to identify them, let alone generalising them. Nevertheless, I think it is a rather safe bet to claim that words which originally mean "head" have a certain tendency to shift their meaning to denote "(the) first, the boss" or "the upper part, the top", while the opposite shift (that words which mean "boss" or "top" will be used to denote "head") is very unlikely to happen. Finally, in grammatical change, or, to be more precise, in grammaticalization (the process by which languages acquire new grammatical categories) directionality is one of the most important constraints (Haspelmath 2004).

Linguists usually know these tendencies very well, and they use them in their daily work, be it when trying to reconstruct the original pronunciation of words in unattested ancestral languages, when deciphering historical documents, or when tracing the semantic development of words through history. Directional changes are also important in evolutionary biology. Ratchet-like (that means: unidirectional) processes serve as a major explanans for constructive neutral evolution (Gray et al. 2010), direction is at the core of lateral gene transfer, and — as David mentioned in an earlier post — the usage of directional (non-reversible) models in phylogenetic reconstruction even provides an elegant way to root a tree (see also Huelsenbeck et al. 2002).

Given the active transfer of ideas from the biological to the linguistic domain in the last two decades, and the important role that directional processes play in both domains, it is surprising to me that methodological transfer has so far been almost exclusively limited to time-reversible models. The only approach known to me that explicitly makes use of linguistic knowledge of directions is that of Baxter (2006). In this paper, Baxter analysed phonological mergers in Chinese dialects within a framework of Camin-Sokal parsimony (Camin and Sokal 1965).
Phonological merger is a specific systemic process in language evolution. When sounds change (and they always change in some way), it may happen that two formerly distinct sounds are pronounced in the same way. As a result, words that formerly sounded different may suddenly sound alike, such as English *write* and *right*, which remain different only in their spelling not pronunciation. Mergers are a prototypical irreversible process. Once a merger has happened, speakers cannot go back, unless they recorded the original distinction and artificially tuned their language. But even this may be less easy than it seems — it is always easy to reduce distinctions. For example, most English speakers wouldn't have many difficulties in artificially pronouncing all instances of *s* as *sh* during a conversation. But being asked to pronounce a randomly chosen set of words with *s* as *sh* will turn out to be much more difficult. For this reason, mergers are an ideal data type for directional models of language change. Their drawback is, however, that they are difficult to determine, which may also be the reason why Baxter's approach has never been tested on other language families since then.

It may be justified to use time-reversible models for analyses that use lexical data, especially cognate sets, as in the approaches following Gray and Atkinson (2003), since it is difficult to determine the impact of directional processes on lexical replacement. Furthermore, due to the specific way the data is sampled, it is extremely difficult to determine directions. Yet in many other approaches that use different types of data, especially in those cases that model sound change processes (Hruschka et al. 2015, Wheeler and Whiteley 2015) or grammatical change (Longobardi et al. 2013), it might have a substantial impact on the results if directionality was explicitly modeled. What does this mean for the directions for the New Year? I keep being surprised by the similarities between evolutionary biology and historical linguistics, be it the organization of information in genomes and languages, the processes that drive evolution, the philosophical questions underlying our investigations, or the quarrels among scholars in their fields. Unfortunately, much of the transfer from the biological to the linguistic domain is still very simplistic, often ignoring the specific differences between the two domains. On the other hand, many fruitful analogies are still out there but have not yet been properly investigated. So, as a direction for those who work in interdisciplinary domains in this New Year, I think we should try to avoid reinventing the wheel, and we should also pay attention to not putting wheels on sledges.

References


In an earlier blogpost I mentioned the now largely abandoned discipline of lexicostatistics that was in vogue in the 1950s, originally initiated by Morris Swadesh (1909-1967; Swadesh 1950, 1952, 1955), but abandoned in the 1960s and henceforth often labeled as some kind of a failed theory that was explicitly proven to be wrong.

The crucial idea of Swadesh was to investigate lexical change from the perspective of the meaning of words. This perspective is contrasted with the perspective which takes similar (cognate) word forms in different languages as a starting point and compares to which degree they differ in their meanings. Swadesh's perspective, instead, starts from a set of meanings and investigates by which word forms they are expressed, and is also called an onomasiological perspective (which "names" are assigned to concepts?), while the other perspective is called a semasiological perspective (which "meanings" can words have?).

From a semasiological perspective, we would start from a set of related words and investigate their meanings. In this way, we could compare English head with German Hauptstadt "capital city" or English cup with German Kopf "head". Through such an analysis, we would learn that there was a semantic shift from the German word Haupt, which originally meant "head", to a more abstract meaning that is now probably best translated as "capital" or "main", and only occurs in compounds, such as Hauptstadt "capital city", Hauptsache "main reason", etc.

From an onomasialogical perspective, we would start from a set of meanings and investigate which words are used in order to express them in different languages:

<table>
<thead>
<tr>
<th>No.</th>
<th>Items</th>
<th>German</th>
<th>English</th>
<th>Dutch</th>
<th>Russian</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>hand</td>
<td>Hand</td>
<td>hand</td>
<td>hand</td>
<td>ruka</td>
</tr>
<tr>
<td>2</td>
<td>arm</td>
<td>Arm</td>
<td>arm</td>
<td>arm</td>
<td>ruka</td>
</tr>
<tr>
<td>3</td>
<td>mainly</td>
<td>hauptsächlich</td>
<td>mainly</td>
<td>hoofdzakelijk</td>
<td>glavny</td>
</tr>
<tr>
<td>4</td>
<td>head</td>
<td>Kopf, (Haupt)</td>
<td>head</td>
<td>hoofd, kop</td>
<td>golova</td>
</tr>
<tr>
<td>5</td>
<td>cup</td>
<td>Tasse</td>
<td>cup</td>
<td>kop</td>
<td>stakan</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

When looking at specific meanings in this way, one can find interesting patterns within one and the same language whenever a language uses the same or similar words to express what are different concepts in other languages. Russian thus uses the same word for "hand" and "arm", Dutch shows the
same word for "head" and "cup", and Russian, Dutsch, and German have similar forms for "mainly" and "head". These patterns can be historically interpreted by reconstructing patterns of semantic shift. In the case of English cup, German Kopf, and Dutch kop, for example, the original meaning of the words was "vessel" or "cup". Later on, the word changed its meaning and came to denote "head" in German. The transition is still reflected in Dutch, where the word can denote both meanings.

We can model this situation by assuming that every word in a language has a certain reference potential (Schwarz 1996: 175; Allwood 2003; List 2014: 21f, 36). This means that every word has the potential to denote different things in the world, due to the concept it denotes primarily. In List (2014: 21), I have tried to depict this as follows:

Reference Potential of the Linguistic Sign

In this visualization, a word form refers to a meaning, and the meaning itself has the potential to denote various things in the world, but with different probabilities. A word that primarily means "head", for example, may likewise be used to denote the "first person", as in the "head of a group", and a word that primarily means "melon" may also be used to denote a "head", due to the similarity in form. We can investigate the reference potential of words by simply looking at different translations in dictionaries. As an example (from List 2014: 36), when looking at our three words English cup, Dutch kop, and German Kopf, we find the following rough arrangement with respect to the reference potential of the word (the thickness of the arrows indicating differences in denotation probability):
Why do I mention all of this? First, I wanted to show that lexical change, no matter which perspective we take, is a very complex phenomenon. In a simplifying model, we could think of a lexicon as a bipartite network consisting of nodes that represent word forms in a language and nodes that represent meanings, and weighted links between word forms and meanings denoting the frequency by which a word is used to denote a given meaning. In such a network representation, lexical change could be
modelled as the re-arrangement of the edges between word forms and meanings. If a word form looses all its edges, this word is lost from the language, but we could also think of new words entering the language, be it that they are borrowed, or created from the language itself. Such a model would be very simplistic, ignoring aspects like word compounding, by which new words are created from existing ones. But it would be much more realistic than the idea that lexical change is just about the gain and loss of words, as assumed in the quasi-standard model of lexical change in phylogenetic reconstruction.

This brings us to my second point. When Swadesh introduced lexicostatistics, and his very specific onomasiological perspective on lexical change, he established a model of lexical change that would deliberately ignore all interesting processes underlying the phenomenon. Since then, we have been looking through a glass darkly. This is like a crime inspector having no other means but watching potential suspects through the windows of their apartments, noticing changes, like the differently coloured words in state A and state B in the Figure below, but never knowing what was really going on inside those flats (state C).

Trough a Glass Darkly: The lexicostatistic perspective on lexical change (A, B), and what is really going on (C).

Yet, when being honest with oneself, the problem of looking through a glass darkly does not pertain to the lexicostatistic perspective alone, but effectively applies to all of our research on language change. It is just the size and the number of windows that we survey, and the cleanliness of the glasses, that may make a little difference.
References


---

Another early tree in linguistics

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

The air is getting thin for those who thought that tree-thinking in linguistics was just a cheap copy from Darwin's family tree schema (1859). David clarified this in many earlier blog posts, and especially in one post about an early language tree (with reticulations) from the 19th century by Felix Gallet (1800; compare with Auroux 1990), and a later post on an early network from the 17th century by Georg Stiernhielm (1671; compare with Sutrop 2012).

More by chance than by actively searching for it, I stumbled upon another hint regarding an even earlier language phylogeny than the one we thought was the earliest so far. This phylogeny (or whatever it is) is mentioned in a recent article by Zeige (2015), which was published in a special issue of the Zoologischer Anzeiger (A Journal of Comparative Zoology) in which the topic of morphology across different sciences was discussed. Note that “morphology” in linguistics refers to the way words are composed from other words, or words are modified by means of inflection or derivation. So, although the term originally stems from biology, it has started to live a life of its own in linguistics.

The phylogeny that Zeige mentions in the article is about the Germanic languages in a broad sense, and was proposed by Justus Georg Schottel (1612-1676) in his lengthy treatment of the German HaupbtSprache (Schottel 1663). In the first volume of the book, Schottel gives 10 laudations on the German language, and in the tenth laudation, we find the schema provided on the next page (the whole book is available in digital form from the Bayerische StaatsBibliothek digital). Since the book is written in both Latin and German, it also contains the same schema in a Latin version.

Schottel has classified the Germanic languages and dialects. That we have a branching scheme here is obvious — that is, a nested set of groups, which could be represented as a dichotomous tree. His schema does not coincide with our modern phylogenetic classification of the Germanic languages, but it comes surprisingly close to it. The key question, however, as David pointed out in an email to me, is whether the classification was intended to represent the development of the languages.

Here, we have a general problem in linguistics, namely that linguists often did not and still do not distinguish between a classification that is intended to represent some observed similarities (which we would call a "synchronic classification" in linguistics), and a genealogical classification that is intended to represent the historical dynamics (which we would call "diachronic classification"). This is also mentioned in Zeige's (2015) article in the context of Schottel's classification; and in an earlier blog post on the Wave theory of linguistic development, we saw how linguists tried to establish an alternative to the family tree but replaced the historical tree by a static, synchronic schema that was no longer genealogical. Schottel published his book in 1663, more than 150 years before Rasmus Rask (1818), Jacob Grimm (1822), and Franz Bopp (1816) began to systematize language comparisons, and when reading Schottel's book one can easily see that he lacks the systematic understanding of language change as a regular process, which layed the foundation for historical linguistics as a scientific
discipline in the 19th century. For this reason, it is difficult to tell exactly what Schottel wanted to show with his schema of the Germanic languages and dialects.

Yet, it is obvious in his book that Schottel had some idea of language diversification as a historical process, and (in my opinion) also in his classification schema. He writes, for example, that Germanic languages like Norwegian, Danish, and Gothic are only remotely "Teutsch" (Germanic), due to the blurred pronunciations ("unkentlich Machung") and introduction of foreign words ("Einmengung der frömden Wörter), thus pointing to processes by which the languages diverged from the Germanic "ideal". Similarly, he mentions that the old German pronunciation is more easily perceivable in the Lower German and Lower Saxon languages ("darin die alte Teutsche Ausrede mehr zu spüren"). Moreover, on page 152 in the laudation, Schottel mentions explicitly a split of the former Germanic language into a High German and a German branch ("Teutsche und the-ho-uetsche (Hochteutsche) Sprache"), and even mentions the sound change from [t] to [z] (compare German zwei vs. English two) that reflects this split.

So, even if the schema reflects the typical uncertainty between static classification and dynamic genealogy, Schottel's work clearly shows tendencies of historical thinking. And for this reason, I would say that the current score for early phylogenies is 2 for linguistics versus 0 for biology, at least as far as the 17th century is concerned. But I am convinced that the last word on this "battle" for priority between biology and humanities has not yet been spoken!
References


Gallet, F. (1800) Arbre Généalogique des langues mortes et vivantes [The Genealogical Tree of Living and Dead Languages]. Illustration.


Rask, R. (1818) Undersøgelse om det gamle Nordiske eller Islandske sprogs oprindelse [Investigation of the origin of the Old Norse or Icelandic language]. Gyldendalske Boghandlings Forlag: Copenhagen.


Monogenesis, polygenesis, and militant agnosticism

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

When playing the cognate hunting game or the etymology identification game in historical linguistics, there are many different rules that one needs to keep in mind. Words that look similar are not necessarily related — they could be simple look-alikes (Trask 2000:202). If words are too similar, they could be borrowings. If we quote colleague X from the camp of linguists believing in theory $t_1$ we should make sure that we also quote colleague Y from the camp of linguists believing in the theory $t_2$, especially if we do not know the peer reviewers, etc.

A particularly important rule that is often surprising for biologists is the rule that says we can only compare languages that we know are related. We could, of course, compare all languages in the world (and people do compare all languages in the world), but the point is that we are not allowed to compare languages historically unless we know whether they share a common origin. This rule is reflected in a long-standing debate regarding the question of how we can prove that two languages are related. Here, we have basically two opposing camps, one claiming that only grammar can prove language relationship, and one claiming that only the lexicon is suitable for that task (Dybo and Starostin 2008, Campbell and Poser 2008).

That we have to prove that two or more languages are related before we can start to compare them is in strong contrast to biology. The idea of multiple origins as an alternative to a single origin itself has also been discussed in evolutionary biology (David has shown this in an earlier blogpost dealing with networks with multiple roots). In linguistics, however, we are largely agnostic regarding the common origin of all languages, and the degree of agnosticism may go even so far that it acquires a missionary zeal. Attempts to explain how language evolved, that is, how language originated as a means for communication, always run the danger of being ridiculed by the linguistic community. Under very bad circumstances, they can even cast a very dark shadow on the linguistic reputation of those who proposed them.

Affirming our disinterest in the origin of language has a long tradition. In its Statuts from 1866 (published in 1871), the Société de Linguistique de Paris declared that it would not support any research on the origin of language. Even August Schleicher, the father of the language tree, affirmed this attitude in a letter to Ernst Haeckel (Schleicher 1863: 22), where he wrote:

It is impossible to presuppose a material descent of all languages from a single proto-language. (My translation, original text: "Eine so zu sagen materielle Abstammung aller Sprachen von einer einzigen Ursprache können wir also unmöglich vorseussen.")

Although it is not explicitly spelled out nowadays, these statutes are still active in most linguistic institutes.

Being agnostic about the origin of language means that we cannot exclude the possibility that two
languages, like, say, Chinese and English, are ultimately not related at all. And if they are ultimately not related, it would be futile to compare them with the hope to find linguistic material that goes back to their common ancestor. Biologists, who usually take the Tree of Life for granted (albeit a bush in the end), might ask themselves for the reasoning behind this agnosticism in linguistics. The reasons are rather simple to state: If we make the very conservative assumption, based on archeological records, that human language originated about 100,000 years ago (Dediu and Levinson 2013), and contrast it with the first written records of languages (about 5,000 years ago), and the presumed time depths of our current comparative method (Meillet 1925, Weiss 2014), which optimistically allows us to reach out 10,000 years back in time, we simply do not have the means to make any qualified linguistic hypothesis regarding the origin of all those 7,000 and more languages spoken today (count based on Hammarström et al. 2015).

The reasons why linguists prefer to maintain an agnostic attitude are completely comprehensible for me. Whether it is good to be agnostic, is another question. And whether it is good to be as militant as are some linguists regarding the question of language origin is yet another one. For the context of evolutionary biology, for example, a little bit of agnosticism regarding the Tree of Life might bring up interesting dynamics. The same could be said about a little bit of "faith" in linguistics, be it that one believes that language originated independently in multiple places at the same or different times, or be it that one supports a monophyletic origin of a "Language of Eden". Neither of the theories has immediate impact on the way we pursue our historical comparison of languages. Even under a monogenesis assumption we would still need to prove a close affinity between languages before we could start comparing them with our traditional methods.

In the long run, however, it might help us to get some of the tension out of our long-standing debates. If we took monogenesis for granted, for example, people would be less afraid of comparing random pairs of languages, and in the long run we could gain new insights into distant relationships. If we rejected monogenesis, on the other hand, we could try to identify how many times language originated independently.

It is (and here you see my own agnostic attitude) not really important whether we stick to monogenesis or polygenesis in the end. What is important is that we are clear about the consequences that either of these two theories might have on our research in the future. Agnosticism is a useful attitude as long as it does not prevent us from asking questions. Following up on David's earlier blogpost, it seems clear to me that especially linguists might profit a lot from rooted network approaches that allow us to keep our agnosticism without suppressing our curiosity.

References

I am not a hard-core science fiction fan. I have not even watched the new Star Wars movie yet. But I am quite interested in all kinds of issues involving artificial intelligence, duels between humans and machines, and also the ethical implications as they are discussed, for example, in the old Blade Runner movie. It is therefore no wonder that my interest was caught by the recent Go-Game human-machine challenge.

Silver et al. (2016) reported in an article about a new Go program, called AlphaGo, that defeated other Go programs with a rate of 99.8%, and finally also defeated the European Go champion, Fan Hui, in 5 matches with 5 to 0. They proudly report in their paper (p. 488):

This is the first time that a computer Go program has defeated a human professional player, without handicap, in the full game of Go — a feat that was previously believed to be at least a decade away.

The secret of the success of the new Go program seems to lie in a smart workflow by which the neural networks of the program were trained. As a result, the program could afford to calculate "thousands of times fewer positions than Deep Blue did in its chess match against Kasparov" (Silver et al. 2016: 489).

I should say that I was never really interested in the Go-game before. My father played it once in a while when I was a child, but I never understood what one actually needs to do. From the articles in the media in which this fight between man and machine was reported, I learned, however, that the Go-Game was apparently considered to be much more challenging than the Chess Game, due to an increased number of positions and moves, and that nobody was expecting the time to be already ripe for machines to beat humans in this task.

When reading the article and reflecting about it, I wondered how complicated the task of finding homologous words in linguistic datasets might be compared to the Go-Game. I know quite a few colleagues who consider this task as impossible to model; and I know that they have not only good reasons, but also a lot of experience in language comparison, so they would not say this without having given it some serious thoughts. But if it is impossible for computer programs to compete with humans in language comparison, does this mean that the Go-Game is a less challenging task?

On the other hand, I know also quite a few colleagues who consider automatic data-driven approaches in historical linguistics to be generally superior to the classical manual workflow of the comparative method (Meillet 1925). In fact, the algorithms for cognate detection that I developed during my PhD (List 2014) are often criticized as lacking the stochastic or the machine-learning component, since they are based on a rather explicit attempt to model how historical linguists compare languages.

Among many classically oriented linguists there is a strong mistrust regarding all kind of automated approaches in historical linguistics, while among many computationally oriented linguists and linguistically oriented computer scientists there is a strong belief that enough data will sooner or later
solve the problems, and that all explicit frameworks with hard-coded parameters are inferior to data-driven frameworks. While classical linguists usually emphasize that the processes are just too complex to be modeled with simple approaches as they are used by computational linguists, the computational camp usually emphasizes the importance of letting "the data decide", or that "the data is robust enough to find signal even with simple models".

Given the success of AlphaGo, one could argue that the computational camp might be right, and that it will be just a matter of time until manual language comparison will be done in a fully automated manner. Our current situation in historical linguistics is somewhat similar to the situation in evolutionary biology during the 1960s and 1970s, when quantitative scholars prophesied (incorrectly, so far) that most classical taxonomists would soon be replaced by computers (Hull 1990: 121f).

However, since we are scientists, we should be really careful with any kind of orthodoxy, and I consider as problematic both the blind trust in machine learning techniques as well as the blind trust in the superiority of human experts over quantitative analyses. The problem with human experts is that they are necessarily less consistent and efficient than machines when it comes to tasks like counting and repeating. Given the increasing amount of digitally available data in historical linguistics, we simply lack the human resources to pursue classical research without trying to automatize at least parts of it.

The problem of computational approaches, and especially machine-learning techniques, however, is that they only provide us with a result of our analysis, not with an explanation that would tell us why the result was preferred over alternative possibilities. Apparently, Go players now have this problem with AlphaGo, since in many cases they do not know why the program made a certain move, they only know that it turned out to be successful. This black-box aspect of many computational approaches does not necessarily constitute a problem in practical applications: When designing an application for automatic speech recognition, the users won't care how the application recognizes speech as long as it understands their demands and acts accordingly. In science, however, it is not just the results that matter, but the explanation.

This is especially important in the historical sciences, where we investigate what happened in the past, and we constantly revise our knowledge about the past events by adjusting our theories and our interpretation of the evidence. If a machine tells me that two words in different languages are homologous, it is not the statement which is interesting but the explanation. Without the explanation, the statement itself is worthless. Since we are dealing with statements about the past, we can never really prove any statement that has been made. But what we can do is investigate explanations and compare the evolution of explanations in the past, thereby selecting those explanations that we prefer, perhaps because they are more probable, more general, or less complicated. A black-box method for word homology prediction would only make sense if we could evaluate the prediction — but if we could evaluate the prediction, we would not need the black-box method any more.

This does not mean that black-box methods are generally useless. A well-trained homology prediction machine could still speed up the process of data annotation, or assist linguists by providing them with
initial hints regarding remotely related language families. But as long as black-box methods remain black boxes, they won't be able to replace the only ones who could still interpret what they produce.

References


Alignments and phylogenetic reconstruction in linguistics and biology

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

In a very interesting article from 2009 (Morrison 2009), David discusses the question of why phylogeneticists would "ignore computerized sequence alignment". This article was really interesting to me for two reasons: First, the article provides some interesting statistics regarding the degree to which biologists manually adjust the alignments that were automatically produced by software. Second, the article points to the seemingly strange situation in biology in which tree-building is considered to be a task that can be entirely carried out by machines, while the majority of scholars would not trust their final sequence alignments to a computer (Morrison 2009: 150).

This situation finds a direct analogon in historical linguistics. Phylogenetic reconstruction is gaining more and more ground, with many scholars applying (mostly Bayesian) phylogenetic tools to analyze their data (Indo-European: Bouckaert et al. 2012, Tupí-Guaraní (South America): Michael et al. 2015, Japonic: Lee and Hasegawa 2011, Pama-Nguyan (Australian): Bowern and Atkinson 2012, Semitic: Kitchen et al. 2009, Bantu: Grollemund et al. 2015, etc.). Fully automated workflows involving automatic sequence comparison are also practiced (Holman et al. 2011, Jäger 2015, Wheeler 2015), but many linguists remain sceptical regarding their results.

One major difference between biology and linguistics is the selection of comparanda. Biological methods usually derive phylogenetic trees from multiply aligned sequences. Linguistic methods derive trees from sets of homologous (cognate) words (cognate sets) distributed across languages whose evolution is modeled as a process of word-gain and word loss (similar to gene-family gain-loss-studies in biology). While biologists fiddle with their alignments, linguists fiddle with their cognate sets. Cognate identification is exclusively done manually at the moment, and scholars use all kinds of information about word relations that they can get, be it etymological dictionaries, which have been published for more than 200 years, or the intuition of the expert who is annotating the data for cognacy.

Identification of cognate sets in linguistics is essentially a task of sequence comparison (List 2014), and algorithmic as well as manual procedures involve the multiple and the pairwise alignment of words (even if it is done only implicitly by human experts). Compared to biology, sequence comparison in historical linguistics is exacerbated by two factors:

- alphabets (phoneme systems) in linguistics are themselves mutable (Geisler and List 2013), so that when aligning two words we need to find both a mapping between the two alphabets, translating one alphabet into the other, plus a scoring function by which we can score the alignment,

- regular sound change (the process by which the phoneme system is changed) and sporadic sound change (the process by which a sound is sporadically assimilated, lost, or added) are not the only processes that contribute to change of words in the lexicon, and morphological change (by which whole blocks of meaningful parts of a word are re-arranged, exchanged, lost, or added) yields patterns that are essentially unalignable.
The problem of finding the correct mapping between two alphabets in linguistics is further exacerbated by language contact: If languages exchange words on a large scale, then this may have a huge impact on the system of the languages, and it may even introduce new sounds to a language that were not there before (thanks to English, German has now the sound [dʒ], as in journalist or job). If borrowing is frequent enough, it may get close to impossible to judge from comparing the words alone, whether two words in different languages have been transferred directly (vertically) from an ancestral language, or laterally.

As a result, it is probably understandable why linguists often refuse to carry out full alignments of the words in their data. An alignment itself does not necessarily tell us much, compared to all of those processes that an expert infers when comparing language data, which are not alignable.

As an example, let us consider the word for "sun" in six Indo-European languages. Since "sun" is a very basic concept, probably fundamental for all human cultures, experts assume that this word was present as *sēh₂el- in Indo-European (an asterisk indicates that the word is not reflected in written sources), and that it was retained as Russian солнце [sənʲtsə], Polish słoneczko [swɔɲtʃɛk], French soleil [sɔlɛl̻], Italian sole [sole], German Sonne [sɔnə], and Swedish sol [suːl] (Wodtko et al. 2008). An obvious alignment, reflecting the surface similarity between all of these words, would be the following one (taken from List 2014: 135):

<table>
<thead>
<tr>
<th>Language</th>
<th>Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Russian</td>
<td>s - ó n t ə -</td>
</tr>
<tr>
<td>Polish</td>
<td>s w ó n j t ə -</td>
</tr>
<tr>
<td>French</td>
<td>s - ó l - ɛ j</td>
</tr>
<tr>
<td>Italian</td>
<td>s - o l - ɛ -</td>
</tr>
<tr>
<td>German</td>
<td>s - o n - ə -</td>
</tr>
<tr>
<td>Swedish</td>
<td>s - u: l - - -</td>
</tr>
</tbody>
</table>

Alignment based on sequence similarity.

This alignment, however, is by no means correct. Russian [sɔnʲtsə] and Polish [swɔɲtʃɛk], for example, share a common suffix, which is reflected as [ntʃə] in Russian and as [nʃɛk] in Polish, and which was innovated in the the common ancestor of Russian and Polish, but is not present in either of the four other languages. So the [n] in German [sɔnə] is essentially not homologous with the [n] in Russian or the [n] in Polish. The same applies to the [ɛ] in French [sɔlɛj] which reflects a diminutive suffix in Latin sol-ículus "small sun", the regular ancestor form of French soleil. Furthermore, the [w] in the Polish word regularly corresponds to the [l] in French, Italian, and Swedish, but it reflects a swap (metathesis) in the order of the vowel and the consonant in Polish — [sɔl] became [sɬ] which became [swɔ]).

Taking all (and more) of this into account, we need to modify our alignment to account more closely for the processes that experts have inferred from intensive language comparison, as shown in the next figure below (taken from List 2014: 135). In this alignment, the swap in Polish is reflected by the white font of the sounds involved, and gray-shaded columns are supposed to reflect the oldest layer of homology.
Historically informed alignment.

However, even this alignment is essentially misleading. The Indo-European word for "sun" supposedly had a complex paradigm in which the word's stem was alternating in the nominative (and accusative) case and the other cases (oblique cases). So, nominative and accusative used the stem *sōh₂wel-, while the other cases used the stem *sh₂ěn-. The Russian, Polish, French, Italian, and the Swedish form go back to the former, while the German form goes back to the latter, since it is further assumed (or it can be assumed) that the alternation was still preserved in the ancestor of Swedish and German.

This means, however, that our alignment above shrinks to an alignment in which only the first letter, the s, is still reflected in all languages! The following graphic (taken from List 2016) illustrates the processes that led to the current situation for four of our six languages:

<table>
<thead>
<tr>
<th>Language</th>
<th>Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Russian</td>
<td>sōh₂wel-</td>
</tr>
<tr>
<td>Polish</td>
<td>sōh₂wel-</td>
</tr>
<tr>
<td>French</td>
<td>sh₂ěn-</td>
</tr>
<tr>
<td>Italian</td>
<td>sh₂ěn-</td>
</tr>
<tr>
<td>German</td>
<td>sh₂ěn-</td>
</tr>
<tr>
<td>Swedish</td>
<td>sh₂ěn-</td>
</tr>
</tbody>
</table>

Morphological processes of lexical change.

What does this example tell us? On the one hand, it gives some explanation for why linguists do not really want to align words (although the first alignments go back to the early 20th century, cf. Dixon and Kroeber 1919). It also explains, why classical linguists have a very sceptical attitude towards the computerization of word comparisons, based on the (partially justified) assumption that computers could not handle the complex patterns that are so characteristic of language change.

On the other hand, comparing the situation with biology as reported in Morrison (2009), we can find an
interesting parallel between the two disciplines: both linguists and biologists do not really trust machines for comparing their sequences (albeit at different levels of analysis), but they do not seem to have many problems in trusting machines to reconstruct their trees.

However, especially this last point, the fact that we trust machines to grow our trees, while we distrust them to prepare the seeds, should ring an alarm bell. First, we seem to lack clear guidelines (at least in linguistics) regarding the way the manual adjustment (of alignments in biology and cognate sets in linguistics) should be carried out, which has a clear impact on repeatability. Second, if we have processes in both fields that yield essentially unalignable patterns, such as duplications and other molecular processes in biology (Morrison 2009: 156), and morphological processes in linguistics, how can we assume that a phylogenetic tree analysis can sufficiently cope with them, even if we manually adjust everything?

References


Can biologists learn from linguists?

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

Of course they can. Biologists who know nothing about linguistics can learn a lot about linguistics from linguists, including the most nerdy, the most boring, and the most interesting things.

However, it is obvious that the question in the title of this post implies a different object of learning, and a more precise title would have been "Can biologists learn about evolution from linguists?" As a linguist, I would of course also provide an affirmative answer, but I doubt that most biologists would agree. At the moment, we have a situation in which the majority of interdisciplinary papers state that linguists can learn from biologists. The opposite, that biologists can learn from linguistics, can rarely be found.

Biology to linguistics

An abundance of analogies between biology and linguistics has been noticed so far, and new analogies are regularly being proposed. When looking at the analogies that have been made so far, we find that most of them have never been really followed up. Languages, for example, have been compared with organisms (Schleicher 1848: 16f), species (Pagel 2009), microbes (Nelson-Sathi et al. 2011, List et al. 2014), mutualist symbionts (van Driem 2004), and populations (Mufwene 2001). Words have been compared with cells (Schleicher 1863: 23f), amino-acids (Zwick 1978), codons (Enguix et al. 2012, Jakobson 1973) and genes (Pagel 2009). Sounds (phonemes) have been compared with nucleic bases (Hruschka et al. 2015, Enguix et al. 2012) and atoms (Zwick 1978). Only a small number of these analogies have received broader attention, many have been rejected quickly after they were first proposed, and only recently has an explicit transfer of methods and models been initiated (Atkinson and Gray 2005).

The tenor of most recent studies, especially in the literature published during the past one to two decades, is often that we finally realize that language evolution is largely the same as biological evolution, surprisingly (for a recent account in this direction, see Pagel 2016). As a result, it is claimed that we can easily use biological methods to study language evolution. We need to use them, since linguistics is in a poor state with no methods of its own, and linguists have never quantified what they know about the history of their languages. Then, finally, with these new methods developed in biology, we see light at the end of the tunnel, and we can draw nice trees of our languages and see how they evolved into their current shape.

I am in complete favour of increasing the objectivity in historical linguistics, making it a more data-driven and a more transparent discipline. I also advocate interdisciplinary transfer of methods and models, and there are quite a few things we can actually learn from biologists in linguistics. What I do not like is this tone, which suggests that biology is the discipline that saved linguistics, waking it up from its 200-year-long sleep in the ivory tower. At the same time, I also do not like the horror-scenarios
in traditional linguistics, which state that quantitative approaches would deprive our discipline of all its wit (see the figure below as a not too serious attempt to visualize these two perspectives). In this context, it is quite interesting to look back in history and to recapitulate what actually happened.

The biological storm of bits and bytes: Will it destroy the ivory tower of historical linguistics or ultimately help it to shine with a new gloss?

The discipline of historical linguistics is about 200 years old, starting with the legendary scholarly work of people like Rasmus Rask (Rask 1818), Jakob Grimm (Grimm 1822), and Franz Bopp (Bopp 1816). Using family trees to model language history goes back to the 17th century, pre-dating the first networks in biology by one century (see David's overview in Morrison 2016). The first explicit alignments showing homologous sounds across words occur at least as early as the beginning of the 20th century (Dixon and Kroeber 1919), cladistic frameworks date back to the second half of the 19th century (Brugmann 1886), and even algorithms for tree reconstruction based on distance data occur back in the 1960s (Dyen's comment in Hymes 1960).

The discipline of historical linguistics can look back on a remarkable history of excellent scholarship. Thanks to this scholarship, we have gained invaluable insights, not only into the history of the world's languages, but also into the mechanisms that trigger linguistic diversity. It is undeniable that methods from evolutionary biology have given us some fresh insights during the past 20 years, but their actual influence is often exaggerated. On the one hand, our experience (since the quantitative turn in historical linguistics) shows that in most cases we cannot use biological methods to analyze our data directly. Instead, we need to carefully adapt them to our needs in order to get the best out of them (as I have tried to show in more detail in List 2014).

On the other hand, there is no example during the past 20 years, that I would know of, where the modern biological methods have really revolutionized our insights into language history. They have
undeniably shifted our attention towards data and quantification. They have exposed weak spots, in our argumentation, and they have forced us to restate questions that we had forgotten to ask. But no new language family has been detected, no deeper genealogies between existing languages have been proposed, and no deeper insights into human prehistory have been achieved by the use of biological methods alone. Historical linguistics has profited from evolutionary biology, but not as a small oasis in the desert that was given water and seeds by the lords of bits and bytes, but as a discipline in which scholars learned to make active and critical use of interdisciplinary approaches.

**Linguistics to biology**

This brings us back to the question of the title. Can biology learn from linguistics? It has done so undoubtedly in the past. Tree-drawing in biology, for example, was popularized by Ernst Haeckel who himself became influenced by the linguist August Schleicher (Sutrop 2012: 300). In the early days of genetics, a multitude of metaphors were borrowed from linguistics to describe biological phenomena with words like "alphabet", "word" (Gamov 1954), or "translation" (Crick 1959).

While not all biologists have been in favor of this tendency (see, for example, Shannon 1978), and the borrowing of terms does not necessarily imply methodological transfer, we also find examples for the explicit transfer of methods and theories from the linguistic to the biological domain. As an example, consider the theory of formal grammar (Chomsky 1959) which still plays a very important role in addressing certain problems in bioinformatics (Searls 1997), like RNA folding and protein structure analysis. Biological textbooks on sequence comparison still tend to include a chapter on formal grammars and their application in biology (Durbin et al. 1998).

Biology could also profit from linguistic insights in the future, and this becomes a bit clearer when we recall, what Schleicher mentioned 150 years ago (and what has been obviously forgotten since then):

> Observing how new forms descend from old ones can be done more straightforwardly and in a larger scale in linguistics than in biology. For once, the linguists have an advantage over the natural scientists. (Schleicher 1863: 18, my translation)

The advantage of linguistics, which Schleicher points out, is the availability of very concrete, very detailed, very valuable data in linguistics. This data allows us to see evolutionary forces in a detailed way of which biologists can only dream. Written sources allow us to trace the history of whole language families like Romance (and to some extent also Chinese dialects) from their ancestral speech varieties down to today. Language change is fast enough to allow us to investigate it in action. Recent topics in biology, like the importance of invoking a system perspective in evolution, have been long since debated and discussed in linguistics (Tynjanow and Jakobson 1928, since they are so much easier to detect.

In the past, when I worked intensively on the implementation of the *Minimal Lateral Network* method (Dagan and Martin 2007, Dagan et al. 2008) on linguistic data (List et al. 2014, List 2015), I stumbled upon numerous examples showing the limits of tree topology as a predictor for lateral transfer events. Given that the same necessarily also holds for lateral gene transfer, I was asking myself whether these false positives and the false negatives in the analyses would simply not matter due to the large amount
of data in biology, or whether it was ignored due to the lack of good data for algorithmic evaluation. Later, when I read David's post on Tardigrades and phylogenetic networks, where he pointed to two analyses on the same data that explained them once with lateral gene transfer (Boothby et al. 2015) and once with errors in the data (Koutsovoulos 2015), I became aware of the strong advantage of my linguistic data, since I could test it against written records, tracing the history of words through centuries, thus being able to spot errors immediately when looking up a data point.

The detail of our data in linguistics is both a blessing and a curse. It enables us to write detailed word histories without ever having heard of tree reconciliation methods. On the other hand, it seduces us to get lost in details, forgetting about the bigger picture, and the bigger questions that we could ask, if this data was properly digitized and formalized. In this regard, historical linguistics still needs to learn from biology, as we have failed to turn historical linguistics into a modern, data-driven discipline. With more and more detailed data becoming available, however, the day will come when Schleicher is proven right, and when biologists can learn from linguists about evolution.

References


More on analogies between biological and linguistic evolution

Johann-Mattis List

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

Analogies between biological and linguistic evolution have been discussed before on this blog. Last month, I asked whether biologists could learn from linguists; a bit earlier, I proposed to distinguish fruitful from unfruitful analogies; and David has written a very long and interesting blog post on false analogies between anthropology and biology.

In contrast to the discussion of similarities in many articles that have been published, most of these posts were rather sceptical and reserved, emphasizing the importance of being extremely careful when using analogies to justify methodological transfer across disciplines. Despite this general scepticism, that I mentioned myself, I am still convinced that methodological transfer can be fruitful when carefully adapting methods to the needs of the target discipline — and we know that this has been done in both directions in the fields of biology and linguistics.

Apart from the problem of adapting methods from other disciplines, one important question is, how to identify fruitful analogies in the first place. As a visiting post-doc in the bioinformatics research group Adaptation, Integration, Reticulation and Evolution, led by Eric Bapteste and Philippe Lopez (UPMC Paris), I have discussed this question a lot during the past one and a half years.

We came up with the idea that it might be useful to restrict the range of potential analogies one might draw between biology and linguistics by concentrating on analogies between processes. Taking processes, rather than research objects, as a starting point comes closer to general approaches to analogy, which usually claim that the core of analogy is similarities of functions (Gentner 1983). By applying this principle to compare aspects of linguistic and biological evolution, we were able to identify some potentially fruitful analogies that could lead to novel approaches, not only in linguistics but potentially also in biology.

Among these are specific processes of divergence (like incomplete lineage sorting in biology, which is very similar to dialect chain dissolution in linguistics), specific introgressive processes (like protein assembly, which shows some striking similarities with word formation), and specific systemic processes (like constructive neutral evolution in biology, providing an explanation for convergent evolution in languages resulting from common descent, also called drift or Sapir's drift). On the other hand, we also found that many processes are most likely to be unique to one of the disciplines, including such processes as sound change in linguistics and natural selection in biology.

These reflections have been summarized in a paper titled "Unity and disunity in evolutionary sciences"which was published at the beginning of this week (List et al. 2016, PDF here). I will not go into further detail of the specific new analogies we proposed, but instead recommend those who are interested in the issue to read our paper (and potentially discuss the issue of analogies further with us).
Since the identification of potentially fruitful, new analogies between biology and linguistics is just a starting point for a closer investigation of the suitability of methodological transfer in practice, I am quite optimistic that I will follow up on the new analogies mentioned above in more detailed future blog posts.

References


Inheritance in cultural evolution

Johann-Mattis List

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

I recently reviewed a book anthology devoted to the application of phylogenetic methods in archaeology (see List 2016, PDF here). This book, entitled Cultural Phylogenetics: Concepts and Applications in Archaeology, edited by Larissa Mendoza Straffon (2016), assembles eight articles by scholars who discuss or illustrate the application of phylogenetic approaches in different fields of anthropology and archaeology.

The volume presents a rich collection of different approaches, covering various topics ranging from the evolution of skateboards (Prentiss et al.) to the spread of the potter's wheel (Knappett). The articles dealing with theoretical questions range from historical accounts of tree-thinking in biology and anthropology (Kressing and Krischel) to an overview of the impact of Darwinian thinking on archaeology and anthropology (Rivero). Although I missed a golden thread when reading the eight articles of the volume, it is definitely worth a read for those interested in evolutionary approaches in a broader sense, as most articles explicitly reflect differences and commonalities between biological and cultural evolution, providing concrete insights into the challenges that archaeologists face when trying to promulgate quantitative approaches.

It is clear that evolution in the general sense is much broader than merely evolution in biology, as I have often tried to illustrate in this blog when showing how phylogenetic approaches can be applied in linguistics. Provided that descent with modification holds — in a broader sense — also for cultural artifacts, it is obvious to search for fruitful analogies between biological and cultural evolution, in order to profit from methodological transfer in disciplines like anthropology and archaeology. It is also clear, however, that certain analogies between biological evolution and evolution in other fields should be considered with great care. Even in linguistics, this is clearly evident, and I have pointed to this problem in the past (see Productive and unproductive analogies...). The goal cannot be a to try to press biological methods into the anthropological template. Instead, we have to rigorously test our proposed analogies, and adapt the biological methods to our needs if necessary.

What surprised me when reading the book was that the majority of the articles did not really seem to care about the crucial differences between biological and cultural evolution, but rather tried to fit the feet and heels of cultural evolution into biology's shoes. Tree thinking dominated most of the articles (with Knappett as a notable exception), and the scholars tried hard to find a clear distinction between vertical and lateral inheritance in cultural evolution. While it is clear that this distinction is the basis for phylogenetic tree applications, where patterns that do not fit a tree are explained as instances of homoplasy or lateral transfer, it is by no means clear why one would go through all the pain to identify these patterns in cultural evolution.

Consider, as an example, the evolution of skateboards. At some point in the history of mankind (some late point!), people decided to put wheels on a board and to do artistic tricks with it. Later, other people
merchandised this idea, and started to sell those boards with wheels. Later on, other companies jumped on the bandwagon and started to produce their own brands, thus instigating a fight for the "best" model for a certain kind of clientel. In all of these cases, ideas for design were clearly taken among groups of people, further modified by specific needs or trends, until the current variety of skateboards arose. But which of these ideas were transferred vertically, and which ideas were transferred laterally? Can we identify processes of "speciation" in skateboard evolution, during which new brands were born?

In biology and linguistics we have the clear-cut criteria of interfertility and intelligibility. They cause us enough problems, given that we have ring species in biology and dialect chains in linguistics, but at least they give us some idea how to classify a given exemplar as belonging to a certain group. But what is the counterpart in the evolution of skateboards? Their brand? Their shape? Their users? The analogy simply does not hold. We have neither vertical nor lateral transfer in topics such as skateboard evolution. All we have is a before and an after—a complex network in which objects were constantly recreated and modified, be it based on ideas that were inspired by other objects or people, or independently developed. It seems completely senseless to search for a distinction between vertical and lateral patterns here, as it is not even clear to what degree we are actually dealing with decent with modification.

It seems to me that the problem of inheritance needs to be addressed in cultural evolution before any further quantitative applications using tree-building methods are carried out. Given that ideas can easily be develop independently, the crucial question for studies of cultural evolution is whether similar ideas can be shown to share a common history. It is (as David mentioned in earlier in a blog post on False analogies between anthropology and biology) the general problem of homology that does not seem to be solved in most studies on cultural evolution. Here, linguistics has generally fewer problems, given that linguists have developed methods to test whether two words are homologous. In cultural evolution, however, the assessment of homology is far from being obvious.

I think that cultural evolution studies such as the ones presented in the book would generally profit from network approaches. By network approaches, I do not necessarily mean evolutionary networks (in the sense of Morrison 2011), as the problem of inheritance is difficult to solve. Instead, I am thinking of exploratory data analysis using phylogenetic networks (Morrison 2011), or some version of similarity networks (Bapteste et al. 2012). Phylogenetic network approaches are frequently used in biology, and are now also very popular in linguistics. Similarity networks are more common in biology, but we have carried out some promising studies of linguistic data (List et al. 2016). As all of these approaches are exploratory and very flexible regarding the data that is fed to them, they might offer new possibilities for exploratory studies on cultural evolution.

References


---

I have been discussing the peculiarities of sound change in linguistics in a range of blog posts in the past (see Alignments and Phylogenetic Reconstruction, Directional Processes in Language Change, Productive and Unproductive Analogies). My core message was that it is really difficult to find an analogy with biology, as sound change is not the simple mutation of one sound in a certain word, but the regular modification of all sounds of all words in the lexicon which occur in a specific contextual slot.

Scholars have tried to model this as concerted evolution (Hruschka et al. 2015). But the analogy with biology does not sound very convincing, as the change concerns the production of speech rather than its product. By this, I mean that sound change concerns the abstract system by which speakers produce the words of their language. Think of speakers in comic books who lose a tooth in some fight. Often, in order to show how their speech suffers from this loss, writers illustrate this by replacing certain "s" sounds in the speech of the victims with a "th" (in German, it would be an "f"). They do this in order to illustrate that with a lost tooth, it is "very difficult to thpeak". In the same way, writers imitate speech of people suffering from speech impediments like sigmatism (lisp). The loss of a tooth changes all "s"es in a person's language. Sound change, at least one type of sound change, is identical with this.

In a recent talk I gave with Nathan Hill at a conference in Poznań, we found a way to demonstrate this on actual language data. In this talk, we used data from eight Burmish languages (a language family spoken mainly in the South-West of China and in Myanmar), which we coded for partial cognates (as these languages contain many compounds). We aligned these cognate sets automatically, and then searched for recurring patterns in the alignments. One needs to keep in mind that our words in linguistics are extremely short, and we have no more than five sounds per alignment in our data, which translates to five sites in an alignment in biology.

While biology knows certain contextual patterns like hydrophilic stretches in alignments (as already demonstrated in the famous ClustalW software, compare Thompson et al. 1994), the context in which a sound occurs in language evolution is even more important. We can, for example, say, that the beginning of a word or morpheme is usually the most stable part, where sounds change much more slowly than in the other parts (in the end of a word or of a syllable). We thus concentrated only on the first sound of each word and looked at the patterns of sounds we could find there.

Those patterns in our data usually look like this:
Note that the symbol "Ø" in this context denotes missing data, as we did not find a cognate set in the given language. As always, most of our data is patchy, and we have to deal with that. You can see that when looking only at the first sound in each alignment, we find quite a degree of variation; and if you look at all the data, you can see some things that seem to structure, but the amount of complexity is still immense. You may see this from the following plot, showing only some 100 of the more than 300 patterns we created (coloured cells represent not necessarily the same sound, but one of ten different sound classes to which the more than 50 different sounds in our data belong):

Sound patterns (initial consonant) in the aligned cognates sets of the Burmish languages

Interestingly, however, most of the variation can be reduced quite efficiently with help of network techniques. Since we are dealing with systemic evolution, it is straightforward to group our more than 300 alignments into groups that evolve in an identical manner. At least this is what our linguistic theory predicts, and what linguists have been studying for the last 200 years. When looking at the patterns I gave above, you can see that we can easily group the four sounds into two groups:

<table>
<thead>
<tr>
<th>Cognate set</th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
<th>L4</th>
<th>L5</th>
<th>L6</th>
<th>L7</th>
<th>L8</th>
</tr>
</thead>
<tbody>
<tr>
<td>word 1</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>Ø</td>
<td>f</td>
<td>f</td>
<td>Ø</td>
<td>p</td>
</tr>
<tr>
<td>word 2</td>
<td>p</td>
<td>Ø</td>
<td>p</td>
<td>p</td>
<td>Ø</td>
<td>f</td>
<td>p</td>
<td>p</td>
</tr>
<tr>
<td>word 3</td>
<td>k</td>
<td>Ø</td>
<td>tɕ</td>
<td>k</td>
<td>s</td>
<td>k</td>
<td>Ø</td>
<td>k</td>
</tr>
<tr>
<td>word 4</td>
<td>Ø</td>
<td>k</td>
<td>tɕ</td>
<td>Ø</td>
<td>s</td>
<td>Ø</td>
<td>s</td>
<td>k</td>
</tr>
</tbody>
</table>
 Essentially, the two groups reflect only two patterns, if we disregard the gaps and merge them into one row each:

<table>
<thead>
<tr>
<th>Cognate set</th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
<th>L4</th>
<th>L5</th>
<th>L6</th>
<th>L7</th>
<th>L8</th>
</tr>
</thead>
<tbody>
<tr>
<td>word 1 / p</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>f</td>
<td>f</td>
<td>p</td>
<td>p</td>
<td></td>
</tr>
<tr>
<td>word 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>word 3 / k</td>
<td>k</td>
<td>tɕ</td>
<td>k</td>
<td>s</td>
<td>k</td>
<td>s</td>
<td>k</td>
<td></td>
</tr>
<tr>
<td>word 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

What is important when grouping two alignments into one pattern is to make sure that they do not contain any conflicting positions. This can be checked in a rather straightforward manner by constructing a network from the data. In this network, the nodes are the alignment sites (word 1, word 2, etc. in our examples), and links are drawn between nodes if two sites are not in conflict with each other. If we use this criterion of compatibility on our data, we receive following network:

![Compatibility network of the sites in our aligned cognate sets](attachment:compatibility_network.png)

In the network, I further coloured the nodes according to the overall similarity of sounds present in them. The legend gives capital letters for major sound classes, in order to facilitate seeing the structure. This network itself, however, does not tell us how to group the data into classes that correspond to one identical process of systemic evolution, as we can still see many conflicts. In order to solve this, we need to carry out a specific partitioning analysis that cuts the network into an ideally minimal number of cliques. Why cliques? Because a clique will represent patterns in our data that do not show any conflicts in their sounds, and this is exactly what we want to see: those patterns that behave identically,
The problem of finding the minimal clique partition of a network is, unfortunately, a hard one (see Bhasker and Samad 1991), so we needed to use some approximate shortcuts. Nevertheless, with a very simple procedure of clique partitioning, we succeeded at reducing the 317 cognate sets that we selected for our study down to 35 groups that covered 74% of the data (234 cognate set), with a minimal size of 2 alignments per group. The "manual" inspection by the Burmish expert in our team (that is Nathan Hill) showed that many of these patterns correspond to what experts assume was one single sound in the ancestral Proto-Burmish language.

But to just illustrate more closely what I mean by reducing patterns to unique groups, look at the following pattern, which shows different nasal sounds in the data:

Nasal sounds in the Burmish data

And then at another pattern, showing s-sounds:

S-sounds in the Burmish data

I think (at least I hope) that the amount of regularity we find here is enough to demonstrate what is meant by the *regularity of sound change* in linguistics: sound change is in some sense just like losing a tooth, but for a complete population of speakers, not just one speaker, as the population starts to change all sounds occurring in a certain environment to some other sound.

Our results are not perfect: the 26% of unique patterns, for example, are something we will need to look into in more detail in the near future. A quick check showed that they may result from errors in the
cognate annotation, but also from peculiarities in the data, and even simply from sounds that are rare in the languages under investigation.

We are currently looking into these issues, trying to refine our approach. I realized, for example, that the minimal clique coverage problem has been studied before by other researchers, and I found a rather large amount of Russian literature on the topic (see, for example, Bratceva and Čerenin 1994 and Ryzhkov 1975), but those approaches do not seem to have been thoroughly studied in the Western literature. We also know that at some point we need to relax our approach, allowing for some exceptions — we know that systemic sound change processes are easily overridden by language-specific factors, be it lateral transfer, or pragmatics in a larger sense (think of Bob Dylan, talking of "the words I never KNOWED" in order to make sure the word rhymes with "ROAD", or the form "wanna" as a shortcut for "want to").

Not all cases in which speakers changed the pronunciation of sounds have systemic reasons, and we are still far from actually understanding the systemic reasons that lead to the regular aspects of sound change. What we can show, however, is that sound change is really something peculiar in language evolution, with no real counterpart in biology. At least, I do not know of any case where a set of 300 alignments could be reduced to some 35 largely identical patterns. This shows, on the other hand, that the classical biological approaches that try to model each site of an alignment independently are definitely not what we need in order to model sound change realistically. The assumption of independence of sites in an alignment is already problematic in biology. In linguistics, at least in the cases illustrated above, it seems to be just as useless as tossing a coin to predict the weather in a desert: it is too much of an effort with very poor results to be expected.

References


Once more on artificial intelligence and machine learning

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

In an earlier blog post, I expressed my scepticism regarding the scientific value of non-transparent machine learning approaches, which only provide a result but no transparent explanation of how they arrive at their conclusion. I am aware that I run the risk of giving the impression of abusing this blog for my own agenda, against artificial intelligence and machine learning approaches in the historical sciences, by bringing the problem up again. However, a recent post in Nature News (Castelvecchi 2016) further substantiates my original scepticism, providing some interesting new perspectives on the scientific and the practical consequences, so I could not resist mentioning it in my post for this month.

Deep learning approaches in research on artificial intelligence and machine learning go back to the 1950s, and have now become so successful that they are starting to play an increasingly important role in our daily lives, be it that they are used to recommend to us yet another book that somebody has bought along with the book we just want to buy, or that they allow us to take a little nap while driving fancy electronic cars and saving carbon footprints for our next round-the-world trip. The same holds, of course, also for science, and in particular for biology, where neural networks have been used for tasks like homolog detection (Bengio et al. 1990) or protein classification (Leslie et al. 2004). This is true even more for linguistics, where a complete subfield, usually called natural language processing, has emerged (see Hladka and Holub 2015 for an overview), in which algorithms are trained for various tasks related to language, ranging from word segmentation in Chinese texts (Cai and Zhao 2016) to the general task of morpheme detection, which seeks to find the smallest meaningful units in human languages (King 2016).

In the post by Castelvecchi, I found two aspects that triggered my interest. Firstly, the author emphasizes that answers that can be easily and often accurately produced by machine learning approaches do not automatically provide real insights, quoting Vincenco Innocente, a physicist at CERN, saying:

As a scientist ... I am not satisfied with just distinguishing cats from dogs. A scientist wants to be able to say: “the difference is such and such.” (Vincenco Innocente, quoted by Castelvecchi 2016: 22)

This expresses precisely (and much more transparently) what I tried to emphasize in the former blog post, namely, that science is primarily concerned with the questions why? and how?, and only peripherally with the question what?

The other interesting aspect is that these apparently powerful approaches can, in fact, be easily betrayed. Given that they are trained on certain data, and that it is usually not known to the trainers what aspects of the training data effectively trigger a given classification, one can in turn use algorithms to train data that will betray an application, forcing it to give false responses. Castelvecchi mentions an experiment by Mahendran and Vedaldi (2015) which illustrates how "a network might see wiggly lines and classify them as a starfish, or mistake black-and-yellow stripes for a school bus"
Putting aside the obvious consequences that arise from abusing the neural networks that are used in our daily lives, this problem is surely not unknown to us as human beings. We can likewise be easily betrayed by our expectations, be it in daily life or in science. This, finally, brings us back to networks and trees, as we all know how difficult it is at times to see the forest behind the tree that our software gives us, or the tree inside the forest of incompletely sorted lineages.

**References**


Isogloss maps are hypergraphs are bipartite networks

Johann-Mattis List

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

Linguists are a very special people. They are very proud, especially when biologists tell them how to do phylogenetic analyses; but their pride is often also justified, as many phylogenetic concepts were initially or independently developed by linguists, be it the family tree model, proposed years before Darwin's (1859) tree by Čelakovský (1853), or even the cladistic principle of synapomorphies, which are called "exclusively shared innovations" in linguistics (see Brugmann 1884).

Linguists also invented one interesting kind of data-display which so far has never been used by biologists (at least as far as I know): maps of isogloss boundaries. The term "isogloss" is an unfortunate term, as it has multiple usages in linguistics, and its history seems to go back to a naive borrowing from chemistry (but I have not really followed the literature here). On most occasions, it just means "shared trait". That is, it denotes a features shared between two or more languages; and given that languages may share many different features, isoglosses for a group of related languages may yield a very complex type of data. Isoglosses are somehow related to the wave theory, the arch-enemy of the family tree in linguistics, which I described as a mystical theory some time ago, since it never really made it to a clear-cut model that could be formalized (The Wave Theory: the predecessor of network thinking in historical linguistics).

Some linguists, nevertheless, insist that the waves that are the core of the wave theory are nothing other than isoglosses. More specifically, the waves represent innovations that contribute to the separation of languages (a change in pronunciation of a word here, a change in grammar there), but which are not transmitted vertically — they spread across the speakers of a language and may even cross linguistic borders. One early visualization of these waves can be found in Bloomfield (1933), as shown here:
What Bloomfield essentially does here is pick certain traits of Indo-European languages, calling them *isoglosses*, and arrange them on a quasi-geographic map of Indo-European languages in such a way that all languages sharing a trait are inside one of these isogloss boundaries.

Only recently, I realised, what this actually means, when I found the "Bible of Network Theory" by Newman (2010) and started reading at a random page, which — as it turned out — treated *hypergraphs*. Hypergraphs, as I learned from Newman, are graphs in which one edge can connect to more than one node, and Newman used exactly the same visualization for these hyperedges as Bloomfield had done in 1933, without knowing that it was actually a rather complex network structure he was proposing.

Even more interesting than the complex graph structure is that hypergraphs can be likewise displayed as *bipartite networks*, in which we distinguish two fundamental kinds of nodes, and in which connections are only allowed between nodes of different kinds, without losing any information. In order to do so, one just converts all hyperedges into a node that connects to all nodes (languages in our case) to which the edges connect in the hypergraph. In the same way that Bloomfield labeled the hyperedges in his legend, we can label the isogloss nodes that connect to the languages. The following image shows the resulting bipartite network for Bloomfield's hypergraph:

![Bipartite Network Diagram](image)

If you now ask what this tells us after all, I will disappoint you — so far it does not tell us anything, it is just a display of data in a different fashion. Note, however, that hypergraph visualization is not a trivial problem, and if you have enclaves not sharing a trait, it may even be impossible to visualize hypergraphs in a two-dimensional space by just using one line that connects to all nodes. Bipartite networks are easier to handle in this regard. Even more importantly, however, bipartite graphs are also easy to handle algorithmically, and biologists are currently developing new methods to handle them (Corel et al. 2016).

If we visualize the Bloomfield data in a bipartite network using network visualization software such as Cytoscape, we can conveniently explore the data, and arrange the nodes in order to search for patterns in the isoglosses. The following visualization, for example, shows that Bloomfield chose the data well...
In order to illustrate the amount of conflicting, apparently non-tree-like, signal in Indo-European languages (remember that linguists tend to dislike trees, but not necessarily in a productive way), as the data describes more of a circular structure than a strict hierarchy.

In order to really interpret this kind of data, however, we should not forget that this is still a data-display network. It is by no means a phylogenetic analysis, as we only show how a certain amount of data selected by a scholar and distributed over the given language groups. A true phylogenetic analysis will need to interpret these data, making bold claims about the history of those shared traits.

The existence of sibilants (s-like sounds, like [s, z, ʃ, ʒ]) for certain velar sounds (k-like sounds, like [k, g, x]), for example, is a trait shared by Balto-Slavic, Indo-Iranian, Armenian, and Albanian, but this does not mean that they all inherited it from a common ancestor, as the process of palatalization, by which velar sounds turn into affricates and fricatives (compare French cent, which was pronounced kentum in Latin), is very frequent in the languages of the world, and may well reflect independent evolution.

Apart from independent development, which would actually force us to revise our network, deleting the respective edges because they are not homologous in the strict sense means that we may also have to deal with differential loss. This quite likely happened with the shared feature labeled as "past e-" in the network, referring to the past tense in Ancient Greek and Indo-Iranian, which was augmented by the prefix e-.

A further reason for those commonalities labelled as isoglosses by linguists may also be simple lateral transfer due to language contact.

Proponents of the wave theory have taken this kind of data as proof that the family tree model is essentially wrong. While I would agree that the family tree model shows only a certain aspect of language evolution, and may therefore be boring at times (and even wrong, if we do not manage to correctly interpret the nature of shared traits), I have a hard time understanding why linguists still insist that isogloss maps are an alternative model of language evolution. They are surely not, in the same way in which splits graphs are not phylogenetic networks, as David emphasized in a recent blogpost.

Unless we add the missing time dimension and analyse how the shared traits originated, isogloss maps and hypergraphs will remain nothing more than an interesting form of data visualization. Given the
recent research on bipartite networks, however, we may have some hope that the mysterious waves in historical linguistics may not only find a formal model of representation, but even bring us to the point where we gain new insights into the history of our languages.

References
