

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

2017

Contents

Similarities and language relationship	1
Models and processes in phylogenetic reconstruction	6
Why we need alignments in historical linguistics	10
The siteswap annotation in juggling, and the power of annotation and modeling	14
Killer arguments and the nature of proof in historical sciences	17
Trees do not necessarily help in linguistic reconstruction	20
More on similarities in linguistics	23
Unattested character states	26
Arguments from authority, and the Cladistic Ghost, in historical linguistics . . .	29
“Man gave names to all those animals”: cats and dogs (with G. Grimm)	35
“Man gave names to all those animals”: goats and sheep (with G. Grimm and C. Anderson)	42
The art of doing science: alignments in historical linguistics	47

Similarities and language relationship

Johann-Mattis List

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

There is a long-standing debate in linguistics regarding the best proof deep relationships between languages. Scholars often break it down to the question of *words vs. rules*, or *lexicon vs. grammar*. However, this is essentially misleading, since it suggests that only one type of evidence could ever be used, whereas most of the time it is the accumulation of multiple pieces of evidence that helps to convince scholars. Even if this debate is misleading, it is interesting, since it reflects a general problem of historical linguistics: the problem of similarities between languages, and how to interpret them.

Unlike (or like?) biology, linguistics has a serious problem with *similarities*. Languages can be strikingly similar in various ways. They can share similar words, but also similar *structures*, similar ways of expressing things.

In Chinese, for example, new words can be easily created by *compounding* existing ones, and the word for 'train' is expressed by combining *huǒ* 火 'fire' and *chē* 車 'wagon'. The same can be done in languages like German and English, where the words *Feuerwagen* and *fire wagon* will be slightly differently interpreted by the speakers, but the constructions are nevertheless valid candidates for words in both languages. In Russian, on the other hand, it is not possible to just put two nouns together to form a new word, but one needs to say something as *огненная машина* (*огнюппаа mašina*), which literally could be translated as 'firy wagon'.

Neither German nor English are historically closely related to Chinese, but German, English, and Russian go back to the same relatively recent ancestral language. We can see that whether a language allows compounding of two words to form a new one or not, is not really indicative of its history, as is the question of whether a language has an article, or whether it has a case system.

The problem with similarities between languages is that the apparent similarities may have different sources, and not all of them are due to historical development. Similarities can be:

1. coincidental (simply due to chance),
2. natural (being grounded in human cognition),
3. genealogical (due to common inheritance), and
4. contact-induced (due to lateral transfer).

As an example for the first type of similarity, consider the Modern Greek word [θεός](#) [θeos] 'god' and the Spanish [dios](#) [diɔs] 'god'. Both words look similar and sound similar, but this is a sheer coincidence. This becomes clear when comparing the oldest ancestor forms of the words that are

reflected in written sources, namely Old Latin *deivos*, and Mycenaean Greek *thehós* ([Meier-Brügger 2002: 57f](#)).

As an example of the second type of similarity, consider the Chinese word *māmā* 媽媽 'mother' vs. the German *Mama* 'mother'. Both words are strikingly similar, not because they are related, but because they reflect the process of language acquisition by children, which usually starts with vowels like [a] and the nasal consonant [m] ([Jakobson 1960](#)).

An example of genealogical similarity is the German *Zahn* and the English *tooth*, both going back to a Proto-Germanic form **tanθ-*. Contact-induced similarity (the fourth type) is reflected in the English *mountain* and the French *montagne*, since the former was borrowed from the latter.

We can display these similarities in the following decision tree, along with examples from the lexicon of different languages (see [List 2014: 56](#)):

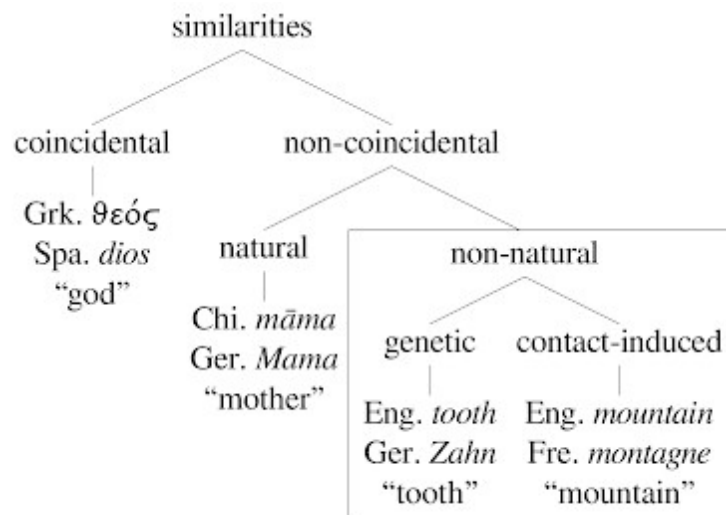


Figure 1: Four basic types of similarity in linguistics

In this figure, I have highlighted the last two types of similarity (in a box) in order to indicate that they are *historical similarities*. They reflect individual language development, and allow us to investigate the evolutionary history of languages. Natural and coincidental similarities, on the other hand, are not indicative of history.

When trying to infer the evolutionary history of languages, it is thus crucial to first rule out the non-historical similarities, and then the contact-induced similarities. The non-historical similarities will only add noise to the historical signal, and the contact-induced similarities need to be separated from the genealogical similarities, in order to find out which languages share a common origin and which languages have merely influenced each other some time during their history.

Unfortunately, it is not trivial to disentangle these similarities. Coincidence, for example, seems to

be easy to handle, but it is notoriously difficult to calculate the likelihood of chance similarities. Scholars have tried to model the probability of chance similarities mathematically, but their models are far too simple to provide us with good estimations, as they usually only consider the first consonant of a word in no more than 200 words of each language ([Ringe 1992](#), [Baxter and Manaster Ramer 2000](#), [Kessler 2001](#)).

The problem here is that everything that goes beyond word-initial consonants would have to take the probability of word structures into account. However, since languages differ greatly regarding their so-called *phonotactic structure* (that is, the sound combinations they allow to occur inside a syllable or a word), an account on chance similarities would need to include a probabilistic model of possible and language-specific word structures. So far, I am not aware of anybody who has tried to tackle this problem.

Even more problematic is the second type of similarity. At first sight, it seems that one could capture *natural similarities* by searching for similarities that recur in very diverse locations of the world. If we compare, for example, which languages have tones, and we find that tones occur almost all over the world, we could argue that the existence of tone languages is not a good indicator of relatedness, since tonal systems can easily develop independently.

The problem with independent development, however, is again tricky, as we need to distinguish different aspects of independence. Independent development could be due to: *human cognition* (the fact that many languages all over the world denote the bark of a tree with a compound *tree-skin* is obviously grounded in our perception); or due to *language acquisition* (like the case of words for 'mother'); but potentially also due to *environmental factors*, such as the size of the population of speakers ([Lupyan et al. 2010](#)), or the location where the languages are spoken (see [Everett et al. 2015](#), but also compare the critical assessment in [Hammarström 2016](#)).

Convergence (in linguistics, the term is used to denote similar development due to contact) is a very frequent phenomenon in language evolution, and can happen in all domains of language. Often we simply do not know enough to make a qualified assessment as to whether certain features that are similar among languages are inherited/borrowed or have developed independently.

Interestingly, this was first emphasized by Karl Brugmann (1849-1919), who is often credited as the "father of cladistic thinking" in linguistics. Linguists usually quote his paper from [1884](#), in order to emphasize the crucial role that Brugmann attributed to *shared innovations* (synapomorphies in the cladistic terminology) for the purpose of subgrouping. When reading this paper thoroughly, however, it is obvious that Brugmann himself was much less obsessed with the obscure and circular notion of shared innovations (which also holds for cladistics in biology; see [De Laet 2005](#)), but with the fact that it is often impossible to actually *find* them, due to our incapacity to disentangle independent development, inheritance and borrowing.

So far, most linguistic research has concentrated on the problem of distinguishing borrowed from inherited traits, and it is here that the fight over *lexicon* or *grammar* as primary evidence for relatedness primarily developed. Since certain aspects of grammar, like case inflection, are rarely transferred from one language to another, while words are easily borrowed, some linguists claim that only grammatical similarities are sufficient evidence of language relationship. This argument is not necessarily productive, since many languages simply lack grammatical structures like inflection, and will therefore not be amenable to any investigation, if we only accept inflectional morphology (grammar) as rigorous proof (for a full discussion, see [Dybo and Starostin 2008](#)). Luckily, we do not need to go that far. [Aikhenvald \(2007: 5\)](#) proposes the following *borrowability* scale:

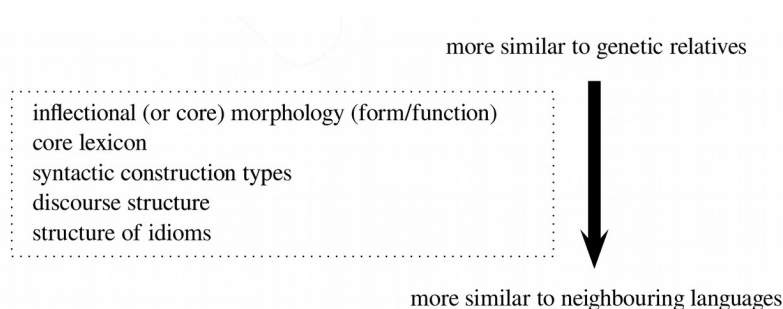


Figure 2: Aikhenvald's (2007) scale of borrowability

As we can see from this scale, core lexicon (basic vocabulary) ranks second, right behind inflectional morphology. Pragmatically, we can thus say: if we have nothing but the words, it is better to compare words than anything else. Even more important is that, even if we compare what people label "grammar", we compare *concrete* form-meaning pairs (e.g., concrete plural-endings), and we never compare abstract features (e.g., whether languages have an article). We do so in order to avoid the "homoplasmy problem" that causes so many headaches in our research. No biologist would group insects, birds, and bats based on their wings; and no linguist would group Chinese and English due to their lack of complex morphology and their preference for compound words.

Why do I mention all this in this blog post? For three main reasons. First, the problem of similarity is still creating a lot of confusion in the interdisciplinary dialogues involving linguistics and biology. David is right: similarity between linguistic traits is more like similarity in morphological traits in biology (phenotype), but too often, scholars draw the analogy with genes (genotype) ([Morrison 2014](#)).

Second, the problem of disentangling different kinds of similarities is not unique to linguistics, but is also present in biology ([Gordon and Notar 2015](#)), and comparing the problems that both disciplines face is interesting and may even be inspiring.

Third, the problem of similarities has direct implications for our null hypothesis when considering certain types of data. David asked in a recent blog post: ["What is the null hypothesis for a](#)

[phylogeny?](#)" When dealing with observed similarity patterns across different languages, and recalling that we do *not* have the luxury to assume [monogenesis in language evolution](#), we might want to know what the null hypothesis for these data should be. I have to admit, however, that I really don't know the answer.

References

- Aikhenvald, A. (2007): Grammars in contact. A cross-linguistic perspective. In: Aikhenvald, A. and R. Dixon (eds.): *Grammars in Contact*. Oxford University Press: Oxford. 1-66.
- Baxter, W. and A. Manaster Ramer (2000): Beyond lumping and splitting: Probabilistic issues in historical linguistics. In: Renfrew, C., A. McMahon, and L. Trask (eds.): *Time depth in historical linguistics*. McDonald Institute for Archaeological Research: Cambridge. 167-188.
- Brugmann, K. (1884): Zur Frage nach den Verwandtschaftsverhältnissen der indogermanischen Sprachen [Questions regarding the closer relationship of the Indo-European languages]. *Internationale Zeitschrift für allgemeine Sprachwissenschaft* 1. 228-256.
- De Laet, J. (2005): Parsimony and the problem of inapplicables in sequence data. In: Albert, V. (ed.): *Parsimony, phylogeny, and genomics*. Oxford University Press: Oxford. 81-116.
- Dybo, A. and G. Starostin (2008): In defense of the comparative method, or the end of the Vovin controversy. In: Smirnov, I. (ed.): *Aspekty komparativistiki.3*. RGGU: Moscow. 119-258.
- Everett, C., D. Blasi, and S. Roberts (2015): Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. *Proceedings of the National Academy of Sciences* 112.5. 1322-1327.
- Gordon, M. and J. Notar (2015): Can systems biology help to separate evolutionary analogies (convergent homoplasies) from homologies?. *Progress in Biophysics and Molecular Biology* 117. 19-29.
- Hammarström, H. (2016): There is no demonstrable effect of desiccation. *Journal of Language Evolution* 1.1. 65-69.
- Jakobson, R. (1960): Why 'Mama' and 'Papa'?. In: *Perspectives in psychological theory: Essays in honor of Heinz Werner*. 124-134.
- Kessler, B. (2001): *The significance of word lists. Statistical tests for investigating historical connections between languages*. CSLI Publications: Stanford.
- List, J.-M. (2014): *Sequence comparison in historical linguistics*. Düsseldorf University Press: Düsseldorf.
- Lupyan, G. and R. Dale (2010): Language structure is partly determined by social structure. *PLoS ONE* 5.1. e8559.
- Meier-Brügger, M. (2002): *Indogermanische Sprachwissenschaft*. de Gruyter: Berlin and New York.
- Morrison, D. (2014): Is the Tree of Life the best metaphor, model, or heuristic for phylogenetics?. *Systematic Biology* 63.4. 628-638.
- Ringe, D. (1992): On calculating the factor of chance in language comparison. *Transactions of the American Philosophical Society* 82.1. 1-110.

Cite as: List, Johann-Mattis (2017): Similarities and language relationship. *The Genealogical World of Phylogenetic Networks* 6:1, 1-5, URL: <http://phylonetworks.blogspot.com/2017/01/similarities-and-language-relationship.html>.

Models and processes in phylogenetic reconstruction

Johann-Mattis List

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

Since I started interdisciplinary work (linguistics and phylogenetics), I have repeatedly heard the expression "model-based". This expression often occurs in the context of parsimony vs. maximum likelihood and Bayesian inference, and it is usually embedded in statements like "the advantage of ML is that it is model-based", or "but parsimony is not model-based". By now I assume that I get the gist of these sentences, but I am afraid that I often still do not get their point. The problem is the ambiguity of the word "model" in biology but also in linguistics.

What is a model? For me, a model is usually a formal way to describe a process that we deal with in our respective sciences, nothing more. If we talk about the phenomenon of lexical borrowing, for example, there are many distinct processes by which borrowing can happen.

A clearcut case is Chinese *kāfēi* 咖啡 "coffee". This word was obviously borrowed from some Western language not that long ago. I do not know the exact details (which would require a rather lengthy literature review and an inspection of older sources), but that the word is not too old in Chinese is obvious. The fact that the pronunciation comes close to the word for coffee in the largest European languages (French, English, German) is a further hint, since the longer a new word has survived after having been transplanted to another language, the more it resembles other words in that language regarding its phonological structure; and the syllable *kā* does not occur in other words in Chinese. We can depict the process with help of the following visualization:

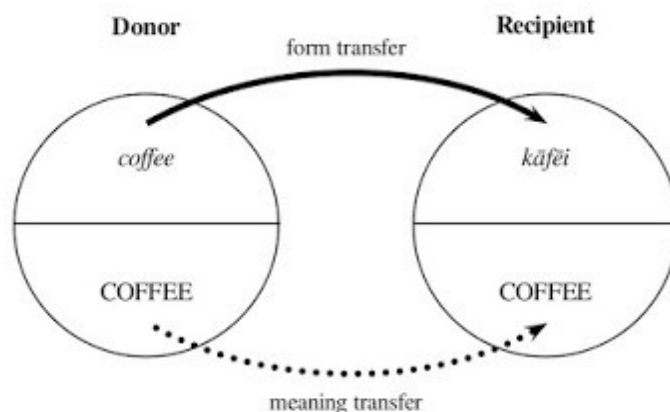


Figure 1: Lexical borrowing: direct transfer

The visualization tells us a lot about a very rough and very basic idea as to how the borrowing of words proceeds in linguistics: Each word has a *form* and a *function*, and *direct borrowing*, as we could call this specific subprocess, proceeds by transferring both the form and the function from the

donor language to the target language. This is a very specific type of borrowing, and many borrowing processes do not directly follow this pattern.

In the Chinese word *xǐnǎo* 洗脑 "brain-wash", for example, the form (the pronunciation) has not been transferred. But if we look at the morphological structure of *xǐnǎo*, being a compound consisting of the verb *xǐ* "to wash" and *nǎo* "the brain", it is clear that here Chinese borrowed only the meaning. We can visualize this as follows:

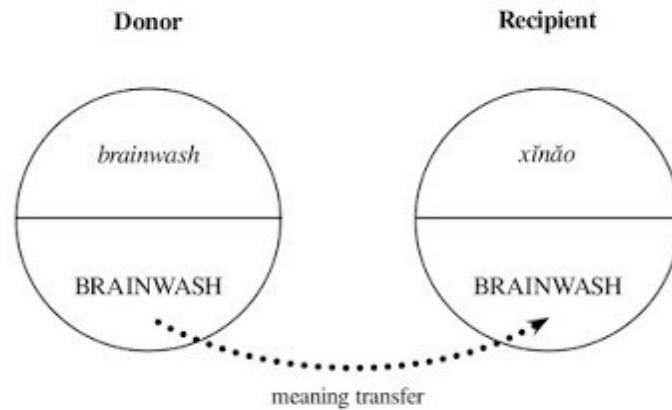


Figure 2: Lexical borrowing: meaning transfer

Unfortunately, I am already starting to simplify here. Chinese did not simply borrow the meaning, but it borrowed the expression, that is, the *motivation* to express this specific meaning in an analogous way to the expression in English. However, when borrowing meanings instead of full words, it is by no means straightforward to assume that the speakers will borrow exactly the same structure of expression they find in the donor language. The German equivalent of *skyscraper*, for example, is *Wolkenkratzer*, which literally translates as "cloudscraper".

There are many different ways to coin a good equivalent for "brain-wash" in any language of the world but which are *not* analogous to the English expression. One could, for example, also call it "head-wash", "empty-head", "turn-head", or "screw-mind"; and the only reason we call it "brain-wash" (instead of these others) is that this word was chosen at some time when people felt the need to express this specific meaning, and the expression turned out to be successful (for whatever reason).

Thus, instead of just distinguishing between "form transfer" and "meaning transfer", as my above visualizations suggest, we can easily find many more fine-grained ways to describe the processes of lexical borrowing in language evolution. Long ago, I took the time to visualize the different types of borrowing processes mentioned in the work of ([Weinreich 1953\[1974\]](#)) in the following graphic:

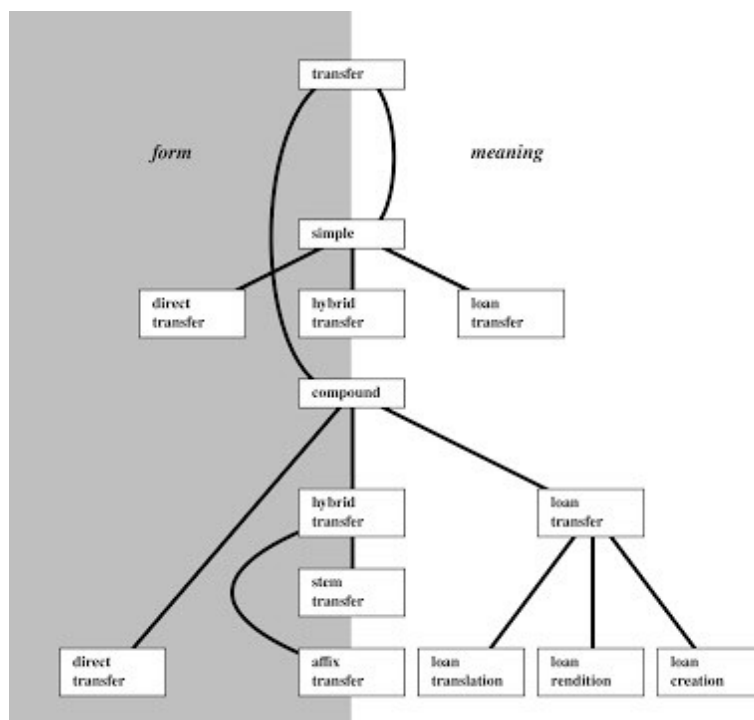


Figure 3: Lexical borrowing: hierarchy following Weinreich (1953[1974])

From my colleagues in biology, I know well that we find a similar situation in bacterial evolution with different types of lateral gene transfer ([Nelson-Sathi et al. 2013](#)). We are even not sure whether the account by Weinreich as displayed in the graphic is actually exhaustive; and the same holds for evolutionary biology and bacterial evolution.

But it may be time to get back to the models at this point, as I assume that some of you who have read this far have begun to wonder why I am spending so many words and graphics on borrowing processes when I promised to talk about models. The reason is that in my usage of the term "model" in scientific contexts, I usually have in mind exactly what I have described above. For me (and I suppose not only for me, but for many linguists, biologists, and scientists in general), models are attempts to formalize processes by classifying and distinguishing them, and flow-charts, typologies, descriptions and the identification distinctions are an informal way to communicate them.

If we use the term "model" in this broad sense, and look back at the discussion about parsimony, maximum likelihood, and Bayesian inference, it becomes also clear that it does not make immediate sense to say that parsimony lacks a model, while the other approaches are *model-based*. I understand why one may want to make this strong distinction between parsimony and methods based on likelihood-thinking, but I do not understand why the term "model" needs to be employed in this context.

Nearly all recent phylogenetic analyses in linguistics use *binary characters* and describe their

evolution with the help of simple *birth-death processes*. The only difference between parsimony and likelihood-based methods is how the birth-death processes are modelled *stochastically*. Unfortunately, we know very well that neither lexical borrowing nor "normal" lexical change can be realistically described as a birth-death process. We even know that these birth-death processes are essentially misleading (for details, see [List 2016](#)). Instead of investing our time to enhance and discuss the stochastic models driving birth-death processes in linguistics, doesn't it seem worthwhile to have a closer look at the real processes we want to describe?

References

- List, J.-M. (2016) Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction. *Journal of Language Evolution* 1.2. 119-136.
- Nelson-Sathi, S., O. Popa, J.-M. List, H. Geisler, W. Martin, and T. Dagan (2013) Reconstructing the lateral component of language history and genome evolution using network approaches. In: : Classification and evolution in biology, linguistics and the history of science. Concepts – methods – visualization. Franz Steiner Verlag: Stuttgart. 163-180.
- Weinreich, U. (1974) Languages in contact. With a preface by André Martinet. Mouton: The Hague and Paris.

Cite as: List, Johann-Mattis (2017): Models and processes in phylogenetic reconstruction. *The Genealogical World of Phylogenetic Networks* 6:2, 6-9, URL: <http://phylonetworks.blogspot.com/2017/02/models-and-processes-in-phylogenetic.html>.

Why we need alignments in historical linguistics

Johann-Mattis List

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

Alignments have been discussed [quite a few times](#) in this blog. They are so extremely common in molecular biology that I doubt that there are any debates about their usefulness, apart from certain attempts to improve the modelling, especially in cases of non-colinear patterns ([Kehr et al. 2014](#)), or to speed up computation ([Mathura and Adlakha 2016](#)). In linguistics, on the other hand, alignments are rarely used, although initial attempts to arrange homologous words in a matrix go back to the early 20th century, as you can see from this example taken from [Dixon and Koerber \(1919: 61\)](#):

		<i>31. Nails</i>				
W	N		kaha	i		
	C		k'a	i		
	SE		tc'a	i		
	SW		tca	i		
Md	NW	tsi'	bi		(Metathesis)	
	NE, S		bi	tsi		
Y	B	go	teo	yi	-c	
Mw	P, L		ti			
	CO	pi	tei		(Borrowed?)	
C	J, CR		tu	r		
	CL		tu	r	-em	
	B		tu	r	-is	

Figure 1: Early alignment from Dixon and Kroeber (1919)

This example is rather difficult to read for those not familiar with the annotation. The authors group homologous words across different indigenous languages from California. The group labels of the languages under investigation are given in abbreviated form at the very left of the matrix, and the actual varieties are listed in the next column. What follows is the actual *alignment*, along with comments in the last column. Regarding the alignments, the authors note on page 55:

A number of sets of cognates have been taken from their numbered place in this list and put at the end to allow of their being printed in columnar form, with a view to bringing out parallelisms that otherwise might fail to impress without detailed analysis and discussion. (Dixon and Kroeber 1919: 55)

In my opinion, this expresses nicely *why* alignments should be used more often in linguistics — due to the problem that our "alphabets" (the sound systems of languages) are undergoing constant change (see [this](#) earlier post for details regarding this claim), we need to infer both the scoring function between different sounds across different languages, and the alignment at the same time. If we look at the similarities the authors spotted, it should become obvious what I mean.

I am not yet sure how to interpret the data exactly, but if I am not mistaken, the authors claim that each of the column contains homologous material. So, they find a similarity between *kaha* in the

first row (the language is Northern Wintun, according to the key to abbreviations in the book), and *tu* in the last row (Monterey Costanoan). The last column shows suffixes, which I think the authors exclude from their analysis, but I could not find additional information confirming this in their book.

The comment column illustrates another problem of representation, namely that the authors do not know how to handle cases of metathesis (or transpositions) consistently. The transposition of the parts of words is a process that is quite frequent in language evolution. It is very frequent in compounds consisting of modifier and modified, such as *milk coffee* in English, where *milk* modifies the *coffee*, while French, for example, puts the modifier after the main noun, expressing this as *café au lait*.

Nowadays, we can handle these cases consistently in linguistics, both in our data annotation and in the alignments, and we can even search for the structures automatically (see [List et al. 2016](#)). One hundred years ago, when Dixon and Kroeber worked out their comparison of the languages in California, they were pioneers who tried to increase the transparency of our discipline, and it is clear that their solutions are not completely satisfying from today's perspective.

It is extremely surprising for me that, despite these early attempts to make our homology judgments in linguistics more transparent, the practice of phonetic alignments is still rarely used by historical linguists. Indeed, the majority of them even think that it is a waste of time, or only useful for the purpose of teaching.

I was reminded of this when I looked at a recent proposal by Bengtson ([2017](#), see also [this blog](#) for details) for deep genetic connections between Basque and North Caucasian languages. Note that the Basque language is traditionally considered as an *isolate*, i.e. a language whose nearest relatives we cannot find among the languages in the world. Many linguists have attempted to solve this puzzle by proposing various hypotheses (see [Forni 2013](#) for an example of attempting to link Basque with Indo-European). Bengtson proposes various types of evidence, which I cannot really judge, as I do not know the languages under comparison, but finally, he also shows a list with potential homologs between Basque and North Caucasian varieties, which you find below.

(gloss)	Basque	Chechen	Avar	Lak / Dargi	Lezgi	Proto-West Caucasian	Proto-North Caucasian
die	*hil	=al-	=al'	L =i=č'a D =ibk'	q'i-	*λə- / *λa-	*=iwlĕ
dog	*hor	pħu 'male dog'	hoy	D χa	χOF (Budukh)	*ł wə	*χHwəy-rV-
ear	*be=lafi	ler-g		D liħi		*łA-	*łĕHi
fire	*šu	ts'e	ts'a	L ts'u D ts'a	ts'ay	*mA=çwə	*çāyĭ
horn	*a=daf	kur	tʃ:ar		firi 'mane'		PEC *ΔwĭrV
I	*ni			L na D nu			PEC *nĭ
know	*e=akin	χ-aʔa	=eq' - (Akwakh)	L =aya- 'hear' D =aq' - / =iq' - 'hear'		*q:łwA 'to hear, to be heard'	*=łqĕ
thou	*hi	ħo		D hu	(Nidzh Udi) hu-n		PEC *hwV
tongue	*minhi	mott	matš:	L maz D mez	mez	*beʔA	*mĕłĭ
tooth	*horc		gožó 'tooth, fang, beak'	L k:arč:i 'tooth' D k:anži 'fang, canine tooth'	(Aguł) gwarž 'prong (of rake)'		PEC *gĕllĭʒwĕ
two	*bi		k'i-go	L k'i-a D k'wi	q'we-d (Udi) p:q	*tqł:wA	* (t)qHwā
what?	*se-r	stĕ-(n)-	s:u-n-	L s:a- D s:e		*sA	*sāy

Figure 2: Potential homologs between Basque and North Caucasian languages (Bengtson 2017)

If you are not a trained historical linguist, and thus do not know what to do with this table, be assured that many historical linguists will feel similarly. As a rough explanation: the concepts are supposed to be very, very stable, being drawn from [Sergey Yakhontov's list of 35 ultra-stable concepts](#), and I think that all words in one row are supposed to be etymologically related — that is, they should be potential homologs across all of the languages. If word forms are preceded by the asterisk symbol (*), this means that they are reconstructed, i.e. not reflected in written sources. But that is all I can tell you for the moment. Where I should start the comparison between the words remains a mystery for me, as I do not know which parts are supposed to be similar. Alignments would help us to see immediately *where* the author thinks that the historical similarities can be found — that is, we would see, which parts of the words are supposed to be homologous.

At this point in the post, I originally planned to provide you with an alignment of Bengtson's table, in order to illustrate the benefits of alignment in linguistics. Unfortunately, I had to admit to myself that I cannot do this, as I simply do not know where to align the words (apart from some rare trivial cases in the table).

I really hope that this will change in the future. Too often, our hypotheses in linguistics suffer from insufficient transparency with regards to the "proofs" and the evidence. I agree that it is very difficult to come up with good alignments in linguistics, especially if one regards cases of metathesis, unrelated parts, and general uncertainty. However, instead of giving in to the problem, we should follow the pioneering work of Dixon and Kroeber, and try to improve the way we present our data to both our colleagues and a broader public.

Theories such as the link between Basque and the North Caucasian languages are usually highly disputed in historical linguistics, and I do not know of any *long range* proposal that has gained broad acceptance during the last 50 years. Yet, maybe this is not because the proposals are not

valid, but simply because those who are proposing these theories have failed to present their findings in a transparent and testable way.

References

- Bengtson, J. (2017) The Euskaro-Caucasian Hypothesis. Current model. [PDE](#).
- Dixon, R. and A. Kroeber (1919) *Linguistic families of California*. University of California Press: Berkeley.
- Forni, G. (2013) Evidence for Basque as an Indo-European language. *The Journal of Indo-European Studies* 41.1 & 2: 1-142.
- Kehr, B., K. Trappe, M. Holtgrewe, and K. Reinert (2014) Genome alignment with graph data structures: a comparison. *BMC Bioinformatics* 15.1: 99.
- List, J.-M., P. Lopez, and E. Baptiste (2016) Using sequence similarity networks to identify partial cognates in multilingual wordlists. In: *Proceedings of the Association of Computational Linguistics 2016 (Volume 2: Short Papers)*. Association of Computational Linguistics, pp. 599-605.
- Mathur, R. and N. Adlakha (2016) A graph theoretic model for prediction of reticulation events and phylogenetic networks for DNA sequences. *Egyptian Journal of Basic and Applied Sciences* 3.3: 263-271.

Cite as: List, Johann-Mattis (2017): Similarities and language relationship. <i>The Genealogical World of Phylogenetic Networks</i> 6:3, 10-13, URL: http://phylonetworks.blogspot.com/2017/03/why-we-need-alignments-in-historical.html .
--

The siteswap annotation in juggling, and the power of annotation and modeling

Johann-Mattis List

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

I have been a juggler for more than 20 years now. It started when I was thirteen, and primarily interested in doing magic tricks, but I quickly realized that there are more transparent ways of presenting ones manipulation skills. About 15 years ago, when I was starting my studies in Berlin, there was a booming juggling scene in that city, with many young people, including many geeks who studied mathematics, programming, or physics. I, myself, was studying Indo-European linguistics by then, a field deprived of formalisms and formulas, devoted to the implicit as reflected in scientific prose that is not amenable to formalization, modeling, or transparent annotation.

It was at that time that some jugglers began to develop an annotation system for juggling patterns. The system was very simple, using numbers to denote the height and the direction of balls (or other objects) flying around from hand to hand. The 1 denoted the transfer of one ball from hand to hand without tossing it, the 2 denoted to hold one ball in one hand, the 3 to throw it from one hand to the other with a height required to juggle three balls, the 4 to throw one ball up in the air so that one would catch it with the same hand, and the 5 denoted the crossing from one hand to the other, but this time slightly higher, as required when juggling five balls. Some of these numbers are indicated in these animated GIFs.



The people called this system [siteswap](#), and they claimed that it was a good idea to formalize juggling to increase creativity, since one was not required to throw all of the balls with the same number, but one could combine them, following some basic mathematical ideas.

When people told me about this, I was extremely skeptical, probably due to my classical education, which gave me the conviction that juggling is an art, and an art cannot be describe in numbers. When people tried to teach me siteswaps, I ridiculed them, showing them some complicated

Cite as: List, Johann-Mattis (2017): The siteswap annotation in juggling, and the power of annotation and modeling. *The Genealogical World of Phylogenetic Networks* 6:4, 14-16, URL: <http://phylonetworks.blogspot.com/2017/04/the-siteswap-annotation-in-juggling-and.html>.

patterns involving body movements (see the next GIFs), and told them they would never be able to describe all the creativity of all the jugglers in the world in numbers.



Only a couple of years later, I realized that the geeks had proven me wrong, when, after a longer break, I was again participating in one of the many juggling conventions that take place throughout the year, in different locations in Europe and the whole world. I saw people doing tricks with three balls that I had never thought of before, and I asked them what they were doing. They answered, that these were siteswaps, and they were juggling patterns they called 441, or 531, respectively, as shown in these GIFs.



I gave in completely, when I saw how they applied the same logic to routines with five and more balls, which they called 654, 97531, or 744, respectively. Especially the 97531 fascinated me. During this routine, all of the balls end up in one vertical line in the air, for just a moment, but enough even for laymen to see the vertical line, which then immediately breaks down to a normal five-ball pattern, as shown here.



I realized, how wrong it was to take the un-annotability of something for granted. But even more importantly, I also understood that models, as restrictive as they may seem to be at first sight, may open new pathways for creativity, showing us things we had been ignoring before.

Only recently, when I promised colleagues to juggle during a talk on linguistics, I detected the parallel with my own studies in historical linguistics. For a long time, the field has been held back by people claiming that things could not be handled formally, for various reasons.

But I am realizing more and more that this is not true. We just need to start with something, some kind of model, which may not be as ideal and as realistic as we might wish it to be, but that may eventually help us to detect things we did not see before. We just need to start doing it, walking in baby-steps, improving our models and our annotation, as well as improving our understanding of the limits and the chances of a given formalization.

Needless to say, the patterns that I deemed to be un-annotatable 10 years ago in juggling can now easily be handled by my colleagues. They did not stop with the normal number system, but kept (and keep) developing it, and they take a lot of inspiration from this.

Cite as: List, Johann-Mattis (2017): The siteswap annotation in juggling, and the power of annotation and modeling. *The Genealogical World of Phylogenetic Networks* 6:4, 14-16, URL: <http://phylonetworks.blogspot.com/2017/04/the-siteswap-annotation-in-juggling-and.html>.

Killer arguments and the nature of proof in historical sciences

Johann-Mattis List

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

Some long time ago, somebody told me this joke, which I just found again on the internet in an English version (following jokes.cc.com, with modifications based on my memory):

Teacher: "Four crows are on the fence. The farmer shoots one. How many are left?"

Little Johnny: "None."

Teacher: "Listen carefully: Four crows are on the fence. The farmer shoots one. How many are left?"

Little Johnny: "None."

Teacher: "Can you explain that answer?"

Little Johnny: "One is shot, the others fly away. There are none left."

Teacher: "Well, that isn't the correct answer, but I like the way you think."

Little Johnny: "Teacher, can I ask a question?"

Teacher: "Sure."

Little Johnny: "There are three women in the park. The first one reads a love novel, the second one reads the newspaper, and the third one updates her FaceBook profile, which one of them is married?"

Teacher: "The one reading the newspaper?"

Little Johnny: "No. The one with the wedding ring on, but I like the way you think."

Given the title of this post, you may wonder why I tell you that joke. The reason is that for me, the essence of the joke is expressing the situation we often have in the historical sciences when we talk about "proof", be it of the closer relationship of different species, or the ultimate relationship of languages. Given the evidence we are given, we can reach an awful lot of conclusions in order to arrive at a convincing story, but if we see the wedding ring on somebody's hand, we know the true story no matter what other evidence we are given. The wedding ring in the joke serves as a killer argument — no matter what other evidence we consider, it is much more likely that the person who is married is the one with the ring than anybody else.

We often face similar situations in the historical sciences where we seek some kind of true story behind a couple of facts, when we are given external evidence that is just pointing to the right answer, or — let's be careful — the most probable answer, independent of where the other evidence might point to. We can think of similar situations in crime investigations, where we may think that a large body of evidence convicts some person as a murderer until we see some video proof that reveals the real offender.

That crime investigations have a lot in common with research in the historical sciences has been noted before by many people, notably the famous Umberto Eco (1932-2016), who edited a whole anthology on the role of circumstantial evidence in linguistics, semiotics, and philosophy ([Eco and Sebeok 1983](#)) where scholars compared the work of Sherlock Holmes with the work of people in the historical sciences. What Sherlock Holmes and historical linguists (and also evolutionary biologists) have in common is the use of *abduction* as their fundamental mode of reasoning. The term itself goes back to Charles Sanders Peirce (1839-1914), who distinguished it from *deduction*

Cite as: List, Johann-Mattis (2017): Killer arguments and the nature of proof in historical sciences. *The Genealogical World of Phylogenetic Networks* 6:5, 17-19, URL: <http://phylonetworks.blogspot.com/2017/05/killer-arguments-and-nature-of-proof-in.html>.

and *induction*:

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](#))

Our problem in the historical sciences is that we are searching an original situation: what was the case a long time ago, based on general knowledge about (evolutionary or historical) processes and the results of this situation. When Sherlock Holmes looks at a crime scene, he sees the results of an action and uses his knowledge of human behaviour to find the one who was responsible for the crime. When doctors listen to the heartbeat of patients who are short of breath, they try to find out what causes their disease by making use of their knowledge about symptoms and the diseases that could have caused them. When linguists look at words from different languages, they make use of their knowledge of processes of language change and language contact in order to work out why those languages are so similar.

As do medical practitioners or crime investigators, we have our general schema, our protocol, which we use to carry out our investigations. Biologists search for similar DNA sequences, linguists look for similar sound sequences. In most cases, this works fine, although we are usually left with uncertainties and things that do not really seem to add up. As long as we can quietly follow the protocol, we are fine; and even if the results of our research do not necessarily last for a long time, being superseded by more recent research, we usually have the impression that we did the best we could, given the complex circumstances with their complex circumstantial evidence. But once in a while, we uncover evidence similar to video proofs in crime investigation, or wedding rings as in the Little Johnny joke — evidence that is so striking that we have to put our protocol to one side and just accept that there is only one solution, no matter what the rest of our evidence or our protocol might point to.

In 1879, Ferdinand de Saussure (1857-1913) predicted two consonantal sounds in Proto-Indo-European based on circumstantial evidence ([Saussure 1879](#)). In 1927, Jerzy Kuryłowicz (1895-1978) could show that one of the sounds was still pronounced in Hittite, an Indo-European language that was not known during Saussure's time ([Lehmann 1992: 33](#)), and had just been deciphered. While Saussure followed protocol in his investigation, Kuryłowicz provided the video proof, and only since then, Saussure's hypothesis has become *communis opinio* in historical linguistics.

I assume that nobody will doubt the existence of different kinds of proof, different qualities of proof, in historical disciplines. If we are left with nothing else but our protocol, we can derive certain conclusions, but we can easily abandon our protocol once we have been presented with those killer arguments, that specific kind of proof that is so striking that we do not need to bother to have a look at any alternative facts again. I do not know of any similar examples in biology, but in linguistics (and in crime investigation, at least judging from the criminal novels I have read), it is

obvious that our evidence cannot only be ranked, but that we also have a huge incline between the standard evidence we use to make most of our arguments and those killer arguments that are so striking that no doubt is left.

In the short story *The Adventure of the Beryl Coronet*, Sherlock Holmes says:

[When] you have excluded the impossible, whatever remains, however improbable, must be the truth.

But this is only partially true, as in Sherlock Holmes' cases the truth is usually (but not always!) presented in such a form that it does not leave any place for doubt. Sherlock Holmes is a genius at finding the wedding rings on the fingers of his witnesses. As historical scientists, we are often much less lucky, but probably also less talented than Mr. Holmes. We are thus left with the fundamental problem of not knowing how to find the killer evidence, or how to quantify the doubt in those cases where we just follow the general protocol of our discipline.

References

- Eco, U. and T. Sebeok (1983) *The Sign of Three. Dupin, Holmes, Peirce*. Indiana University Press: Bloomington.
Lehmann, W. (1992) *Historical linguistics. An Introduction*. Routledge: London.
Peirce, C. (1931/1958) *Collected Papers of Charles Sanders Peirce*. Harvard University Press: Cambridge, Mass.
Saussure, F. (1879) *Mémoire sur le système primitif des voyelles dans les langues indo-européennes*. Teubner: Leipzig.

Cite as: List, Johann-Mattis (2017): Killer arguments and the nature of proof in historical sciences. *The Genealogical World of Phylogenetic Networks* 6:5, 17-19, URL: <http://phylonetworks.blogspot.com/2017/05/killer-arguments-and-nature-of-proof-in.html>.

Trees do not necessarily help in linguistic reconstruction

Johann-Mattis List

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

In historical linguistics, "linguistic reconstruction" is a rather important task. It can be divided into several subtasks, like "lexical reconstruction", "phonological reconstruction", and "syntactic reconstruction" — it comes conceptually close to what biologists would call "ancestral state reconstruction".

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

In a narrow sense, linguistic reconstruction only points to phonological reconstruction, which is something like the holy grail of computational approaches, since, so far, no method has been proposed that would convincingly show that one can do without expert insights. [Bouchard-Côté et al. \(2013\)](#) use language phylogenies to climb a language tree from the leaves to the root, using sophisticated machine-learning techniques to infer the ancestral states of words in Oceanic languages. [Hruschka et al. \(2015\)](#) start from sites in multiple alignments of cognate sets of Turkish languages to infer both a language tree, as well as the ancestral states along with the sound changes that regularly occurred at the internal nodes of the tree. Both approaches show that phylogenetic methods could, in principle, be used to automatically infer which sounds were used in the proto-language; and both approaches report rather promising results.

None of the approaches, however, is finally convincing, both for practical and methodological reasons. First, they are applied to language families that are considered to be rather "easy" to reconstruct. The tough cases are larger language families with more complex phonology, like Sino-Tibetan or any of its subbranches, including even shallow families like Sinitic (Chinese), or Indo-European, where the greatest achievements of the classical methods for language comparison have been made.

Second, they rely on a wrong assumption, that the sounds used in a set of attested languages are necessarily the pool of sounds that would also be the best candidates for the *Ursprache*. For example, [Saussure \(1879\)](#) proposed that Proto-Indo-European had at least two sounds that did not survive in any of the descendant languages, the so-called *laryngeals*, which are nowadays

Cite as: List, Johann-Mattis (2017): Trees do not necessarily help in linguistic reconstruction. *The Genealogical World of Phylogenetic Networks* 6:6, 20-22, URL: <http://phylonetworks.blogspot.com/2017/06/trees-do-not-necessarily-help-in.html>.

commonly represented as h_1 , h_2 , and h_3 , and which leave complex traits in the vocalism and the consonant systems of some Indo-European languages. Ever since then, it has been a standard assumption that it is always possible that none of the ancestral sounds in a given proto-language is still attested in any its descendants.

A third interesting point, which I consider a methodological problem of the methods, is that both of them are based on language trees, which are either given to the algorithm or inferred during the process. Given that most if not all approaches to ancestral state reconstruction in biology are based on some kind of phylogeny, even if it is a rooted evolutionary network, it may sound strange that I criticize this point. But in fact, when linguists use the classical methods to infer ancestral sounds and ancestral sound systems, phylogenies do not necessarily play an important role.

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

Sound change in this sense is simply not well treated in any paradigm that assumes some kind of parsimony, as it simply occurs too often independently. The question is less acute with vowels, where scholars have observed cycles of change in ancient languages that are attested in written sources. Even more problematic is the change of tones, where scholars have even less intuition regarding preference directions or preference transitions; and also because ancient data does not describe the tones in the phonetic detail we would need in order to compare it with modern data. In contrast to consonant reconstruction, where we can do almost exclusively without phylogenies, phylogenies may indeed provide some help to shed light on open questions in vowel and tone change.

But one should not underestimate this task, given the systemic pressure that may crucially impact on vowel and tone systems. Since there are considerably fewer empty spots in the vowel and tone space of human languages, it can easily happen that the most natural paths of vowel or tone development (if they exist in the end) are counteracted by systemic pressures. Vowels can be more easily confused in communication, and this holds even more for tones. Even if changes are "natural", they could create conflict in communication, if they produce very similar vowels or tones

that are hard to distinguish by the speakers. As a result, these changes could provoke mergers in sounds, with speakers no longer distinguishing them at all; or alternatively, changes that are less "natural" (physiologically or acoustically) could be preferred by a speech society in order to maintain the effectiveness of the linguistic system.

In principle, these phenomena are well-known to trained linguists, although it is hard to find any explicit statements in the literature. Surprisingly, linguistic reconstruction (in the sense of phonological reconstruction) is hard for machines, since it is easy for trained linguists. Every historical linguist has a catalogue of existing sounds in their head as well as a network of preference transitions, but we lack a machine-readable version of those catalogues. This is mainly because transcriptions systems widely differ across subfields and families, and since no efforts to standardize these transcriptions have been successful so far.

Without such catalogues, however, any efforts to apply vanilla-style methods for ancestral state reconstruction from biology to linguistic reconstruction in historical linguistics, will be futile. We do not need the trees for linguistic reconstruction, but the network of potential pathways of sound change.

References

- Bouchard-Côté, A., D. Hall, T. Griffiths, and D. Klein (2013): Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences* 110.11. 4224–4229.
- Hruschka, D., S. Branford, E. Smith, J. Wilkins, A. Meade, M. Pagel, and T. Bhattacharya (2015): Detecting regular sound changes in linguistics as events of concerted evolution. *Current Biology* 25.1: 1-9.
- Saussure, F. (1879): *Mémoire sur le système primitif des voyelles dans les langues indo-européennes*. Teubner: Leipzig.

Cite as: List, Johann-Mattis (2017): Trees do not necessarily help in linguistic reconstruction. *The Genealogical World of Phylogenetic Networks* 6:6, 20-22, URL: <http://phylonetworks.blogspot.com/2017/06/trees-do-not-necessarily-help-in.html>.

More on similarities in linguistics

Johann-Mattis List

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

In an [earlier blogpost](#) I discussed various reasons for similarity of certain traits in languages. I emphasized four major reasons for similarities, for example, in the lexicon of languages: *coincidence*, *natural reasons*, *inheritance*, and *contact* (see also [List 2014: 55f](#) and [Aikhenvald 2007: 5](#)). Despite the problems of distinguishing inherited from borrowed traits, which I called *historical reasons for similarity*, controlling for coincidence and history can often be done in a rather straightforward way. Coincidence can be called by applying a frequency criterion: if certain similarities are extremely spurious, they are usually due to chance. Historical similarities can be detected with the help of classical methods for language comparison. If, using these methods, we know, for example, that two or more languages are genetically related or have been developing in close contact with each other, then we will usually assume that shared traits among them are due to their shared history.

The third group of similarities, on the other hand, which I called *natural*, is a bit more difficult to interpret, since it is not entirely clear what "natural" means in this context. My earlier example was the word for "mother", which in many languages is expressed as "mama", similar to "father", which is often expressed as "papa", even in languages where we know that they are not related, or only extremely distantly related (if we assume that language was only invented once), and will thus be acquired rather early by children.

In the case of "mama" and "papa", we can blame our articulatory apparatus, which makes sounds like [m], [p], and [a] very easy to pronounce for all humans, no matter where and in which time they are born. Calling this "nature" is probably justified, given that pronouncability is not *per se* characteristic for language as a general means of complex communication. In sign languages, for example, pronouncability does not play any role, as those languages are never pronounced, but expressed with the help of gestures. But even in sign languages, we also find cross-linguistic similarities, which seem to be independent of coincidence or history: body parts, for example, are often expressed iconically, e.g., by pointing to them (see [Woodward 1993](#) for details).

However, not all of those similarities between languages that are **not** due to history or coincidence are necessarily due to our articulation apparatus. We can think of many different reasons for cross-linguistic similarities, such as, for example, innate settings of the human brain, or global similarities of the environment in which humans live. In the past, colleagues have occasionally pointed out to me the heterogeneity of this class of "natural" similarities. When trying to further subdivide them, the former could be called "similarities due to cognition", while the latter could be called "similarities due to environment". But neither of these two groups seems to be quite satisfying, as

we do not really know the relation between environment and cognition. We may also assume that there is a certain influence between the two, and depending on where we draw the border, we would either subscribe to a predominantly Aristotelian viewpoint, where we assign the predominant role to the environment, or a Platonic viewpoint, where we assign it to the innate "ideas" which are given to us along with our brain.

As an example for the difficulty of distinguishing different sources of "natural" similarity, let us have a look at how languages of the world express a fixed set of concepts. In a very simplistic view, given only two things we want to express, for instance the concept "hand" and the concept "arm", we can ask whether a given language will use the same or different words as a rule. English, for example, uses two different words, namely *hand* and *arm*, and so does German (*Hand* and *Arm*), while Russian uses only one word, *ruka*, to refer to both concepts in most situations (in Russian, there is another word *kist'*, which can be used to denote "hand", but it is rarely used). We can say that Russian *ruka* is *polysemous*, since the word form has at least two meanings. A better way of expressing this is to say that Russian colexifies "hand" and "arm" ([François 2008](#)), since the term *polysemy* has a specific usage in linguistics, referring to words expressing multiple meanings that should be "conceptually close" or "developed from semantic change", which is an extremely vague definition that further requires us to know the history of a given word form and the development of its meanings.

Cross-linguistically, the colexification of "arm" and "hand", i.e. that many languages tend to use a single word to denote both concepts, occurs extremely often in the languages of the world; so often that we can rule out that the use of one word for two concepts is due to coincidence (compare the colexifications of "arm" in the [CLICS](#) database by [List et al. 2014](#) through [this link](#)). Given that the colexification recurs also in different language families spoken in different regions of the world, we can further rule out historical reasons. This leaves us with the heterogeneous class of "natural reasons for similarities". But what kind of natural similarities are we dealing with here? Are they cognitive? They surely are in some sense, as we can say that humans have good reasons to consider the hand and the arm as one continuous part of their body.

But this continuity is also given by the structure of our body, which itself is given independently of our perception. One could argue that our perception grounds in our bodily experience, but if we look further into other frequent colexifications, e.g. between "dark" and "black" (this occurs in more than 20 language families, see [here](#)), as well as "bright" and "white" (occurs in three language families, see [here](#)), our perception is less dependent on our body but more on the environment in which we experience darkness and brightness, since most humans have eyesight and do not live entirely in caves.

It is some kind of the egg-hen problem of who was there first, and the more I think about it, I prefer to avoid giving any clear-cut preference to either the egg nor the hen. We can obviously try to make

a more fine-grained distinction between different kinds of non-historical and non-coincidental similarities between languages, but unless psychologists and cognitive scientists solve general problems of perception and environment, it seems that, at least for the moment, "natural similarities" is explicit enough as a term to describe universal patterns in the languages of the world.

References

- François, A. (2008) Semantic maps and the typology of colexification: intertwining polysemous networks across languages. In: Vanhove, M. (ed.): *From polysemy to semantic change*. Benjamins: Amsterdam. 163-215.
- List, J.-M., T. Mayer, A. Terhalle, and M. Urban (eds.) (2014) CLICS: Database of Cross-Linguistic Colexifications. *Forschungszentrum Deutscher Sprachatlas*: Marburg. <http://www.webcitation.org/6ccEMrZYM>.
- List, J.-M., M. Cysouw, and R. Forkel (2016) Concepticon. A resource for the linking of concept lists. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, 2393-2400.
- Woodward, J. (1993) Lexical evidence for the existence of South Asian and East Asian sign language families. *Journal of Asian Pacific Communication* 4.2: 91-107.

Cite as: List, Johann-Mattis (2017): More on similarities in linguistics. *The Genealogical World of Phylogenetic Networks* 6:7, 23-25, URL: <http://phylonetworks.blogspot.com/2017/07/more-on-similarities-in-linguistics.html>.

Unattested character states

Johann-Mattis List

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

In an earlier post from [January 2016](#), I argued that it is important to account for directional processes when modeling language history through character-state evolution. In previous papers ([List 2016](#); [Chacon and List 2015](#)), I tried to show that this can be easily done with asymmetric step matrices in a parsimony framework. Only later did I realize that this is nothing new for biologists who work on morphological characters, thus supporting David's claim that we should not compare linguistic characters with the genotype, but with the phenotype ([Morrison 2014](#)). Early this year, a colleague introduced me to *Mk*-models in phylogenetics, which were first introduced by [Lewis \(2001\)](#) and allow analysis of multi-state characters in a likelihood framework.

What was surprising for me is that it seems that *Mk*-models seem to outperform parsimony frameworks, although being much simpler than elaborate step-matrices defined for morphological characters ([Wright and Hillis 2014](#)). Today, I read that a recent paper by [Wright et al. \(2016\)](#) even shows how *asymmetric transition rates* can be handled in likelihood frameworks.

Being by no means an expert in phylogenetic analyses, especially not in likelihood frameworks, I tend to have a hard time understanding what is actually being modeled. However, if I correctly understand the gist of the Wright et al. paper, it seems that we are slowly approaching a situation in which more complex scenarios of lexical character evolution in linguistics no longer need to rely on parsimony frameworks.

But, unfortunately, we are not there yet; and it is even questionable whether we will ever be. The reason is that all multi-state models that have been proposed so far only handle transitions between *attested* characters: *unattested* characters can neither be included in the analyses nor can they be inferred.

I have pointed to this problem in some previous blogposts, the last one published in [June](#), where I mentioned [Ferdinand de Saussure](#), (1857-1913), who postulated two unattested consonantal sounds for Indo-European ([Saussure 1879](#)), of which one was later found to have still survived in Hittite, a language that was deciphered and shown to be Indo-European only about 30 years later ([Lehmann 1992: 33](#)).

The fact that it is possible to use our traditional methods to infer unattested sounds from circumstantial evidence, but not to include our knowledge about them into phylogenetic analyses, is a huge drawback. Potentially even greater are the situations where even our traditional methods do not allow us to infer unattested data. Think, for example, of a word that was once present in some

language but was later completely lost. Given the ephemeral nature of human language, we have no way to know this, but we know very well that it easily happens when just thinking of some terms used for old technology, like *walkman* or soon even *iPod*, which the younger generations have never heard about.

Colleagues with whom I have discuss my concerns in this regard are often more optimistic than I am, saying that even if the methods cannot handle unattested characters they could still find the major signal, and thus tell us at least the general tendency as to how a language family evolved. However, for classical linguists, who can infer quite a lot using the laborious methods that still need to be applied manually, it leaves a sour taste, if they are told that the analysis deliberately ignored crucial aspects of the processes and phenomena they understand very well. For example, if we detect that some intelligence test is right in about 80% of all cases, we would also abstain from using it to judge who we allow to take up their studies at university.

I also think that it is not a satisfying solution for the analysis of morphological data in biology. It is probably quite likely that some ancient species had certain traits which later evolved into the traits we observe which are simply no longer attested anywhere, either in fossils or in the genes. I also wonder how well phylogenetic frameworks generally account for the fact that what the evidence we are left with may reflect much less of what was once there.

In [Chacon and List \(2015\)](#), we circumvent the problem by adding ancestral but unattested sounds to the step matrices in our parsimony analysis. This is of course not entirely satisfactory, as it adds a heavy bias to the analysis of sound change, which no longer tests for all possible solutions but only for the ones we fed into the algorithm. For sound change, it may be possible to substantially expand the character space by adding sounds attested across the world's languages, and then having the algorithms select the most probable transitions. But given that we still barely know anything about general transition probabilities of sound change, and that databases like [Phoible \(Moran 2015\)](#) list more than 2,000 different sounds for a bit more than 2,000 languages, it seems like a Sisyphean challenge to tackle this problem consistently.

What can we do in the meantime? Not very much, it seems. But we can still try to improve our methods in baby steps, trying to get a better understanding of the major and minor processes in linguistic and biological evolution; and not forgetting that, although I was only talking about phylogenetic tree reconstruction, in the end we also want to have all of this done in network approaches.

References

- Chacon, T. and J.-M. List (2015) Improved computational models of sound change shed light on the history of the Tukanoan languages. *Journal of Language Relationship* 13: 177-204.
- Lehmann, W. (1992) *Historical linguistics. An Introduction*. Routledge: London.

- Lewis, P. (2001) A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology* 50: 913-925.
- List, J.-M. (2016) Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction. *Journal of Language Evolution* 1: 119-136.
- Moran, S., D. McCloy, and R. Wright (eds) (2014) *PHOIBLE Online*. Max Planck Institute for Evolutionary Anthropology: Leipzig.
- Morrison, D.A. (2014) Are phylogenetic patterns the same in anthropology and biology? [bioRxiv](https://doi.org/10.1101/007811).
- Saussure, F. (1879) *Mémoire sur le système primitif des voyelles dans les langues indo-européennes*. Teubner: Leipzig.
- Wright, A. and D. Hillis (2014) Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. *PLoS ONE* 9.10. e109210.
- Wright, A., G. Lloyd, and D. Hillis (2016) Modeling character change heterogeneity in phylogenetic analyses of morphology through the use of priors. *Systematic Biology* 65: 602-611.

Cite as: List, Johann-Mattis (2017): Unattested character states. *The Genealogical World of Phylogenetic Networks* 6:8, 26-28, URL: <http://phylonetworks.blogspot.com/2017/08/unattested-character-states.html>.

Arguments from authority, and the Cladistic Ghost, in historical linguistics

Johann-Mattis List

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

Arguments from authority play an important role in our daily lives and our societies. In political discussions, we often point to the opinion of trusted authorities if we do not know enough about the matter at hand. In medicine, favorable opinions by respected authorities function as one of [four levels of evidence](#) (admittedly, the lowest) to judge the strength of a medicament. In advertising, the (at times doubtful) authority of celebrities is used to convince us that a certain product will change our lives.

Arguments from authority are useful, since they allow us to have an opinion without fully understanding it. Given the ever-increasing complexity of the world in which we live, we could not do without them. We need to build on the opinions and conclusions of others in order to construct our personal little realm of convictions and insights. This is specifically important for scientific research, since it is based on a huge network of trust in the correctness of previous studies which no single researcher could check in a lifetime.

Arguments from authority are, however, also dangerous if we blindly trust them without critical evaluation. To err is human, and there is no guarantee that the analysis of our favorite authorities is always error proof. For example, famous linguists, such as [Ferdinand de Saussure](#) (1857-1913) or [Antoine Meillet](#) (1866-1936), revolutionized the field of historical linguistics, and their theories had a huge impact on the way we compare languages today. Nevertheless, this does not mean that they were right in all their theories and analyses, and we should never trust any theory or methodological principle *only* because it was proposed by Meillet or Saussure.

Since people tend to avoid asking why their authority came to a certain conclusion, arguments of authority can be easily abused. In the extreme, this may accumulate in totalitarian societies, or societies ruled by religious fanaticism. To a smaller degree, we can also find this totalitarian attitude in science, where researchers may end up blindly trusting the theory of a certain authority without further critically investigating it.

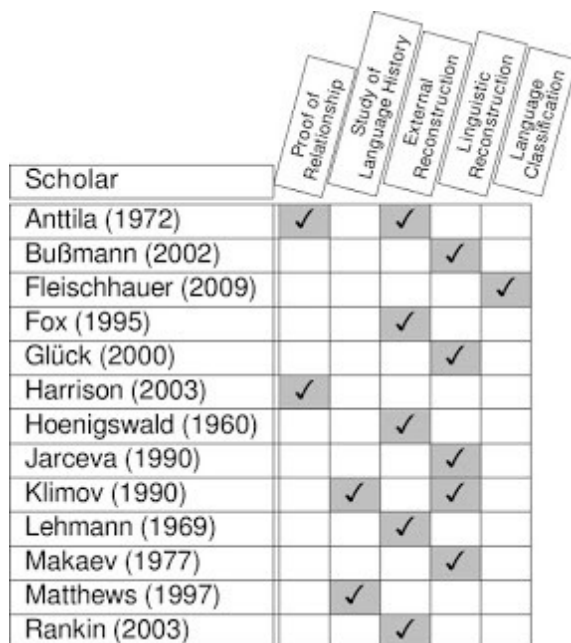
The comparative method

The authority in this context does not necessarily need to be a real person, it can also be a theory or a certain methodology. The [financial crisis from 2008](#) can be taken as an example of a methodology, namely classical "economic forecasting", that turned out to be trusted much more than it deserved. In historical linguistics, we have a similar quasi-religious attitude towards our traditional [comparative method](#) (see [Weiss 2014](#) for an overview), which we use in order to

Cite as: List, Johann-Mattis (2017): Arguments from authority, and the Cladistic Ghost, in historical linguistics. *The Genealogical World of Phylogenetic Networks* 6:8, 29-34, URL: <http://phylonetworks.blogspot.com/2017/09/arguments-from-authority-and-cladistic.html>.

compare languages. This "method" is in fact no method at all, but rather a huge bunch of techniques by which linguists have been comparing and reconstructing languages during the past 200 years. These include the detection of cognate or "homologous" words across languages, and the inference of regular sound correspondence patterns (which I discussed in a [blog from October last year](#)), but also the reconstruction of sounds and words of ancestral languages not attested in written records, and the inference of the phylogeny of a given language family.

In all of these matters, the comparative method enjoys a quasi-religious authority in historical linguistics. Saying that they do not follow the comparative method in their work is among the worst things you can say to historical linguists. It hurts. We are conditioned from when we were small to feel this pain. This is all the more surprising, given that scholars rarely agree on the specifics of the methodology, as one can see from the table below, where I compare the key tasks that different authors attribute to the "method" in the literature. I think one can easily see that there is not much of an overlap, nor a pattern.



Scholar	Proof of Relationship	Study of Language History	External Reconstruction	Linguistic Reconstruction	Language Classification
Anttila (1972)	✓		✓		
Bußmann (2002)				✓	
Fleischhauer (2009)					✓
Fox (1995)			✓		
Glück (2000)				✓	
Harrison (2003)	✓				
Hoenigswald (1960)			✓		
Jarceva (1990)				✓	
Klimov (1990)		✓		✓	
Lehmann (1969)			✓		
Makaev (1977)				✓	
Matthews (1997)		✓			
Rankin (2003)			✓		

Figure 1: Varying accounts on the "comparative methods" in the linguistic literature

It is difficult to tell how this attitude evolved. The foundations of the comparative method go back to the early work of scholars in the 19th century, who managed to demonstrate the genealogical relationship of the Indo-European languages. Already in these early times, we can find hints regarding the "methodology" of "comparative grammar" (see for example [Atkinson 1875](#)), but judging from the literature I have read, it seems that it was not before the early 20th century that people began to introduce the techniques for historical language comparison as a methodological framework.

How this framework became **the** framework for language comparison, although it was never really established as such, is even less clear to me. At some point the linguistic world (which was always

characterized by aggressive battles among colleagues, which were fought in the open in numerous publications) decided that the numerous techniques for historical language comparison which turned out to be the most successful ones up to that point are a specific method, and that this specific method was so extremely well established that no alternative approach could ever compete with it.

Biologists, who have experienced drastic methodological changes during the last decades, may wonder how scientists could believe that any practice, theory, or method is everlasting, untouchable and infallible. In fact, the comparative method in historical linguistics is always changing, since it is a label rather than a true framework with fixed rules. Our insights into various aspects of language change is constantly increasing, and as a result, the way we practice the comparative method is also improving. As a result, we keep using the same label, but the product we sell is different from the one we sold decades ago. Historical linguistics are, however, very conservative regarding the authorities they trust, and our field was always very skeptical regarding any new methodologies which were proposed.

[Morris Swadesh \(1909-1967\)](#), for example, proposed a quantitative approach to infer divergence dates of language pairs ([Swadesh 1950](#) and later), which was immediately refuted, right after he proposed it ([Hoijer 1956](#), [Bergsland and Vogt 1962](#)). Swadesh's idea to assume constant rates of lexical change was surely problematic, but his general idea of looking at lexical change from the perspective of a fixed set of meanings was very creative in that time, and it has given rise to many interesting investigations (see, among others, [Haspelmath and Tadmor 2009](#)). As a result, quantitative work was largely disregarded in the following decades. Not many people paid any attention to [David Sankoff's \(1969\)](#) PhD thesis, in which he tried to develop improved models of lexical change in order to infer language phylogenies, which is probably the reason why Sankoff later turned to biology, where his work received the appreciation it deserved.

Shared innovations

Since the beginning of the second millennium, quantitative studies have enjoyed a new popularity in historical linguistics, as can be seen in the numerous papers that have been devoted to automatically inferred phylogenies (see [Gray and Atkinson 2003](#) and *passim*). The field has begun to accept these methods as additional tools to provide an understanding of how our languages evolved into their current shape. But scholars tend to contrast these new techniques sharply with the "classical approaches", namely the different modules of the comparative method. Many scholars also still assume that the only valid technique by which phylogenies (be it trees or networks) can be inferred is to identify *shared innovations* in the languages under investigation ([Donohue et al. 2012](#), [François 2014](#)).

The idea of shared innovations was first proposed by [Brugmann \(1884\)](#), and has its direct counterpart in Hennig's ([1950](#)) framework of *cladistics*. In a later book of Brugmann, we find the following passage on shared innovations (or synapomorphies in Hennig's terminology):

The only thing that can shed light on the relation among the individual language branches [...] are the specific correspondences between two or more of them, the innovations, by which each time certain language branches have advanced in comparison with other branches in their development. (Brugmann 1967[1886]:24, my translation)

Unfortunately, not many people seem to have read Brugmann's original text in full. Brugmann says that subgrouping requires the identification of shared innovative traits (as opposed to shared retentions), but he remains skeptical about whether this can be done in a satisfying way, since we often do not know whether certain traits developed independently, were borrowed at later stages, or are simply being misidentified as being "shared". Brugmann's proposed solution to this is to claim that shared, potentially innovative traits, should be numerous enough to reduce the possibility of chance.

While biology has long since abandoned the cladistic idea, turning instead to quantitative (mostly stochastic) approaches in phylogenetic reconstruction, linguists are surprisingly stubborn in this regard. It is beyond question that those uniquely shared traits among languages that are unlikely to have evolved by chance or language contact are good proxies for subgrouping. But they are often very hard to identify, and this is probably also the reason why our understanding about the phylogeny of the Indo-European language family has not improved much during the past 100 years. In situations where we lack any striking evidence, quantitative approaches may as well be used to infer potentially innovated traits, and if we do a better job in listing these cases (current software, which was designed by biologists, is not really helpful in logging all decisions and inferences that were made by the algorithms), we could profit a lot when turning to *computer-assisted frameworks* in which experts thoroughly evaluate the inferences which were made by the automatic approaches in order to generate new hypotheses and improve our understanding of our language's past.

A further problem with cladistics is that scholars often use the term *shared innovation* for inferences, while the cladistic toolkit and the reason why Brugmann and Hennig thought that shared innovations are needed for subgrouping rests on the assumption that one knows the true evolutionary history ([DeLaet 2005: 85](#)). Since the true evolutionary history is a tree in the cladistic sense, an innovation can only be identified if one knows the tree. This means, however, that one cannot use the innovations to infer the tree (if it has to be known in advance). What scholars thus mean when talking about shared innovations in linguistics are *potentially shared innovations*, that is, characters, which are diagnostic of subgrouping.

Conclusions

Given how quickly science evolves and how non-permanent our knowledge and our methodologies are, I would never claim that the new quantitative approaches are the only way to deal with trees or

networks in historical linguistics. The last word on this debate has not yet been spoken, and while I see many points critically, there are also many points for concrete improvement ([List 2016](#)). But I see very clearly that our tendency as historical linguists to take the comparative method as the only authoritative way to arrive at a valid subgrouping is not leading us anywhere.



Figure 2: Do computational approaches really switch off the light which illuminates classical historical linguistics?

In a recent review, Stefan Georg, an expert on Altaic languages, writes that the recent computational approaches to phylogenetic reconstruction in historical linguistics "switch out the light which has illuminated Indo-European linguistics for generations (by switching on some computers)", and that they "reduce this discipline to the pre-modern guesswork stage [...] in the belief that all that processing power can replace the available knowledge about these languages [...] and will produce 'results' which are worth the paper they are printed on" ([Georg 2017](#): 372, footnote). It seems to me, that, if a discipline has been enlightened too much by its blind trust in authorities, it is not the worst idea to switch off the light once in a while.

References

- Anttila, R. (1972): An introduction to historical and comparative linguistics. Macmillan: New York.
- Atkinson, R. (1875): Comparative grammar of the Dravidian languages. *Hermathena* 2.3. 60-106.
- Bergsland, K. and H. Vogt (1962): On the validity of glottochronology. *Current Anthropology* 3.2. 115-153.
- Brugmann, K. (1884): Zur Frage nach den Verwandtschaftsverhältnissen der indogermanischen Sprachen [Questions regarding the closer relationship of the Indo-European languages]. *Internationale Zeitschrift für allgemeine Sprachwissenschaft* 1. 228-256.
- Bußmann, H. (2002): Lexikon der Sprachwissenschaft. Kröner: Stuttgart.
- De Laet, J. (2005): Parsimony and the problem of inapplicables in sequence data. In: Albert, V. (ed.): Parsimony, phylogeny, and genomics. Oxford University Press: Oxford. 81-116.
- Donohue, M., T. Denham, and S. Oppenheimer (2012): New methodologies for historical linguistics? Calibrating a lexicon-based methodology for diffusion vs. subgrouping. *Diachronica* 29.4. 505–522.
- Fleischhauer, J. (2009): A Phylogenetic Interpretation of the Comparative Method. *Journal of Language Relationship* 2. 115-138.
- Fox, A. (1995): Linguistic reconstruction. An introduction to theory and method. Oxford University Press: Oxford.
- François, A. (2014): Trees, waves and linkages: models of language diversification. In: Bowerman, C. and B. Evans (eds.):

- The Routledge handbook of historical linguistics. Routledge: 161-189.
- Georg, S. (2017): The Role of Paradigmatic Morphology in Historical, Areal and Genealogical Linguistics. *Journal of Language Contact* 10. 353-381.
- Glück, H. (2000): Metzler-Lexikon Sprache . Metzler: Stuttgart.
- Gray, R. and Q. Atkinson (2003): Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426.6965. 435-439.
- Harrison, S. (2003): On the limits of the comparative method. In: Joseph, B. and R. Janda (eds.): The handbook of historical linguistics. Blackwell: Malden and Oxford and Melbourne and Berlin. 213-243.
- Haspelmath, M. and U. Tadmor (2009): The Loanword Typology project and the World Loanword Database. In: Haspelmath, M. and U. Tadmor (eds.): Loanwords in the world's languages. de Gruyter: Berlin and New York. 1-34.
- Hennig, W. (1950): Grundzüge einer Theorie der phylogenetischen Systematik. Deutscher Zentralverlag: Berlin.
- Hoenigswald, H. (1960): Phonetic similarity in internal reconstruction. *Language* 36.2. 191-192.
- Hoijer, H. (1956): Lexicostatistics. A critique. *Language* 32.1. 49-60.
- Jarceva, V. (1990): . Sovetskaja Enciklopedija: Moscow.
- Klimov, G. (1990): Osnovy lingvističeskoj komparativistiki [Foundations of comparative linguistics]. Nauka: Moscow.
- Lehmann, W. (1969): Einführung in die historische Linguistik. Carl Winter:
- List, J.-M. (2016): Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction. *Journal of Language Evolution* 1.2. 119-136.
- Makaev, E. (1977): Obščaja teorija sravnitel'nogo jazykoznanija [Common theory of comparative linguistics]. Nauka: Moscow.
- Matthews, P. (1997): Oxford concise dictionary of linguistics . Oxford University Press: Oxford.
- Rankin, R. (2003): The comparative method. In: Joseph, B. and R. Janda (eds.): The handbook of historical linguistics. Blackwell: Malden and Oxford and Melbourne and Berlin.
- Sankoff, D. (1969): Historical linguistics as stochastic process . . McGill University: Montreal.
- Weiss, M. (2014): The comparative method. In: Bowerman, C. and N. Evans (eds.): The Routledge Handbook of Historical Linguistics. Routledge: New York. 127-145.

Cite as: List, Johann-Mattis (2017): Arguments from authority, and the Cladistic Ghost, in historical linguistics. *The Genealogical World of Phylogenetic Networks* 6:8, 29-34, URL: <http://phylonetworks.blogspot.com/2017/09/arguments-from-authority-and-cladistic.html>.

"Man gave names to all those animals": cats and dogs

Guido Grimm¹ and Johann-Mattis List²

¹Independent researcher, Lyon, ²Max Planck Institute for the Science of Human History

As specialists, we rarely dare to dive into cross-disciplinary research. However, in a small series of posts, we will now try to open a door between linguistics, phylogenetics, biogeography, and molecular genetics (with its various subdisciplines), using the curious cases of domestic animals, such as *cat*, *dog*, *goat*, and *sheep*, and what these are called in various Eurasian languages, with a special focus on Indo-European languages.

Today's post will introduce the little dataset that we have created, and discuss the findings for the names of *cats* and *dogs*. A follow-up post will be devoted to *goats* and *sheep*.

Domesticated animals and their names

Various types of archaeological and biological research revolve around the domestication of animals — [GoogleScholar](#) gives tens of thousands of hits for search items such as "cat domestication"; and we have several blog posts about the need for networks to illustrate the genealogy of domestication. However, linguistic literature on these topics is rather sparse, often related to specific language families, such as domesticated animals in the Indo-European proto-society (Anthony and Ringe 2015).

Nevertheless, many studies mention the potential value of linguistic evidence as some specific kind of *indirect evidence*, which should be considered when carrying out research on domestication (see, for example, Kraft et al. 2015). Furthermore, the public interest in domestic animals such as *cat*, *dog*, *goat* and *sheep*, is reflected by the number of languages in which Wikipedia articles are available: the domestic dog (219 entries), our most trusted companion animal, narrowly beats the cat (211 entries), our least-productive domestic animal but, according to *cliché*, an obligatory accessory for e.g. literates, thinkers, and little old ladies (entry counts include extinct ones like Gothic). Sheep are available for 166 languages, and goats for 142.

One doesn't have to travel far to recognize substantial difference between the four animal names. For example, when Guido moved to Sweden, the most confusing thing was "Fåret Shaun", which he knew as "Shaun, das Schaf" in German, or "Shaun, the sheep" in English. [As an aside, Shaun's name is a pun in English, but not in German or Swedish.] While Swedish and German / English differ greatly in the pronunciation of the words they use to denote "sheep", the Swedish words for "cat" (Swedish *katt*, German *Katze*), "dog" (*hund* vs. *Hund*), and "goat" (*get* vs. *Geiß*) are essentially the same (using Guido's dialect of German). They also are basically the same for many other essential items, such as "house" (*hus* vs. *Haus*), and "hand" (*hand* vs. *Hand*).

Cite as: Grimm, Guido and List, Johann-Mattis (2017): "Man gave names to all those animals": cats and dogs. *The Genealogical World of Phylogenetic Networks* 6:10, 35-41, URL: <http://phylonetworks.blogspot.com/2017/10/man-gave-names-to-all-those-animals.html>.

Since Guido moved to France, he has been watching "Shaun le mouton"; and *Hund* ("dog") has become *chien*. He now needs to look for *chèvre* ("goat") when making choosing his cheeses; but his cats are called *chats*, which is similar in writing (and linguistic history) but phonetically rather different, as the word is pronounced as [ʃa] (*sha*).

When Mattis visited China, he had few problems memorizing the word for "cat", as the Chinese word *māo* is quite similar to the sound which cats are alleged to make in many languages (see the list on [Wikipedia](#) for cross-linguistic similarities of *onomatopoeia*). The words for "sheep" and "goat", on the other hand, were surprisingly the same, the former being called *míanyáng*, which roughly translates as "soft sheep/goat", while the latter is called *shānyáng* which translates to "mountain sheep/goat".

Differences in animal naming

We were intrigued by these differences and similarities of animal names across different languages. So, we decided to investigate this further, by comparing pronunciation differences for "dog", "cat", "goat", and "sheep" across a larger sample of languages. For this purpose, we selected 28 different languages, and searched for the translations as they are given in the different Wikipedia articles. We then manually added the pronunciations, based on different sources, such as [Wiktionary](#), our own knowledge of some of the languages, or specialized sources listing translations and transcriptions (Key and Comrie 2016; Huang et al. 1992).

We then used the overall pronunciation distances for all languages as proposed by Jäger (2015), who applied sophisticated alignment algorithms to a sample of [40 historically stable words](#) per language for a large sample of North Eurasian languages (taken from the [ASJP](#) database). Since our sample contains languages which have never been shown to be historically related, the networks which we inferred from these distances should **not** be interpreted as true phylogenies, but rather as an aid for visualizing overall similarities among them.

To compare the pronunciation differences of our small datasets of animal names, we used the [LingPy](#) software (List and Forkel 2016, <http://lingpy.org>) to cluster the data into preliminary sets of phonetically similar words. As we lack the data to carry out deep inference of truly historical similarities, for this purpose we used the *Sound-Class-Based Phonetic Alignment Algorithm* (for details, see List et al. 2017). This algorithm compares words for shallow phonetic similarity with some degree of historical information. As a result, the inferred clusters do not (as we will see below) reflect true instances of cognacy (homology), but rather serve as a proxy for similarity of pronunciation.

Cats and Dogs

variety would also not necessarily feel obliged to borrow the terms from neighboring language communities.

In Hebrew (not included in Figure 1), the word for cat is כַּתוּל *khatúl*. The Celtic Irish term is *cat*, and even the Basques, with their entirely unrelated language, have the word *katu*, probably a borrowing from the surrounding Romance languages (cf. Spanish *gato*). When the Germanic tribes (BC) and Slavs (AD) arrived on horseback, accompanied by their **hunda-* (Kroonen 2013: 256), or their **pesə* (Derksen 2008: 431), they settled down, started farming, and then took up the **kattōn-* and the **kotə* from the locals. This is interesting, because we have to assume (based on genetics and modern distribution of the wild subspecies of *Felis sylvestris*) that there were always wild cats in the European woods. Either the word for them was lost in surviving languages, or the hunters and gathers living in Europe never bothered to name a small furry animal that – at best – could be just glimpsed.

Notably, the South Asian Indo-European languages and the East Asian Sino-Tibetic languages have their own terms for cats (Figure 1), but the word is globally quite invariable in stark contrast to the terms for "dog".

Where does this lead?

Our graphs are at this point indicate many curiosities. Nevertheless, by mapping words associated with animals (or plants), crucial for the history of human civilisation, we may tap into a complete new data set to discuss different scenarios erected by archaeologists and historians regarding domestication and beyond. While linguists, archaeologists, and geneticists have been working a lot on these questions on their own, examples for a rigorous collaboration, involving larger datasets and common research questions, are – to our current knowledge from sifting the literature – still rather rare. Furthermore, most linguistic accounts are anecdotal. They provide valuable insights, but these insights are not amenable for empirical investigations, as they are only reflected in prose. As a result, recent articles concentrating on archaeogenetic studies often ignore linguistic evidence completely. Given the uncertainty about the origin of domesticated animals and plants, despite advanced methods and techniques in archaeology and genetics, it seems that this strategy of simply putting linguistic evidence to one side deserves some re-evaluation.

It seems to be about time to pursue these questions in data-driven frameworks. When doing so, however, we need to be careful in the way we treat linguistic data as evidence. What we need is a thorough understanding of the processes underlying "naming" in language evolution. We constantly modify our lexicon, be it (i) by no longer using certain words, (ii) by using certain previously unfashionable words more frequently, (iii) by coining new words, or (iv) by borrowing words from our linguistic neighbors. So far, we still barely understand under which conditions societies will tend to keep a certain word against pressure from linguistic neighbors who use a different term, or when they will prefer to coin their own new words for newly introduced techniques, animals, or

plants, instead of taking the words along with the technology.

Linguists can say a few things about this; and etymological dictionaries, some of which we also consulted for this study, offer a wealth of information for some terms. However, without formalizing our linguistic knowledge, providing standardization efforts (compare the [Tsammalex](#) or the [Concepticon](#) projects) and improvement of algorithms for automatic sequence comparison, linguists will have a hard time keeping pace with quickly evolving disciplines like archaeogenetics and archaeology.

References

- Anthony, D. and D. Ringe (2015) The Indo-European homeland from linguistic and Archaeological perspectives. *Annual Review of Linguistics* 1: 199-219.
- Botigue, L., S. Song, A. Scheu, S. Gopalan, A. Pendleton, M. Oetjens, A. Taravella, T. Seregely, A. Zeeb-Lanz, R. Arbogast, D. Bobo, K. Daly, M. Unterlander, J. Burger, J. Kidd, and K. Veeramah (2017) Ancient European dog genomes reveal continuity since the Early Neolithic. *Nature Communications* 8: 16082.
- Derksen, R. (2008) *Etymological dictionary of the Slavic inherited lexicon*. Brill: Leiden and Boston.
- Driscoll, C., D. Macdonald, and S. O'Brien (2009) From wild animals to domestic pets, an evolutionary view of domestication. *Proceedings of the National Academy of Sciences* 106 Suppl 1: 9971-9978.
- Frantz, L.A., V.E. Mullin, M. Pionnier-Capitan, O. Lebrasseur, M. Ollivier, A. Perri, A. Linderholm, V. Mattiangeli, M.D. Teasdale, E.A. Dimopoulos, A. Tresset, M. Duffraisse, F. McCormick, L. Bartosiewicz, E. Gal, É.A. Nyerges, M.V. Sablin, S. Bréhard, M. Mashkour, A. Bălăşescu, B. Gillet, S. Hughes, O. Chassaing, C. Hitte, J.-D. Vigne, K. Dobney, C. Hänni, D.G. Bradley, G. Larson (2016) Genomic and archaeological evidence suggest a dual origin of domestic dogs. *Science* 352: 1228-1231.
- Huáng Bùfán 黃布凡 (1992) *Zàngmiǎn yǔzú yǔyán cíhuì* [A Tibeto-Burman lexicon]. Zhōngyāng Mínzú Dàxué 中央民族大学 [Central Institute of Minorities]: Běijīng 北京.
- Jäger, G. (2015) Support for linguistic macrofamilies from weighted alignment. *Proceedings of the National Academy of Sciences* 112: 12752-12757.
- Key, M. and B. Comrie (2016) *The intercontinental dictionary series*. Max Planck Institute for Evolutionary Anthropology: Leipzig.
- Kraft, K., C. Brown, G. Nabhan, E. Luedeling, J. Luna Ruiz, G. Coppens d'Eeckenbrugge, R. Hijmans, and P. Gepts (2014) Multiple lines of evidence for the origin of domesticated chili pepper, *Capsicum annuum*, in Mexico. *Proceedings of the National Academy of Sciences of the United States of America* 111: 6165-6170.
- Kroonen, G. (2013) *Etymological dictionary of Proto-Germanic*. Brill: Leiden and Boston.
- List, J.-M. and R. Forkel (2016) *LingPy. A Python library for historical linguistics*. Max Planck Institute for the Science of Human History: Jena.
- List, J.-M., S. Greenhill, and R. Gray (2017) The potential of automatic word comparison for historical linguistics. *PLOS ONE* 12: 1-18.
- Pfeifer, W. (1993) *Etymologisches Wörterbuch des Deutschen*. Akademie: Berlin.
- Starostin, S. (2007) Opredelenije ustojčivosti bazisnoj leksiki [Determining the stability of basic words]. In : S. A. Starostin: *Trudy po jazykoznaniju* [S. A. Starostin: Works on linguistics. Languages of Slavic Cultures: Moscow. 580-590.

Final Remark

Given that we had little time to review all of the literature on domestication in these disciplines, we may well have missed important aspects, and we may well have even failed to be original in our claims. We would like to encourage potential readers of this blog to provide us with additional hints

and productive criticism. In case you know more about these topics than we have reported here, please get in touch with us — we will be glad to learn more.

Cite as: Grimm, Guido and List, Johann-Mattis (2017): "Man gave names to all those animals": cats and dogs. *The Genealogical World of Phylogenetic Networks* 6:10, 35-41, URL: <http://phylonetworks.blogspot.com/2017/10/man-gave-names-to-all-those-animals.html>.

"Man gave names to all those animals": goats and sheep

Guido Grimm¹, Johann-Mattis List², and Cormac Anderson²

¹Independent researcher, Lyon, ²Max Planck Institute for the Science of Human History

This is the second of a pair of posts dealing with the names of domesticated animals. In [the first part](#), we looked at the peculiar differences in the names we use for cats and dogs, two of humanity's most beloved domesticated predators. In this, the second part (and with some help from Cormac Anderson, a fellow linguist from the [Max Planck Institute for the Science of Human History](#)), we'll look at two widely cultivated and early-domesticated herbivores: goats and sheep.

Similar origins, but not the same

Both goats and sheep are domesticated animals that have an explicitly economic use; and, in both cases, genetic and archaeological evidence points to the Near East as the place of domestication (Naderi et al. 2007). The main difference between the two is the natural distribution of goats (providing nourishment and leather) and sheep (providing the same plus wool). This distribution is also reflected in the phonetic (dis)similarities of the terms used in our sample of languages (Figures 1 and 2).

Capra aegagrus, the species from which the domestic goat derives, is native to the Fertile Crescent and Iran. Other species of the genus, similar to the goat in appearance, are restricted to fairly inaccessible areas of the mountains of western Eurasia (see Figure 3, taken from Driscoll et al. 2009). On the other hand, *Ovis aries*, the sheep and its non-domesticated sister species, are found in hilly and mountainous areas throughout the temperate and boreal zone of the Northern Hemisphere. Whenever humans migrated into mountainous areas, there was the likelihood of finding a beast that:

Had wool on his back and hooves on his feet,
Eating grass on a mountainside so steep
[Bob Dylan: [Man Gave Names to all those animals](#)].

Goats

Goats were actively propagated by humans into every corner of the world, because they can thrive even in quite inhospitable areas. Reflecting this, differences in the terms for "goat" generally follow the main subgroups of the Indo-European language family (Figure 1), in contrast to "cat", "dog", and "sheep". From the language data, it seems that for the most part each major language expansion, as reflected in the subgroups of Indo-European languages, brought its own term for "goat", and that it was rarely modified too much or borrowed from other speech communities.

There is one exception to this, however. The terms in the Italic and Celtic languages look as though

Cite as: Grimm, Guido, Anderson, Cormac, and List, Johann-Mattis (2017): "Man gave names to all those animals": goats and sheep. *The Genealogical World of Phylogenetic Networks* 6:11, 42-46, URL: <http://phylonetworks.blogspot.com/2017/11/man-gave-names-to-all-those-animals.html>.

and the [z] became a [r] in many Scandinavian words. The fact that both Italian and Danish plus Swedish have cognate terms for "sheep", however, does not mean that their common ancestors used the same term. It is much more likely that speakers in both communities came up with similar ways to name their most important herded animals. It is possible, for example, that this term generically meant "livestock", and that the sheep was the most prototypical representative at a certain time in both ancestral societies.

Furthermore, we see substantial phonetic variation in the Romance languages surrounding the Mediterranean, where both sheep and goats have probably been cultivated since the dawn of human civilization. Each language uses a different word for sheep, with only the Western Romance languages being visibly similar to *ovis*, their ancestral word in Latin, while Italian and French show new terms.

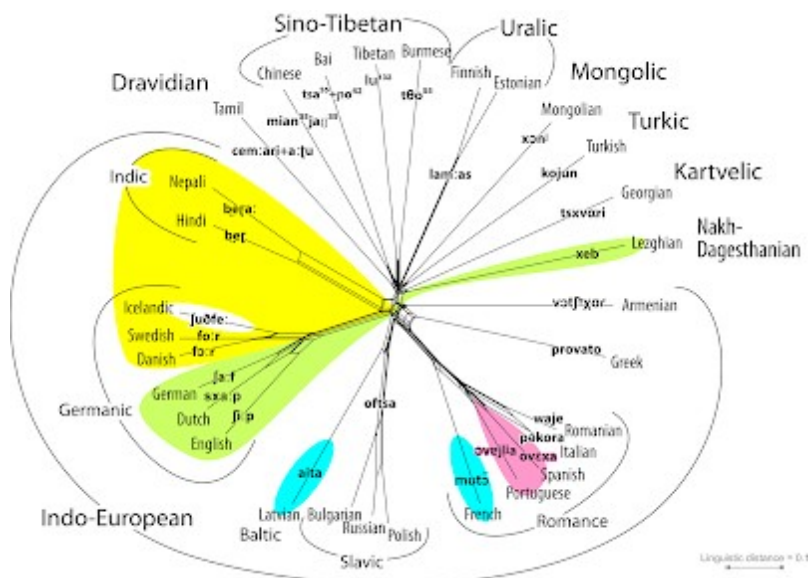


Figure 2: Phonetic comparison of words for "sheep"

More interesting aspects

The wild sheep, found in hilly and mountainous areas across western Eurasia, was probably hunted for its wool long before mouflons (a subspecies of the wild sheep) were domesticated and kept as livestock. The word for "sheep" in Indo-European, which we can safely reconstruct, was *h₂ owis*, possibly pronounced as [xovis], and still reflected in Spanish, Portuguese, Romanian, Russian, Polish. It survives in many more languages as a specific term with a different meaning, addressing the milk-bearing / birthing female sheep. These include English *ewe*, Faroese *ær* (which comes in more than a dozen combinations; Faroes literally means: "sheep islands"), French *brebis* (important to know when you want sheep-milk based cheese), German *Aue* (extremely rare nowadays, having been replaced by *Mutterschaf* "mother-sheep"). In other languages it has been lost completely.

What is interesting in this context is that while the phonetic similarity of the terms for "sheep" resembles the pattern we observe for "dog", the history of the words is quite different. While the

words for "dog" just continued in different language lineages, and thus developed independently in different groups without being replaced by other terms, the words for "sheep" show much more frequent replacement patterns. This also contrasts with the terms for "goat", which are all of much more recent origin in the different subgroups of Indo-European, and have remained rather similar after they were first introduced.

The reasons for these different patterns of animal terms are manifold, and a single explanation may never capture them all. One general clue with some explanatory power, however, may be how and by whom the animals were used. Humans, in particular nomadic societies, rely on goats to colonize or survive in unfortunate environments, even into historic times. For instance, goats were introduced to South Africa by European settlers to effectively eat up the thicket growing in the interior of the Eastern Cape Province. Once the thicket was gone, the fields were then used for herding cattle and sheep.

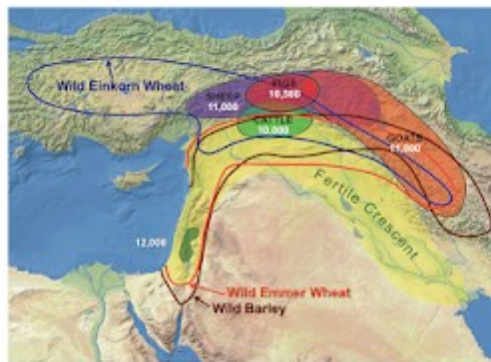


Fig. 1. Map of the Near East indicating the Fertile Crescent (according to ref. 23). Shaded areas indicate the approximate areas of domestication of pig, cattle, sheep, and goats with dates of initial domestication in calibrated years B.P. (after ref. 2). Colored lines enclose the wild ranges of Einkorn wheat, emmer wheat, and barley (after ref. 21). Green-shaded area in southern Levant indicates the region where all 3 grains were first domesticated 12,000 years B.P.

Figure 3: Map from Driscoll et al. (2009)

There are other interesting aspects of the plot.

For example, as mentioned before, in Chinese the goat refers to the "mountain sheep/goat" and the "sheep/goat" is the "soft sheep". While it is straightforward to assume that *yáng*, the term for "sheep/goat", originally only denoted one of the two organisms, either the sheep or the goat, it is difficult to say which came first. The term *yáng* itself is very old, as can also be seen from the Chinese character used, which serves as one of the base radicals of the writing system, depicting an animal with horns: 羊. The sheep seems to have arrived in China rather early (Dodson et al. 2014), predating the invention of writing, while the arrival of the goat was also rather ancient (Wei et al. 2014) (and might also have happened more than once). Whether sheep arrived before goats in China, or vice versa, could probably be tested by haplotyping feral and locally bred populations while recording the local names and establishing the similarity of words for goat and sheep.

While the similar names for goat and sheep may be surprising at first sight (given that the animals do not look all that similar), the similarity is reflected in quite a few of the world's languages, as can

be seen from the [Database of Cross-Linguistic Colexifications](#) (List et al. 2014) where both terms form a cluster.

Source Code and Data

We have uploaded source code and data to [Zenodo](#), where you can download them and carry out the tests yourself (DOI: [10.5281/zenodo.1066534](https://doi.org/10.5281/zenodo.1066534)). Great thanks goes to [Gerhard Jäger](#) (Eberhard-Karls University Tübingen), who provided us with the pairwise language distances computed for his 2015 paper on "Support for linguistic macro-families from weighted sequence alignment" (DOI: [10.1073/pnas.1500331112](https://doi.org/10.1073/pnas.1500331112)).

References

- Dodson, J., E. Dodson, R. Banati, X. Li, P. Atahan, S. Hu, R. Middleton, X. Zhou, and S. Nan (2014) Oldest directly dated remains of sheep in China. *Sci Rep* 4: 7170.
- Driscoll, C., D. Macdonald, and S. O'Brien (2009) From wild animals to domestic pets, an evolutionary view of domestication. *Proceedings of the National Academy of Sciences* 106 Suppl 1: 9971-9978.
- Jäger, G. (2015) Support for linguistic macrofamilies from weighted alignment. *Proceedings of the National Academy of Sciences* 112.41: 12752–12757.
- Kroonen, G. (2013) *Etymological dictionary of Proto-Germanic*. Brill: Leiden and Boston.
- List, J.-M., T. Mayer, A. Terhalle, and M. Urban (eds) (2014) *CLICS: Database of Cross-Linguistic Colexifications*. Forschungszentrum Deutscher Sprachatlas: Marburg.
- Naderi, S., H. Rezaei, P. Taberlet, S. Zundel, S. Rafat, H. Naghash, et al. (2007) Large-scale mitochondrial DNA analysis of the domestic goat reveals six haplogroups with high diversity. *PLoS One* 2.10. e1012.
- Pfeifer, W. (1993) *Etymologisches Wörterbuch des Deutschen*. Akademie: Berlin.
- Wei, C., J. Lu, L. Xu, G. Liu, Z. Wang, F. Zhao, L. Zhang, X. Han, L. Du, and C. Liu (2014) Genetic structure of Chinese indigenous goats and the special geographical structure in the Southwest China as a geographic barrier driving the fragmentation of a large population. *PLoS One* 9.4: e94435.

Final remark

As in the case of cats and dogs, we have reported here merely preliminary impressions, through which we hope to encourage potential readers to delve into the puzzling world of naming those animals that were instrumental for the development of human societies. In case you know more about these topics than we have reported here, please get in touch with us, we will be glad to learn more.

Cite as: Grimm, Guido, Anderson, Cormac, and List, Johann-Mattis (2017): "Man gave names to all those animals": goats and sheep. *The Genealogical World of Phylogenetic Networks* 6:11, 42-46, URL: <http://phylonetworks.blogspot.com/2017/11/man-gave-names-to-all-those-animals.html>.

The art of doing science: alignments in historical linguistics

Johann-Mattis List

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

In the past two years, during which I have been writing for this blog, I have often tried to emphasize the importance of alignments in historical linguistics — alignment involves explicit decisions about which characters / states are cognate (and can thus be aligned in a data table). I have also often mentioned that explicit alignments are still rarely used in the field.

To some degree, this situation is frustrating, since it seems so obvious that scholars align data in their head, for example, whenever they write etymological dictionaries and label parts of a word as irregular, not fulfilling their expectations when assuming regular sound change (in the sense in which I have described it [before](#)). It is also obvious that linguists have been trying to use alignments before (even before biologists, as I tried to show in [this earlier post](#)), but for some reason, they never became systematized.

As an example for the complexity of alignment analyses in historical linguistics, consider the following figure, which depicts both an early version of an alignment (following [Dixon and Kroeber 1919](#)), and a "modern" version of the same data. For the latter, I used the EDICTOR (<http://edictor.digling.org>), a software tool that I have been developing during recent years, and which helps linguists to edit alignments in a consistent way ([List 2017](#)). The old version on the left has been modified in such a way that it becomes clearer what kind of information the authors tried to convey (for the original, see [my older post](#)), while the EDICTOR version contains some markup that is important for linguistics, which I will discuss in more detail below.



Figure 1: Alignments from Dixon and Kroeber (1919) in two flavors

If we carefully inspect the first alignment, it becomes evident that the scholars did not align the data *sound by sound*, but rather *morpheme by morpheme*. Morphemes are those parts in words that are supposed to bear a clear-cut meaning, even when taken in isolation, or when abstracting from multiple words. The plural-ending -s in English, for example, is a morpheme that has the function to indicate the plural (compare *horse* vs. *horses*, etc.). In order to save space, the authors used

Cite as: List, Johann-Mattis (2017): The art of doing science: alignments in historical linguistics. *The Genealogical World of Phylogenetic Networks* 6:12, 47-51, URL:<http://phylonetworks.blogspot.com/2017/12/the-art-of-doing-science-alignments-in.html>.

abbreviations for the language group names and the names for the languages themselves.

The authors have further tried to save space by listing identical words only once, but putting two entries, separated by a comma, in the column that I have labelled "varieties". If you further compare the entries for NW (=North-Western Maidu) and NE/S (=North-Eastern Maidu and Southern Maidu), you can see that the first entry has been swapped: the *tsi'* in *tsi'-bi* in NW is obviously better compared with the *tsi* in NE/S *bi-tsi* rather than comparing *bi* in NE with *tsi* in NE/S. This could be a typographical error, of course, but I think it is more likely that the authors did not quite know how to handle swapped instances in their alignment.

In the EDICTOR representation of the alignment, I have tried to align the sounds in addition to aligning the morphemes. My approach here is rather crude. In order to show which sounds most likely share a common origin, I extracted all homologous morphemes, aligned them in such a way that they occur in the same column, and then stripped off the remaining sounds by putting a checkmark in the IGNORE column on the bottom of the EDICTOR representation. When further analyzing these sound correspondences with some software, like the LingPy library ([List et al. 2017](#)), all sounds that occur in the IGNORE column will be ignored. Correspondences will then only be calculated for the core part of this alignment, namely the two columns that are left over, in the center of the alignment.

In many cases, this treatment of sound correspondences and homologous words in alignments is sufficient, and also justified. If we want to compare the homologous (cognate) parts across words in different languages, we can't align the words entirely. Consider, for example, the German verb *gehen* [ge:ən] and its English counterpart *go* [gəʊ]. German regularly adds the infinitive ending *-en* to each verb, but English has long ago dropped all endings on verbs apart from the *-s* in the third person singular (compare *go* vs. *goes*). Comparing the whole of the verbs would force us to insert gaps for the verb ending in German, which would be linguistically not meaningful, as those have not been "gapped" in English, but lost in a morphological process by which *all endings* of English verbs were lost.

There are, however, also cases that are more complicated to model, especially when dealing with instances of *partial cognacy* (or *partial homology*). Compare, for example, the following alignment for words for *bark (of a tree)* in several dialects of the Bai language, a Sino-Tibetan language spoken in China, whose affiliation with other Sino-Tibetan languages is still unclear (data taken from [Wang 2006](#)).

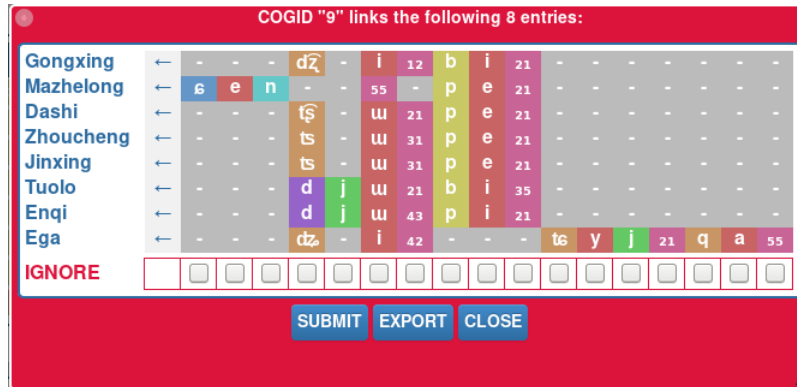


Figure 2: Alignment for words for "bark" in Bai dialects

In this example, the superscript numbers represent tones, and they are placed at the end of each syllable. Each syllable in these languages usually also represents a morpheme in the sense mentioned above. That means, that each of the words is a compound of two original meanings. Comparison with other words in the languages reveals that most dialects, apart from Mazhelong, express *bark* as *tree-skin*, which is a very well-known expression that we can find in many languages of the world. If we want to analyze those words in alignments, we could follow the same strategy as shown above, and just decide for one core part of the words (probably the *skin* part) and ignore the rest. However, for our calculations of sound correspondences, we would lose important information, as the *tree* part is also cognate in most instances and therefore rather interesting. But ignoring only the unalignable part of the first syllable in Mazhelong would also not be satisfying, since we would again have gaps for this word in the *tree* part in Mazhelong which do not result from sound change.

The only consistent solution to handle these cases is to split the words into their morphemes, and then to align all sets of homologous morphemes separately. This can also be done in the EDICTOR tool (but it requires more effort from the scholar and the algorithms). An example is shown above, where you can see how the tool breaks the linear order in the representation of the words as we find them in the languages, in order to cluster them into sets of homologous "word-parts".



Figure 3: Alignments of partial cognates in the Bai dialects

But if we only look at the *tree* part of those alignments, namely the third cognate set from the left, with the ID 8, we can see a further complication, as the gaps introduced in some of the words look a little bit unsatisfying. The reason is that the *j* in Enqi and Tuolo may just as well be treated as a part

of the initial of the syllable, and we could re-write it as *dj* in one segment instead of using two. In this way, we might capture the correspondence much more properly, as it is well known that those affricate initials in the other dialects ([ts, tʃ, dz, dz̥]) often correspond to [dj]. We could thus rewrite the alignment as shown in the next figure, and simply decide that in this situation (and similar ones in our data), we treat the *d* and the *j* as just one main sound (namely the initial of the syllables).

● Cognate set "8" links the following 7 entries:

Gongxing	dʒ̥	i	12
Dashi	tʃ̥	u	21
Zhoucheng	ts	u	31
Jinxing	ts	u	31
Tuolo	dj	u	21
Enqi	dj	u	43
Ega	dz̥	i	42

EDIT ALIGN CLOSE

Figure 4: Revised alignment of "tree" in the sample

Summary and conclusions

Before I start boring those of the readers of this blog who are not linguists, and not particularly interested in details of sound change or language change, let me just quickly summarize what I wanted to illustrate with these examples. I think that the reason why linguists never really formalized alignments as a tool of analysis is that there are so many ways to come up with possible alignments of words, which may *all* be reasonable for any given analysis. In light of this multitude of possibilities for analysis, not to speak of historical linguistics as a discipline that often prides itself by being based on hard manual labor that would be impossible to achieve by machines, I can in part understand why linguists were reluctant to use alignments more often in their research.

Judging from my discussions with colleagues, there are still many misunderstandings regarding the purpose and the power of alignment analyses in historical linguistics. Scholars often think that alignments directly reflect sound change. But how could they, given that we do not have any ancestral words in our sample? Alignments are a tool for analysis, and they can help to identify sound change processes or to reconstruct proto-forms in unattested ancestral languages; but they are by no means the true reflection of what happened and how things changed. They are the starting point, not the end point of the analysis. Furthermore, given that there are many different ways in which we can analyze how languages changed over time, there are also many different ways in which we can analyze language data with the help of alignments. Often, when comparing different alignment analyses for the same languages, there is no simple *right* and *wrong*, just a different

emphasis on the initial analysis and its purpose.

As David wrote in an email to me:

"An alignment represents the historical events that have occurred. The alignment is thus a static representation of a dynamic set of processes. This is ultimately what causes all of the representational problems, because there is no necessary and sufficient way to achieve this."

This also nicely explains why alignments in biology as well, with respect to the goal of representing *homology*, "may be more art than science" ([Morrison 2015](#)), and I admit that I find it a bit comforting that biology has similar problems, when it comes to the question of how to interpret an alignment analysis. However, in contrast to linguists, who have never really given alignments a chance, biologists not only use alignments frequently, but also try to improve them.

If I am allowed to have an early New Year wish for the upcoming year, I hope that along with the tools that facilitate the labor of creating alignments for language data, we will also have a more vivid discussion about alignments, their shortcomings, and potential improvements in our field.

References

- Dixon, R. and A. Kroeber (1919) *Linguistic families of California*. University of California Press: Berkeley.
- List, J.-M. (2017) A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*, pp. 9-12.
- LingPy: A Python library for historical linguistics*. Version 2.6. Max Planck Institute for the Science of Human History: Jena.
- Morrison, D. (2015) Molecular homology and multiple-sequence alignment: an analysis of concepts and practice. *Australian Systematic Botany* 28: 46-62.
- Wang, W.-Y. (2006) Yǔyán, yǔyīn yǔ jìshù \hàna 語言,語音與技術 [*Language, phonology and technology*]. Xiānggǎng Chéngshì Dàxué: Shànghǎi 上海.

Cite as: List, Johann-Mattis (2017): The art of doing science: alignments in historical linguistics. *The Genealogical World of Phylogenetic Networks* 6:12, 47-51,
URL:<http://phylonetworks.blogspot.com/2017/12/the-art-of-doing-science-alignments-in.html>.