Day1: From Cognates to Correspondences

Johann-Mattis List (DLCE, MPI-EVA Leipzig)

1 Introduction

Our goal for this first session in our little course on quantitative and formal methods in historical language comparison is to discuss some major developments which were made in the field of historical linguistics in the past years, and which – in my opinion – may also have some importance for the practitioners of language comparison, especially those who are not independently interested in quantitative methods.

1.1 Modeling and Representation

In the sciences, scholars often talk about *modeling*. Scholars *model* sound change, they *model* language change, and they try to *model* lexical borrowing. It is not always clear what is meant with the term *modeling*, and it seems that scholars use it with varying ideas in mind. If we talk about *modeling* in the context of quantitative and formal approaches to historical language comparison, I use the term *model* in the sense of what Bröker and Ramscar (2021) call an *implemented model*. While a general model can also exist of a prose explanation of the mechanisms underlying a phenomenon, an *implemented model* is a model which can be shown to work in some piece of software and applied to some data.

> To explain why the contributions of representations, algorithms, and computations will only rarely manifest themselves in fully independent ways [...], it is important to recognise that in practice, models in the brain and cognitive sciences are typically presented in one of two distinct ways: either as abstract model descriptions, or as implemented models. Abstract model descriptions typically comprise symbolic (i.e. verbal or algebraic) descriptions of the relationships between what are typically quite loosely defined quantities or entities. Accordingly, while abstract models can appear to be more or less "formal", they typically fail to fully specify representations (what exactly will be counted and in which format) and typically fail to fully specify the algorithms that will transform these representations into predictions (Figure 6). It is in fact only when these latter steps are made, and an abstract model is actually implemented, that it can be considered formal in any meaningful sense. (ibid.: 17/25)

Of crucial importance for implemented models is the way in which data are *represented*, since this determines how the implementation works. In the work I will present, for example, we may conveniently represent language data (words in the lexicon of a language, etc.) in the form of tables. These can be printed to paper, but they can also be typed into spreadsheets on the computer. The representation of data is thus the basis upon which we build our models and implement them in computer code.

> To recapitulate: Representational choices can significantly alter the performance of a model, the predictions it makes and thus the way it is interpreted. (ibid.: 20/25)

The distinction between models, implemented models, and representations, does not define the term "model" itself. Atkinson and Gray (2006: 94) write about models, that they are "lies that lead us to the truth". Is this a useful characterization of models?

1.2 Integrated Data Representations

When working with data, scholars often use very different representations of their data. They may have one file for their syntactic properties they collect, one word document, where they collect their favorite quotes, another spreadsheet where they started to collect sound changes, and some old FileMaker database, which they still use for convenience, to enlarge their personal etymological dictionary of their favorite language. When working with data, scholars also often commit certain common errors in data collection. The most common errors are to extract information from sources without storing a reference to the original sources, or to copy text from some resource into a cell in a spreadsheet and later modify this content manually without keeping the original raw data.

As of now, there are many good guidelines for working transparently with data out in the internet (Perkel 2022), and I recommend that all who feel a bit insecure about how to collect data properly to inform themselves about these resources and generally take much more time in planning or experimenting with different formats of data representation than starting to collect data and eventually destroying information without having intended to do so. I also recommend to think about *integrated data representation*, that is, to think about ways to work on different questions with the same data, and to extract certain important aspects of the annotation of a dataset rather than paste it into a separate data sheet. As an example, scholars may store a dictionary of a given language written in orthography, and additionally they may type off the phoneme inventory of that language from another resource and collect these separately. It would be much better to work on a dictionary in phonetic transcription from which the same information could be derived (the inventory should be extractable from the dictionary). Examples for integrated data handling have been recently published by our group in the form of the Lexibank repository (List et al. 2022a), where we compute several lexical and phonological features of various languages from the wordlists, which we have collected and standardized.

Why is it so important to keep the raw data when collecting data for one's studies?

2 Cognates

The starting point for many analyses in the field of historical language comparison is the identification of cognate words across related languages. It is well known that this is typically done in an iterative fashion when applying the comparative method (Ross and Durie 1996). This means that scholars usually first identify some potential cognates, then search for sound correspondences among those, then identify more cognates, kick out some that turn out to be wrong, etc. In order to formalize this procedure, we do not only need to be clear about the way in which we want to represent words in our data (we will discuss this in detail when discussing phonetic alignments), but also how we want to define *cognacy* in this context. It turns out that this is far more difficult than one might think at first sight.

2.1 Terminology

In order to approach the question of how to model, represent, infer, and annotate relations of cognacy, it is important to get a clearer picture of the terminology we use in the field of historical linguistics and beyond. Although most historical linguistics would probably confirm that they have a very clear idea of the term *cognacy* and its meaning, we can find quite divergent applications in the practice. Historical linguists usually agree that cognacy describes the relation between words that share a common ancestor form and have descended from this common form only via vertical inheritance – as opposed to lateral transfer, which would point to borrowing events. When it comes to identifying cognates in multilingual datasets, however, we can quickly see that there may be various disagreements in practice.

I attribute these problems to the insufficiency of the term *cognacy* to capture fundamental relations we want to handle in our field. As I have shown before, our terminology would profit a lot from being modified and made more precise and clear, reflecting fundamental relations of descent. The following table (taken from List 2016, which is an extension of List 2014), we find terms for cognacy in linguistics (along with suggestions for new terms) contrasted with typical terms for *homology* in evolutionary biology.

<table>
<tr><th colspan="2" rowspan="2">Historical relations</th><th colspan="5">Terminology</th></tr>
<tr><th colspan="2">Biology</th><th colspan="3">Linguistics</th></tr>
<tr><td rowspan="3">Common descent</td><td>Direct</td><td rowspan="3">Homology</td><td>Orthology</td><td rowspan="3">Etymological relation</td><td rowspan="2">Cognacy</td><td>Direct cognacy</td></tr>
<tr><td>Indirect</td><td>Paralogy</td><td>Indirect cognacy</td></tr>
<tr><td>Involving lateral transfer</td><td>Xenology</td><td colspan="2">Indirect etymological relation</td></tr>
</table>

It seems that the linguistic view on relatedness, when comparing it with the view reflected in the terminology of evolutionary biology, is very biased towards borrowing. What reasons may contribute to this view?

2.2 Relations

While the specification of the terminology on cognacy was helpful for some time, it turned out that it was still not capturing all important aspects which we need for a formal representation of cognate sets for the purpose of historical language comparison. What was specifically missing was a dimension that would allow us to determine if a cognate relation can be considered *regular* or not, since this in turn is crucial for the identification of regular sound correspondences. The following table (taken from Schweikhard and List 2020) revises the previous table on cognate relations by adding regularity as a new dimension and asking more clearly how each of the dimension must be specified.

<table>
<tr><th rowspan="2">Relation</th><th rowspan="2">Biological term</th><th rowspan="2">Regularity</th><th>Morphological</th><th>Semantic</th><th>Stratic</th></tr>
<tr><th colspan="3">continuity</th></tr>
<tr><td>traditional notion of cognacy</td><td>-</td><td>+/–</td><td>+/–</td><td>+/–</td><td>+</td></tr>
<tr><td>cognacy à la Swadesh</td><td>-</td><td>+/–</td><td>+/–</td><td>+</td><td>+</td></tr>
<tr><td>direct cognacy</td><td>~~orthology~~</td><td>+/–</td><td>+</td><td>+/–</td><td>+</td></tr>
<tr><td>oblique cognacy</td><td>-</td><td>+/–</td><td>–</td><td>+/–</td><td>+</td></tr>
<tr><td>etymological relation</td><td>homology</td><td>+/–</td><td>+/–</td><td>+/–</td><td>+/–</td></tr>
<tr><td>oblique etymological relation</td><td>xenology</td><td>+/–</td><td>+/–</td><td>+/–</td><td>–</td></tr>
<tr><td>**strict cognacy**</td><td>-</td><td>+</td><td>+/–</td><td>+/–</td><td>+/–</td></tr>
</table>

This table also introduces a new term, following List (2018), *strict cognacy*, which is unfortunately misrepresented in the table, in so far as it must require morphological and stratic continuity, since strict cognates are defined as word forms which differ only with respect to regular sound change and semantics. We will see that these cognates are crucial for alignment analyses and for the identification of regular sound correspondences, which depend on alignments.

How can we deal with cases in which parts of a word are strictly cognate (the root) while other parts aren't?

3 Alignments

Phonetic alignments are a central concept of the formal and quantitative work on historical language comparison which our group pursues. The crucial idea of phonetic alignments is to represent words as a sequence of sounds, which in turn allows us to handle the comparison of words as the problem of comparing two or more sequences, which is most typically done with the help of alignment analyses (List 2014). Implicitly, alignment analyses have been used for a long time in historical linguistics, but explicitly, it has only been recently that they started to trigger the interest of scholars, as reflected in a growing amount of studies devoted to the topic (Kondrak 2000, List 2014, Prokić et al. 2009).

3.1 Representing Words as Sequences of Sounds

Before we can align words, we need to discuss how we want to represent them as sequences of sounds. Here, we can see a lot of variation in scholarly practice, mostly resulting from the fact that scholars insist to stick to traditional orthography or traditional phonological representations when comparing cognate words in their specific language family. Traditions are very strong here, and scholars are often unwilling to give them up. For a proper representation of words as sequences of sounds, however, it is indispensable to have some clear notion of how sounds should be represented and how they should be chained up to form a sequence.

The solution we pursue by now is to use a specific – and much stricter – version of the International Phonetic Alphabet, named B(road)IPA, published as part of the Cross-Linguistic Transcription Systems Reference catalog (`https://clts.clld.org`, Anderson et al. 2018, List et al. 2021), which defines more than 8000 different sound segments and is based on a generative component which allows to define more possible sounds on the fly. We represent words in this transcription, but we do not have any strong opinion on whether the representation should be phonetic or phonological. We tend to recommend phonological representations which do not exaggerate phonological theory (e.g., by proposing that Chinese has only one vowel).

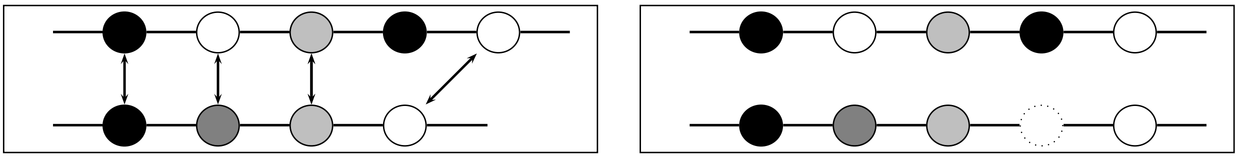
In order to represent a word, distinct *sounds* need to be written with a space as separator, so that we can avoid ambiguities which may result from pre- vs. post-aspirated plosives, or various affricates, such as [ts] or [tʃ]. Morpheme boundaries are written with the help of a plus symbol, thus, German *Hausmeister* "housekeeper" could be represented as [h au s + m ai s t ə r].

In addition, we are currently experimenting with the possibility to *cluster* sounds (which are still listed as separate sounds but marked as belonging together, using a dot instead of a space), if it is known that they tend to evolve together (Hill and List in preparation). Thus one could represent all instances of German [s t] as [s.t], since – as we know – the two sounds tend to behave in a specific manner when it comes to their sound change patterns.

Are there certain sounds which would universally tend to evolve together and thus qualify as candidate for a grouped representation?

3.2 Aligning Sound Sequences

An alignment between a certain number of sequences is a technique by which the sequences are arranged in a matrix in such a way that all corresponding segments in the sequences appear in the same columns, while empty slots, resulting from segments which correspond with not other segment in another sequence are filled with *gap* symbols. We follow the tradition of evolutionary biology in representing gaps with a dash. This is also the reason why we use the plus character for morpheme boundaries. In the following figure, a pairwise alignment of two chains is shown, which illustrates the major idea of alignment analyses, which introduce gaps in order to adjust the lengths of two different sequences.



For phonetic alignments, we do nothing else than that, we introduce gap symbols when we find that a sound in a word from one language has no counterpart in the other language (such as the [n] in German *anderer* “other” not having a counterpart in English *other*).

How can we deal with cases of metathesis when carrying out pairwise or multiple word alignments?

3.3 “Alignability”

In Schweikhard and List (2020) we introduced the term *alignability*, in order to emphasize that not all alignments which we can think of in historical linguistics are also meaningful.

> This distinction accounts for the relation of strict as opposed to loose cognates and embraces the fact that only word forms which are strictly cognate can be aligned in a meaningful way. (ibid.: 9/25)

As an example for problems of alignability, compare the two alignments given in the following table (originally outlined in List 2014).

Language	**Alignment**						
Russian	s	-	ɔ	n	ts	ə	-
Polish	s	w	ɔ	nʲ	ts	ɛ	-
French	s	-	ɔ	l	-	ɛ	j
Italian	s	-	o	l	-	e	-
German	s	-	ɔ	n	-	ə	-
Swedish	s	-	uː	l	-	-	-

(a) Global Alignment

Language	**Alignment**							
Russian	s	ɔ	-	-	n	ts	ə	
Polish	s	-	w	ɔ	nʲ	ts	ɛ	
French	s	ɔ	l	-	-	-	-	ɛj
Italian	s	o	l	-	-	-		e
German	s	ɔ	-	-	-	-		nə
Swedish	s	uː	l	-	-	-	-	

(b) Local Alignment

Here, we have different words for *sun* in various Indo-European languages, which we could align in a naive way as shown on the right, or in a historically more informed way, as shown on the right. It turns out, that in the proposed informed alignment, we are only left with the initial, since we have not (yet) figured out a proper way of handling metathesis (as observed in the Polish word form).

However, a closer look at the word forms also reveals that they may not be *alignable* after all, given that they reflect two distinct Indo-European roots of *sun*, which are thought to reflect a unique (!) root alternation, only observed for the term *sun* in Proto-Indo-European, involving **-l-* and **-n-*. Thus, instead of aligning the Polish form with the German form, which both have inherited different reflexes of the alternating paradigm in Indo-European, it may be better to avoid aligning both words after all, and instead try to align only those word forms which go back to the same original root form.

Can the dot-representation for the grouping of sounds help to resolve some potential problems of multiple alignments in historical language comparison?

4 Correspondences

One of the biggest breakthroughs which we have made in our group during the past years was to find ways to handle correspondence patterns formally. The major ideas was presented in a study by List (2019) which also introduces an algorithm by which correspondence patterns can be inferred from phonetic alignments across multiple languages. Since then, we are exploring correspondence patterns in different language families and try hard to make their handling easier. However, until now, the handling and our understanding of correspondence patterns is still not perfect, and much more work will be needed in the future. In order to understand the major idea behind the notion of correspondence patterns, it is important to go back to the work of Anttila (1972).

4.1 Alignment Sites

The following table shows regular sound correspondences across four Indo-European languages, illustrated with help of alignments along the lines of Anttila (ibid.: 246). In contrast to the original illustration, lost sounds are displayed with help of the dash '-' as a gap symbol, while missing words (where no reflex in Gothic or Latin could be found) are represented by the '∅' symbol.

		A		B			C		D				E					F		
Sanskrit	y	u	g	a	m	dh	u	h	i	(tar)	s	n	u	ṣ	(ā)	-	r	u	dh	(iras)
Greek	z	u	g	o	n	th	u	g	a	(ter-)	-	n	u	-	(os)	e	r	u	th	(rós)
Latin	i	u	g	u	m	Ø	Ø	Ø	Ø	(Ø)	-	n	u	r	(us)	-	r	u	b	(er)
Gothic	j	u	k	-	-	d	au	h	-	(tar)	Ø	Ø	Ø	Ø	(Ø)	Ø	Ø	Ø	Ø	(Ø)
Gloss			'yoke'					'daughter'					'daughter-in-law'					'red'		

This illustration highlights six *alignment sites*, that is, columns of an alignment. These sites are chosen, because they are useful to illustrate the concept of *sound correspondence patterns* or simply *correspondence* patterns, which we have been trying to formalize during the past years. In contrast to an alignment site in isolation, a correspondence pattern is a patterns of sounds across different related languages (one sound per language) which is reflected in *many* alignment sites. When looking at the illustration above, we can identify two major patterns in the six sites, namely AEF on the one hand and CEF on the other hand. A and C cannot form a pattern, since they are in conflict with respect to the sound in Gothic. E and F are compatible with both A and C, because they both have *missing data*, which masks the true structure of the pattern to which they belong. The major task of historical language comparison consists in the identification of correspondence patterns in a language family.

Correspondence patterns have been used to predict word forms for which no cognate reflex had been identified before. How could this work, when considering the example above?

4.2 Correspondence Patterns in the Literature

In the literature, correspondence patterns are – unfortunately – typically only shown in a condensed form, which does not allow for concrete predictions or applications. As an example, consider the following table illustrating correspondence patterns identified by Clackson (2007: 37), taken from List (2019: 142).

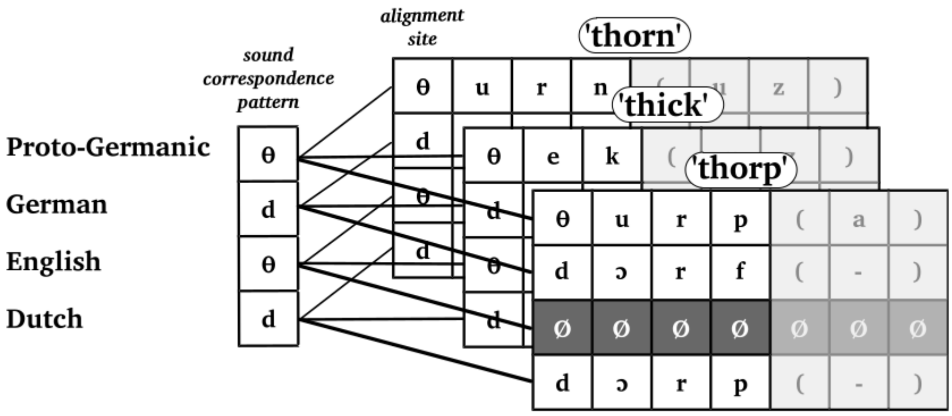
PIE	Hittite	Sanskrit	Greek	Latin	Gothic	...
*p	p	p	p	p	f b	...
*b	b p	b	b	b	p	...
*b^h	b p	b^ɦ/bh	p^h/ph	f b	b	...
*t	t	t	t	t	θ/þ d	...
*d	d t	d	d	d	t	...
*d^h	d t	d^h/dh h	t^h/th	f d b	d	...
...	...	...	...	...	...	...
*k^w	k^w/ku	k c	k p t	k^w/qu	h^w/hw g	...
*g^w	k^w/u	g j	g b d	g^w/gu u	q	...
*g^{wh}	k^w/ku g^w/gu	g^ɦ/gh h	p^h/ph t^h/th k^h/kh	f g^w/gu u	g b	...

The problem of the table is that it does not show how individual patterns look (e.g., if we have two sounds in Hittite as reflexes of Proto-Indo-European *b^h*, are their distributions the same as for the two sounds in Gothic? One can surely argue that this depends on the conditioning context, which is missing here, but one could as well argue that the concrete patterns derived from concrete alignments would give us much more information here, since they would also allow us to assess the frequency of the patterns in the data, etc.

> The representation distinguishes different sound reflexes separated by a space and by a dash, what is the difference here?

4.3 Correspondence Pattern Identification

The following figure (taken from ibid.) summarizes some of the terminology discussed so far. We consider the identification of correspondence patterns in aligned cognate sets as one of the key tasks of historical language comparison, since it is the basis of linguistic reconstruction and crucial for the evaluation of cognate sets and the justification of regular sound change.



In List (ibid.), I proposed an algorithm that helps to identify regular sound correspondence pattern from aligned cognate sets. This method was then applied to *predict* unobserved reflexes of cognate sets of Western Kho-Bwa languages in a field word experiment which were later verified through additional field work (Bodt and List 2022). It was also expanded as a method for automated supervised phonological reconstruction (List et al. 2022b) and later used as a baseline for a shared task on word prediction (List et al. 2022c). But apart from the fact that this method works quite well, I consider it much more important now, that we have established a much stricter *representation* of correspondence patterns than what has been used in the literature so far. This representation is specifically useful in combination with proto-forms, as it allows scholars to check language by language how regular their cognate sets are, and to search for conditioning contexts that explain why one and the same proto-form show more than one reflex in the same language. We will illustrate the manual inspection of correspondence patterns in more detail during the fourth lecture. The following table (taken from List 2019) gives a short idea on the potential of this representation in showing correspondence patterns for Chinese dialects derived from Middle Chinese voiced alveolar initials (**d*) in dependence of the Middle Chinese tone of the syllables in question. As can be seen, we find a strict division in diverging patterns resulting from **d* under the even (P) tone, as opposed to the other three tones (SQR).

#	Cogn.	MC	MC Tones	BJ	SZ	CS	NC	MX	TY	GZ	FZ	TB
30	13	*t	PSQR	t	t	t	t	t	t	t	t	t
41	9	*th	PSQR	t^h	t^h	t^h	t^h	t^h	t^h	t^h	t^h	t^h
747	3	*d	P	t^h	d	t	t^h	t^h	t^h	t^h	t^h	t^h
718	2	*d	P	t^h	d	t	t^h	t^h	t^h	t^h	t	t^h
719	1	*d	P	t^h	d	t	t^h	t	t^h	t^h	t	t^h
1,096	1	*d	P	t^h	Ø	t	t^h	t^h	t^h	t^h	t	t
73	5	*d	QR	t	d	t	t^h	t^h	t^h	t	t	t^h
484	1	*d	R	t	d	t	l	t^h	t^h	t	t	t^h
654	1	*d	S	t	d	t	t^h	t	t^h	t	t	t^h

How can we explain the irregularity of certain patterns, and how can represent irregularity in our data?

References

Anderson, C., T. Tresoldi, T. C. Chacon, A.-M. Fehn, M. Walworth, R. Forkel, and J.-M. List (2018). "A Cross-Linguistic Database of Phonetic Transcription Systems." *Yearbook of the Poznań Linguistic Meeting* 4.1, 21–53.

Anttila, R. (1972). *An introduction to historical and comparative linguistics.* New York: Macmillan.

Atkinson, Q. D. and R. D. Gray (2006). "How old is the Indo-European language family? Illumination or more moths to the flame?" In: *Phylogenetic methods and the prehistory of languages.* Ed. by P. Forster and C. Renfrew. Cambridge, Oxford, and Oakville: McDonald Institute for Archaeological Research, 91–109.

Bodt, T. A. and J.-M. List (2022). "Reflex prediction. A case study of Western Kho-Bwa." *Diachronica* 39.1, 1–38.

Bröker, F. and M. Ramscar (2021). "Representing absence of evidence: Why algorithms and representations matter in models of language and cognition." *Language, Cognition and Neuroscience* 37.1, 1–24.

Clackson, J. (2007). *Indo-European linguistics.* Cambridge: Cambridge University Press.

Kondrak, G. (2000). "A new algorithm for the alignment of phonetic sequences." In: *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference.* (Seattle, 04/29–05/03/2000), 288–295.

List, J.-M. (2014). *Sequence comparison in historical linguistics.* Düsseldorf: Düsseldorf University Press.

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

— (2018). *Regular cognates: A new term for homology relations in linguistics.* Vol. 5. 8.

— (2019). "Automatic inference of sound correspondence patterns across multiple languages." *Computational Linguistics* 45.1, 137–161.

List, J.-M., C. Anderson, T. Tresoldi, and R. Forkel (2021). *Cross-Linguistic Transcription Systems. Version 2.1.0.* Jena: Max Planck Institute for the Science of Human History. URL: `https://clts.clld.org`.

List, J.-M., R. Forkel, S. J. Greenhill, C. Rzymski, J. Englisch, and R. D. Gray (2022a). "Lexibank, A public repository of standardized wordlists with computed phonological and lexical features." *Scientific Data* 9.316, 1–31.

List, J.-M., N. W. Hill, and R. Forkel (2022b). "A new framework for fast automated phonological reconstruction using trimmed alignments and sound correspondence patterns." In: *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change.* (Dublin, 05/26–05/27/2022). Association for Computational Linguistics. Dublin, 89–96.

List, J.-M., E. Vylomova, R. Forkel, N. Hill, and R. D. Cotterell (2022c). "The SIGTYP shared task on the prediction of cognate reflexes." In: *Proceedings of the 4th Workshop on Computational Typology and Multilingual NLP.* "SIGTYP 2022" (Seattle, 07/14/2022). Association for Computational Linguistics. Seattle: Max Planck Institute for Evolutionary Anthropology, 52–62.

Perkel, J. M. (2022). "Six tips for better spreadsheets."

Prokić, J., M. Wieling, and J. Nerbonne (2009). "Multiple sequence alignments in linguistics." In: *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education.* "LaTeCH-SHELT&R 2009" (Athens, 03/30/2009), 18–25. acm: `1642052`.

Ross, M. and M. Durie (1996). "Introduction." In: *The comparative method reviewed. Regularity and irregularity in language change.* Ed. by M. Durie. New York: Oxford University Press, 3–38.

Schweikhard, N. E. and J.-M. List (2020). "Developing an annotation framework for word formation processes in comparative linguistics." *SKASE Journal of Theoretical Linguistics* 17.1, 2–26.