

Computer-Assisted Language Comparison

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

College of Chinese Language and Culture
Nankai University, Tianjin

Winter Term 2017

Contents

1	Introduction	3
	Foundations	3
	Sequence Comparison	14
2	CALC	23
	CLDF	24
	CALC	34

1 Introduction

The introduction comprises two lectures, one devoted to classical and one devoted to more recent computational approaches to historical language comparison.

Foundations of Historical Language Comparison

1 Research Object

1.1 Languages

What are Languages?

What counts as a language, i.e. which tradition of speech we label as language, does not depend on pure linguistic criteria, but also on social and cultural criteria (Barbour und Stevenson 1998: 8). Accordingly, we assume that people in Shànghǎi, Běijīng, and Měixiàn all speak dialects of “Chinese”, while people in Scandinavia speak languages such as “Norwegian”, “Swedish”, or “Danish”. This does not mean that the Chinese varieties show less differences than the Scandinavian ones, as we can see from Table 1:

Běijīng Chinese	1	iou ²¹	i ⁵⁵	xuei ³⁵	pei ²¹ fəŋ ⁵⁵	kən ⁵⁵	t ^h ai ⁵¹ iaŋ ¹¹	t͡ʂəŋ ⁵⁵	tsai ⁵³	naə ⁵¹	t͡ʂəŋ ⁵⁵ luən ⁵¹
Hakka Chinese	1	iu ³³	it ⁵⁵	pai ³³ a ¹¹	pet ³³ fuŋ ³³	t ^h uŋ ¹¹	nit ¹¹ t ^h eu ¹¹	hək ³³	e ⁵³		au ⁵⁵
Shànghǎi Chinese	1	fi ²²		t ^h ä ⁵⁵ tsɿ ²¹	poɿ ³ foŋ ⁴⁴	taɿ ⁵	t ^h a ³³ fiä ⁴⁴	tsəŋ ³³	hɔ ⁴⁴		ləɿ ¹¹ lə ²³ tsa ⁵³
Běijīng Chinese	2	ʒei ³⁵		də ⁵⁵		pən ³⁵	liŋ ²¹	ta ⁵¹			
Hakka Chinese	2	man ³³	pin ¹¹		k ^w ɔ ⁵⁵	vɔi ⁵³					
Shànghǎi Chinese	2	sa ³³	pin ⁵⁵	fiəɿ ²¹		pən ³³	zi ⁴⁴	du ¹³			
Norwegian	1	nu:ravɪnˀŋ	ɔ	su:lŋ						kraŋlɔt	ɔm
Swedish	1	nu:ɖanvɪndən	ɔ	su:lən		tyɪstadə	ən gɔŋ				ɔm
Danish	1	noːʌnvenˀŋ	ʌ	so:lʔn	k ^h ʌm		enˀgəŋ	i sɖɪɔdˀ			ʌmˀ
Norwegian	2	vem	a	dem	sŋ	va:	ɖŋ	stæfˀkæstə			
Swedish	2	vem	av	dɔm	sɔm	va		stɔkɔst			
Danish	2	vemˀ	a	bɐm	ɖ	va	ɖŋ	sɖæˀgæsɖə			

Tabelle 1: “Der Nordwind und die Sonne” in verschiedenen Sprachvarietäten

The table shows phonetic transcriptions of the translation of the sentence “The Northwind and the sun were disputing, who was stronger” in six different linguistic varieties. Unfortunately, there is no further information on the structure of the table. How can we explain it anyway? Which conclusions can be drawn with respect to the classification of Chinese speech varieties into dialects and Scandinavian speech varieties into languages?

Language as a Diasystem

In order to allow linguists to handle the complex, heterogeneous character of languages more realistically, sociolinguistics usually invokes the model of the *diasystem* (Bussmann 1996: 312). According to this model, languages are complex aggregates of different linguistic systems, which ‘coexist and influence each other’ (Coseriu 1973: 40).¹ An important aspect is the existence of a so-called “roof language” (*Dachsprache*), i.e., a language variety which

¹ My translation, original text: “die miteinander koexistieren und sich gegenseitig beeinflussen”

serves as standard for interdialectal communication (Goossens 1973: 11). The linguistic varieties (dialects, sociolects) which are connected by such a standard constitute the “variety space” (*Varietätenraum*) of a language (Oesterreicher 2001), as shown in Figure 1.

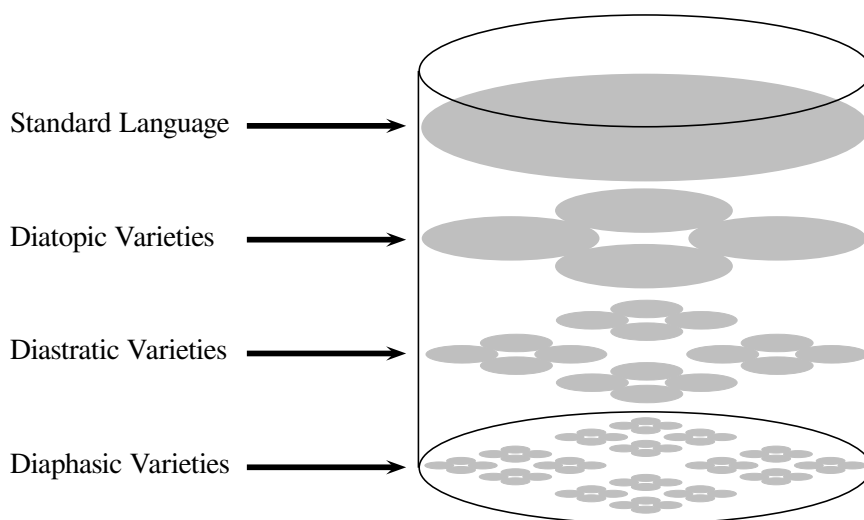


Abbildung 1: Language as a diasystem

How can the model of the diasystem help us to explain the different division of Chinese and Scandinavian speech varieties into dialects and languages?

Simplified Model of Language in Historical Linguistics

In historical linguistics, we use a simplified language model. We are less interested what language is in reality, but more, how language changes and languages change. Language is seen as a system. In a broader sense, a system is a set of elements and a set of relations which hold for the set of elements (Marchal 1975: 462f). For our language model in historical linguistics, this means that linguistic systems contain *sounds* (phones, phonemes) and *signs* (words, morphemes) as elements, as well as *phonotactic* and *syntactic rules* as relations.

Is such a simplified model sufficient for a treatment of the most important problems in historical linguistics?

1.2 Signs

The Classical Sign Model

In historical linguistics, linguistic signs are usually treated in the context of the traditional sign model by Saussure (1916). As Roman Jakobson notes, we distinguish two sides: the form and the content:

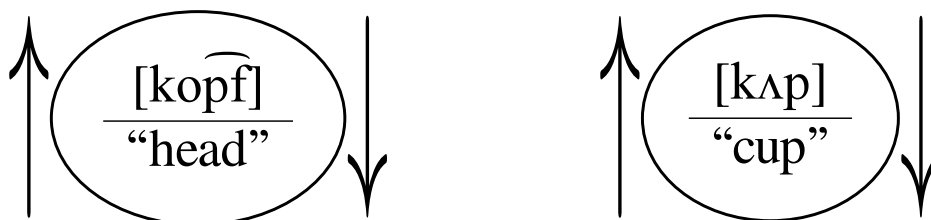
The sign has two sides: the sound, or the material side on the one hand, and meaning, or the intelligible side on the other. Every word, and more generally every verbal sign, is a combination of sound and meaning, or to put it another way, a combination of signifier and signified [...]. (Jakobson 1976 [1978]: 3)

1 Introduction

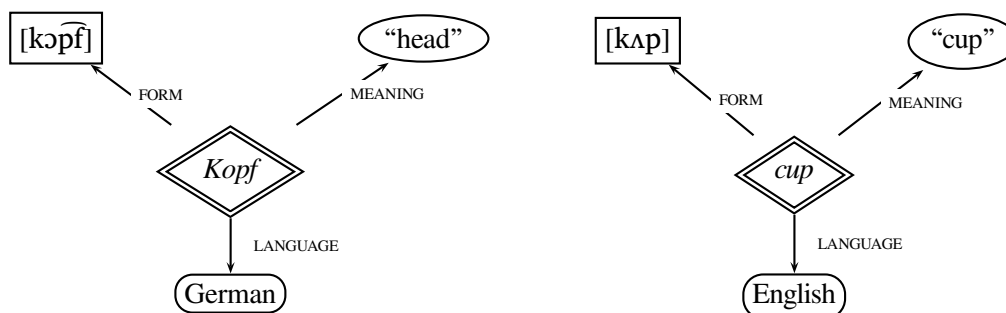
What does Jakobson mean with the words “material” and “intelligible”?

Additional Thought on Signs

Normally, the classical sign model by Saussure is depicted as follows:



Important for the linguistic sign is, however, not only the *form* (signifier) and the *meaning* (signified), but also the linguistic *system* in which the sign is used. A more detailed depiction of the sign model should therefore also include the system as a constitutive aspect of the linguistic sign:



If we look at the structure of sign form and sign meaning, we can find fundamental differences between the two. The sign form is a (phonetic) sequence, that is, a linear arrangement of distinctive sounds. These sounds are material, since they can be measured as waves in the air, or as traces of ink on a sheet of paper. Important for the sign form is furthermore its linearity, since not only the assembly of different sounds is crucial for the distinction between different sign forms, but also the order of elements. We can therefore say that the sign form is (a) substantial, (b) segmentable, and (c) linear. But what about the sign meaning? Fill in the corresponding terms in the right column of the table.

No.	Form	Meaning
(a)	substantial	
(b)	segmentable	
(c)	linear	

2 Change

Change in the Odes

When reading Chinese poems from the *Book of Odes* ((Shījīng 詩經 ca. 1050–600 BCE) in modern pronunciation, it becomes immediately evident that the language has changed, as many instances no longer rhyme. Although it took scholars some time to figure out what happened, and why these passages did not rhyme in the Odes, scholars from the Míng 明 (1368–1644) and Qīng 清 dynasties (1644–1911) eventually realised that it was language change driving these differences (Baxter 1992: 153-157). This is illustrated in the following table (taken from List 2017).

Chinese Text	Translation	RW	Patterns	MCH	OCBS-Rhyme
燕燕於飛	The swallows go flying	<i>fēi</i> 飛	A	*pjij	*-ər
下上其音	falling and rising are their voices;	<i>yīn</i> 音	B	*ʔim	*-əm
之子於歸	This young lady goes to her new home,	<i>guī</i> 歸	A	*kjiw	*-əj
遠送於南	far I accompany her to the south.	<i>nán</i> 南	B	*nom	*-əm
瞻望弗及	I gaze after her, can no longer see her,	[<i>jí</i> 及]	–	[*gip]	[*-əp]
實勞我心	truly it grieves my heart	<i>xīn</i> 心	B	*sim	*-əm

2.1 Preliminary Considerations

Change as Process

That sound change proceeds in an overwhelmingly regular manner can be easily illustrated by looking at words from Latin and the words in its descendant languages (as for example Italian). The data in Table 2 are merely illustrations for this phenomenon, the number of examples can be easily expanded.

Meaning	Latin	Italian	Meaning	Latin	Italian
“feather”	plu:ma	pjuma	“tongue”	lingua	lingwa
“flat”	pla:nus	pjano	“moon”	lu:na	luna
“square”	plate:a	pjats:a	“slow”	lentus	lento

Tabelle 2: Lateinische und Italienische Wörter

If the data in the table are indeed only a small collection and many more examples are available, which fundamental characteristics of sound change can be derived from this?

Change as Law

The overwhelming regularity of sound change was enthusiastically met by scholars from the 19th century. The regularity assumption quickly led to the formulation of *sound laws*. The phenomenon was deliberately compared with the regularity of natural laws. The strongest hypothesis regarding sound change was formulated by the so-called Neogrammarians (*Junggrammatiker*), a group of linguists in Leipzig (Germany) who were much younger on

1 Introduction

average than the rest of the linguists at that time in Germany. The following quote is taken from a text passage which later became famous under the name *Neogrammarian Manifesto*:

All sound change, as long as it proceeds mechanically, follows exceptionless laws, that is, the direction of the sound change is the same with all members of a linguistic society, the only exception being the cases that dialect split occurs, and all words in which the sound which is targeted by a given sound change recurs under the same conditions will be affected by the change without exception. (Osthoff und Brugmann 1878: XIII)²

What follows from this ?

2.2 Sound Change

Sound Change Mechanisms

Based on our knowledge of sound change as a process, we can draw the following two intermediate conclusions:

- Sound change is a *recurrent* process: Not only a few words of a given language are modified sporadically, but many, if not the whole lexicon of a language.
- Sound change is a *contextually restricted* process: The phonetic environment determines whether certain sound changes happen or not.

However, it is by no means clear whether this characterizes all mechanisms of sound change. In the second half of the 20th century, many linguists, especially Chinese scholars, opposed this view and pointed to sound change processes which were not following the description of the Neogrammarians:

When a phonological innovation enters a language it begins as a minor rule, affecting a small number of words [...]. As the phonological innovation gradually spreads across the lexicon, however, there comes a point when the minor rule gathers momentum and begins to serve as a basis for extrapolation. At this critical cross-over point, the minor rule becomes a major rule, and we would expect diffusion to be much more rapid. The change may, however, reach a second point of inflection and eventually taper off before it completes its course, leaving behind a handful of words unaltered. (Chen 1972: 474f)

The opposing views can be characterized as follows:

	Neogramm. S. C.	Lex. Diffusion
lexically	abrupt	gradual
phonetically	gradual	abrupt

Do these two theories really contradict each other, or is it possible that sound change proceeds according to two different mechanisms?

²Aller lautwandel, soweit er mechanisch vor sich geht, vollzieht sich nach ausnahmslosen gesetzen, d.h. die richtung der lautbewegung ist bei allen angehörigen einer sprachgenossenschaft, ausser dem fall, dass dialektspaltung eintritt, stets dieselbe, und alle wörter, in denen der der lautbewegung unterworfen laut unter gleichen verhältnissen erscheint, werden ohne ausnahme von der änderung ergriffen.

Sound Change Types

If we concentrate on the substantial aspects of sound change, we can distinguish different *types of sound change*:

Continuity: A sound does not change.

- Old High German [hant] > Modern German [hant] “hand”

Substitution: A sound is replaced by another sound.

- Old High German [sne:o] > Modern German [ʃne:] “snow”

Deletion: A sound is lost.

- Old High German [aŋst] > Modern German [aŋst] “fear”

Addition: A sound is gained.

- Old High German [joman] > Modern German [je:mant] “somebody”

What about sound changes in which elements are swapped (metathesis)? Shouldn't they also occur in this survey?

2.3 Semantic Change

While sound change proceeds as an *alternation*, that is, each sound change modifies the form of a sign in its entirety, semantic change proceeds primarily in steps of *cumulation* and *reduction*: the meaning of signs is being *expanded* (cumulation) resulting in polysemy, or *reduced*, resulting in a loss of polysemy. Basic examples are shown in the next table:

Language	Form	Meaning
Proto-Germanic	*[kup:az]	“cup / vessel”
Old High German	[kɔpf]	“cup / vessel”, “head”
High German	[kɔpf]	“head”

We can distinguish many different types of semantic change, however, we can summarize most types under two major types, namely *metaphor* and *metonymy*:

metaphor: ancestral meaning and descendant meaning are in a similarity relation.

- “cup” > “head”, “see” > “think”

metonymy: ancestral meaning and descendant meaning are in a close relation of continuity (part vs. whole, person vs. thing).

- “stone (material)” > “stone (object)”, “head” > “person”

Are metonymy and metaphor really enough to summarize all different possible types of semantic change?

2.4 Lexical Change

Lexical change primarily refers to the change of form-meaning relations of the linguistic signs which make up the lexicon of a given language. We can investigate semantic change from the perspective of a set of concrete meanings (“head”, “hand”, “foot”, etc.), of which we assume that they occur in all cultures and across all times (basic vocabulary, 核心词). The fundamental process here is lexical replacement: a form which was primarily used to express only one certain meaning, is no longer used to express this meaning, but replaced by another linguistic sign.

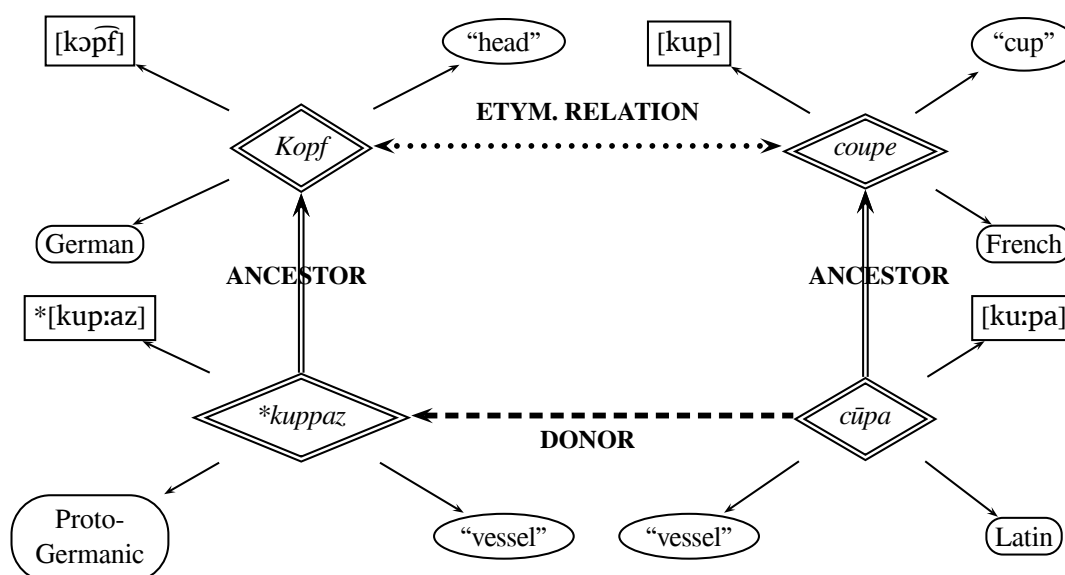
3 Relations

3.1 Sign Relations

We can distinguish many different sign relations in historical linguistics. The following, however, seem to be the most important ones (for more information, see List 2016):

Relation	Characterisation
Ancestor-Descendant	two signs, of which one became the other through a gradual process of change
Cognacy	two signs who have a common ancestor
Donor-Recipient	signs in different languages of which one was transferred into the other language as part of a discrete process

The figure (taken from List (2014)) below try to illustrate the historical scenarios behind the different sign relations. Try to explain what is going on there.

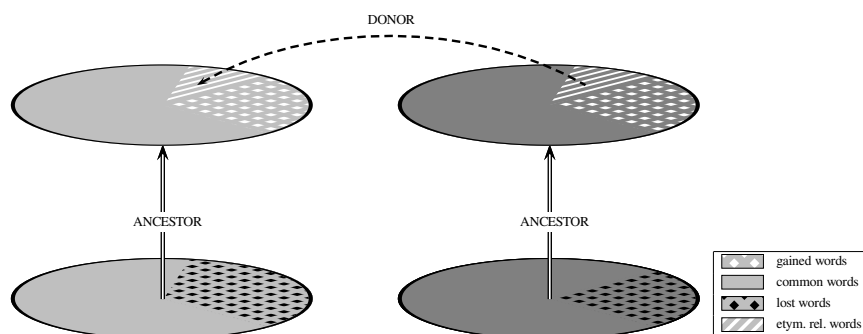


3.2 Language Relations

Language relations are much more complex. Nevertheless, in historical linguistics we have the following three most important relations on which linguists normally concentrate:

Relation	Characterisation
Ancestor-Descendant	holds for two languages of which one became the other through a gradual change process
Genetic Relationship	holds for two languages who share a common ancestor
Contact	holds for languages of which one has somehow influenced the other

The figure below tries to illustrate the three fundamental language relations in historical linguistics. What is the obvious problem of this kind of illustration?



3.3 Family Trees

Traditionally, we use the family tree model to illustrate how languages have developed into their current shape. Among the first linguists to popularize this model was August Schleicher (1821 – 1868). Schleicher, August, used “the image of a branching tree” (Schleicher 1853: 787, my translation) to show how languages diverged from their common ancestor.

4 Similarities

4.1 Form Similarities

If we only regard the sign form, we can distinguish different kinds of similarities:

Substantial Similarities: Direct similarities between sound segments of two sign forms.

- German [pɔst] “post” vs. German [o:pst] “fruits”
- German [flaʃn] “bottle” vs. German [ʃla:fn] “sleep”

Structural Similarities: similarities between the linear structure of two sign forms

- German [tantən] “glass bead” vs. German [kɛrkɛr] “cellar”
- German [mama] “mama” vs. German [papa] “papa”

Substantial-structural Similarities: similarities between sound segments of two sign forms whose and their linear arrangement.

- German [hant] “hand” vs. Englisch [hænd] “hand”
- German [mi:f] “smell” vs. Russian [mi:f] “myth”

What if two words are similar but not exactly identical in their sounds? How could we measure this?

4.2 Systematic Similarities

Systematic similarities are similarities between the lexical systems of different languages. In contrast to formal and functional similarities, systematic similarities deal with **recurrent formal and functional similarities** between the signs of two or more languages.

Meaning	Italian	French	Meaning	Italian	French
“square”	pjats:a	plas	“tear”	lakrima	larm
“feather”	pjuma	plym	“tongue”	lingwa	lāg
“flat”	pjano	plā	“moon”	luna	lyn

$$j = l$$

$$l = l$$

$$pj = pl, l = l$$

What do we need to do in order to detect systematic similarities?

5 Proof

5.1 Mode of Reasoning

Modes of reasoning are important in order to make conclusions or to consider something as proven. They are also important for historical linguistics. The fundamental modes of reasoning are *deduction*, *induction*, and *abduction* and can be best distinguished if we consider them as a combination of rule, event, and result:

deduction: “All bunnies have long ears, and the thing that brings the Easter eggs is a bunny. Therefore, the thing that brings the Easter eggs has long ears.”

induction: “The thing that brings the Easter eggs is a bunny, and the thing that brings the Easter eggs has long ears. Therefore, all bunnies have long ears.”

abduction: “All bunnies have long ears, and the thing that brings the Easter eggs has long ears. Therefore, the thing that brings the Easter eggs is a bunny.”

Abductive reasoning is usually labeled the weakest of the three modes of reasoning. Why is that so, and why are we still using it so abundantly in the historical sciences?

5.2 Evidence

In order to *prove genetic relationship between languages* it is important to separate the different types of evidence strictly. The basic evidence are *similarities in the synchronic structures of two or more languages*. That we can infer the past from the present is a general principle in historical sciences. This inference process is best described as “historical fact abduction” (Schurz 2008) and used in all historical sciences. To avoid wild speculations, such a procedure needs to be based on *cumulative evidence* that is, on a multitude of different types of evidence, which can only be explained by invoking an individual hypotheses.

- The sole widely recognized type of evidence for genetic relationships are *systematic similarities* (regular sound correspondences) between two or more languages.

- Not all forms of systematic similarity are convincing: one must be able to show that the systematic similarities are not *coincidental*, *natural* or *contact-induced*.
- If one can show that systematic similarities can only be explained by postulating a *genetic relationship*, one can assume (for the time being) that this has been proven.

How can we distinguish coincidental and contact-induced from similarities due to common origin?

Literatur

- Barbour, S. und P. Stevenson (1998). *Variation im Deutschen. Soziolinguistische Perspektiven*. Berlin: de Gruyter.
- Baxter, W. H. (1992). *A handbook of Old Chinese phonology*. Berlin: de Gruyter.
- Busmann, H., Hrsg. (1996). *Routledge dictionary of language and linguistics*. A. d. Deutschen übers. von G. Trauth und K. Kazzazi. London und New York: Routledge.
- Chen, M. (1972). "The time dimension. Contribution toward a theory of sound change". *Foundations of Language* 8.4, 457–498. JSTOR: 25000618.
- Coseriu, E. (1973). *Probleme der strukturellen Semantik. Vorlesung gehalten im Wintersemester 1965/66 an der Universität Tübingen*. Tübingen: Narr.
- Goossens, J. (1973). *Niederdeutsch. Sprache und Literatur. Eine Einführung*. Neumünster: Karl Wachholtz.
- Jakobson, R. (1976 [1978]). *Six lectures on sound and meaning*. A. d. Französischen übers. von J. Mephram. Mit einer Einl. von C. Lévi-Strauss. Cambridge und London: MIT Press.
- 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.
 - (2017). "Using network models to analyze Old Chinese rhyme data". *Bulletin of Chinese Linguistics* 9.2, 218–241.
- Marchal, J. H. (1975). "On the concept of a system". English. *Philosophy of Science* 42.4, 448–468. JSTOR: 187223.
- Oesterreicher, W. (2001). "Historizität, Sprachvariation, Sprachverschiedenheit, Sprachwandel". In: *Language typology and language universals. An international handbook*. Hrsg. von M. Haspelmath. Berlin und New York: Walter de Gruyter, 1554–1595.
- Osthoff, H. und K. Brugmann (1878). *Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen*. Bd. 1. Leipzig: Hirzel.
- Saussure, F. de (1916). *Cours de linguistique générale*. Hrsg. von C. Bally. Lausanne: Payot.
- Schleicher, A. (1853). "Die ersten Spaltungen des indogermanischen Urvolkes The first splits of the Indo-European people". *Allgemeine Monatsschrift für Wissenschaft und Literatur* 3, 786–787.

Foundations of Computer-Assisted Sequence Comparison

1 Introductory Remarks

1.1 Automatic Sequence Comparison

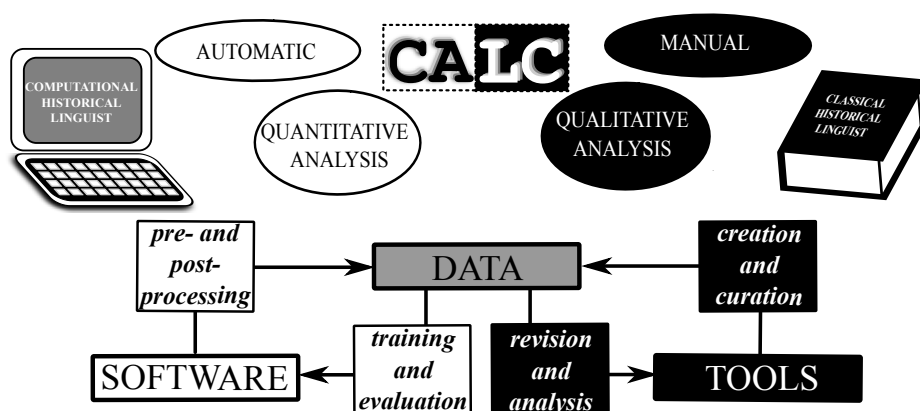
Automatic sequence comparison refers to techniques which can be used in historical linguistics in order to carry out an automatic comparison of words inside and across languages. In times where linguistic data in digital form is constantly increasing, it is important to make use of automatic approaches for the task of word comparison. Most of the work in historical linguistics is still being carried out manually. If we use automatic methods, we can profit from the increased speed that computers provide. The methods are also steadily increasing in accuracy (List u. a. 2017), even if they cannot yet compete with trained experts.

Given what we have learned about sound change: What obvious obstacles have computational approaches to cope with here?

1.2 Computer-Assisted Language Comparison

A problem of computational methods is that they usually do not provide the same accuracy as the analyses provided by human experts. Since human experts are slow in annotation, however, and at times also not very consistent, we are in a dilemma: if we use the computational methods, we will produce many errors in our analyses, but if we annotate data manually, we will be inefficient. This is where the idea of *computer-assisted language comparison* (List 2016) comes into play. If we analyse the data automatically in a first run, nothing prevents us from refining the data afterwards. In addition to pure computational analyses, computer-assisted language comparison comes along with specific *interfaces* which experts can use in order to correct the analyses which have been produced by the computer methods. In this way, we have an integrated workflow in which data is passed back and forth between experts and computers.

The graphic below shows a tentative workflow for computer-assisted language comparison. What do we need to keep in mind if we want to follow this workflow?

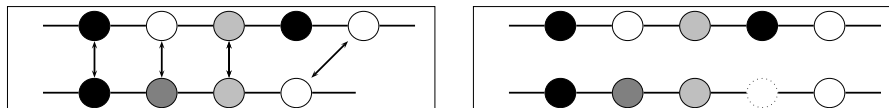


2 Phonetic Alignment

2.1 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?

2.2 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	ts	ə
Polish	s	w	ɔ	nʲ	ts	ɛ
French	s	-	ɔ	l	-	ɛ
Italian	s	-	o	l	-	e
German	s	-	ɔ	n	-	ə
Swedish	s	-	u:	l	-	-

(a) Globale Alinierung

Language	Alignment					
Russian	s	ɔ	-	-	n	ts
Polish	s	-	w	ɔ	nʲ	ts
French	s	ɔ	l	-	-	-
Italian	s	o	l	-	-	-
German	s	ɔ	-	-	-	-
Swedish	s	u:	l	-	-	-

(b) Lokale Alinierung

The table above shows two different kinds of alignments of reflexes of the word Indo-European **séh₂uel-*, one global alignment and a local alignment. What comes to mind when comparing the two alignments? Why are correct alignments so difficult in historical linguistics?

2.3 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” (ebd.: 190). Some terms are furthermore rather “explanative” than “descriptive” because they also denote a reason

1 Introduction

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 und 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 <i>hant</i> > German <i>Hand</i>
Substitution	Ersetzung eines Lauts		Old High German <i>snēo</i> > German <i>Schnee</i> “snow”
Insertion	Gewinn eines Lauts	$\emptyset > y$	Old High German <i>ioman</i> > German “somebody”
	loss of a sound	$x > \emptyset$	Old High German <i>angust</i> > German <i>Angst</i> “fear”
Metathesis		$xy > yx$	Proto-Slavic <i>*žltъ</i> > Czech <i>žlutý</i> “yellow”

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

2.4 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 by Dolgopolsky (1964). 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’” (ebd.: 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, ts, tʃ
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?

2.5 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_l⁵¹ “sun” of the Běijīng word and separates the suffix t^hou¹ “head (suffix)”.

Primary Alignment						Secondary Alignment					
Haikou	z	i	-	t	-	³					
Beijing	z _l	ɿ	⁵¹	t ^h	ou	¹					

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

2.6 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:

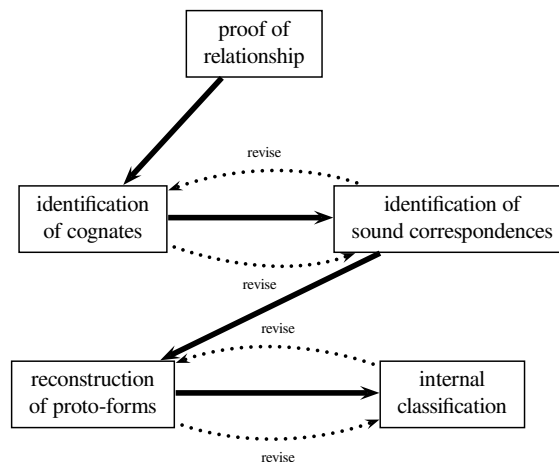
DOCULECT	SEGMENTS	ROOT	STEM	DERIVATION
French	sol+ej	*soh ₂ wl-	*soh ₂ wl + ?	RECTUS DIM
Spanish	sol	*soh ₂ wl-	*soh ₂ wl	RECTUS
German	zɔnɐ	*soh ₂ wl-	*sh ₂ en	OBLIQUUS
Swedish	su:l	*soh ₂ wl-	*soh ₂ wl	RECTUS

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

3 Cognate Detection

3.1 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?

3.2 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?

3.3 Automatic Cognate Detection

Detection of 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”	pjats:a	plas
“feather”	pjuma	plym
“flat”	pjano	plā

Meaning	Italian	French
“tear”	lakrima	larm
“tongue”	lingwa	lāg
“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 ts: a
“feather”	p j u m a
“flat”	p j a n o

“tear”	l a k r i m a
“tongue”	l i ŋ w a
“moon”	l u n a

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?

3.4 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 吴. 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* [tsa:n], Italian *dente* [dente], Dutch *tand* [tand], Russian *zub* [zup], und English *tooth* [tu:θ] (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 **ǵomb^h-o-* “(finger)nail” sind (*zub*) (DERKSEN: 549).

	tsa:n	dente	tand	zup	tu:θ
tsa:n	0.00	0.53	0.35	0.57	0.57
dente	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
tu:θ	0.57	0.52	0.39	0.70	0.00

1 Introduction

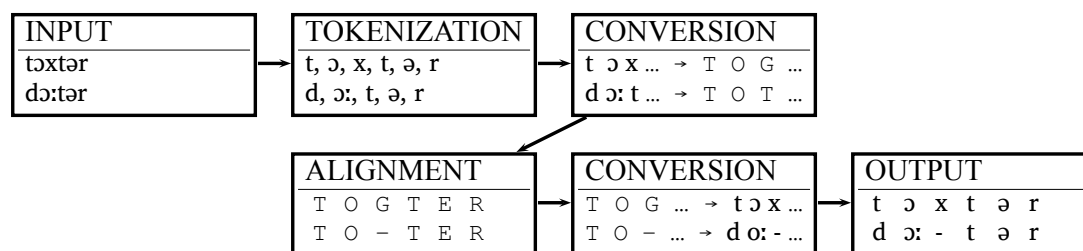
Automatic clustering has the advantage that the evidence which may be missing when comparing only one language pair, can be backed up by additional evidence. This nicely accounts for the use of *cumulative evidence* (Sturtevant 1920: 11), which is a fundamental aspect of the comparative methods for historical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

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

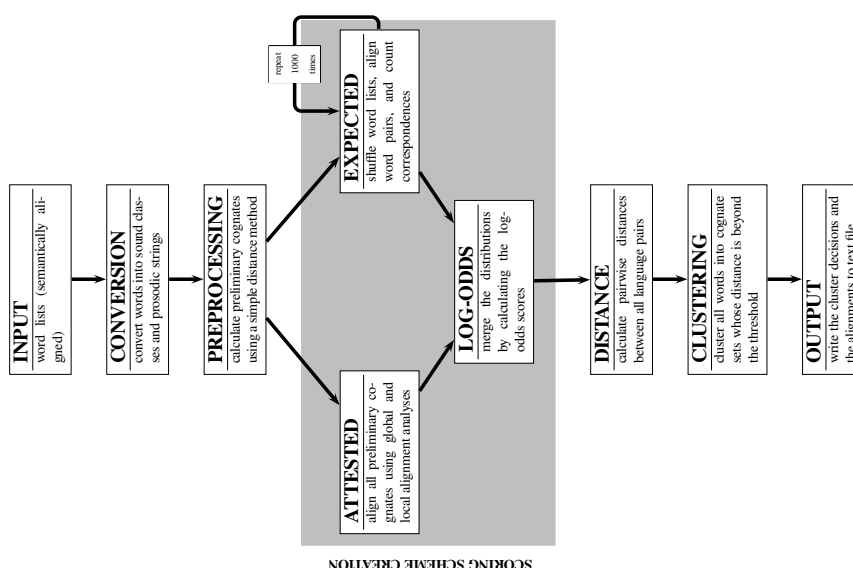
3.5 LexStat

Below is the workflow of the LexStat method for automatic cognate detection (List 2012a). 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.

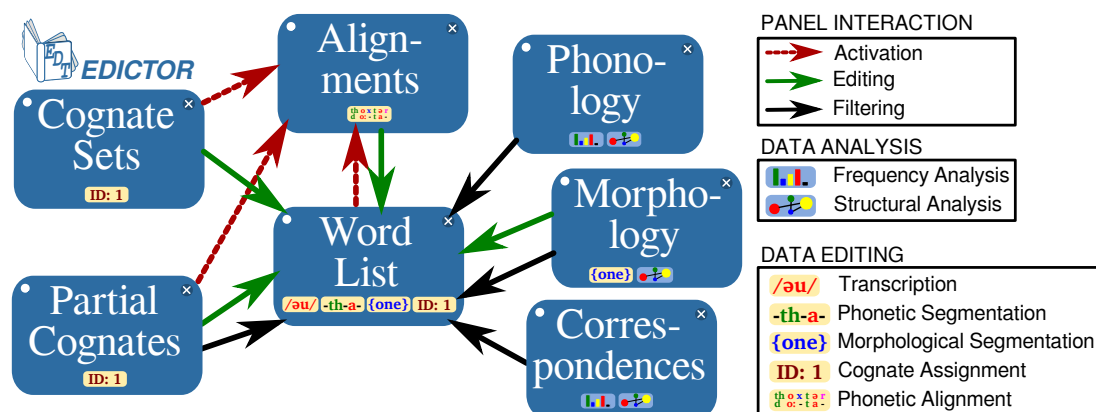


LexStat often has problems to distinguish true cognates from borrowings if borrowings are abundant. Why is that so?



4 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://editor.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 und 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?

Literatur

- Campbell, L. und M. Mixco (2007). *A glossary of historical linguistics*. Edinburgh: Edinburgh University Press.
- Derksen, R., Komp. (2008). *Etymological dictionary of the Slavic inherited lexicon*. Leiden Indo-European Etymological Dictionary Series 4. Leiden und Boston: Brill.
- Dolgopolsky, A. B. (1964). "Gipoteza drevnejšego rodstva jazykovych semej Severnoj Evrazii s verojatnostej točki zrenija [A probabilistic hypothesis concerning the oldest relationships among the language families of Northern Eurasia]". *Voprosy Jazykoznanija* 2, 53–63.
- Gusfield, D. (1997). *Algorithms on strings, trees and sequences*. Cambridge: Cambridge University Press.
- List, J.-M. (2012a). "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, 23.–24. 04. 2012). Stroudsburg, 117–125.
- (2012b). "SCA. Phonetic alignment based on sound classes". In: *New directions in logic, language, and computation*. Hrsg. von M. Slavkovik und D. Lassiter. Berlin und Heidelberg: Springer, 32–51.
- (12/2016). *Computer-Assisted Language Comparison: Reconciling Computational and Classical Approaches in Historical Linguistics*. Jena: Max Planck Institute for the Science of Human History.
- (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. und R. Forkel (2016). *LingPy. A Python library for historical linguistics*. Version 2.5. URL: <http://lingpy.org>.
- List, J.-M., S. J. Greenhill und R. D. Gray (01/2017). "The potential of automatic word comparison for historical linguistics". *PLOS ONE* 12.1, 1–18.
- Sturtevant, E. H. (1920). *The pronunciation of Greek and Latin*. Chicago: University of Chicago Press. Internet Archive: [pronunciationr00unkngoog](https://www.archive.org/details/pronunciationr00unkngoog).
- Trask, R. L., Komp. (2000). *The dictionary of historical and comparative linguistics*. Edinburgh: Edinburgh University Press.

2 CALC

The section on Computer-Assisted Language Comparison comprises two lectures, one devoted to cross-linguistic data formats, and one devoted to examples for computer-assisted language comparison in practice.

Cross-Linguistic Data Formats

1 Introduction

1.1 Data in Linguistics

Linguistics is beyond doubt a data-driven discipline, and most of our daily linguistic work is based on evaluating, creating, and analysing different kinds of data. If I want to investigate grammatical phenomena, I will need grammatical data, normally example sentences drawn from some kind of corpus. If I want to compare typological aspects of different phenomena, I will again need some kind of corpus in which I can find contrastive examples, or I will have to build this corpus myself. Even if I simply want to learn a language which I do not know before, I need data, as I will need some grammatical descriptions with tables, example sentences, as well as a good dictionary which helps me how to translate words from the foreign language into my own mother tongue.

Aren't there any fields in linguistics which are less data-driven in some way, or is all in linguistics about empirical approaches and data collections?

1.2 Data in Historical Linguistics

Historical linguistics is in some sense even more data-hungry than general linguistics, since we cannot invoke our linguistic intuition in order to resolve phenomena involving languages which have long since disappeared. As a result, historical linguistics heavily relies not only on ancient documents, but also on extensive collections of contemporary languages, be it dictionaries, word lists, or grammars. In order to shed light on the past of our languages, linguists sift through dictionaries, hunting for cognate words, and searching for spurious similarities in the grammars of the languages they investigate. Etymological dictionaries, one of the most typical examples for research results in historical linguistics, are a classical example for a database which was built on paper: they represent the results of intensive language comparison with enormous amounts of references to older literature as well as the most recent findings. Etymological dictionaries are no databases, but the way they are created as well the data they are supposed to reflect virtually cries for them to be represented in form of a database rather than in simple prose.

If we look at the multiple references to previous literature and the steady accumulation of new knowledge based on previously created knowledge in etymological dictionaries, what similar data-structure which has recently gained great popularity, comes to mind?

1.3 Problems with Data in Linguistics

We have huge problems with data in both linguistics and historical linguistics. These problems can be summarized under three core aspects, which related to (a) the availability of data, (b) the transparency, and (c) the comparability of data. Availability is a constant nuisance, as it is still rather the exception than the rule that scholars actually share the data they collected and used in order to write an article. It is not uncommon, that scholars even propose a new classification for a certain group of languages without supplementing neither data nor code which was used in order to arrive at the conclusions (Tamburelli and Brasca 2017), not to speak of numerous grammatical descriptions of languages which are provided

without sharing or properly citing their data. The situation which is reported in the following quote, which is taken from a review on a recent handbook on Sino-Tibetan languages, is quite representative for the field of historical and general linguistics in its current state:

It is disappointing that so many among the authors of newly commissioned articles did not cite their data; this failing is particularly perplexing in the case of those authors who benefited from the generosity of agencies that explicitly require archiving in public repositories. The move toward open data is still in its early days. (Hill 2017: 306)

Apart from data *availability*, the field also suffers from a lack of *transparency* of the data that people share. As an example, consider the following table provided by Bengtson (2017) in which the author tries to prove the relationship of Basque with North Caucasian.

(gloss)	Basque	Chechen	Avar	Lak / Dargi	Lezgi	Prot-West-Caucasian	Proto-North-Caucasian
die	*hil	= al-	= al' =	L = ič'a D -ibk'-	q'i-	* ʎə - / * ʎa-	* = iwʎ.E
dog	*hor	phu 'male dog'	hoy	D ɣa	χor (Budukh)	*ŁIwa	*χHwěy-rV-
ear	*be = laŕi	ler-g		D lihi		*ŁA-	*łěHi
f re	*śu	ts'e	ts'a	L ts'u D ts'a	ts'ay	*mA = c w. a	*č äyi
horn	*a = daŕ	kur	tɬ:ar		f ri 'mane'		PEC *ʎ.wĩ rV
I	*ni			L na D nu		*q'lwA 'to hear; to be heard'	* = iŕĕ

Table 1: Supposed cognate sets between Basque and North-Caucasian languages (from Bengtson 2017)

It is not difficult to see that we can barely see *anything* in this representation. Later on, we will see how this kind of data can be represented in a much more efficient and transparent way.

Last not least, comparability of data also poses a significant problem of itself, since scholars often do not pay enough attention when it comes to sharing their data in such a form that it is actually comparable with the data provided by other scholars. It is obvious that not all aspects of languages are comparable cross-linguistically in the end, but it is also clear that many aspects *are*, and as a result, linguists should always try to offer their data in such a form that they maximize the potential synergies with other fields. If we do not try hard to increase the comparability, transparency, and availability of our data, our research will end up being irreproducible, and reproducibility is one of the key aspects of scientific research.

In which cases may it be justified to not share all of ones data in science?

2 The CLDF Initiative

2.1 General Ideas

The *Cross-Linguistic Data Formats* initiative (Forkel et al. 2017, <http://cldf.clld.org>) comes along with: (a) standardization efforts, (b) software APIs which help to test and use the data, and (c) working examples for best practice. (a) points to linguistic meta-databases like Glottolog (Hammarström et al. 2017), Concepticon (List et al. 2016), and CLTS (List 2017b). These databases help scholars to make explicit what data (what languages, what concepts, what sounds) they are working with, and additionally aid them in merging different datasets into larger data collections. They aim, in brief, at increasing the *comparability* of linguistic data. (b) points to software (currently written in Python), which helps users to *test* how well their data conforms to the standards established by the CLDF initiative. The software contributes to the transparency of the data, as it requires data to be presented

in both machine- and human-readable formats. (c) points to existing datasets which have been created by different scholars and try to illustrate how the standards can be used and implemented. These working examples (see, for example, Hill and List 2017) increase both the *availability* of data, as well as contributing to *transparency* and *comparability*.

What is the advantage of using existing meta-data-bases like Glottolog or Concepticon for data collection and data annotation?

2.2 Technical Aspects

The details can be assessed from the CLDF website (<http://cldf.clld.org>) where apart from the specification along with working examples an ontology can be found which explains certain core aspects of different data types which can be used in wordlists, collections of grammatical features, or dictionary data. As a general format for the machine-readable specification we use CSV with metadata in JSON, following the W3C's Model for Tabular Data and Metadata on the Web (W3C Consortium 2015). Our CLDF ontology builds and expands upon the *General Ontology for Linguistic Description* (GOLD, GOLD Community 2010). The `pycldf` API in Python (<https://github.com/glottobank/pycldf>) is close to its first release and can be used to test how well datasets conform to CLDF.

Aren't there better formats for lexical data around, such as TEI or simple XML?

2.3 Standards

CLDF can be divided into different *modules* in which specific standards for frequently recurring tabular datatypes are defined. Currently, CLDF features three main modules, one for (a) *wordlists*, one for (b) *dictionaries*, and one for (c) *features*. The CLDF wordlist standard is integrated in different tools, like LingPy (List and Forkel 2016) and Beastling (Maurits et al. 2017), and draws conceptually a lot from the annotation practices for historical language comparison developed for the EDICTOR (List 2017a) tool. The CLDF dictionary standard will serve as the basic format for the Dictionaria project (<http://dictionaria.clld.org>), and the feature standards, which define basic ways to handle grammatical features in cross-linguistic datasets, will be available for both long established typological databases like WALS (Dryer and Haspelmath 2011), and for the Grambank database (<http://glottobank.org>), which is currently being assembled by colleagues from the MPI (Jena).

All standards make use of the core meta-data-bases which were mentioned before. Linguists should try to link all their language varieties to Glottolog, and if word lists or other forms of questionnaires are being used, the concepts should be linked to the Concepticon. If phonetic transcriptions are being used, data can further be tested whether it is compliant with the CLTS standard, which is currently being developed. It is important to note that neither of the meta-data-bases is considered to be a fixed system that could no longer be modified. All of them are flexible, considered as community effort, easy to expand, and actively developed. Every year, the maintainers publish a new release in which they try to account for all requests, complaints, and modification requests which colleagues brought up since the last release. This guarantees that users who want to model their data in CLDF can actively participate in further advancing the core meta-data-bases which are used in CLDF.

Why do we need standards for phonetic transcriptions? Isn't there the IPA standard?

3 Concepticon¹

3.1 Introduction

In 1950, Morris Swadesh (1909 – 1967) proposed the idea that certain parts of the lexicon of human languages are universal, stable over time, and rather resistant to borrowing. As a result, he claimed that this part of the lexicon, which was later called *basic vocabulary*, would be very useful to address the problem of subgrouping in historical linguistics:

[...] it is a well known fact that certain types of morphemes are relatively stable. Pronouns and numerals, for example, are occasionally replaced either by other forms from the same language or by borrowed elements, but such replacement is rare. The same is more or less true of other everyday expressions connected with concepts and experiences common to all human groups or to the groups living in a given part of the world during a given epoch. (Swadesh 1950: 157)

He illustrated this by proposing a first *list of basic concepts*, which was, in fact, nothing else than a collection of concept labels, as shown below:²

I, thou, he, we, ye, one, two, three, four, five, six, seven, eight, nine, ten, hundred, all, animal, ashes, back, bad, bark, belly, big, [...] this, tongue, tooth, tree, warm, water, what, where, white, who, wife, wind, woman, year, yellow. (ibid.: 161)

In the following years, Swadesh refined his original concept lists of basic vocabulary items, thereby reducing the original test list of 215 items first to 200 (Swadesh 1952) and then to 100 items (Swadesh 1955). Scholars working on different language families and different datasets provided further modifications, be it that the concepts which Swadesh had proposed were lacking proper translational equivalents in the languages they were working on, or that they turned out to be not as stable and universal as Swadesh had claimed (Alpher and Nash 1999, Matisoff 1978). Up to today, dozens of different concept lists have been compiled for various purposes.

Who was one of the earliest Chinese scholars to propose a specific concept list?

3.2 Concept Lists

Concept lists are simply speaking collections of concepts which scholars decided to compile at some point. In an ideal concept list, concepts would be described by a *concept label* and a short *definition*. Most published concept lists, however, only contain a concept label. On the other hand, certain concept lists have been further expanded by adding structure, such as *rankings*, *divisions*, or *relations*.

Concept lists are compiled for a variety of different *purposes*. The purpose for which a given concept list was originally defined has an immediate influence on its *structure*. Given the multitude of use cases in both synchronic and diachronic linguistics, it is difficult to give an exhaustive and unique classification scheme for all concept lists which have been compiled in the past. In the following table, we have nevertheless tried to distinguish eight basic types of concept lists and give one list for each of the types as a prototypical example.³

¹ This part is based on List et al. (2016)

² This list contains 123 items in total. According to Swadesh, these items occurred both in his original test list of English items, and in the data on the Salishan languages, which he employed for his first glottochronological study.

³ For further information regarding these concept lists, just click on the links in the “Example” field of the table.

Type	Example	Purpose
basic vocabulary list (“Swadesh list”)	Swadesh 1952 / 200 items	subgrouping
subdivided concept list	Yakhontov 1991 (Starostin 1991) / 35 + 65 items	genetic relationship, layer identification
“ultra-stable” concept list	Dolgopolsky 1964 / 15 items	genetic relationship
questionnaire	Allen 2007 / 500 items	dialect / language comparison
ranked list	Starostin 2007 / 110 items	subgrouping, layer identification
list of concept relations	DatSemShift, Bulakh et al. 2013 / 2424 items	representation of concept relations
special-purpose concept list	Matisoff 1978 / 200 items	subgrouping of Tibeto-Burman languages
historical concept list	Leibniz 1768 / 128 items	language comparison

Table 2: Examples for different types of concept list as they can be found in the literature

3.3 Linking Concept Lists

While all the concept lists which have been published so far constitute language resources with rich and valuable information, we lack guidelines, standards, best practices, and models to handle their interoperability. Language diversity is often addressed with region- or language-specific questionnaires. This makes it difficult to integrate and compare these resources.

The Concepticon is an attempt to overcome these difficulties by linking the many different concept lists which are used in the linguistic literature. In order to do so, we offer open, linked, and shared data in collaborative architectures. Our data is curated openly on GitHub (<https://github.com/clld/concepticon-data>). The Concepticon itself is published as Linked Open Data (<http://concepticon.clld.org>) within the CLLD framework, which allows us to reuse tools built on top of the CLLD API, in particular the `clldclient` package (<https://github.com/clld/clldclient>).

In our Concepticon, all entries from concept lists are partitioned into sets of labels referring to the same concept – so called *concept sets*. Each concept set is given a unique identifier (Concepticon ID), a unique label (Concepticon Gloss), a human-readable definition (Concepticon Definition), a rough semantic field, and a short description regarding its *ontological category*. Based on the availability of resources, we further provide metadata for concept sets, including links to the Princeton WordNet (University 2010), OmegaWiki (OmegaWiki 2005) and BabelNet (Navigli and Ponzetto 2012), and links to norm data bases, like SimLex-999 (Hill et al. 2015), the MRC Psycholinguistic database (Wilson 1988), and the Edinburgh Associative Thesaurus (Kiss et al. 1973).

A concept list is a collection of concepts that is deemed interesting by scholars. Minimally, it consists of an *identifier* for each concept which the lists contains, and a *label* by which the concept is referenced. The creator of a concept list is called a *compiler*. Each concept list is tied to one or more *sources*, it is given in one or more *source languages* and was compiled for one or more *target languages*. A *description* gives further information on each concept list in human-readable form, and tags are used to provide information regarding some basic characteristics of the concept list. The following figure illustrates how concept hierarchies are superimposed on our concept sets.

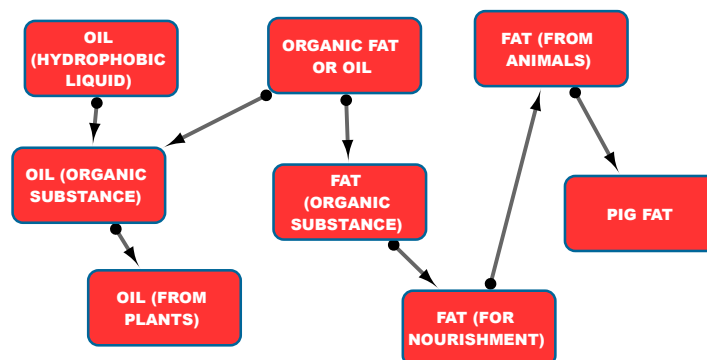


Figure 1: Concept relations between 'oil', and 'fat'

What is the concept from the semantic field for 'fat' which we would expect in a Chinese questionnaire?

3.4 Examples

As a simple example for typical problems involving the linking of concept lists, consider the concepts given in the table below. Here, the four lists apparently intend to denote the same concept 'dull'. From the Chinese terms used in the lists by Ben Hamed and Wang (2006) and Chén (1996), however, we can clearly see that the intended meaning is not 'dull' in the sense of 'being blunt (of a knife)', but 'stupid'. Given that both authors originally wanted to render Swadesh's original concept lists in their research, this shows that we are dealing with a translation error here which may well result from the fact that in many concept lists, only 'dull' is used as a concept label, without further specification.

Compiler	Label	Concepticon
Blust (2008)	dull, blunt	DULL
Chén (1996)	呆, 笨 / dull	STUPID
Comrie & Smith (1977)	dull	DULL
Wang (2006)	笨 (不聪明) / dull	STUPID
Swadesh 1952	dull (knife)	DULL

Table 3: Erroneous translations in concept lists

What other errors in translations can be possible, when considering Swadesh's original list of 200 concepts?

3.5 Outlook

The forthcoming version of the Concepticon will feature a great number of new concept lists. In addition, it will also offer a full-fledged software API which is already available online for testing and which offers new and improved algorithms for an automatic preliminary linking of concept lists. These algorithms are quite powerful, given that they make use in all the concept sets which we have linked so far. In some sense, the algorithm is learning from each new concept list we add and offers all feasible solutions to link a certain concept in a given concept list to our Concepticon.

What is the specific challenge of designing an algorithm for automatic concept set linking?

4 Cross-Linguistic Transcription Systems

4.1 Introduction

Many linguists think that the International Phonetic Alphabet as defined by the International Phonetic Association is a clear-cut standard that does not leave any doubt and just has to be taken seriously by linguists (*IPA Handbook* 1999). However, if we look at the ways in which linguists produce linguistic data, we can first see, that the IPA is not the only phonetic transcription system currently in use. In addition, there is also the *North American Phonetic Alphabet* which is inconsistently and differently used by authors working chiefly on North American languages. There is the *Uralic Phonetic Alphabet*, which is often used but has also never been rigorously standardized (Sovijärvi and Peltola 1970). There is the *Lautschrift der Theutonista* (Wiesinger 1964) which was chiefly used to transcribe German dialect varieties, and there are the specific but largely regular idiosyncrasies of Chinese dialectologists who still keep using an older IPA version from the 1970ies.

Does it really make a difference, which transcription systems linguists use?

4.2 Problems

As a result of this high number of different transcription systems, we encounter many problems when trying to make our data cross-linguistically comparable. Essentially, if linguists say that their data has “IPA inside” this may mean different things depending on the linguists. In addition, the IPA itself creates ambiguities and does not consider itself as a standard in the common sense, but more as a set of suggestions that should help linguists carrying out phonetic transcriptions. Unfortunately, linguists even disregard the suggestions made by the IPA, not to speak of many pitfalls resulting from the Unicode standard and its use (Moran and Cysouw 2017).

Why does the IPA not want to be a standard?

4.3 Comparative Databases

As of now, there are many comparative databases which offer interesting cross-linguistic data, mainly for phoneme inventories in the languages of the world, but sometimes even containing lexical descriptions. The following table gives an overview on some larger datasets:

Dataset	Transcr. Syst.	Sounds
GLD (Ruhlen 2008)	NAPA (modified)	600+ (?)
Phoible (Moran et al. 2014)	IPA (specified)	2000+
GLD (Starostin 2015)	UTS	?
ASJP (Wichmann et al. 2016)	ASJP Code	700+
PBase (Mielke 2008)	IPA (specified)	1000+
Wikipedia	IPA (unspecified)	?
JIPA	IPA (norm?)	800+

Table 4: Cross-linguistic datasets with different transcription systems

What is the JIPA?

4.4 Objective of CLTS

The goal of CLTS is to provide a standard for phonetic transcription for the purpose of cross-linguistic studies by offering standardized ways to represent sound values serve as "comparative concepts" in the sense of Haspelmath (2010). Similar to the Concepticon, we want to allow to register different transcription systems but link them with each other by linking each transcription system to unique sound segments. In contrast to Phoible or other databases which list solely the inventories of languages, CLTS is supposed to serve as a standard for the handling of lexical data in the CLDF framework, as a result, not only sound segments need to be included in the framework, but also ways to transcribe lexical data consistently.

What consequences does it have if CLTS is supposed to serve for phonetic transcription of lexical entries?

4.5 Strategy

We register transcription systems by linking the sounds to phonetic feature bundles which serve as identifiers for sound segments. When being given a form that is supposed to be presented in a given transcription system, we apply a three-step normalization procedure that goes from (1) NFD-normalization (Unicode decomposed characters), via (2) Unicode confusables normalization (<http://unicode.org/cldr/utility/confusables.jsp>), to (3) dedicated *alias symbols*. We divide sounds in different sound classes (currently *vowel*, *consonant*, *diphthong*, *cluster*, *click*, *tone*) to define specific rules for their respective feature sets. Additionally, we allow for a quick expansion of the set of features and the sound segments for each alphabet by applying a procedure that tries to guess unknown sounds by decomposing them into base sounds and diacritics.

On top of the different sounds we can register in this way, we link the feature bundles with datasets, like Phoible, LingPy's sound class system, Wikipedia's sound descriptions, or the binary feature systems published along with PBase (see above for references). Our feature system is not ambitious, as it is neither minimal, nor ordered, nor exclusive, nor binary, as in features systems that have been proposed in the past (Chomsky and Halle 1968). They merely serve as a means of description, following the IPA as closely as possible. The following two tables illustrate how characters are analysed in CLTS.

Input	NFD	Confus.	Alias	Out
ã (U+00E3)	a (U+0061) ã (U+0303)			ã
a (U+0061) : (U+003a)		a (U+0061) : (U+02d0)		a:
ts (U+02a6)		t (U+0074) s (U+0073)		ts

Table 5: Three-step normalization in CLTS.

Sound	Identifier
ã	nasalized unrounded open front vowel
a:	long unrounded open front vowel
ts	voiceless alveolar affricate consonant

Table 6: Identifiers for sounds.

Wouldn't it be sufficient to go for simple NFD normalization, given that Unicode is a real standard?

4.6 API, Online Demo, and Statistics

The API is similar to the one which is shipped with the Concepticon and offers easy ways for experienced Python users to use the data for automatic analyses. In addition, we are

working on an online demo, which currently exists as a prototype and can be accessed via <http://calc.digling.org/clts/>.

Our current statistics are constantly changing in this stage, and we expect to expand the data quickly during the next months. Currently, we have registered two transcription systems, B(road)IPA and ASJP, as well as two meta-data-sets (Phoible and PBase). The following table shows, how many sounds of Phoible and Pbase we already cover:

Dataset	Matched	Generated	Missed	Perc.
Phoible	613	616	772	61%
PBase	496	265	521	59%

Table 7: Current coverage of CLTS

What problems can be expected when trying to link all of the sounds in Phoible and Pbase?

4.7 Outlook

In the future, we plan to add four more transcription systems (UPA, NAPA, GLD-UTS, X-SAMPA), more more metadata (Index Diachronica, Ruhlen's Database, sound examples, examples from the JIPA), we want to enhance the Python API to work on all platforms, and all Python versions (2 and 3), and we want to enhance the web-application (allow to select between different transcription systems, translate between systems, etc.).

All nice, but what do you think can be done with all this ``normalized`` data? Why do we even need unified transcription systems?

References

- Allen, B. (2007). *Bai Dialect Survey*. SIL International. PDF: <http://www.sil.org/silesr/2007/silesr2007-012.pdf>.
- Alpher, B. and D. Nash (1999). "Lexical replacement and cognate equilibrium in Australia". *Australian Journal of Linguistics: Journal of the Australian Linguistic Society* 19.1, 5–56.
- Ben Hamed, M. and F. Wang (2006). "Stuck in the forest: Trees, networks and Chinese dialects". *Diachronica* 23, 29–60.
- Bengtson, J. D. (2017). *The Euskaro-Caucasian Hypothesis. Current model. A proposed genetic relationship between Basque (Vasconic) and the North Caucasian language family*. Ed. by A. for the Study of Language in Prehistory.
- Bulakh, M., D. Ganenkov, I. Gruntov, T. Maisak, M. Rousseau, and A. Zalizniak, eds. (2013). *Database of semantic shifts in the languages of the world*. URL: <http://semshifts.iling-ran.ru/> (visited on 11/04/2014).
- 陈保亚, C. B. (1996). *Lùn yǔyán jiēchù yǔ yǔyán liánméng* 论语言接触与语言联盟 [Language contact and language unions]. Běijīng 北京: Yǔwén 语文.
- Chomsky, N. and M. Halle (1968). *The sound pattern of English*. New York, Evanston, and London: Harper and Row.
- Dolgopolsky, A. B. (1964). "Gipoteza drevnejšego rodstva jazykovych semej Severnoj Evrazii s verojatnostej točki zrenija [A probabilistic hypothesis concerning the oldest relationships among the language families of Northern Eurasia]". *Voprosy Jazykoznanija* 2, 53–63; English translation: Dolgopolsky, A. B. (1986). "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*. Ed. and trans. from the Russian by V. V. Shevoroshkin. Ann Arbor: Karoma Publisher, 27–50.
- Dryer, M. S. and M. Haspelmath, eds. (2011). *The World Atlas of Language Structures online*. Munich: Max Planck Digital Library.
- Forkel, R., S. Greenhill, and J.-M. List (2017). *Cross-Linguistic Data Formats (CLDF)*. Jena: Max Planck Institute for the Science of Human History.
- GOLD Community (2010). *General Ontology for Linguistic Description (GOLD)*. Ontology. Department of Linguistics (The LINGUIST List), Indiana University.
- Hammarström, H., R. Forkel, and M. Haspelmath (2017). *Glottolog*. Version 3.0. URL: <http://glottolog.org>.
- Haspelmath, M. (2010). "Comparative concepts and descriptive categories". *Language* 86.3, 663–687.
- Hill, F., R. Reichart, and A. Korhonen (2015). "SimLex-999: Evaluating semantic models with (genuine) similarity estimation". *Computational Linguistics* 41.4, 665–695.
- Hill, N. W. (2017). "The State of Sino-Tibetan". Review of Thurgood and Lapolla (2017) *The Sino-Tibetan Languages*. Second Edition. *Archiv Orientalní* 85, 305–315.
- Hill, N. W. and J.-M. List (2017). "Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages". *Yearbook of the Poznań Linguistic Meeting* 3.1, 47–76.

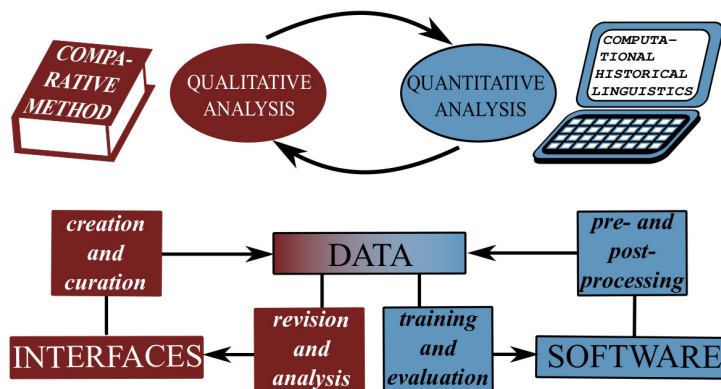
- IPA Handbook (1999). *Handbook of the International Phonetic Association. A guide to the use of the international phonetic alphabet*. Cambridge: Cambridge University Press.
- Kiss, G., C. Armstrong, R. Milroy, and J. Piper (1973). "An associative thesaurus of English and its computer analysis". In: *The computer and literary studies*. Ed. by A. Aitken, R. Bailey, and N. Hamilton-Smith. Edinburgh: Edinburgh University Press, 153–165.
- Leibniz, G. W. von (1768). "Desiderata circa linguas populorum, ad Dn. Podesta [*Desiderata regarding the languages of the world*]". In: *Godefridi Guilielmi Leibnitii opera omnia, nunc primum collecta, in classes distributa, praefationibus et indicibus exornata* [Collected works of Gottfried Wilhelm Leibniz, now first collected, divided in classes, and enriched by introductions and indices]. Ed. by L. Dutens. Vol. 6. 2. Geneva: Fratres des Tournes, 228–231.
- List, J.-M. (2017a). "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.
- (2017b). *Establishing a cross-linguistic database of phonetic notation systems*. Paper, presented at the workshop "Phonological Representation in the Quantitative Era of Comparative Linguistics, organized as part of PLM 2017" (Poznan, 09/18/2017).
- List, J.-M. and R. Forkel (2016). *LingPy. A Python library for historical linguistics*. Version 2.5. URL: <http://lingpy.org>.
- 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. "LREC 2016"* (Portorož, 05/23–05/28/2016). Ed. by N. C. C. Chair, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis. European Language Resources Association (ELRA), 2393–2400.
- Matisoff, J. A. (1978). *Variational semantics in Tibeto-Burman. The 'organic' approach to linguistic comparison*. Institute for the Study of Human Issues.
- Maurits, L., R. Forkel, G. A. Kaiping, and Q. D. Atkinson (08/2017). "BEASTling: A software tool for linguistic phylogenetics using BEAST 2". *PLOS ONE* 12.8, 1–17.
- Mielke, J. (2008). *The emergence of distinctive features*. Oxford: Oxford University Press.
- Moran, S. and M. Cysouw (02/2017). *The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles*. Zürich: Zenodo.
- Moran, S., D. McCloy, and R. Wright, eds. (2014). *PHOIBLE Online*. URL: <http://phoible.org/>.
- Navigli, R. and S. P. Ponzetto (2012). "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network". *Artificial Intelligence* 193, 217–250.
- OmegaWiki (2005). *OmegaWiki: A dictionary in all languages*. URL: <http://www.omegawiki.org/>.
- Ruhlen, M. (2008). *A global linguistic database*. Moscow: RGGU.
- Sovijärvi, A. and R. Peltola (1970). *Suomalais-Ugrilainen Tarkekirjoitus Uralic Phonetic Alphabet. Transcription System*. Helsinki: University of Helsinki.
- Starostin, G. S. and P. Krylov, eds. (2011). *The Global Lexicostatistical Database. Compiling, clarifying, connecting basic vocabulary around the world: From free-form to tree-form*. URL: <http://starling.rinet.ru/new100/main.htm>.
- Starostin, S. A. (1991). *Altajskaja problema i proischozhenije japonskogo jazyka* [*The Altaic problem and the origin of the Japanese language*]. Moscow: Nauka.
- Swadesh, M. (1950). "Salish internal relationships". *International Journal of American Linguistics* 16.4, 157–167. JSTOR: 1262898.
- (1952). "Lexico-statistic dating of prehistoric ethnic contacts. With special reference to North American Indians and Eskimos". *Proceedings of the American Philosophical Society* 96.4, 452–463. JSTOR: 3143802.
- (1955). "Towards greater accuracy in lexicostatistic dating". *International Journal of American Linguistics* 21.2, 121–137. JSTOR: 1263939.
- Tamburelli, M. and L. Brasca (2017). "Revisiting the classification of Gallo-Italic: a dialectometric approach". *Digital Scholarship in the Humanities* fqx41.
- University, P. (2010). *WordNet. A lexical database for English*. Online Resource. Princeton.
- W3C Consortium (12/17/2015). *Model for Tabular Data and Metadata on the Web*. W3C Recommendation. W3C.
- Wichmann, S., E. W. Holman, and C. H. Brown (2016). *The ASJP database*. Jena: Max Planck Institute for the Science of Human History.
- Wiesinger, P. (1964). "Das phonetische Transkriptionssystem der Zeitschrift "Teuthonista". Eine Studie zu seiner Entstehung und Anwendbarkeit in der deutschen Dialektologie mit einem Überblick über die Geschichte der phonetischen Transkription im Deutschen bis 1924". *Zeitschrift für Mundartforschung* 31.1, 1–20.
- Wilson, M. D. (1988). "The MRC psycholinguistic database: Machine readable dictionary. Version 2." *Behavioural Research Methods, Instruments and Computers* 20.1, 6–11.

Computer-Assisted Language Comparison

1 Introduction

Traditional methods in historical linguistics are based on manual data annotation. With more data available, they reach their practical limits. Computational methods which have been proposed during the last three decades are fast and efficient, but they are not very accurate. As a result, they cannot replace experience and intuition of experts. Since experts are slow compared to computers, while computers are not very accurate compared to experts, we need *combined* frameworks which reconcile classical and computational approaches. Such a framework of *computer-assisted language comparison* (CALC, List 2016b) may drastically increase the consistency of expert annotation while correcting for the lack of accuracy in computational analyses.

The graphic below shows the general work flow of CALC. What is crucial for the data in such a framework?



2 LingPy: Cognate Detection and More

2.1 Introduction

Software plays a crucial role in the CALC framework, as it helps us to speed up the comparative methods. Generally, software in historical linguistics has taken a lot of inspiration from evolutionary biology in the past, and will continue to do so in the future (List et al. 2016a). However, following the agenda outlined in List (2014), it is important to note that software should not be blindly transferred from applications designed for other fields of science, but rather carefully adapted to our specific linguistic needs.

Among the most crucial tasks for quantitative historical linguistics are the detection of cognate words (*automatic cognate detection*) as well as the task of *ancestral state reconstruction* which proposes concrete scenarios showing how a given number of traits evolved along a given reference phylogeny. Both approaches can be carried out with help of the LingPy Python package for quantitative tasks in historical linguistics (List and Forkel 2016) as well as with custom plugins for the package. It would go too far to explain all the details of these algorithms in this context. Therefore, I will content myself in showing what *can* be

done with the software, rather than showing *how* one can do it. For these purposes, I refer the readers to the extensive documentation of the LingPy project which can be found at <http://lingpy.org>.

Why is it important to adapt the software to our linguistic needs?

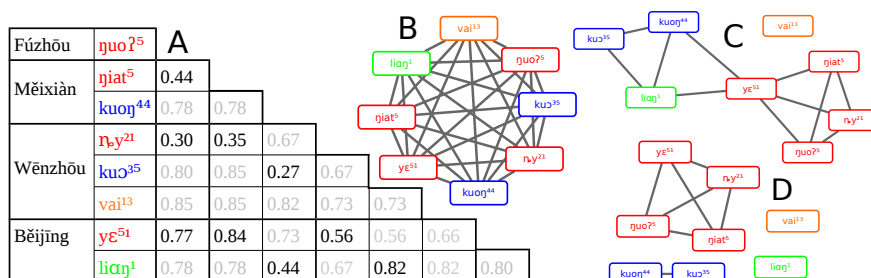
2.2 Cognate Detection

Cognacy is similar to the concept of *homology* in biology (Haggerty et al. 2014), denoting a relation between words which share a common history (List 2014). Quantitative approaches additionally distinguish cognates which have retained, and cognates which have shifted their meaning (Starostin 2013). Further aspects of cognacy are rarely distinguished, although they are obvious and common. Words which go back to the same ancestor form can for example have been morphologically modified, such as French *soleil* which does not go directly back to Latin *sōl* ‘sun’ but to *sōliculus* ‘small sun’ which is itself a derivation of *sōl* (REW). Another problem are words which have been created from two or more morphemes via processes of *compounding*. While these cases are rather rare in the core vocabulary of Indo-European languages, they are very frequent in South-East Asian language families like Sino-Tibetan or Austro-Asiatic. In 200 basic words across 23 Chinese dialects (Ben Hamed and Wang 2006), for example, almost 50% of the nouns and more than 30% of all words consist of two or more morphemes. This is illustrated in the following table (originally from List et al. 2016b).

Variety	Form	Character	Cognacy
Fúzhōu	ŋuoʔ ⁵	月	1
Měixiàn	ŋiat ⁵ kuon ⁴⁴	月光	1 2
Wēnzhōu	ny ²¹ ku ³⁵ vai ¹³	月光佛	1 2 3
Běijīng	ye ⁵¹ lian ¹	月亮	1 4

While algorithms for cognate detection are quite reliable by now, reaching levels of accuracy close to 90% for shallow language families (List et al. 2017a), we still face considerable problems in identifying partial cognates, as illustrated for the Chinese dialects above. In List et al. (2016b), we have proposed a new approach that can be applied to search for cognate sets in Chinese data, assigning syllables to cognate sets, rather than full words. The workflow of this approach is slightly more complicated than the normal workflow for cognate detection, but the LingPy package offers a rather stable implementation, and the results are comparable in accuracy to the ones which can be reported for normal cognate detection algorithms. Furthermore, with help of the EDICTOR tool (List 2017a) described below, linguists can easily refine the findings manually, which makes partial cognate detection a prime example for a CALC workflow.

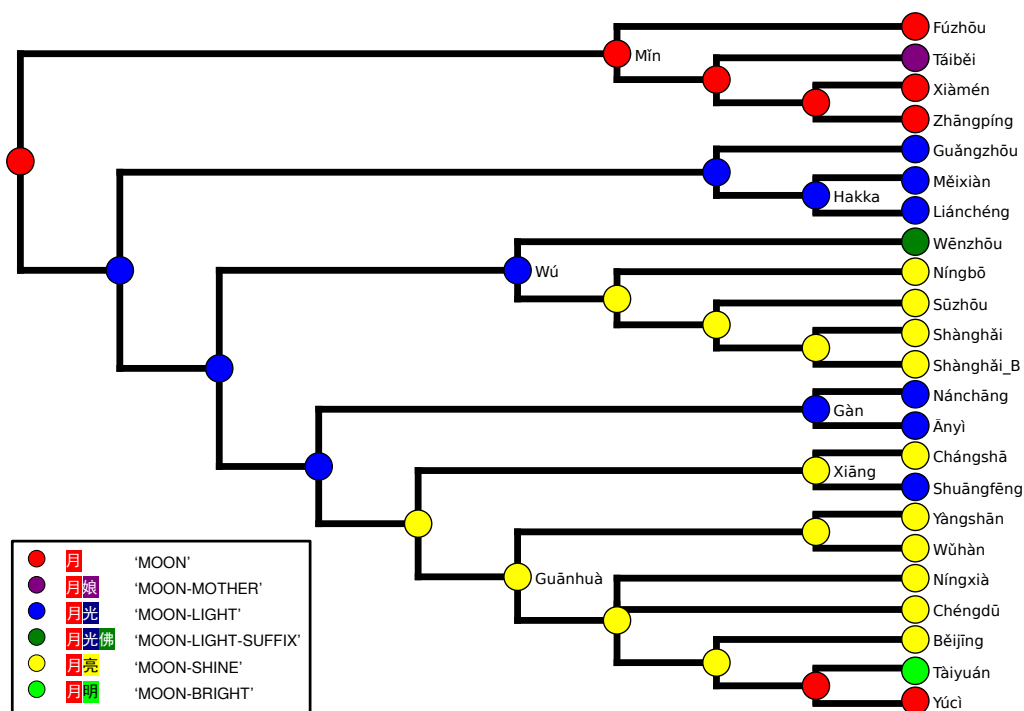
The graphic below shows the workflow of the algorithm for partial cognate detection. What are the crucial steps of this workflow?



2.3 Ancestral State Reconstruction

Ancestral state reconstruction plays a crucial role in evolutionary biology but is still less common in historical linguistics, although most linguists apply similar procedures in their heads when it comes to assessing which words were present in a given proto-language. If we deal with questions such as the development of the lexicon in Chinese dialect history, the problem of partial cognacy becomes again extremely difficult to handle, as we often lack the sufficient intuition regarding major tendencies of language change when it comes to compounds in Chinese dialects. We know for sure that at some point words were mono-morphemic in Ancient Chinese, but we have a hard time in saying exactly when the transition to multi-morphemic words started. An experimental plugin to the LingPy package (List 2016a) can be used to investigate lexical evolution in Chinese dialects in which partial cognates are modeled as multiple states of the same character, and evolutionary scenarios for lexical change are inferred within a parsimony framework. Despite the well-known and obvious shortcomings of parsimony analyses, this framework offers first insights into more detailed evolutionary scenarios of lexical evolution, which is indispensable for a more thorough investigation of lexical change in South-East Asian languages.

The algorithm which was used to infer evolutionary scenarios for partial cognates across Chinese dialects cannot (yet) handle borrowing events. Does this problem show up in the visualization of the scenario inferred for ‘‘moon’’?



3 CLICS: Cross-Linguistic Colexifications

3.1 Polysemy, Homophony, and Colexification

Polysemy and *homophony* are two seemingly contrary concepts in linguistics. However, in the end they describe both the same phenomenon, namely that a *word form* in a given

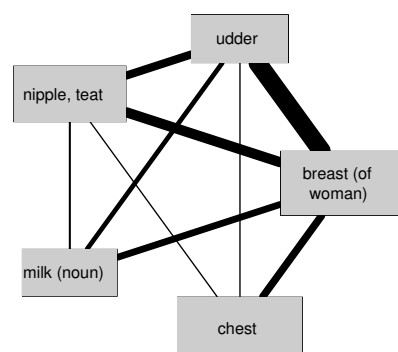
language can have multiple *meanings*. François (2008) therefore suggests to replace the two *interpretative* terms by the *descriptive* term *colexification*. Colexification in this context only means that an individual language ‘is said to colexify two functionally distinct senses if, and only if, it can associate them with the same lexical form’ (ibid.: 171).

How can the distinction between interpretative and descriptive terminology be understood?

3.2 Colexification Networks

If one has enough data, it is considerably simple to construct concept networks from cross-linguistic colexifications. The starting point are semantically aligned word lists for a large amount of different languages from different language families. If, for example, we consider the word list for Russian and German below, we can see that the data contains two “polysemies”, namely Russian *derevo* which can refer to both “tree” and “wood”, and German *Erde* which can refer to both “earth, land” and “ground, soil”. If we now assemble all these connections in a single network, where the nodes represent concepts and the edges represent polysemies, we have created a concept network, or more specifically, a *colexification network*. In order to make the visualization and the analysis more powerful, we can further add *weights* to the edges, representing how many times a certain colexification is encountered in our data.

Key	Concept	Russian	German	...
1.1	world	mir, svet	Welt	...
1.21	earth, land	zemlja	Erde , Land	...
1.212	ground, soil	počva	Erde , Boden	...
1.420	tree	derevo	Baum	...
1.430	wood	derevo	Wald	...
...



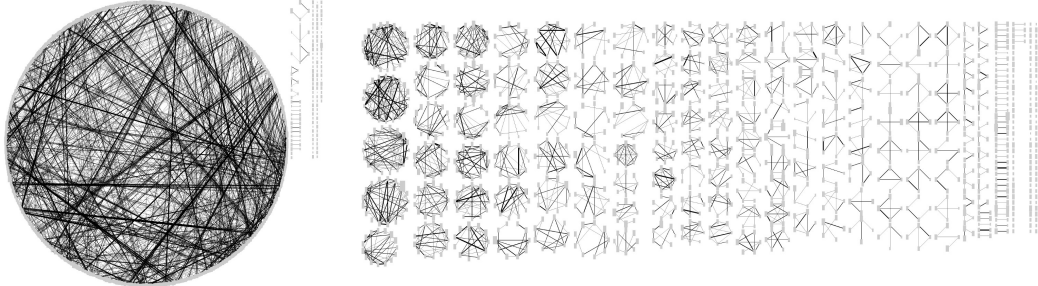
When looking at the original definition of “polysemy” in the linguistic literature, what could become a problem related to colexification networks?

3.3 Analysing Colexification Networks

Taking a colexification network alone does not necessarily help us in answering questions regarding semantic change or human cognition. This is due to the increasing complexity of colexification networks, the more concepts and languages we add. The graphic below, for example, shows a network which has been constructed from an analysis of 195 languages covering 44 language families (List et al. 2013).

What we need is a network analysis which uses specific algorithms to analyse the structure of the network more properly. In concrete, analyses for *community detection* can help us to partition the networks into groups which correspond to important *semantic fields*. The term *community* was first coined in social network analysis, where it was used to identify communities of people in social networks. In a broader sense, a community refers to “groups of vertices within which the connections are dense but between which they are sparser” (Newman 2004: 4). In List et al. (2013), we used the algorithm by Girvan and Newman

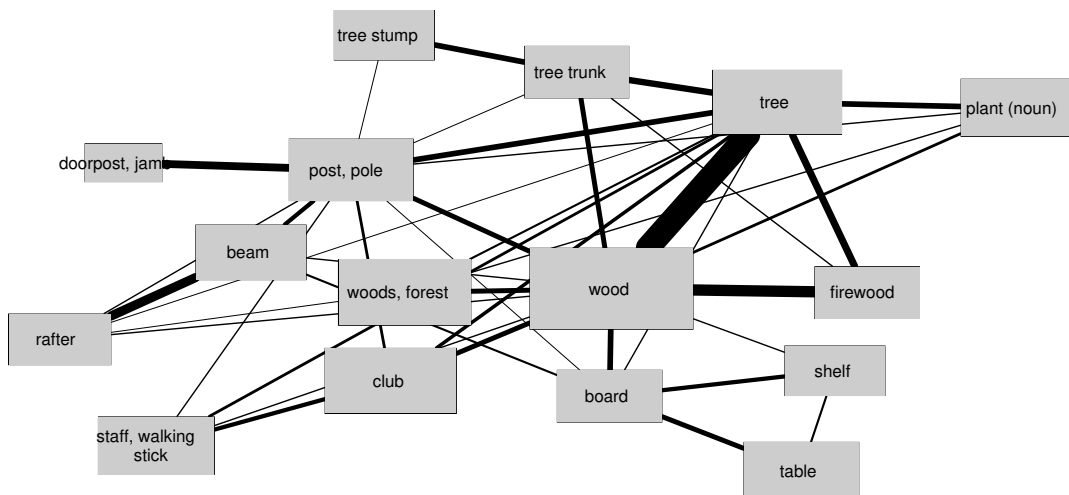
(2002) to analyse the network on the left. The result is given in the graphic on the right, where the originally almost completely connected network has been partitioned into 337 communities, with 104 being relatively big (5 and more nodes, covering a rather large parts of the 1289 concepts in our original database (879, 68%).



(a) complete networks

(b) analysed network

Below a community from the network is shown, in which meanings which center around ``tree`` and ``wood`` have been grouped together. What can we learn from the network? What can't we learn?



3.4 CLICS

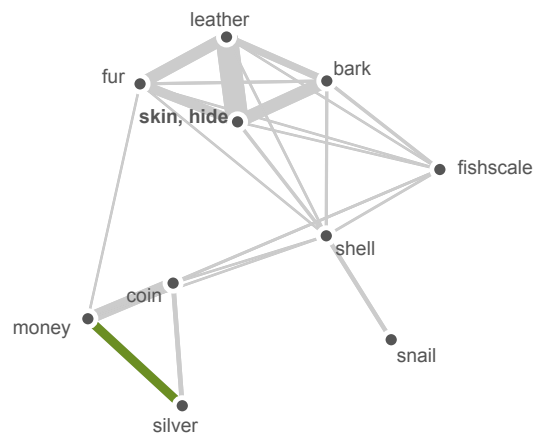
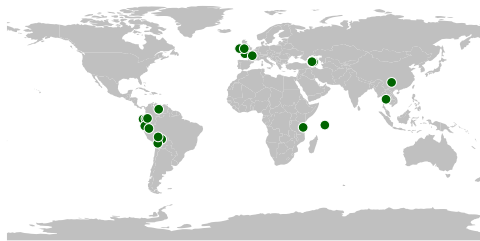
CLICS (List et al. 2014, <http://clics.lingpy.org>) is an online database of synchronic lexical associations ("colexifications") in currently 221 language varieties of the world. Large databases offering lexical information on the world's languages are already readily available for research in different online sources. However, the information on tendencies of meaning associations they enshrine is not easily extractable from these sources themselves.

As CLICS comes along with a powerful visualization suite (Mayer et al. 2014), it is very convenient to query the information regarding meaning associations. CLICS thus also serves as an example for computer-assisted language comparison, in so far as it illustrates how analyses created by machines can be made accessible to the detailed inspection by researchers.

Below is an example visualization from the CLICS database. How can the global patterning be interpreted?

49 links for "silver" and "money":

Language	Family	Form
1. Ignaciano	Arawakan	ne
2. Aymara, Central	Aymaran	kul ^h ki
3. Tsafiki	Barbacoan	ka'la
4. Seselwa Creole French	Creole	larzan
5. Miao, White	Hmong-Mien	nyiaj
6. Breton	Indo-European	arhant
7. French	Indo-European	argent
8. Gaelic, Irish	Indo-European	airgead
9. Welsh	Indo-European	arian
10. Cofán	Isolate	koriΦiʔdi



4 EDICTOR: Cognate Annotation and More¹

4.1 Introduction

The Etymological DICTIONARY ediTOR (EDICTOR) is a free, interactive, web-based tool designed to aid historical linguists in creating, editing, analysing, and publishing etymological datasets. The EDICTOR offers interactive solutions for important tasks in historical linguistics, including facilitated input and segmentation of phonetic transcriptions, quantitative and qualitative analyses of phonetic and morphological data, enhanced interfaces for cognate class assignment and multiple word alignment, and automated evaluation of regular sound correspondences. As a web-based tool written in JavaScript, the EDICTOR can be used in standard web browsers across all major platforms. Due to the simplicity of its format requirements and the strictness of its machine- and human-readable annotation, the tool is ideal for computer-assisted workflows in historical linguistics in which linguists first use existing software packages to analyse their data automatically and then manually correct the results.

4.2 Partial Cognate Annotation

The manual annotation of partial cognates is tedious. In order to ease the task, a partial cognate editor was included in the most recent version of the EDICTOR tool, which greatly facilitates the annotation task. All that is required is that the data are morphologically segmented by the user. Once this is done, users can load their data into the EDICTOR tool and indicate which morphemes in a set of pre-defined words (usually translations of the same comparison concept) are cognate. Since this can be done in a simple drag-and-drop fashion, by which the user selects and deselects the words which are grouped into one partial cognate set, the annotation can be carried out quickly and is also less prone to error than the use of spreadsheet software not designed for this task.

¹ See List (2017a) and Hill and List (2017) for details.

The graphic below shows an example of partial cognate annotation with the EDICTOR. As well as this seems to work, which cases can surely NOT be handled with this method?

DOCULECT	CONCEPT	TOKENS	ID-1044	ID-1043	ID-1046	ID-1045	ID-2074
Old_Burmese	the feather	a ¹⁰⁴⁴ m ⁵⁵ u ⁵⁵ j ¹⁰⁴³	a ³¹	m ⁵⁵ u ⁵⁵ j ⁵⁵			
Bola	the feather	a ³¹ 1044 m ⁵⁵ a ³⁵ u ¹⁰⁴³	a ³¹	m ⁵⁵ a ³⁵ u ³⁵			
Achang_Longchuan	the feather	a ³¹ 1044 m ⁵⁵ u ³¹ 1043	a ³¹	m ⁵⁵ u ³¹			
Atsi	the feather	f ²¹ 1046 m ⁵⁵ a ⁵⁵ u ¹⁰⁴³		m ⁵⁵ a ⁵⁵ u ⁵⁵	f ²¹		
Lashi	the feather	s ⁵⁵ 1046 m ⁵⁵ o ⁵⁵ u ¹⁰⁴³		m ⁵⁵ o ⁵⁵ u ⁵⁵	s ⁵⁵		
Maru	the feather	f ³⁵ 1046 m ⁵⁵ u ⁵⁵ k ¹⁰⁴³		m ⁵⁵ u ⁵⁵ k ⁵⁵	f ³⁵		
Rangoon	the feather	ŋ ⁴ 1045 m ⁵⁵ w ⁵⁵ e ¹⁰⁴³		m ⁵⁵ w ⁵⁵ e ⁵⁵		ŋ ⁴ 55	

4.3 Partial Colexifications

What is often being ignored in phylogenetic analyses is the importance of *internal reconstruction* (Anttila 1972: 264-273). All languages make a considerable re-use of their word form material throughout their lexicon. *Word families* play a crucial role in lexical organisation (Brysbaert et al. 2016). Knowing which words inside a language belong to the same word family or share material from the same family is crucial for both historical language comparison as well as phylogenetic reconstruction. In order to facilitate the annotation of word families, the EDICTOR tool contains a *morpheme annotation module* that allows one to inspect automatically created bipartite networks for individual languages and to annotate compounds in a meaningful way (Hill and List 2017). The general idea behind this compound structure analysis is to annotate compounds in a way similar to how linguists annotate sentences in inter-linear glossed text. For each word in the data, we provide a language-internal analysis that reveals the motivation of compound formation. Essentially, this yields a language-internal word family analysis, as it allow us to identify cognates within the same language.

The table below gives an example for compound analysis and morpheme annotation in the EDICTOR tool. When comparing this with similar structures that could be found for Chinese dialects, which are the obvious drawbacks of this analysis?

ID	COGID	CONCEPT	MORPHEMES	TOKENS
3368	400	the river	water - mo-suffix	v u l i 51 m o 55
3535	425	the sea	water - sea	v u l i 51 m i ŋ 21
4868	409	the water	water	v u l i 51
598	619	to buy	buy	v u l i 51

5 Rhyme Analysis and the Reconstruction of Old Chinese²

5.1 Introduction

The analysis of rhyme patterns is one of the core methods for the reconstruction of Old Chinese phonology. It emerged when scholars of the Suí 隋(581–618) and Táng 唐(618–907) dynasties realized that old poems, especially those in the Book of Odes (Shījīng 詩經 ca. 1050–600 BCE), were full of inconsistencies regarding the rhyming of words. While the first reaction was to attribute inconsistencies to a different, less strict attitude towards rhyming

²See List (2017b) and List et al. (2017b) for details.

practiced by the ancestors (as advocated by Lù Dēmíng 陸德明, 550–630), or to a habit of the elders to switch the pronunciation in certain words in order to make them rhyme (a practice called *xiéyīn* 諧音 ‘sound harmonization’, Baxter 1992:153). Later scholars from the Míng 明 (1368–1644) and Qīng 清 dynasties (1644–1911) realized that the inconsistencies in the rhyme patterns reflect the effects of language change (Baxter 1992:153–157). The following table illustrates this in more detail:

Chinese Text	Translation	RW	Patterns	MCH	OCBS-Rhyme
燕燕於飛	The swallows go flying	<i>fēi</i> 飛	A	*pjij	*-ər
下上其音	falling and rising are their voices;	<i>yīn</i> 音	B	*ʔim	*-əm
之子於歸	This young lady goes to her new home,	<i>guī</i> 歸	A	*kjiw	*-əj
遠送於南	far I accompany her to the south.	<i>nán</i> 南	B	*nom	*-əm
瞻望弗及	I gaze after her, can no longer see her,	[<i>jí</i> 及]	–	[*gip]	[*-əp]
實勞我心	truly it grieves my heart	<i>xīn</i> 心	B	*sim	*-əm

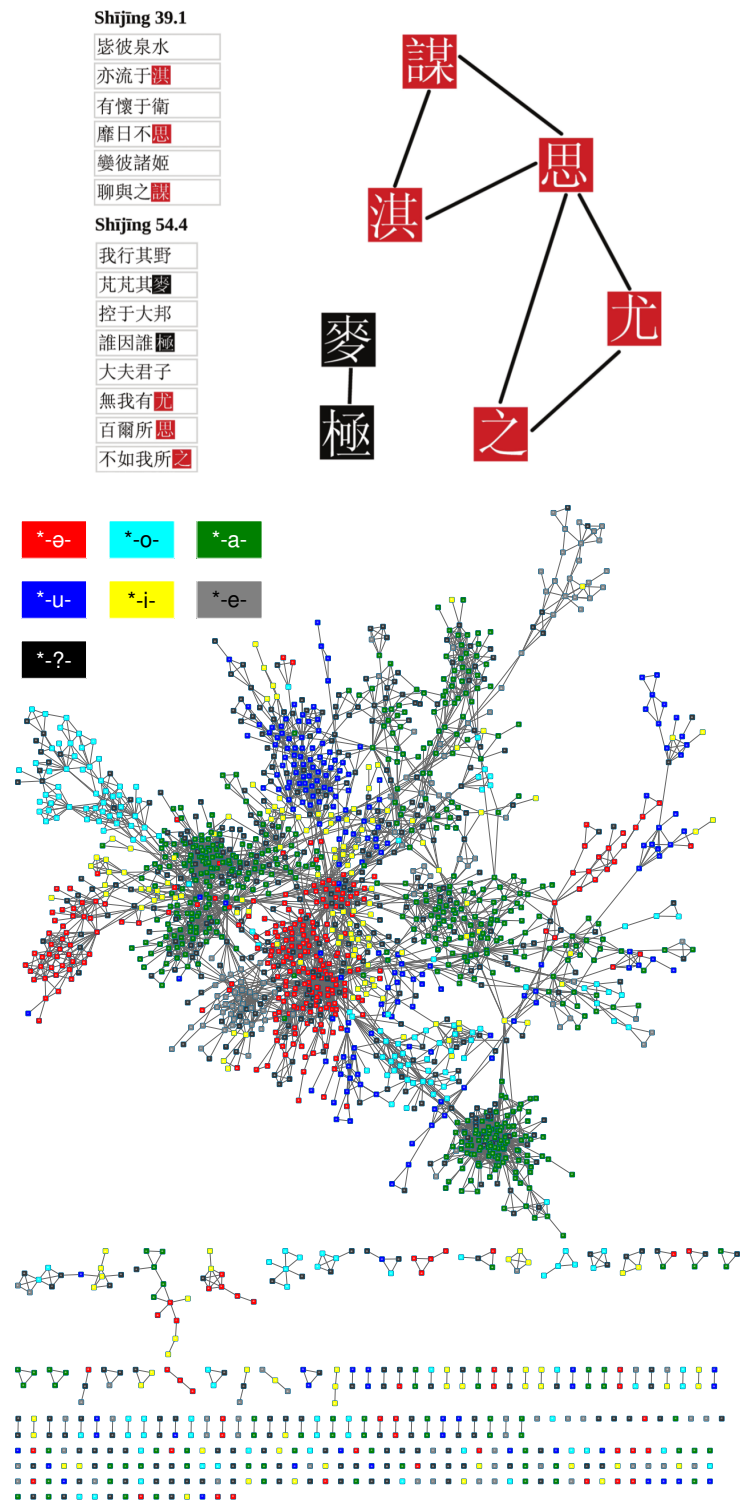
Assuming that rhyming was originally rather consistent, with rhyme words being mostly identical in the pronunciation of nucleus and coda, the analysis of rhyme words makes it not only possible to establish rhyme categories but also to interpret them further phonetically or phonologically. The classical approach for rhyme analysis, which is called *sīguàn shéngqiān fǎ* 絲貫繩牽法 ‘link-and-bind method’ (Gěng 2004), or *yùnjiǎo xìlián fǎ* 韻腳系聯法 ‘rhyme linking method’ (Lǚ 2009), consists of roughly two steps: In a first step, groups of Old Chinese words, mostly represented by one Chinese character and identified to rhyme with each other in a given text are collected. In a further step, these groups are compared with each other. If identical words are found in different groups, those groups can be combined to form larger groups. This procedure is then repeated until categories of rhymes can be identified that ideally do not show any more transitions among each other. This approach is essentially similar to the ‘linking method’ *xilián fǎ* 系聯法 see Liú 2006:56–67), first proposed in Chén Lǐ’s 陳禮 (1818–1882) *Qièyùnkǎo* 切韻考 (1848), by which characters used in *fǎnqiè* 反切 readings in rhyme books are clustered into groups of supposedly common pronunciations for initials and rhymes. In both approaches, similarities in pronunciation are indirectly inferred by spinning a web of direct links between characters.

The figure below illustrates the linking method for the zhǐ 之 group in the Book of Odes. What is the obvious drawback of this method?

27.3.A			sī 絲							
30.2.A	lái 來	sī 思								
33.3.A	lái 來	sī 思								
39.1.A		sī 思								
54.4.B		sī 思				zhī 之			yóu 尤	
58.1.A			sī 絲	qī 淇	móu 謀					
58.6.B		sī 思				zāi 哉		qī 期		
59.1.A		sī 思		qī 淇	móu 謀					
66.1.A	lái 來	sī 思				zāi 哉		qī 期		
130.1.A						zāi 哉			méi 梅	
204.4.A				qī 淇			zhī 之		méi 梅	yóu 尤
227.2.A						zāi 哉				

5.2 Network Approach to Rhyme Analysis

The crucial idea of our computer-assisted approach to rhyme analysis is to construct a *network of rhyme patterns* in which nodes represent rhyme words and connections between nodes represent how often those rhymes co-occur in the Book of Odes. The following graphic illustrates this procedure for two stanzas of the Shijing:



The major advantage of this representation is that we can apply various methods for network analysis to data which was assembled in this form. As a result, we can investigate the rhyme network and test to which degree different reconstruction systems offer a consistent view on Old Chinese rhyming. As a very simple test, we can check whether a given reconstruction system conforms to the principle of *vowel purity* (Ho 2016) which expects words with similar vowels to rhyme more often than words with different vowels. Our test, which is reported in List et al. (2017b) could show that most of the Old Chinese reconstruction systems which postulate 6 vowels correspond more closely to vowel purity than other reconstruction systems with more or less vowels. Even by eyeballing the figure above, in which vowel quality is reflected with help of colors following the OC reconstruction system by Baxter and Sagart (2014), one can see that words rhyming with each other tend to have the same vowel.

If six-vowel reconstruction systems perform better on vowel purity, does this automatically mean that they are better in general?

5.3 The Shijing Rhyme Browser

In order to make it more convenient for the readers to investigate the data underlying this paper in full detail, an interactive web-based application was created. This freely available Shijing Browser (<http://digling.org/shijing/>) lists all potential rhyme words in tabular form along with additional information including the *pīnyīn* transliteration, the Middle Chinese reading, the reconstruction by Baxter and Sagart (ibid.), the reading by Pān (2000), the GSR index (Karlgren 1957), and the number of poem, stanza, and section. With help of interactive search fields, the data can quickly be filtered, enabling the users to search for specific poems, for specific characters, or for specific readings. When clicking on the “Poem” field in the application, a window pops up and shows the whole poem, in which all rhyme words are highlighted. In certain cases, where potential alternative rhymes were identified, this is marked in an additional column. In a recently modified version, we contrast rhyme annotations by Wáng (1980 [2006]) with those given in Baxter (1992) (<http://digling.org/shijing/wangli/>, List 2017c). The table below gives an example on the organization of the interface.

Text	Stanza	MCH	Pān Wúyùn	OCBS	Wáng Lì	Starostin	Rhyme	Group
遵彼汝墳，伐其條枚	1.AB	mwoj	muwul	m [˥] əj	muəi	māj	A	微
未見君子，惄如調飢	1.CD	tsiX	kril	Cə.kə[j]	kiei	krəj	A	脂
遵彼汝墳，伐其條肆	2.AB	sijH	ph-ljwds	s-ləp-s	jiet	slhəps	B	质
既見君子，不我遐棄	2.CD	khijjH	khids	[k ^h][t]-s	khiet	khijS	B	质
魴魚赭尾	3.A	mj+ǰX	mwul?	[m]əj?	miuəi	məj?	C	微
王室如燬	3.B	xjweX	qh ^w ral?	[ŋ](r)aj?	xiuəi	h ^w ej?	C	微
雖則如燬	3.C	xjweX	qh ^w ral?	[ŋ](r)aj?	xiuəi	h ^w ej?	C	微
父母孔迺	3.D	nyeX	mljel?	n[ə][r]?	njiei	n(h)ej?	C	脂

What could be the problem of comparing rhymes in books other than the Book of Odes?

References

- Anttila, R. (1972). *An introduction to historical and comparative linguistics*. New York: Macmillan.
- Baxter, W. H. (1992). *A handbook of Old Chinese phonology*. Berlin: de Gruyter.
- Baxter, W. H. and L. Sagart (2014). *Old Chinese. A new reconstruction*. Oxford: Oxford University Press.
- Ben Hamed, M. and F. Wang (2006). "Stuck in the forest: Trees, networks and Chinese dialects". *Diachronica* 23, 29–60.
- Brysbaert, M., M. Stevens, P. Mandera, and E. Keuleers (2016). "How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age". *Frontiers in Psychology* 7, 1116.
- François, A. (2008). "Semantic maps and the typology of colexification: intertwining polysemous networks across languages". In: *From polysemy to semantic change*. Ed. by M. Vanhove. Amsterdam: Benjamins, 163–215.
- Girvan, M. and M. E. Newman (2002). "Community structure in social and biological networks". *Proceedings of the National Academy of Sciences of the United States of America* 99.12, 7821–7826.
- Haggerty, L. S., P. A. Jachiet, W. P. Hanage, D. A. Fitzpatrick, P. Lopez, M. J. O'Connell, D. Pisani, M. Wilkinson, E. Baptiste, and J. O. McInerney (2014). "A pluralistic account of homology: adapting the models to the data". *Molecular Biology and Evolution* 31.3, 501–516.
- Hill, N. W. and J.-M. List (2017). "Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages". *Yearbook of the Poznań Linguistic Meeting* 3.1, 47–76.
- Ho, D.-a. (2016). "Such errors could have been avoided. Review of "Old Chinese: A new reconstruction". by William H. Baxter and Laurent Sagart". *Journal of Chinese Linguistics* 44.1, 175–230.
- Karlgren, B. (1957). "Grammata serica recensa". *Bulletin of the Museum of Far Eastern Antiquities* 29, 1–332.
- List, J.-M., T. Mayer, A. Terhalle, and M. Urban, eds. (2014). *CLICS: Database of Cross-Linguistic Colexifications*. Version 1.0. Archived at: <http://www.webcitation.org/6ccEMrZYM>. URL: <http://clics.lingpy.org>.
- List, J.-M. (2014). *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.
- (2016a). "Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction". *Journal of Language Evolution* 1.2, 119–136.
 - (12/2016b). *Computer-Assisted Language Comparison: Reconciling Computational and Classical Approaches in Historical Linguistics*. Jena: Max Planck Institute for the Science of Human History.
 - (2017a). "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.
 - (2017b). "Using network models to analyze Old Chinese rhyme data". *Bulletin of Chinese Linguistics* 9.2, 218–241.
 - (05/2017c). *Vertikale und laterale Aspekte der chinesischen Dialektgeschichte* [Vertical and lateral aspects of Chinese dialect history]. Jena: Max Planck Institute for the Science of Human History.
- List, J.-M. and R. Forkel (2016). *LingPy. A Python library for historical linguistics*. Version 2.5. URL: <http://lingpy.org>.
- List, J.-M., A. Terhalle, and M. Urban (2013). "Using network approaches to enhance the analysis of cross-linguistic polysemies". In: *Proceedings of the 10th International Conference on Computational Semantics – Short Papers*. "IWCS 2013" (Potsdam, 03/19–03/22/2013). Association for Computational Linguistics. Stroudsburg, 347–353. PDF: <http://aclweb.org/anthology-new/W/W13/W13-0208.pdf>.
- List, J.-M., J. S. Pathmanathan, P. Lopez, and E. Baptiste (2016a). "Unity and disunity in evolutionary sciences: process-based analogies open common research avenues for biology and linguistics". *Biology Direct* 11.39, 1–17.
- List, J.-M., P. Lopez, and E. Baptiste (2016b). "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. Berlin, 599–605.
- List, J.-M., S. J. Greenhill, and R. D. Gray (01/2017a). "The potential of automatic word comparison for historical linguistics". *PLOS ONE* 12.1, 1–18.
- List, J.-M., J. S. Pathmanathan, N. W. Hill, E. Baptiste, and P. Lopez (2017b). "Vowel purity and rhyme evidence in Old Chinese reconstruction". *Lingua Sinica* 3.1, 1–17.
- Mayer, T., J.-M. List, A. Terhalle, and M. Urban (2014). "An interactive visualization of cross-linguistic colexification patterns". In: *Visualization as added value in the development, use and evaluation of Linguistic Resources. Workshop organized as part of the International Conference on Language Resources and Evaluation*, 1–8. URL: <http://www.lrec-conf.org/proceedings/lrec2014/workshops/LREC2014Workshop-VisLR%20Proceedings.pdf>.
- Meyer-Lübke, W., comp. (1911). *Romanisches etymologisches Wörterbuch*. Sammlung romanischer Elementar- und Handbücher 3.3. Heidelberg: Winter.
- Newman, M. E. J. (2004). "Analysis of weighted networks". *Physical Review E* 70.5, 056131.

- Starostin, G. S. (2013). "Lexicostatistics as a basis for language classification". In: *Classification and evolution in biology, linguistics and the history of science. Concepts – methods – visualization*. Ed. by H. Fangerau, H. Geisler, T. Halling, and W. Martin. Stuttgart: Franz Steiner Verlag, 125–146.
- 潘悟云, P. W. (2000). *Hànyǔ lìshǐ yīnyùnxué* 汉语历史音韵学 [Chinese historical phonology]. Shànghǎi 上海: Shànghǎi Jiàoyù 上海教育.
- 王力, W. L. (1980 [2006]). *Hànyǔ shǐgǎo* 漢語史稿 [History of the Chinese language]. Repr. Běijīng 北京: Zhōnghuá Shūjú 中华书局.