Sequence Comparison

Johann-Mattis List (University of Passau)

1 Background

Many structures we are dealing with on a daily basis can be modeled as sequences. Movies are sequences of pictures, songs are sequences of sounds, and recipes are sequences of instructions. What they all have in common is that they can be seen as ordered chains of objects whose 'identity is a product of their *order* and their *content*' (List 2014: 63). Due to the pervasiveness of sequences in our lives, sequence *comparison* is an important topic across many scientific disciplines. Especially in biology and computer science, many general problems can be reduced to the comparison of sequences. Several solutions to common problems in the field of sequence comparison have been developed so far. As a result, when trying to develop new methods for the field of comparative computational linguistics, it is useful to start from reviewing those methods that are already available and which have been discussed and reviewed in due detail.

Can recipes always be reduced to sequences of instructions?

Discreteness and Continuity

Objects modeled as sequences are not always *discrete* but may also appear as representing some a function of a continuous variable, such as space or time (Kruskal 1983: 130). If we treat them as sequences, it means we have to make them discrete before investigating them. In linguistics, we have a long tradition of making the continuous discrete, as can be prominently seen from the way we handle the speech signal. While speech is something continuous, and 'neither the movements of the speech organs nor the acoustic signal offers a clear division of speech into successive phonetic units' (IPA 1999: 5), humans have for a very long time been treating speech as something that can be segmented into certain units, be they alphabetic, segmenting speech up to the level of distinct sounds, or 'morpheme-syllabic' (Chao 1968: 108), such as the Chinese writing system, segmenting speech into blocks that are supposed to represent meaningful elements of speech.

Chinese Traditional Phonology, an early linguistic discipline in China, did not distinguish entire sounds, as we do in alphabetic writing systems, but rather made a distinction between 'initials' and 'finals' of a syllable, that is, the starting sound (the onset) and the final sounds (the rhyme). Would this be a suitable way to handle German speech?

Defining a Sequence

We can define a sequence as follows (taken from List 2014: 65).

> Given an alphabet (a non-empty finite set, whose elements are called characters), a sequence is an ordered list of characters drawn from the alphabet. The elements of sequences are called segments. The length of a sequence is the number of its segments, and the cardinality of a sequence is the number its unique segments. (cf. Böckenbauer and Bongartz 2003: 30f)

Additionally, we can define certain properties or relations of sequences (taken from List 2014: 65f):

(a) t is a subsequence of s, if t can be derived from s by deleting some of the segments of s without changing the order of the remaining segments,

(b) t is a substring of s, if t is a subsequence of s and the derivation of t from s can be carried out by deleting only elements from the beginning and the end of s,

(c) t is a prefix of s, if t is a substring of s and the derivation of t from s can be carried out by deleting only elements from the end of s,

(d) t is a suffix of s, if t is a substring of s and the derivation of t from s can be carried out by deleting only elements from the beginning of s.

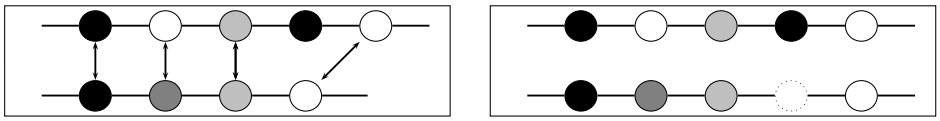
Why is it important to distinguish a subsequence of a substring, and what would be a general term for both suffix and prefix?

2 Phonetic Alignment

Alignment Analyses in General

Alignments are the most popular way to compare differences in sequences. We can define an alignment of two sequences as follows:

> An alignment of n ($n > 1$) sequences is a matrix of n rows in which all sequences are arranged in such a way that all segments which correspond to each other are placed in the same column, while segments not corresponding to other segments in a given sequence are represented with help of gap symbols in the sequence which lacks the given segment. (Gusfield 1997: 216)



The Levenshtein distance between two sequences S_1 and S_2 is defined as the number of edit operations needed to convert S_1 into S_2. With help of alignments, this can be easily handled and illustrated. How exactly?

Phonetic Alignment Analyses in Specific

Although alignment analyses are a very general way to compare sequences, they are not frequently being used in historical linguistics. Obviously, historical linguists align words in their heads, because without alignments, we could never identify regular sound correspondences, but most of the time, these comparisons are carried out implicitly, and they are rarely visualized. In addition, we often have problems when comparing words, since not all elements in historically related words are necessarily *alignable*.

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

(a) Globale Alinierung

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

(b) Lokale Alinierung

The table above shows two different kinds of alignments of reflexes of the word Indo-European *séh$_2$u̯el-, one global alignment and a local alignment. What comes to mind when comparing the two alignments? Why re correct alignments so difficult in historical linguistics?

Types of Sound Change

There is a long tradition of classifying specific sound changes into different types in historical linguistics. Unfortunately, the terminology is not very neat, ranging from very specific terms up to very abstract ones. We thus find terms like "rhotacism" (Trask 2000: 288), which refers to the change of [s] to [r], but also terms like *lenition*, which is a type of change "in which a segment becomes less consonant-like than previously" (ibid.: 190). Some terms are furthermore rather "explanative" than "descriptive" because they also denote a reason why a change happens, Thus, *assimilation* is often not only described as "[a] change in which one sound becomes more similar to another", but it is instead also emphasized that this happens "through the influence of a neighboring, usually adjacent, sound" (Campbell and Mixco 2007: 16).

The following table lists five more or less frequent types of sound change, by simply pointing to the relation between the source and the target, which serves as the sole criterion for the classification:

Typ	**Description**	**Notation**	**Example**
Continuation		$x > x$	Old High German *hant* > German *Hand*
Substitution	Ersetzung eines Lauts		Old High German *snēo* > German *Schnee* "snow"
Insertion	Gewinn eines Lauts	$\varnothing > y$	Old High German *ioman* > German "somebody"
	loss of a sound	$x > \varnothing$	Old High German *angust* > German *Angst* "fear"
Metathesis		$xy > yx$	Proto-Slavic **žьltъ* > Czech *žlutý* "yellow"

The table contains missing examples. Can you fill them out?

Sound Classes

We need to keep in mind that substantial differences between sounds (like between [p] and [b] or [f]) do not necessarily allow us to conclude that the words are not related, as sound change often follows certain general preferences. On the other hand, surface similarity between sounds does not prove anything in historical linguistics, unless we can show that this similarity is also regular (in terms of recurrent sound correspondences). Nevertheless, if we want to find cognate words, or get an idea on how to align two words we have not seen before, it is useful to turn to surface similarities to guide our first analysis. We thus need a heuristics that enables us to search for *probably* corresponding elements.

To account for this, we can make use of the concept of *sound classes* which was first proposed byDolgopolsky ("Gipoteza drevnejšego rodstva jazykovych semej Severnoj Evrazii s verojatnostej točky zrenija"). The basic idea is that sound which often occur in correspondence relation across the languages of the world can be divided in classes such that "phonetic correspondences inside a ‚type' are more regular than those between different ‚types'" (ibid.: 35).

No.	Cl.	Description	Examples
1	P	labial obstruents	p, b, f
2	T	dental obstruents	d, t, θ, ð
3	S	sibilants	s, z, ʃ, ʒ
4	K	velar obstruents, dental and alveolar affricates	k, g, ʦ, ʧ
5	M	labial nasal	m
6	N	remaining nasals	n, ɲ, ŋ
7	R	liquids	r, l
8	W	voiced labial fricative and initial rounded vowels	v, u
9	J	palatal approximant	j
10	ø	laryngeals and initial velar nasal	h, ɦ, ŋ

The table above shows Dolgopolsky's original sound class scheme. What comes to mind when comparing the reflexes of the words for "sun" in Indo-European with these classes?

Morphemes and Secondary Structures

Words can be segmented into sounds, but they can also be secondarily segmented, for example into syllables or morphemes. The morpheme structure of words plays a crucial role in phonetic alignment, since it governs the way we compare words. In der phonetischen Alinierungen kommt die wichtigste Rolle dabei der

The table below gives an example for the differences between a naive primary alignment and an informed secondary alignment While the primary alignment infers a wrong correspondence between final [t] and initial [t^h], the secondary alignment correctly matches only the first morpheme $z\chi^{51}$ "sun" of the Běijīng word and separates the suffix t^hou^1 "head (suffix)".

Primary Alignment						
Haikou	z	i	-	t	-	^{3}
Beijing	z$_ɻ$	ʅ	^{51}	t^h	ou	^{1}

Secondary Alignment							
Haikou	z	i	t	^{3}	-	-	-
Beijing	z$_ɻ$	ʅ	-	^{51}	t^h	ou	^{1}

What is the general problem with morpheme structure in languages other than the ones from South-East Asia?

Alignability

Not all aspects of language are completely sequential. We also find many hierarchical aspects. Word formation, for example, is often hierarchic, resembling syntax. If we want to compare sound sequences which have an underlying hierarchical structure, a normal alignment can only be used if the underlying structures are similar enough. If this is not the case, an alignment of entire words does not make sense. Instead, we need to identify and annotate those elements which *are* alignable. A more proper rendering of the structure of words for "sun" for example, can be found here:

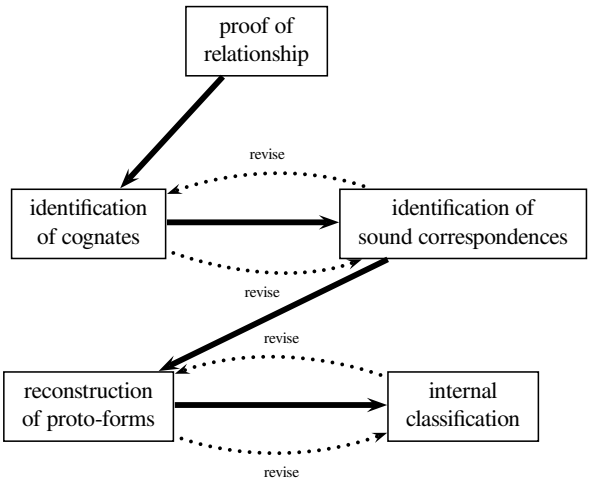
DOCULECT	SEGMENTS	ROOT	STEM	DERIVATION
French	sol←ej	*soh$_2$wl-	*soh$_2$wl + ?	RECTUS DIM
Spanish	sol	*soh$_2$wl-	*soh$_2$wl	RECTUS
German	zɔnɛ	*soh$_2$wl-	*sh$_2$en	OBLIQUUS
Swedish	su:l	*soh$_2$wl-	*soh$_2$wl	RECTUS

What are the obvious problems we encounter when trying to model the data as shown in the table above?

3 Cognate Detection

The Comparative Method

The comparative method, as the "fundamental method" for the identification of sound correspondences and the reconstruction of proto-languages, has many different definitions in the literature. I see the core of the classical workflow of historical language comparison as shown on the figure on the right. The dashed lines indicate that each step of this workflow is iterative and interacts with other steps.



Die komparative Methode wird oft als iteratives Verfahren beschrieben, wobei der iterative Charakter als eine große Stärke der Methode hervorgehoben wird. Was bedeutet "iterativ" überhaupt, und warum sollte das eine Stärke sein?

Traditional Approaches to Cognate Detection

If we look at the traditional procedure for cognate detection which is usually practiced in historical linguistics (often summarized under the term "comparative method"), we can describe this procedure as follows:

- Assemble a list of potential cognate sets.
- Align the words in your cognate list.
- Extract a list of potential sound correspondences from the alignments.
- Improve the cognate list and the correspondence list by:
  - Adding and removing correspondences from the correspondence list.
  - Adding and removing cognates from the cognate list.
- Stop, when the results are satisfying and ready for publication.

The iterative character applies to the whole workflow of the comparative method. How can we describe the dependency between the reconstruction of proto-forms and internal classification?

4 Automatic Cognate Detection

Quantifying Sound Correspondences

In bioinformatics, it is important to compute the probability of correspondences in DNA and protein alignment. This is done by comparing an *attested* with an *expected* distribution. Transferred to linguistics, this means that we compare a list of corresponding sounds with a distribution which we would

expect if the languages were not genetically related. In order to substantiate this, linguists usually show long lists of potential cognates, as shown in the list below:

Meaning	Italian	French
“square”	pjatsːa	plas
“feather”	pjuma	plym
“flat”	pjano	plɑ̃

Meaning	Italian	French
“tear”	lakrima	laʁm
“tongue”	liŋgwa	lɑ̃g
“moon”	luna	lyn

However, in the end, it is not only lists of words which are interesting for us, but lists of *aligned* words. Without alignments, we cannot properly construct our list of sound correspondences.

“square”	p j a ʦ: a p l a s –	
“feather”	p j u m a p l y m –	
“flat”	p j a n o p l ɑ̃ – –	

“tear”	l a k r i m a l ɑ – ʁ – m –	
“tongue”	l i ŋ w a l ɑ̃ – g –	
“moon”	l u n a l y n –	

Quantifying sound correspondences now only requires to count. For this, we construct a simple matrix, in which we mark down all co-occurrences of all sound combinations we encounter. The problem is, that we will miss context-dependent similarities when doing so. In order to account for this, we can use a rough notion of context by adding sonority context (rising sonority, falling sonority, etc.). Based on this, we can even with our manual method see, how cognates could be easily identified automatically.

	p	j	a	l	...
p	3	0	0	0	...
l	0	3	0	3	...
a	0	0	1	0	...
...	...	...	...	...	...

	p / #	j / C	a / C	l / C	...
p / #	3	0	0	0	...
l / #	0	0	0	3	...
l / C	0	3	0	0	...
a / V	0	0	1	0	...
...	...	...	...	...	...

Is the integration of phonetic context really important for cognate detection?

Clustering

Clustering is the process by which objects are divided into groups. If we talk about the Wú dialects in China, for example, we talk about a clustering of the Chinese dialects into one group which we call Wú 吳. Cognate detection is also a clustering procedure, as we divide words into groups, and we assume that words inside a group go back to a common ancestor. The words German *Zahn* [tsaːn], Italian *dente* [dɛnte], Dutch *tand* [tand], Russian *zub* [zup], und English *tooth* [tʊːθ] (all meaning “tooth”) can be clustered into different groups. Some go back to Proto-Indo-European **deh₃nt-* „toth” sind (*Zahn*, *dente*, *tand* und *tooth*), and one goes back to Proto-Indo-European **ǵombʰ-o-* “(finger)nail” sind (*zub*) (DERKSEN: 549).

	tsaːn	dɛnte	tand	zup	tʊːθ
tsaːn	0.00	0.53	0.35	0.57	0.57
dɛnte	0.53	0.00	0.10	0.97	0.52
tand	0.35	0.10	0.00	0.86	0.39
zub	0.57	0.97	0.86	0.00	0.70
tʊːθ	0.57	0.52	0.39	0.70	0.00

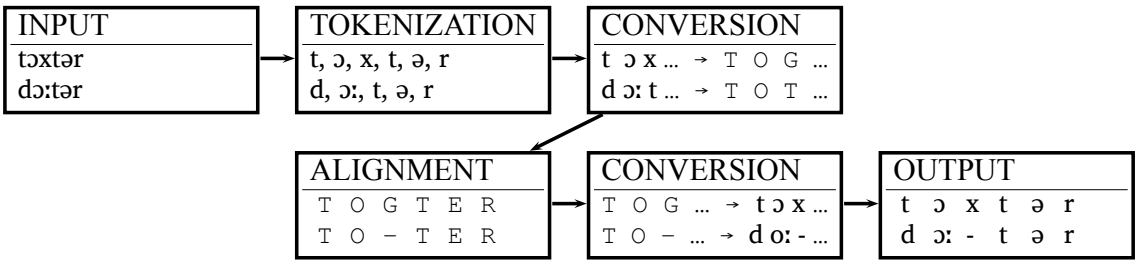
Automatic clustering has the advantage that the evidence which may be missing when comparing only one language pair, can be backed up by additional evidence. This nicely accounts for the use of *cumulative evidence* (Sturtevant 1920: 11), which is a fundamental aspect of the comparative methods for historical language comparison.

The table shows pairwise sequence distances which have been computed with help of the SCA alignment algorithm (List 2012) for the five words for "tooth" mentioned above. How would a possible cluster look like?

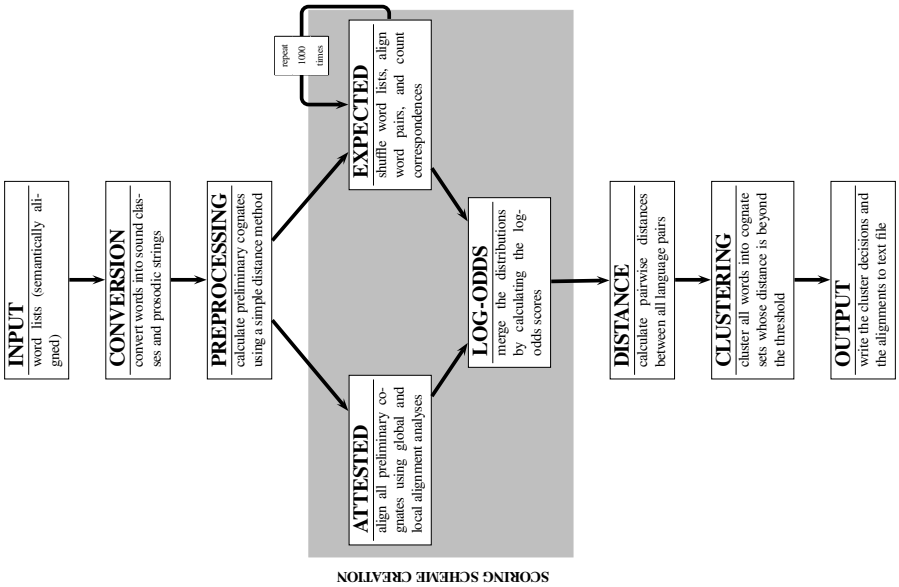
LexStat

Below is the workflow of the LexStat method for automatic cognate detection (ibid.). This method cumulates the aforementioned ideas for automatic cognate detection and assigns them to a common framework which comes close to the basic ideas of the "comparative method". Phonetic alignment plays a two-fold role: first it is used as initial heuristic to find the best candidates when being used to analyse multiple languages. Second, it is used as final procedure to infer the distances between all strings which are then fed to a cluster algorithm that finally partitions the data into groups of supposedly cognate words.

The phonetic alignment algorithm is based on sound classes. It does not align phonetic sequences directly, but rather modifies IPA characters to the simpler sound classes first, and later converts them back, as illustrated in the second figure below.

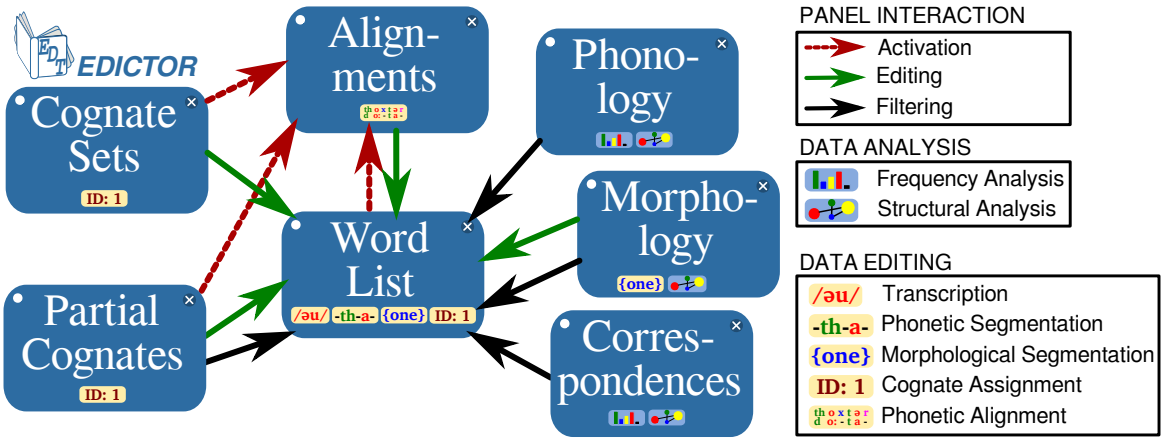


LexStat often has problems to distinguish true cognates from borrowings if borrowings are abundant. Why is that so?



5 Cognate Annotation

The computer-assisted framework requires that linguists can easily access the data which was analysed by a computer program in order to refine them. This can be easily done with help of the EDICTOR tool (List 2017) which is freely available at `http://edictor.digling.org` and can be used to annotate and refine cognate judgments. The LexStat algorithm, as it is implemented in the LingPy software package (List and Forkel 2016), creates the data automatically in a format which can be easily edited with the EDICTOR. In this way, the data is both accessible in human- and machine-readable form.



The figure above shows the basic modules of the EDICTOR. One module is named “partial cognates”. What does this mean?

References

Böckenbauer, H.-J. and D. Bongartz (2003). *Algorithmische Grundlagen der Bioinformatik*. German. Stuttgart, Leipzig, and Wiesbaden: Teubner.

Campbell, L. and M. Mixco (2007). *A glossary of historical linguistics*. Edinburgh: Edinburgh University Press.

Chao, Y. (1968). *A grammar of spoken Chinese*. Berkeley, Los Angeles, and London: University of California Press.

Derksen, R., comp. (2008). *Etymological dictionary of the Slavic inherited lexicon*. Leiden Indo-European Etymological Dictionary Series 4. Leiden and Boston: Brill.

Dolgopolsky, A. B. “Gipoteza drevnejšego rodstva jazykovych semej Severnoj Evrazii s verojatnostej točky zrenija [A probabilistic hypothesis concering the oldest relationships among the language families of Northern Eurasia].” *Voprosy Jazykoznanija* 2 (1964), 53–63; English translation: Dolgopolsky, A. B. “A probabilistic hypothesis concerning the oldest relationships among the language families of northern Eurasia.” In: *Typology, relationship and time. A collection of papers on language change and relationship by Soviet linguists. Typology, Relationship and Time. A collection of papers on language change and relationship by Soviet linguists*. Ed. and trans. from the Russian by V. V. Shevoroshkin. Ann Arbor: Karoma Publisher, 1986, 27–50.

Gusfield, D. (1997). *Algorithms on strings, trees and sequences*. Cambridge: Cambridge University Press.

IPA, ed. (1999). *Handbook of the International Phonetic Association. A guide to the use of the international phonetic alphabet*. Cambridge: Cambridge University Press.

Kruskal, J. B. (1983). “An overview of sequence comparison. Time warps, string edits, and macromolecules.” *SIAM Review* 25.2, 201–237. JSTOR: `2030214`.

List, J.-M. (2012). “LexStat. Automatic detection of cognates in multilingual wordlists.” In: *Proceedings of the EACL 2012 Joint Workshop of Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources*. “LINGVIS & UNCLH 2012” (Avignon, 04/23–04/24/2012). Stroudsburg, 117–125.

— (2014). *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.

— (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*. Valencia: Association for Computational Linguistics, 9–12.

List, J.-M. and R. Forkel (2016). *LingPy. A Python library for historical linguistics*. Version 2.5. URL: `http://lingpy.org`.

Sturtevant, E. H. (1920). *The pronunciation of Greek and Latin*. Chicago: University of Chicago Press. Internet Archive: `pronunciationgr00unkngoog`.

Trask, R. L., comp. (2000). *The dictionary of historical and comparative linguistics*. Edinburgh: Edinburgh University Press.