Multilingual Computational Linguistics

Lecture, presented at the University of Passau (Winter Term 2022-2023)

Johann-Mattis List
mattis.list@uni-passau.de

Chair of Multilingual Computational Linguistics
University of Passau

2023

Contents

Introduction

Johann-Mattis List (University of Passau)

While the discipline of computational linguistics mostly deals with the modeling and the investigation of individual languages (often “big” languages such as English, German, Arabic, or Chinese), Multilingual Computational Linguistics focuses on the comparison of languages, trying to develop new methods and techniques by which languages can be compared automatically or in a computer-assisted manner. The comparison itself follows different perspectives (maintaining a historical, typological, or areal viewpoint). In this scientific practice course, we will take a closer look at basic theories and methods which are relevant for the discipline of Multilingual Computational Linguistics. We will look at large corpora with multiple languages of the world as well as data from individual languages and language families. If wanted, we can focus on specific language families, which are relevant for the studies of the participants. Thematically, we want to look at the inference of cognates, the detection of borrowings, the reconstruction of phylogenies, and the modeling of semantic change and sound change. If participants are specifically interested in topics that we could additionally cover, they should write a short email before the end of January, to give us time to check if we can include those topics in the course.

1 Multilingual Computational Linguistics

Given the multitude of multilingual applications in the field of computational linguistics, it may not be easy to give an exact definition of the field, since it will inevitably depend on the individual researchers' background and scientific preferences, how they fill “multilingual computational linguistics” with life. Given my specific background as a historical and comparative linguists working on computational applications that help us to increase the efficiency and accuracy of historical and typological language comparison, my major approach towards multilingual computational linguistics does not have a lot to do with the typical NLP applications that translate across a couple of well-understood languages with huge corpora. What I offer instead are a couple of novel methods and techniques by which we can compare languages synchronically and diachronically with the help of computational methods. While this may seem very narrow-minded at first sight, the scientific focus is much broader, since it touches upon a large range of topics ranging from classical linguistic typology and classical historical linguistics via more recent corpus-based approaches in linguistic typology and phylogenetic approaches in historical linguistics up to topics in psycholinguistics that try to learn more about human cognition through a close investigation of linguistic diversity.

Since the field of multilingual computational linguistics represented in this course is still in its infancy, with most major applications having only been made in the last ten years, we cannot make use of off-the-shelf tools for computational comparative linguistics but are instead in a situation where we need to design these tools and often create them from scratch. As a result, our work allows to gain concrete insights into interdisciplinary work, since we need to check with the methodology of many different disciplines (ranging from classical language comparison via bioinformatics up to computer science and digital humanities) in order to handle the problems we face in our research. As a result, we pay specific attention to *scientific problem solving*, to *open data* and *open science* in general, as well as to traditional methodologies applied in many disciplines of the humanities before the arrival of computational methods.

2 Course Organization

The course is divided into 15 sections distributed over four days. On each day, we start with a rather short morning section in which smaller topics and questions are introduced, followed by two larger sections and one practice section in the evening. On the third day, there is no practice section, and we have only one larger section before a concluding section. We use the practice sections specifically to address individual problems brought to the course by course members. Thus, although some guidelines for these topics are provided, we assume a lot of flexibility on the concrete questions here and will generally split into groups rather than working in a big group with all course members.

3 Course Plan

In the following, a detailed course plan is given.

Day 1: Introductory Topics

10–10:30: Introduction of group participants, which can be extended to the coffee break.

11–12:30: Background on Comparative Linguistics

14–15:30: Scientific Problem Solving

16–16:30: Practice Session: Discuss Unsolved Problems in Smaller Groups

Day 2: Modeling and Standards

10–10:30 Cross-Linguistic Data Formats

11–12:30 Reference Catalogs

14–15:30 Standardized Data Collections in Multilingual Computational Linguistics

16–16:30 Practice Session: Standardize and Retrostandardize Data, Parse Texts

Day 3: Inference

10-10:30 Computer-Assisted Language Comparison

11-12:30 Sequence Comparison

14-15:30 Semantic Networks

16-16:30 Practice Session: Workflow Development and Testing

Day 4: Analyzing

10-10:30 Chinese Computational Linguistics

11-12:30 Computer-Assisted Text Analysis

14-15:00 Final Discussion

Background on Comparative Linguistics

Johann-Mattis List (University of Passau)

1 Preliminary Considerations

What is a Language?

What counts as a languages, 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 and 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}	tʰai^{51}iaŋ^{11}	ʈ͡ʂəŋ^{55} tsai^{53}	naɚ^{51}	ʈ͡ʂəŋ^{55}luən^{51}
Hakka Chinese	1	iu^{33}	it^{55}	pai^{33}a^{11}	pet^{33}fuŋ^{33}	tʰuŋ^{11}	ɲit^{11}tʰeu^{11}	hɔk^{33}	e^{53}	au^{55}
Shànghǎi Chinese	1	ɦi^{22}		tʰã^{55} tsɿ^{21}	poʔ^{3}foŋ^{44}	taʔ^{5}	tʰa^{33}ɦiã^{44}	tsəŋ^{33} hɔ^{44}		ləʔ^{1}lə^{23}tsa^{53}
Běijīng Chinese	2	ʂei^{35}		də^{55}		pən^{35} liŋ^{21}	ta^{51}			
Hakka Chinese	2	man^{33}	ɲin^{11}		kʷɔ^{55}	vɔi^{53}				
Shànghǎi Chinese	2	sa^{33}	ɲiŋ^{55}	ɦəʔ^{21}		pəŋ^{33} zɿ^{44}	du^{13}			
Norwegian	1	nuːɾɑʋinˑn̩	ɔ	suːln̩					kɾɑŋlət	ɔm
Swedish	1	nuːɖanvɪndən	ɔ	suːlən		ty̥ɪstadə	ən gɔŋ			ɔm
Danish	1	noʌ̯ʌnvenˀn̩	ʌ	soːl̩ˀn	kʰʌm		eŋɡ̊ɑŋ	i sd̥ʁiðˀ		ʌmˀ
Norwegian	2	ʋem	ɑ	dem	sm̩	ʋɑː	ɖɳ̩	stæɽ̥kəstə		
Swedish	2	vɛm	ɑv	dɔm	sɔm	vɑ		staɹkast		
Danish	2	vɛmˀ	a	b̥m̩	d̥	vɑ	d̥n̩	sd̥æʌ̯ɡ̊əsd̥ə		

Table 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 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.

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

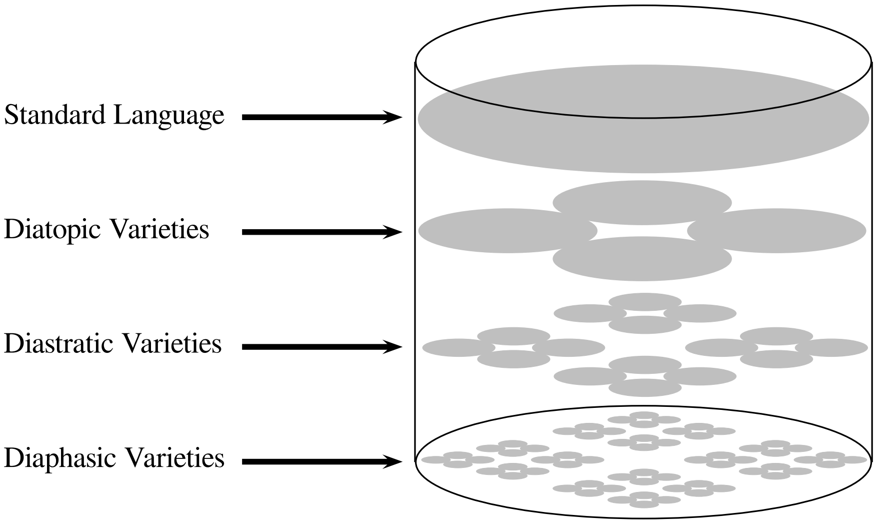


Figure 1: Language as a diasystem

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

What is a Linguistic Sign?

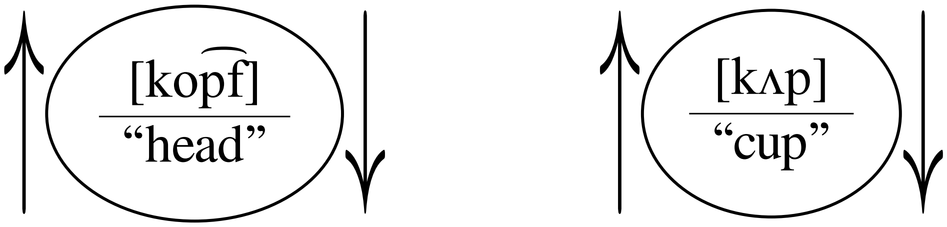
In historical linguistics, linguistic signs are usually treated in the context of the traditional sign model by Saussure (*Cours de linguistique générale*). As Roman Jakobson notes, we distinguish two sides: the form and the content:

> 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)

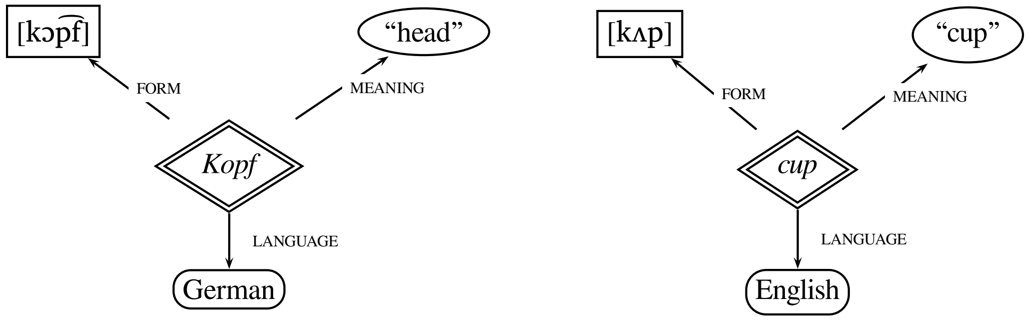
What does Jakobson mean with the words “material” and “intelligible”?

An Extended Sign Model for Comparative Linguistics

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	

How do we Compare Languages?

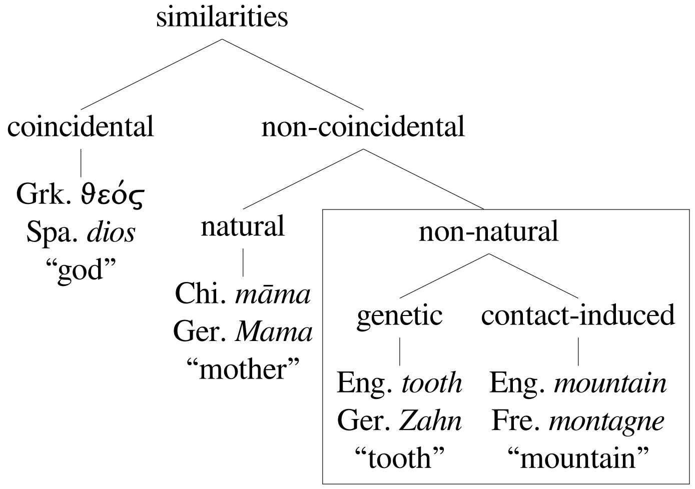
In a very simple model, we can say that a language consists of a certain number of words (or linguistic signs, as we have seen before) and a certain number of syntactic rules by which these words can be combined to form phrases. In spoken languages, the words themselves are formed from a fixed number of sounds which can be combined according to a fixed number of phonotactic rules.

While this model of language as a bag of words may seem very simple, it is effectively the model that was underlying most of the quantitative comparative analyses that have been published so far. Additionally one should say, that even classical linguists who do not work in a quantitative framework tend to use this model in their analyses.

When comparing languages, we need to identify a *tertium comparationis*, that is, we need to find aspects according to which we compare languages. Similar to comparing two objects, for example, two bicycles, we will try to break down the comparison to certain *features*, such as the wheels of our bikes, or their saddle. By comparing the characteristics of these features, e.g., the size of the wheels, or their thickness, we can then start to draw certain conclusions.

As a very simple conclusion, we could try to determine if the bikes are from the same brand. But we can also ask, whether they have been built for the same purpose, or whether they are used in similar environments. These three factors do not need to coincide, and one may need to be an expert in bike construction to learn more about it, but whenever we compare objects with each other, we essentially (1) identify certain similarities based on certain *comparative concepts* (Haspelmath 2010) which serve as the basis of our comparison, and we can then (2) seek explanations for the similarities between the objects.

When only considering similarities between words, we can see four different kinds of similarities presented in the following figure (based on List 2014). How do these similarities relate to our bicycle example, and how do they relate to comparative linguistics and its sub-disciplines?



2 Historical Linguistics

Objective

One of the core objectives of investigating languages from a historical viewpoint is to find out how they *evolved* into their current shape. Similarities of interest for historical linguistics are therefore always those similarities that can be shown to be a result of common ancestry. Since language change goes peculiar pathways, it may not always be easy to find a proper *tertium comparationis* in historical linguistics. What surfaces as an article in one language may well go back to an older demonstrative and surface as a copula in another language. For this reason, the primary focus of historical linguists in identifying historical similarities between languages is not the function or the meaning of a given word or morpheme in a given language, but the sounds from which these are built. Although sounds also change their shape, it has been convincingly shown that they do so in a rather systematical manner. Therefore, when finding the patterns underlying the correspondences of sounds across different languages, it is often rather easy to determine if the languages are historically related and how closely.

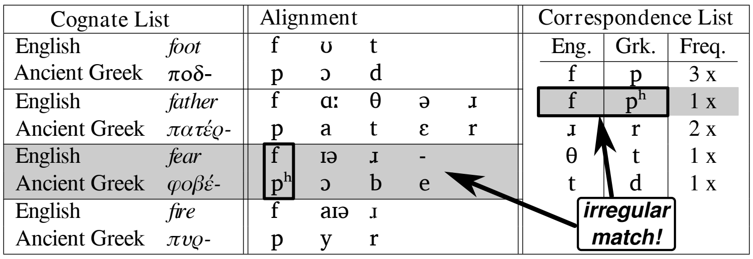
The description of objectives given above does not provide any further information on the areas where historical linguists investigate language evolution. Which ones are probably the most important areas (or aspects of language) in which historical linguists investigate how change proceeds?

Methods

The apparently most important method employed in historical linguistics is the so-called *comparative method*. The comparative method is an overarching framework that historical linguists use to study language history. The application of the framework is tedious, involving many iterative steps. Scholars start by comparing words from different languages in order to identify sets of potentially related words (*cognates*). They then set up lists of sound correspondences and use this information to revise their initial list of cognates (see Table 3). This new information is again used to revise the list of corresponding segments, and so on, until the results can no longer be refined. By applying this method to two or more languages, linguists assemble *cognate words* and *correspondence patterns*, which are then used to infer change scenarios that explain the different correspondence patterns by invoking an ancestral language from which the sounds in the descendant languages (the reflex sounds) can be derived in the most convincing fashion.

Apart from the comparative method, historical linguists have developed and are developing additional methods to handle different topics, such as, for example, semantic change (which we will discuss in Session 3), but also the topic of *phylogenetic reconstruction* enjoys some prominence, although some scholars subsume the classical, non-computational techniques under the framework of the comparative method itself.

The table below gives an example with respect to the detection of sound correspondences between English and Ancient Greek. How can the principle be handled for more than one language?



Cognate List		Alignment					Correspondence List		
English	*foot*	f	ʊ	t			Eng.	Grk.	Freq.
Ancient Greek	*ποδ-*	p	ɔ	d			f	p	3 x
English	*father*	f	ɑː	θ	ə	ɹ	f	p^h	1 x
Ancient Greek	*πατέϱ-*	p	a	t	ɛ	r	ɹ	r	2 x
English	*fear*	f	ɪə	ɹ	-		θ	t	1 x
Ancient Greek	*φοβέ-*	p^h	ɔ	b	e		t	d	1 x
English	*fire*	f	aɪə	ɹ					
Ancient Greek	*πυϱ-*	p	y	r					

irregular match!

Detecting regular sound correspondences in classical historical language comparison.

Models

Scholars like Jacob Grimm had a rather fuzzy understanding of the historical relatedness of languages, and many scholars kept thinking that contemporary languages could be directly "derived" from each other. This changed in the mid of the 19th century, when scholars started to take the idea that languages seem to evolve in tree-like patterns more seriously. While this idea had been around for some time before the advent of "modern" historical linguistics (List et al. 2016), it was not until scholars like August Schleicher (1821-1868) started to propagate the idea not only in words, but also in illustrations (Schleicher 1853, Schleicher 1861), that the family tree model of language history was accepted as something useful to discuss in historical linguistics.

By now, the family tree can be seen as one of the most influential models in historical linguistics. Although it has been challenged, language evolution can hardly be studied without it. The same cannot be said about models for sound change or semantic change. While these models exist, they are much more detailed and specific and rarely gain such a huge acceptance as the tree model of language diversification.

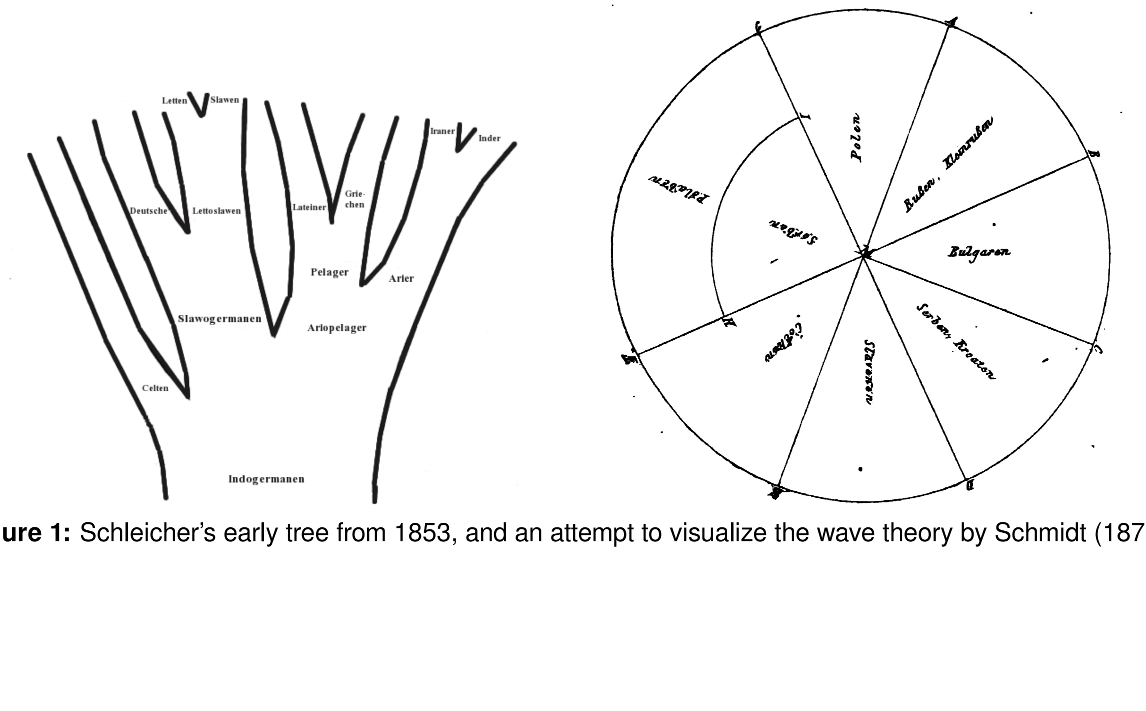


Figure 1: Schleicher's early tree from 1853, and an attempt to visualize the wave theory by Schmidt (1875).

If you compare Schleicher's early tree drawing from 1853 with modern phylogenetic trees, they will look quite different, in terms of abstraction. What could this reflect about the thoughts of the authors?

3 Linguistic Typology

Objective

While historical linguistics deals with the development of particular languages or language families, linguistic typology focuses on those aspects of languages which surface independently of individual language histories. While historical linguistics concentrates on those similarities among languages which are due to change among particular languages, linguistic typology seeks to identify those similarities which have developed independently from a languages' descent. Following our comparison with bicycles, linguistic typology would be interested in the various types of bikes which are being produced (e.g., mountain bikes, road bikes, etc.), while historical linguistics is interested in brands.

At times it appears that linguistic typology deals with synchrony while historical linguistics deals with diachrony. Is this reasonable?

Methods

There are multiple ways of comparing languages, and there is a large number of aspects for which languages can be compared. Given that – unlike historical linguistics – typology deals with more abstract similarities that are not due to common descent, it is more difficult to find suitable *tertia comparationis*, or *comparative concepts*, as they are called by Haspelmath (2010). In typology and in linguistics in general, there is a rather heated debate about the nature of the comparative concepts that linguists define and select in order to compare different languages with each other. A concept like *case*, for example, can be interpreted in multiple ways, and it is not always clear how case should be understood. The confusion also arises from tradition. The Latin *ablative* case, for example, is not a true ablative in the original sense of the word, denoting a case that indicates the starting point of a departure, answering the question "from where", as it is still the predominant usage of the ablative case in Sanskrit. Instead, the Latin ablative shares many properties with the Russian *instrumental* case, which itself is not a true instrumental anymore, as it is again used to express many additional functions that are not predominantly related to the instrumental use of a given object, answering the question "with what?". When starting from the semantics, on the other hand, for example from the questions which are taught in school times in order to deal with case in inflecting languages like Latin, it is clear that languages use different strategies to encode the relevant information, and some could belong to some general grammatical notion of *case*, while other strategies are also available and actively used by many of the world's languages.

But the debate goes beyond pure terminology, since typologists often do not agree with respect to the reality behind the comparative concepts they use. Some linguists say they reflect (or should reflect) some deep innate properties that might find their direct reflection in our brains, some say they are mere tools for comparison, which may be practically defined, but do not need to have a clear relation to any deeper reality, and some scholars take an intermediate position, emphasizing that some of the concepts by which linguists compare languages are useless, but that there should be some deeper value to them. Haspelmath (2018), for example, emphasizes that there is a crucial distinction between language-specific categories, such as the *ablative* in Latin, and cross-linguistic comparative concepts, but that linguists often confuse the two, since they wrongly assume that linguistic categories would have a direct manifestation similar to the idea of *natural kinds* in physics and chemistry. Bond (2019) and other proponents of *Canonical Typology*, on the other hand, argue that cross-linguistic comparison can be carried out by relying on the notion of a *canon*, that is, a "logically motivated archetype from which attested and unattested patterns are calibrated" (ibid.: 83).

No matter how typologists motivate their comparative concepts in the end, it seems clear that the techniques which have been developed to compare languages typologically have greatly improved during the last decades and centuries. As a result, language comparison is nowadays much less biased towards classical European languages and Sanskrit than it was before.

Why does semantics play such an important role in typological language comparison?

Models

While historical linguistics has a standard model of language evolution, we do not find comparable standard models of language typology in the field of linguistic typology. The reason for a lack of unified models is that it is extremely likely that there is no unique reasons for similarity across languages which are not due to contact or common descent, but rather an interaction of multiple factors. Common factors mentioned and investigated by linguists include (1) efficiency of coding (Nettle 1995), (2) climate (Everett et al. 2015), (3) population size (Bromham et al. 2015), or (4) social structure (Lupyan and Dale 2010).

Judging from the short list of only four factors mentioned here, why is it clear that these are not necessarily competing models of linguistic typology?

4 Areal Linguistics

Objective

While languages can be similar due to common descent or due to general properties that all human languages share, there is a third non-trivial reason why languages can exhibit similarities: language contact. In contact situations, when there is a sufficient number of bilingual speakers, not only words but also structures can be easily transferred from one language to another. To identify which material can be transferred during contact, and under which circumstances and with which dynamics language contact occurs can be seen as the primary objective of *areal linguistics*.

In the bicycle example above, it was mentioned that bikes can be similar when they are used in similar environments. Does this reflect a situation similar to language contact?

Methods

We have already seen that it is rather difficult to say exactly what the methods are which are used in linguistic typology, which is why we looked at the selection of comparanda, or comparative concepts, rather than discussing specific methodological frameworks. In areal linguistics, we have similar problems, since it is difficult to identify a unified methodological framework. Instead, scholars use different shortcuts in order to distinguish borrowed from non-borrowed traits (see the short overview in List 2019).

Could the above-mentioned comparative method be used for lexical comparison in the realm of areal linguistics?

Models

At times, scholars contrast the model of a family tree in historical linguistics with the wave model in areal linguistics. The major idea is that innovations, that is, novel ways of speaking, can expand across dialect continua and contact areas in form of waves that may not reach all corners of a given area. What a wave cannot model that well, however, is the direction of influence, and specifically in those cases where we can find many borrowings between languages in well-known contact areas, such as South-East Asia, we find that languages do not influence each other mutually, but that often one language may exhibit more influence over another language. Here, a model of a directed network seems to be much more useful to model contact phenomena.

What is a directed network?

References

Barbour, S. and P. Stevenson (1998). *Variation im Deutschen. Soziolinguistische Perspektiven*. Berlin: de Gruyter.

Bond, O. (2019). “Canonical Typology.” In: *The Oxford handbook of morphological theory*. Ed. by J. Audring and F. Masini. Oxford: Oxford University Press, 409–431.

Bromham, L., X. Hua, T. G. Fitzpatrick, and S. J. Greenhill (2015). “Rate of language evolution is affected by population size.” *Proceedings of the National academy of Sciences of the United States of America* 112.7, 2097–2102.

Bussmann, H., ed. (1996). *Routledge dictionary of language and linguistics*. Trans. from the German by G. Trauth and K. Kazzazi. London and New York: Routledge.

Coseriu, E. (1973). *Sincronía, diacronia e historia. El problema del cambio lingüístico* [Synchrony, diachrony, and history. The problem of linguistic change]. Madrid: Biblioteca Románica Hispánica.

Everett, C., D. E. Blasi, and S. G. Roberts (2015). “Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots.” *Proceedings of the National Academy of Sciences of the United States of America* 112.5, 1322–1327.

Goossens, J. (1973). *Niederdeutsch. Sprache und Literatur. Eine Einführung*. Neumünster: Karl Wachholtz.

Haspelmath, M. (2010). “Comparative concepts and descriptive categories.” *Language* 86.3, 663–687.

— (2018). In: *Aspects of linguistic variation*. Ed. by D. V. Olmen, T. Mortelmans, and F. Brisard. Berlin and New York: De Gruyter Mouton, 83–113.

Jakobson, R. (1976 [1978]). *Six lectures on sound and meaning*. Trans. from the French by J. Mepham. With an intro. by C. Lévi-Strauss. Cambridge and London: MIT Press.

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

— (2019). “Automated methods for the investigation of language contact situations, with a focus on lexical borrowing.” *Language and Linguistics Compass* 13.e12355, 1–16.

List, J.-M., J. S. Pathmanathan, P. Lopez, and E. Bapteste (2016). “Unity and disunity in evolutionary sciences: process-based analogies open common research avenues for biology and linguistics.” *Biology Direct* 11.39, 1–17.

Lupyan, G. and R. Dale (2010). “Language structure is partly determined by social structure.” *PLoS ONE* 5.1, e8559.

Nettle, D. (1995). “Segmental inventory size, word length, and communicative efficiency.”

Oesterreicher, W. (2001). “Historizität, Sprachvariation, Sprachverschiedenheit, Sprachwandel.” In: *Language typology and language universals. An international handbook*. Ed. by M. Haspelmath. Berlin and New York: Walter de Gruyter, 1554–1595.

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.

— (1861). *Compendium der vergleichenden Grammatik der indogermanischen Sprache*. Vol. 1: *Kurzer Abriss einer Lautlehre der indogermanischen Ursprache*. Weimar: Böhlau.

Schmidt, J. (1875). *Zur Geschichte des indogermanischen Vocalismus. Zweite Abteilung*. Weimar: Hermann Böhlau.

Scientific Problem Solving

Johann-Mattis List (University of Passau)

1 Open Problems

Problems

When working every day on very detailed scientific problems, one always runs danger of loosing track of the broader challenges of one's field. That these challenges exist, and that we often still lack sufficient answers to certain problems becomes specifically clear when listening to the questions which laypeople or scientists from other fields ask with respect to one's area of expertise. In linguistics, for example, people are usually very surprised that the question of how language evolved the first time, the question regarding the origin of language, has been officially banned from the agenda of linguistics already in the 19th century, in the often-quoted *statuts* of the *Société de Linguistique de Paris*:

> La Société n'adment aucune communication concernant, soit l'origine du langage, soit la création d'une langue universelle. ("Statuts" 1871: III)

That there are in fact good reasons to avoid these questions becomes obvious when having a look at the large amount of speculative accounts on the origin of language, ranging from Herder's 1778 onomatopoetic speculation of early human beings running through the woods and imitating the sounds of the things surrounding them, or to recent mystic accounts, which have so far been ignored by a larger public:

> The Proto-Sapiens grammar was so simple that the sporadic references in previous paragraphs have essentially described it. The prime importance of sound symbolism for the people of nature should be noted again before we further detail that the vowel "E" was felt as indicating the 'yin" element, passivity, femininity etc., while "O" indicated the "yang" element, activeness, masculinity etc.; "A" was neutral or spiritual, indicating things conceived by the mind and emotions rather than with the physical senses. (Papakitsos and Kenanidis 2018: 8)

But at times, we may forget that there are valid problems in our field which we do not address, because we focus too much on the hard problems of the mainstream, or on tiny problems for which we know we might never find a sufficient answer. These problems may become evident when talking with laypeople, who may at times simply ask a question that would appear silly for a trained linguist. An example for such a question is the number of words that a language disposes of. While this sounds silly for linguists at the first sight, the question is in fact important for our science in multiple ways. It is important for the field of didactics, where it could help us to provide more efficient lessons on the most important words, it is important for historical linguistics, as it would allow us to measure how many of the words we can actually trace back in history, and it would be important for cognitive research, as it would allow us to assess the amount of information individuals can make use of when speaking.

In a paper on similarities between linguistic and biological evolution, we circumvented the question by giving a simple assessment on the words one needs in order to reach a level of proficiency according to different didactic studies (List 2016). But in the same year, Brysbaert et al. (2016) proposed a way to measure the amount of words that an English speaking person knows:

> Based on an analysis of the literature and a large scale crowdsourcing experiment, we estimate that an average 20-year-old native speaker of American English knows 42,000 lemmas and 4,200 non-transparent multiword expressions, derived from 11,100 word families. (ibid.: 1)

Starostin (1989) argues that every language has about 1000 roots which reflect its ancestry. Does this hold cross-linguistically and how much variation should we expect when comparing the languages of the world?

Hilbert and Hilpert Problems

At the and of the last year, inspired by a discussion I had with students who asked me about the biggest challenges for computational historical linguistics, I decided to sit down and make a short list of tasks that I consider challenging, but of which I think that they could still be solved some time in the nearer or further future. The idea to make such a list of questions is not new to mathematicians, who have their well-known Hilbert Problems, proposed by David Hilbert in 1900 (published in Hilbeert 1902). In linguistics, I first heard about them from Russell Gray, who himself was introduced to this by a talk of the linguist Martin Hilpert, who gave a talk on challenging questions for linguistics in 2014, called “Challenges for 21st century linguistics”. Russell Gray since then has emphasized the importance to propose “Hilbert” questions for the fields of linguistic and cultural evolution, and has also presented his own big challenges in the past.

Due to my methodological background, the problems I identified and assembled are by no means big and in some sense also not necessarily extremely challenging (at least on first sight). Instead, the problems I decided for, when being asked, are problems I would like to see tackled, since I think they could help us to further advance our knowledge indirectly, by giving us the possibility to use the solutions of the problems to then answer deeper question on problems in multilingual computational linguistics. One further aspect of the problems that I selected is that these challenges can all be solved by algorithms or workflows. Even when being “small” in some sense, this does not mean, of course, that these problems are not challenging in the big sense. It also does not automatically mean that they can be solved in the near future. But given that the work in the field of computational and computer-assisted language comparison, progresses steadily, at times even at an impressive paste, I have some trust that these problems will indeed be solvable within the next 5-10 years.

What problems in your discipline do you consider unsolvable?

Ten Open Problems for Multilingual Computational Linguistics

When writing down my ten open problems for multilingual computational linguistics, I announced this in a blog post with the blog *The genealogical world of phylogenetic networks*, edited by David Morrison (`http://phylonetworks.blogspot.com/`), in January, with the plan of discussing each of the problems in detail in monthly blog posts throughout the year. At the end of 2019, ten problems had been discussed, and I later decided to elaborate further on them in order to write a small book. Until now, however, I have not found time to finish it or to make any significant progress in this regard.

The 10 problems, which are listed in Table **??** can be further classified into three different groups, which roughly correspond to three different categories important for research in general, namely *modeling*, *inference*, and *analysis*. This trias, inspired by Dehmer et al. (2011: XVII), follows the general idea that scientific research in the historical disciplines usually starts from some kind of idea we have about our research object (the *model* stage), and based on which we then apply methods to infer the phenomena in our data (the *inference* stage). Having inferred enough examples for the phenomenon, we can then *analyze* it qualitatively or quantitatively (the *analysis* stage) and use this information to update our model.

The first group in my list of problems deals with questions of *inference*, including the *detection of morpheme boundaries* (# 1), the *induction of sound laws* (# 2), the *detection of borrowings* (# 3), and *phonological reconstruction* (# 4). What all these problems have in common is that they deal with inference in the sense described above, in so far as they start from linguistic data in some specific form, and the task is to find specific patterns in the data, which have not been annotated in the data beforehand.

The second group of problems deals with questions of *modeling*, including the *simulation of lexical change*, i.e., the design of consistent models that describe how the lexemes of a language change over time, the *simulation of sound change*, i.e., the simulation of the sound-change process by which sounds in a language change in dependence of the context in which they occur, and *the statistical proof of language relatedness*. While the simulation problems are clear problems of *modeling*, given that a simulation requires a model to be then applied to some artificial or existing datasets, the statistical proof or language relationship is a specific case, since it requires a model of language relatedness in order to test this model against a random model in which languages are thought to be unrelated. While there are numerous attempts in the literature to come up with a convincing statistical model to prove genetic relationship (Baxter and Manaster Ramer 2000, Kassian et al. 2015, Kessler 2001, Mortarino 2009, Ringe 1992), none of the attempts which have been proposed so far deals with lexical comparisons in all their complexity. Either, scholars only compare initial consonants with each other (Kessler 2001, Ringe 1992), or they resort to sound classes (Baxter and Manaster Ramer 2000, Kassian et al. 2015), and even if scholars compute random models for whole alignments of potentially related words (List 2014a), they have the problem of not accounting for the factor of closeness due to borrowing.

The last group of problems all have *typology* in their title, and belong to the class of *analysis* problems, dealing with the analysis of *semantic change*, *semantic promiscuity*, and *sound change*. What is meant by *typology* in this context is a data-driven estimate of the overall cross-linguistic frequency of these phenomena. Since we lack consistent accounts on the general tendencies of these processes and phenomena when excluding areal and genetic factors, the task is simply to come up with a consistent estimate on each of them. While semantic change and sound change are probably self-explaining in this context, the question of semantic promiscuity deserves some more attention. What is essentially meant by this term is the degree to which certain words, due to their original meanings, are re-used or re-cycled in the human lexicon.

While the term *promiscuity* has been used before in other contexts in linguistics, the specific usage of promiscuity to denote what one could also call *semantic productivity* or *concept productivity* was first proposed in List et al. (2016b), where biological and linguistic processes were consistently compared with each other, and semantic promiscuity was identified as a phenomenon similar to *domain promiscuity* in protein evolution in biology, with an explicit analogy being identified between the processes of *word formation* in linguistics and *protein assembly* in biology (ibid.: 5). For further elaborations of the concept of *semantic promiscuity*, compare List (2018) and Schweikhard (2018). Nowadays, I have again changed the terminology and no longer use the term *semantic promiscuity*, but rather *lexical root productivity* (List 2023).

Number	**Problem**	**Class**
1	automatic morpheme segmentation	inference
2	automatic sound law induction	inference
3	automatic borrowing detection	inference
4	automatic phonological reconstruction	inference
5	simulating lexical change	modeling
6	simulating sound change	modeling
7	statistical proof of language relatedness	modeling
8	typology of semantic change	analysis
9	typology of semantic promiscuity	analysis
10	typology of sound change	analysis

Does it seem useful to change one's terminology so often, and what may the rapid change in terminology for the same phenomenon reflect?

2 Computer-Assisted Strategies for Problem Solving

The way in which we carry out multilingual computational linguistics in this course is to follow a computer-assisted paradigm in which we try to design targeted methods that aid humans to do some boring tasks in a very accurate and efficient manner or to help humans to detect patterns in larger datasets. This means that we often have to develop new methods from scratch. In order to address the open problems in our field, some basic strategies for problem solving are helpful and important.

The framework for computer-assisted problem solving which I try to pursue in my own research and which I try to propagate does not neglect the possibility of using machine-learning techniques to tackle specific problems, but it does also not necessarily require that they be used exclusively. We do not naively accept machine learning solutions, but start instead from a careful inspection of the problems we actually want to solve. In many cases, a complex solution involving neural networks or Bayesian inference techniques may actually not be needed, since there are smart heuristics, or even complete solutions that do not require any stochastic component. In the same way in which we would not use a machine learning method to tackle the problem of multiplication, it is futile to have an algorithm searching for sound correspondences without any underlying model of sequence comparison or alignments.

That does not mean that machine learning solutions should be excluded per se, and in fact, many of the algorithms for cognate detection, which scholars call *supervised* or based on *linguistic knowledge*, make use of classical techniques, like random works, in specific stages of their workflow. But the decision when to use a specific technique is usually always based on some explicit reasoning that takes the phenomenon to be investigated into account, as well as the existing qualitative solutions that were developed within the field itself, and actual solutions in computer science or similar disciplines, such as bioinformatics, which are consulted to provide inspiration for possible solutions.

The current strategy, which has been applied to propose automatic solutions for various aspects of historical linguistics (List 2014b, List 2019) starts from a detailed investigation (also in collaboration with experts on the topic) of the existing qualitative solutions to a given problem in historical linguistics. As a second step, we try to describe the task in a clear way, by naming explicitly the input data and the output data we expect from the automatic method. We then try to model the process, while at the same time being prepared to further modify the requirements regarding the input data. The solution for the problem is then sought by looking at neighboring disciplines and topics, specifically graph theory, sequence comparison techniques in computer science and bioinformatics, in order to come up with a solution to the problem.

Does your discipline tend to use computer-based or computer-assisted approaches to tackle the major problems?

3 Modeling, Representation, and Implementation

Modeling and Representation

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

> To explain why the contributions of representations, algorithms, and computations will only rarely manifest themselves in fully independent ways [...], it is important to recognise that in practice, models in the brain and cognitive sciences are typically presented in one of two distinct ways: either as abstract model descriptions, or as implemented models. Abstract model descriptions typically comprise symbolic (i.e. verbal or algebraic) descriptions of the relationships between what are typically quite loosely defined quantities or entities. Accordingly, while abstract models can appear to be more or less "formal", they typically fail to fully specify representations (what exactly will be counted and in which format) and typically fail to fully specify the algorithms that will transform these representations into predictions [...]. It is in fact only when these latter steps are made, and an abstract model is actually implemented, that it can be considered formal in any meaningful sense. (Bröker and Ramscar 2021: 17/25)

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

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

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

Integrated Data Representations

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

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

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

4 Modeling, Inference, and Analysis

Modeling

The models that are used so far in computational historical linguistics are all rather simple. While this may at times be surprising for classical linguists, who have a very complex idea of change process and also very detailed knowledge of the complex range of what is possible in language change, reducing the complexity of models is a necessary step in all scientific research. Rather then trying to establish the most complex models before we start to infer something, we should investigate how far we can go with a simplifying model and where its specific limits lie.

Crucial aspects for the models in multilingual computational linguistics are the concept of *language*, *word* (or linguistic sign), *word form*, and *word meaning*. Higher dimensions relevant for questions of language use, such as the speaker-listener interaction, are usually disregarded in the initial stages of investigation. The most common model for a language ist to treat a given language as a bag of words (or a bag of linguistic signs). Depending on the perspective, one can invoke a set of grammatical rules by which these signs are combined to form sentences. The linguistic sign itself follows the basic idea of Saussure (*Cours de linguistique générale*) with the modification that the sign is not seen as a duplet of *form* and *meaning*, but a triplet of form, meaning, and the *language* to which the sign belongs (List 2014b).

The sign form is usually modeled as a *sequence of sounds*, which implies that we can *segment* each word into a certain number of sounds. The sequences are constructed or constrained by *phonotactic rules*. If needed, one can add an additional layer of segmentation, dependent on the research question (e.g., one could look at a word consisting of morphemes consisting of sound segments, or a word consisting of syllables consisting of sound segments). These *secondary sequence structures* are of a certain importance in modern approaches for sequence comparison (List 2014b, List et al. 2016b), but they are often also deliberately disregarded. While the sign form is best treated as a sequence of sounds, the sign meaning is usually handled as a *network of senses*.

While this model of language as a bag of words may seem very simply, it is effectively the model that was underlying most of the phylogenetic analyses that have been published so far. Additionally one should say, that even classical historical linguists tend to use this model in their analyses. When needed, throughout this course, we will discuss more complex models in due time.

To address the problem that we face a drastic lack of comparability with respect to the data that has been produced in multilingual computational linguistics, the Cross-Linguistic Data Initiative (`https://cldf.clld.org`, Forkel et al. 2018) has published a set of recommendations for unified data standards in diversity linguistics, which are now gaining more and more popularity among scholars. These recommendations build more or less directly on the above-mentioned language model, and the current plan is to expand these further, based on the need and the availability of more complex models. As a very important aspect of standardization, CLDF comes along with *reference catalogs*, which are basically meta-datasets, that offer standards for the handling of languages (Glottolog, `https://glottolog.org`, Hammarström et al. 2018), concepts (Concepticon, `https://concepticon.clld.org`, List et al. 2016a), and sounds in transcription (CLTS, `https://clts.clld.org`, Anderson et al. 2018).

In addition to the modeling of the data, the modeling of the processes, which has been not mentioned here, is of great importance. What models can you think of that would explain, for example, the process of sound change, or the process of lexical change?

Inference

As mentioned before, the inference of dated language phylogenies is by far the most popular of the computational methods proposed so far in the field of computational historical linguistics. Discussing the details of these approaches would, unfortunately, go beyond the scope of this session, but good review literature that provides some basic insights is now readily available (Greenhill 2015). What seems important to mention in this context is that the bag-of-words model mentioned before can be seen as the standard model that is essentially used to search for a language phylogeny. When discussing the simulation of language change in a later session, we will discuss more complex ways to simulate language change, which in theory also allow to handle the interaction between speaker and listener.

Second in popularity are methods for automated sequence comparison, which are very popular in dialectology, where methods for phonetic alignment are used to compute aggregate distances between dialect varieties, based on pronunciation distances derived from pre-selected lists of words (Nerbonne et al. 2011). In addition, methods for phonetic alignments are also used for the task of automated cognate detection, which tries to infer which words in a multi-lingual wordlist go back to the same ancestor. Techniques for automated cognate detection are quite well-developed by now, and have been shown to work surprisingly well, with accuracy scores of up to 90% on shallower language families (List et al. 2017), while the accuracy usually drops to around 60%-70% when dealing with larger datasets (Jäger et al. 2017). Further aspects of inference include automated borrowing detection (Mennecier et al. 2016), the detection of sound correspondences and sound correspondence patterns (List 2019), and also the automated prediction of so far unobserved words (Bodt and List 2019), which is specifically useful to support fieldworkers working on small groups of related languages.

How can automated word prediction be useful for linguistic field work?

Analysis

As it was mentioned briefly before, the distinction between what counts as inference and what counts as analysis are not always easy to draw. Intuitively, analysis should involve g-linguistic questions in the sense discussed in the first session, but it is clear that there is no formal justification for it, and it seems to depend more on the workflow, whether a certain step (such as – for example – phylogenetic inference) is labeled as part of the inference or the analysis step. An example for such a borderline case is the *Database of Cross-Linguistic Colexifications* (CLICS, `https://clics.clld.org`, Rzymski et al. 2020), which offers cross-linguistic accounts on polysemies, which are displayed in form of a network analysis that provides information on the relative cross-linguistic closeness of more than 1500 different concepts, reflected in more than 1000 of the world's languages. The more classical analyses which are usually presented, however, try to test certain theories by analysing the data which has been inferred previously. In these cases, the large-scale cross-linguistic databases, which are increasingly produced, play an important role, as they allow scholars to test their hypotheses on a global scale, allowing them, for example, to test hypotheses regarding the transmission of Creole languages (Blasi et al. 2017), the evolution of syntax (Widmer et al. 2017), or the impact of our diet on evolution of our speech sounds (Blasi et al. 2019).

What hypotheses can be derived from historical linguistics that could be tested with the help of cross-linguitic approaches?

References

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

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

Baxter, W. H. and A. Manaster Ramer (2000). “Beyond lumping and splitting: Probabilistic issues in historical linguistics.” In: *Time depth in historical linguistics*. Ed. by C. Renfrew, A. McMahon, and L. Trask. Cambridge: McDonald Institute for Archaeological Research, 167–188.

Blasi, D. E., S. M. Michaelis, and M. Haspelmath (2017). “Grammars are robustly transmitted even during the emergence of creole languages.” *Nature Human Behaviour* 1, 723–729.

Blasi, D. E., S. Moran, S. R. Moisik, P. Widmer, D. Dediu, and B. Bickel (2019). “Human sound systems are shaped by post-Neolithic changes in bite configuration.” *Science* 363.1192, 1–10.

Bodt, T. A. and J.-M. List (2019). “Testing the predictive strength of the comparative method: An ongoing experiment on unattested words in Western Kho-Bwa languages.” *Papers in Historical Phonology* 4.1, 22–44.

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.

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

Dehmer, M., F. Emmert-Streib, A. Graber, and A. Salvador, eds. and introd. (2011). Weinheim: Wiley-Blackwell.

Forkel, R., J.-M. List, S. J. Greenhill, C. Rzymski, S. Bank, M. Cysouw, H. Hammarström, M. Haspelmath, G. A. Kaiping, and R. D. Gray (2018). “Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics.” *Scientific Data* 5.180205, 1–10.

Greenhill, S. (2015). “Evolution and Language: Phylogenetic Analyses.” In: *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*. Ed. by J. D. Wright. Second Edition. Oxford: Elsevier, 370 –377.

Hammarström, H., R. Forkel, and M. Haspelmath (2018). *Glottolog*. Version 3.3. URL: `http://glottolog.org`.

Herder, J. G. (1778). *Abhandlung über den Ursprung der Sprache, welche den von der königl. Academie der Wissenschaften für das Jahr 1770 gesetzten Preis erhalten hat. Welche den von der Königl. Academie der Wissenschaften für das Jahr 1770 gesetzten Preis erhalten hat*. Berlin: Christian Friedrich Voß. Google Books: `QP4TAAAAQAAJ`.

Hilbeert, D. (1902). “Mathematical problems.” *Bulletin of the New York Mathematical Society* 8.1, 437–479.

Jäger, G., J.-M. List, and P. Sofroniev (2017). “Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists.” In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Long Papers*. “EACL 2017”. Valencia: Association for Computational Linguistics, 1204–1215.

Kassian, A., M. Zhivlov, and G. S. Starostin (2015). “Proto-Indo-European-Uralic comparison from the probabilistic point of view.” *The Journal of Indo-European Studies* 43.3-4, 301–347.

Kessler, B. (2001). *The significance of word lists. Statistical tests for investigating historical connections between languages*. Stanford: CSLI Publications.

List, J.-M. (2014a). “Investigating the impact of sample size on cognate detection.” *Journal of Language Relationship* 11, 91–101.

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

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

— (2018). “Von Wortfamilien und promiskuitiven Wörtern [Of word families and promiscuous words].” *Von Wörtern und Bäumen* 2.10. URL: `https://wub.hypotheses.org/464`.

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

— (2023). *Inference of Partial Colexifications from Multilingual Wordlists*.

List, J.-M., M. Cysouw, and R. Forkel (2016a). “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. Luxembourg: European Language Resources Association (ELRA), 2393–2400.

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

List, J.-M., S. J. Greenhill, and R. D. Gray (2017). “The potential of automatic word comparison for historical linguistics.” *PLOS ONE* 12.1, 1–18.

List, J.-M., P. Lopez, and E. Bapteste (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.

Mennecier, P., J. Nerbonne, E. Heyer, and F. Manni (2016). “A Central Asian language survey.” *Language Dynamics and Change* 6.1, 57–98.

Mortarino, C. (2009). “An improved statistical test for historical linguistics.” *Statistical Methods and Applications* 18.2, 193–204.

Nerbonne, J., R. Colen, C. Gooskens, P. Kleiweg, and T. Leinonen (2011). “Gabmap – A web application for dialectology.” *Dialectologia* Special Issue II, 65–89.

Papakitsos, E. C. and I. K. Kenanidis (2018). “Going to the root: Paving the way to reconstruct the language of homo-sapiens.” *International Linguistics Research* 1.2, 1–16.

Perkel, J. M. (2022). “Six tips for better spreadsheets.” *Nature* 608, 229–230.

Ringe, D. A. (1992). “On calculating the factor of chance in language comparison.” *Transactions of the American Philosophical Society*. New Series 82.1, 1–110. JSTOR: `1006563`.

Rzymski, C. et al. (2020). “The Database of Cross-Linguistic Colexifications, reproducible analysis of cross- linguistic polysemies.” *Scientific Data* 7.13, 1–12. URL: `https://clics.clld.org`.

Saussure, F. de. *Cours de linguistique générale*. Ed. by C. Bally. Lausanne: Payot, 1916; German translation: — . *Grundfragen der allgemeinen Sprachwissenschaft*. Trans. from the French by H. Lommel. 2nd ed. Berlin: Walter de Gruyter & Co., 1967.

Schweikhard, N. E. (2018). “Semantic promiscuity as a factor of productivity in word formation.” *Computer-Assisted Language Comparison in Practice* 1.11.

Starostin, S. A. “Sravnitel’no-istoričeskoe jazykoznanie i leksikostatistika [Comparative-historical linguistics and lexicostatistics].” In: *Lingvističeskaja rekonstrukcija i drevnejšaja istorija Vostoka* [Linguistic reconstruction and the oldest history of the East]. Vol. 1: *Materialy k diskussijam na konferencii* [Materials for the discussion on the conference]. Ed. by S. V. Kullanda, J. D. Longinov, A. J. Militarev, E. J. Nosenko, and V. A. Shnirelḿan. Moscow: Institut Vostokovedenija, 1989, 3–39; English translation: — . “Comparative-historical linguistics and lexicostatistics.” In: *Time depth in historical linguistics*. Trans. from the Russian by I. Peiros. Vol. 1. Papers in the prehistory of languages. Cambridge: McDonald Institute for Archaeological Research, 2000, 223–265.

“Statuts” (1871). “Statuts. Approuvés par décision ministérielle du 8 Mars 1866.” *Bulletin de la Société de Linguistique de Paris* 1, III–IV.

Widmer, M., S. Auderset, J. Nichols, P. Widmer, and B. Bickel (2017). “NP recursion over time: Evidence from Indo-European.” *Language* 93.4, 799–826.

Problems and Solutions Across Disciplines

Johann-Mattis List (University of Passau)

1 Question for the Practice Section

What are the most important solvable but unsolved problems in your discipline or sub-field that have so far not been solved?

What are the most important problems you would like to solve in your own work?

What are the unsolvable problems in your discipline?

Which scientific subfields could help you in addressing the problems in your discipline?

Which problems in your discipline have been solved in the last ten years?

2 Tasks for the Practice Section

Take your favorite three problems and try to design a workflow to solve them.

- Identify data and representations.
- Look into the models.
- Identify the kind of the problem (modeling, inference, analysis).
- Try to find scientific disciplines that might help you with your problems.
- Try to find people in the seminar who might help you with your problem.

Cross-Linguistic Data Formats

Johann-Mattis List (University of Passau)

1 Introduction

Data in Linguistics

Linguistics is a discipline in which data play an important role. The main part of the work of many linguists consists in the inspection of data, in the curation of data, in the analysis of data, or in the correction of data. We need grammatical data to investigate grammatical phenomena. These include, among others, example sentences from larger corpora, usually presented in the form of interlinear-glossed text (Lehmann 2004). We need typological data in order to investigate questions on the structure of the languages in the world. These data are typically larger collections of phenomena extracted from individual grammars. If we want to investigate the lexicon of languages, we need wordlists or dictionaries. If data are not available, one needs to create one's own datasets, for example, by going to the field and searching for informants of a given language variety, or by inspecting secondary sources from which data could be extracted.

Are there any fields of linguistics in which data do not play a role?

Data in Comparative Linguistics

People working in the field of language comparison are traditionally even hungrier for data than people working on one particular language's syntax. When comparing languages, we cannot create the data in our heads through introspection. In order to investigate phenomena like language change, we need to compare different data points on the same or different language varieties, and these data points cannot be generated in our heads, they need to be collected. The process of data collection in the field of comparative linguistics may turn out to be quite tedious. Comparative linguists – specifically those working in traditional paradigms – sift through dictionaries, word lists, historical documents, grammars, they interview informants in order to gain more and fresh data on particular language varieties that are not very well documented, and they normally spend a much larger time of their research on the collection of data than on anything else. The results of studies on comparative linguistics can be shared in multiple forms. Etymological dictionaries, for example, are considered to be the "king's discipline" in historical linguistics, because they allow us to see the development of one particular language or an entire language family. In linguistic typology, the major research output are books devoted to specific specific topics of grammar that can then be investigated in the form of a survey, such as, for example, "number" (Corbett 2004), have for a long time been the major research output. Nowadays, with the advent of larger online collections that can be searched on the internet, another major research output are typological databases, which are typically collected by individuals reading the grammars for particular languages in order to extract certain aspects of information. A famous example for this kind of data is the *World Atlas of Language Structures Online* (Dryer and Haspelmath 2013). The problem of etymological dictionaries is that they are still delivered in the form of a book. Although knowledge has been collected in a systematic manner in order to compile them, the knowledge is no longer available in a systematic form, once the dictionary has been compiled. On the contrary, in order to work with etymological dictionaries, the only way to use them in many cases is to inspect them manually, reading individual entries and digesting their content. While typological databases allow us to search quickly for one specific phenomenon, they often go too far in the way in which the original data has been converted to fit the format of the target database. As a result, it is not always useful to rely on the information blindly, and those who have been working with these databases know very well, that there is often no way around reading the original literature from which these collections have been compiled.

Etymological dictionaries are often based on older literature, which is frequently quoted, remixed, and modified. Where do we also find this attempt to cumulatively bring the knowledge about some topic to perfection?

Data Problems in Comparative Linguistics

There are numerous problems resulting from the way in which data is managed and organized in the field of comparative linguistics. We can distinguish three major problems. The problem of (a) availability, (b) transparency, and (c) comparability.

The lack of availability is very annoying, not in the sense that we have no access to a given article in the form of a scan or a book, but rather because many authors collect data, write articles about them, but then do not share their data officially. It is still not surprising that articles are being published in which new ideas are postulated or new conclusions are being made, but in which scholars do not share the data upon which they base their conclusions openly (Tamburelli and Brasca 2017). The same holds for many grammatical descriptions, in which scholars extract individual sentences from their personally collected private corpus but never reference them sufficiently, nor offer the full corpus. This can be seen from the following quote taken from a review of a handbook on Sino-Tibetan languages.

> It is disappointing that so many among the authors of newly commissioned articles did not cite their data; this failing is particuarly 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 the availability we also face the problem of data *transparency*. As an example, see Bengtson (2017), where the author tries to show the readers that Basque and North Caucasian are related.

(gloss)	Basque	Chechen	Avar	Lak / Dargi	Lezgi	Prot-West-Caucasian	Proto-North-Caucasian
die	*hil	=al-	=al’=	L =ič’a D -ibk’-	q’i-	* ƛ̣ə - / *ƛ̣a-	* =iwƛ̣E
dog	*hor	pħu ‘male dog’	hoy	D χa	χor (Budukh)	*ŁIwa	*χHwĕy-rV-
ear	*be=lař	ler-g		D liħi		*ŁA-	*łĕHi
f re	*śu	ts’e	ts’a	L ts’u D ts’a	ts’ay	*mA=c w̩ a	*c̣ ăyɨ
horn	*a=dař	kur	tł:ar		f ri ‘mane’		PEC *ƛwɨ̆ rV
I	*ni			L na D nu		*q́:IwA ‘to hear; to be heard’	* =ĭq̇Ē

Here, it is incredibly hard to interpret or understand the similarities which the author claims to have detected.

As a last problem, we have the problem of comparability of research data. Here, we often find the situation that scholars do not pay attention to sharing their data in such a form that they could be easily compared with other, often similar data, published in independent studies. It is clear that comparability of data is hard to achieve, but some basic aspects of comparability, like a consistent indication of the origin of data, a unified phonetic transcription, consistent standards in naming language varieties and concepts, all of this is indispensable if we want to contribute with our data to science in general. Comparability is unfortunately mostly ignored in comparative linguistics, although many scholars appreciate large data collections in which data have been made comparable. The lack of comparability also contributes to the increasing problem that studies in comparative linguistics can often not be reproduced.

In which cases would it be justified or even important *not* to share research data?

2 The CLDF Initiative

General Ideas

The *Cross-Linguistic Data Formats*-Initiative (CLDF, Forkel et al. 2018, `http://cldf.clld.org`) has the following goals:

(a) working toward the standardization (and retro-standardization) of cross-linguistic research data,

(b) establishing software APIs that help us to check if data conform to these standards and to make use of the data in one's research, and

(c) providing examples for *best-practice*.

In order to address (a), CLDF proposes to make use of metadata bases (*reference catalogs*) like Glottolog (Hammarström et al. 2021), Concepticon (List et al. 2022b), and CLTS (List et al. 2021a). These metadata collections help scholars to make explicit what kind of data they use (which language varieties, which concepts, which sounds). Their goal is to contribute to increasing the *comparability of research data* in comparative linguistics.

In order to address (b), CLDF provides software packages (typically written in Python) that can be used to access data coded in CLDF (CL Toolkit, `https://pypi.org/project/cltoolkit`, List and Forkel 2021), to convert existing data to CLDF (CLDFBench, `https://pypi.org/project/cldfbench`, Forkel and List 2020), or to check if a given dataset conforms to the standards outlined by CLDF (PyCLDF, `https://pypi.org/project/pycldf`, Forkel et al. 2021b). The software in this contexts makes sure that data are both machine- and human-readable at the same time.

In ordert to accomplish (c), CLDF propogates collections of existing datasets coded in CLDF. These collections can be used and inspected by users interested to present their own data in CLDF. They give concrete examples of problem-handling within the CLDF framework and serve as a practial knowledge base where users can take inspiration for their own work. The by now largest collection of individual CLDF datasets, all prepared with the help of the CLDFBench package is the Lexibank repository, offering more than 100 datasets consisting of CLDF wordlists, covering several thousand of the worlds' languages and several dozens of the world's language families (List et al. 2022a).

What is the advantage of using metadata collections like Glottolog when collecting data transparently?

Technical Aspects

The technical aspects of CLDF can be retrieved from the project website (`http://cldf.clld.org`), where one finds a specification and individual examples of the underlying ontology. Currently, CLDF offers three major datatypes, namely Wordlist, Structure Dataset, and Dictionary. The general format in which tabular data are shared is CSV (comma-separated value) with an additional metadata file in JSON format that explains how the CSV data should be interpreted and which columns are linked with each other, following the W3C recommendations for tabular data and metadata on the web (W3C

Consortium 2015, `https://csvw.org`). The CLDF ontology builds on the *General Ontology for Linguistic Description* (GOLD, Community 2010). The `pycldf` Python package (`https://github.com/glottobank/pycldf`, Forkel et al. 2021b) provides the possibility to read and write CLDF data, and also includes commandline facilities to check of a dataset conforms to the CLDF requirements as well as to convert a CLDF dataset into SQLITE format (a very common format for databases that can be read from normal files). The `CLDFBench` package (Forkel and List 2020), allows to convert data to CLDF in a convenient way, using the commandline and standardized Python code. CLDFBench has been extended with `PyLexibank` (Forkel et al. 2021a), a Python package dedicated to the creation of CLDF Wordlists used for the creation of the Lexibank repository (List et al. 2022a).

Why use tabular formats if you could use TEI or plain XML?

Standards in CLDF

CLDF consists of different modules in which specific standard requirements for certain data types are stored. As of now, there are three main modules (a) Wordlist, (b) Dictionary, and (c) Structure Dataset. Additional examples exist that show how more complex data types can also be represented in CLDF, including interlinear-glossed text (List et al. 2021b), and combined datasets in which a wordlist is accompanied by a structure dataset or in which particular structural datasets, like phoneme inventories are handled in a similar form, which could later on be modeled in their own module (Anderson et al. 2021).

In order to convert one's data to CLDF, the first step is to select the appropriate data model (the module). If no model fits a given requirement, one can also use a Generic module that has minimal basic requirements. Most linguistic data come along in the form of *triples*, consisting of a language (variety), an parameter (the question that a dataset asks), and a value (the answer regarding the question). Thus, if one creates a dataset that asks whether a language has an article or not, one would start from a list of individual language varieties, then ask the question (the parameter) "has article?", and then provide the answer "yes / no / dunno". This triplet structure could in theory be rendered by a simple table, rendering this triple structure.

Language_ID	**Parameter_ID**	**Value**
German	has article?	yes
English	has article?	yes
Chinese	has article?	no

However, since we may want to provide additional information on the languages in our sample, we'd prefer to add an individual table for the languages, where this information is stored. Additionally, we may want to add more information on the parameter (or the collection of multiple parameters), and this information would then also better be stored in a specific parameter table. Finally, if one wants to store the sources (e.g., the grammar from which one has taken the information on the article status) one would want to provide them as well in a separate file.

As a result, a typical Structure Dataset in CLDF can consist of a language table, a parameter table, and a value table, and a list of sources (in BibTeX format) which are linked with each other via identifiers.

The same model can be used – with slight modifications – to account for a word list, where we have again one table for the languages, one table for the parameters (the concepts in this specific case) and one table for the values (the word forms, called form table in CLDF).

Language_ID	Parameter_ID	Form
German	HAND	hant
English	HAND	hænd
Chinese	HAND	ʃɔu^{214}
German	FOOT	fuːs
English	FOOT	hænd
Chinese	FOOT	tsu^{35}

What is the difference between a word list and a dictionary?

References

Anderson, C., T. Tresolid, S. J. Greenhill, R. Forkel, R. Gray, and J.-M. List (2021). *Measuring variation in phoneme inventories.*

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.

Community, G. (2010). *General Ontology for Linguistic Description (GOLD).* Ontology. Department of Linguistics (The LINGUIST List), Indiana University.

Corbett, G. G. (2004). *Number.* Cambridge: Cambridge University Press.

Dryer, M. S. and M. Haspelmath, eds. (2013). *WALS Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Forkel, R., S. J. Greenhill, H.-J. Bibiko, C. Rzymski, T. Tresoldi, and J.-M. List (2021a). *PyLexibank. The python curation library for lexibank [Software Library, Version 2.8.2].* Geneva: Zenodo.

Forkel, R. and J.-M. List (2020). "CLDFBench. Give your Cross-Linguistic data a lift." In: *Proceedings of the Twelfth International Conference on Language Resources and Evaluation.* "LREC 2020" (Marseille). Luxembourg: European Language Resources Association (ELRA), 6997–7004.

Forkel, R., J.-M. List, S. J. Greenhill, C. Rzymski, S. Bank, M. Cysouw, H. Hammarström, M. Haspelmath, G. A. Kaiping, and R. D. Gray (2018). "Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics." *Scientific Data* 5.180205, 1–10.

Forkel, R., C. Rzymski, and S. Bank (2021b). *PyCLDF (Version 1.18.0).* Jena: Max Planck Institute for the Science of Human History.

Hammarström, H., M. Haspelmath, R. Forkel, and S. Bank (2021). *Glottolog. Version 4.4.* Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: `https://glottolog.org`.

Hill, N. W. (2017). "The State of Sino-Tibetan." Review of Thurgood and Lapolla (2017) The Sino-Tibetan Languages. Second Edition. *Archiv Orientální* 85, 305–315.

Lehmann, C. (2004). "Interlinear morphemic glossing." In: *Morphology. An international handbook.* Ed. by G. E. Booij, C. Lehmann, J. Mugdan, and S. Skopeteas. Vol. 2. Berlin and New York: De Gruyter, 1834–1857.

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

List, J.-M. and R. Forkel (2021). *CL Toolkit. A Python Library for the Processing of Cross-Linguistic Data [Software Library, Version 0.1.1].* Geneva: Zenodo.

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

List, J.-M., N. A. Sims, and R. Forkel (2021b). "Towards a sustainable handling of interlinear-glossed text in language documentation." *ACM Transactions on Asian and Low-Resource Language Information Processing* 20.2, 1–15.

List, J.-M., A. Tjuka, C. Rzymski, S. J. Greenhill, and R. Forkel (2022b). *CLLD Concepticon [Dataset, Version 3.0.0].* Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: `https://concepticon.clld.org/`.

Tamburelli, M. and L. Brasca (2017). "Revisiting the classification of Gallo-Italic: a dialectometric approach." *Digital Scholarship in the Humanities* fqx41.

W3C Consortium (12/17/2015). *Model for Tabular Data and Metadata on the Web.* W3C Recommendation. W3C.

Reference Catalogs

Johann-Mattis List (University of Passau)

1 Background

Reference catalogs – in the sense in which we use them in CLDF – are understood as bigger metadata collections that are curated by a team of dedicated scholars independently of individual data collections. These catalogs offer definitions of common linguistic constructs (in the sense in which the term *construct* is used in psychology, see Cronbach and Meehl 1955) or *comparative concepts* in the sense of Haspelmath (2010). The structure of these metadata collections is crucially dependent on their general nature, and it is therefore not possible to provide a definition of this structure beforehand. However, what is possible is to say that reference catalogs tend to provide an *identifier* for a given construct, such as, a given *language* or a given *concept* or a given *speech sound*, along with (a) datasets that might make use of this construct, and (b) references that might define this construct. As a result, a reference catalog links an identifier to additional resources, which may refer to some literature references (typically modeled in BibTeX) or to some additional datasets (which could also be referred to with the help of URLs or DOIs). Major reference catalogs used in CLDF are (a) Glottolog, the reference catalog for language identifiers (Hammarström et al. 2021), (b) Concepticon, a reference catalog for concepts (List et al. 2022b), and (c) CLTS, a reference catalog for speech sounds. The major advantage of these reference catalogs is that they outsource the business of providing a consistent standard. Linguists making use of them do no longer need to define the constructs themselves, instead, they can link – where available – to the reference catalogs which take care of “the rest”, by providing additional resources and by also taking the blame if the information they offer is wrong.

Why is it important to model concepts as constructs in linguistic research?

2 Glottolog

Glottolog (`https://glottolog.com`, Hammarström et al. 2021) is a reference catalog for language varieties and offers not only the identifiers for more than 7000 language varieties, but also an extensive bibliography that characterizes these language varieties. In this form, Glottolog is an excellent starting point for those who want to learn more about a particular language variety, since alone the bibliography delivers almost exhaustively all information that is available for individual languages.

In addition, Glottolog offers a preliminary classification of the language varieties in the form of language trees. This classification is close to the communis opinio in the field, but given that there are many different opinions here, no phylogeny can ever be perfect, and one should rather use the phylogeny offered by Glottolog as a convenient reference, rather than ground truth.

Glottolog also offers geolocations for most of the varieties the catalog contains. This is extremely convenient, since it means one can plot languages easily on a map, when having obtained their *Glottocodes*, the unique identifiers offered by the reference catalog, consisting of four letters and four numbers, derived based on the following criteria, outlined in Forkel and Hammarström (2022: 918)

- An ID specifically designed for machine readability, not confusable with an informal or human-directed identifier
- An ID type oblivious to level of linguistic abstraction (idiolect, sociolect, dialect, language, subfamily, family, etc.)

- An ID system for languages that improves on the ISO 639-3 language identifiers in terms of quality, transparency and anchoring

In addition, Glottolog can be accessed through a powerful Python API that offers users the possibility to search for language varieties, to extract trees in standardized formats, and to query all information also displayed on the website.

What is the difference between a language and a language variety?

3 Concepticon

Concepticon (List et al. 2016, List et al. 2022b) ist ein Katalog von sogenannten Concept Sets, einer Verlinkung von Questionnaires, wie Swadesh-Listen, etc., die für inzwischen mehr als 2900 Konzepte Definitionen und Links zu existierenden Questionnaires liefert. Das Concepticon ist essentiell für die Aggregierung von Daten, aber auch aus historischer Perspektive interessant. Erhältlich ist das Concepticon unter `http://concepticon.clld.org`. Wir werden uns in einer Sitzung den theoretischen Grundlagen des Concepticons widmen und in einer weiteren Sitzung lernen, wie wir die Software verwenden können.

Kann man Konzepte überhaupt definieren?

Background

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:[1]

> 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?

[1] 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.

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.[2]

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

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).

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

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.

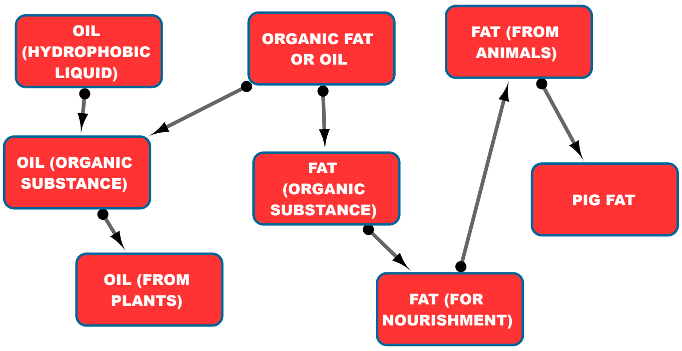


Figure 1: Concept relations between 'oil', and 'fat'

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

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（ 不聡明 ）/ 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?

4 Cross-Linguistic Transcription Systems (CLTS)

CLTS (List et al. (2021), `https://clts.clld.org`) is a reference catalog for speech sounds, offering more than 8000 consistently defined speech sounds, which link to several databases. CLTS is our standard for phonetic transcriptions in CLDF, specifically in the Lexibank collection (List et al. 2022a). It comes along with a rather powerful Python API that allows to conduct several operations with sounds, comparing sounds for their similarity based on distinctive features, or offering strategies to translate the representation of sounds across different transcription systems.

How is it possible that there are more than 8000 different sounds in human languages?

Background

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 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?

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 2018).

Why does the IPA not want to be a standard?

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. 2019)	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?

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?

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) ◌̃ (U+0303)			ã
a (U+0061) : (U+003a)		a (U+0061) ː (U+02d0)		aː
ʦ (U+02a6)		t (U+0074) s (U+0073)		ts

Table 5: Three-step normalization in CLTS.

Sound	Identifier
ã	nasalized unrounded open front vowel
aː	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?

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?

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?

5 Norms, Ratings, and Relations

A very recent reference catalog, called NoRaRe, the Database of Cross-Linguistic Norms, Ratings, and Relations for Words and Concepts (`https://norare.clld.org`, Tjuka et al. 2022) collects conceptual metadata from psycholinguistic datasets for individual words and concepts across different languages. The major idea of the NoRaRe collection was to offer a way to consistently compare conceptual metadata collected in the context of psychology and psycholinguistics, but also in the context of computational linguistics across datasets. As of now, NoRaRe offers data for 113 datasets, from which 713 variables are derived. These variables can often be compared across languages. Thus, one can find frequency information not only for Spanish, but also for English, German, etc. The underlying concepts, for which these variables are defined, are consecutively linked to the Concepticon. As a result, data that links to Concepticon can also make active use of the norms, ratings, and relations offered in NoRaRe.

There is not a clear-cut distinction between words and concepts in the NoRaRe database, where words are often thought of as representing individual concepts. Is this handling of the data a problem, or can it be justified in some way?

6 Future Reference Catalogs

More reference catalogs may be produced in the future. Since the creation of reference catalogs is tedious, however, it is hard to tell what reference catalog will come next. Candidates are a reference catalog for the glosses used in interlinear-glossed text (e.g., in the form of a Grammaticon), or senses that are used to define morphemes in words (some kind of a Morphemicon), or an extended collection of metadata for individual speech sounds (some Phoneticon).

Why is there not yet a reference catalog for bibliographic entries?

References

Allen, B. (2007). *Bai Dialect Survey*. Dallas: 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.

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).

Chén, B. 陈保亚. (1996). *Lùn yǔyán jiēchù yǔ yǔyán liánméng* 论语言接触与语言联盟 *[Language contact and language unions]* [Language contact and language unions]. Běijīng 北京: Yǔwén □□.

Chomsky, N. and M. Halle (1968). *The sound pattern of English*. New York, Evanston, and London: Harper and Row.

Cronbach, L. J. and P. E. Meehl (1955). “Construct validity in psychological tests.” *Psychological Bulletin* 52, 281–302.

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

Forkel, R. and H. Hammarström (2022). “Glottocodes: Identifiers linking families, languages and dialects to comprehensive reference information.” *Semantic Web* 13.6. Ed. by J. Bosque-Gil, M. Dojchinovski, P. Cimiano, J. Bosque-Gil, P. Cimiano, and M. Dojchinovski, 917–924.

Hammarström, H., M. Haspelmath, R. Forkel, and S. Bank (2021). *Glottolog. Version 4.4*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: `https://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.

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

List, J.-M., 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. Luxembourg: European Language Resources Association (ELRA), 2393–2400.

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

List, J.-M., A. Tjuka, C. Rzymski, S. J. Greenhill, and R. Forkel (2022b). *CLLD Concepticon [Dataset, Version 3.0.0]*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: `https://concepticon.clld.org/`.

Matisoff, J. A. (1978). *Variational semantics in Tibeto-Burman. The ”organic” approach to linguistic comparison*. Philadelphia: Institute for the Study of Human Issues.

Mielke, J. (2008). *The emergence of distinctive features*. Oxford: Oxford University Press.

Moran, S. and M. Cysouw (2018). *The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles*. Berlin: Language Science Press.

Moran, S. and D. McCloy, eds. (2019). *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History.

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*.

Ruhlen, M. (2008). *A global linguistic database*. Moscow: RGGU.

Sovijärvi, A. and R. Peltola (1970). *Suomalais-Ulgrilainen Tarkekirjoitus* UralicPhoneticAlphabet. 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 proischoždenije 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.

— (1955). “Towards greater accuracy in lexicostatistic dating.” *International Journal of American Linguistics* 21.2, 121–137. JSTOR: `1263939`.

Tjuka, A., R. Forkel, and J.-M. List (2022). “Linking norms, ratings, and relations of words and concepts across multiple language varieties.” *Behavior Research Methods* 54.2, 864–884.

University, P. (2010). *WordNet. A lexical database for English*. Online Resource. Princeton.

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.

Standardized Data Collections in Multilingual Computational Linguistics

Johann-Mattis List (University of Passau)

1 Background

Since we started the CLDF initiative in 2014, many datasets have been converted to CLDF, including word lists, structure datasets, and dictionaries. It is difficult to give a concrete number on the individual datasets that have been created, but it is quite likely that they exceed 200 or even 300 now. In order to increase the *findability* of CLDF datasets in the sense of the F in FAIR data (Wilkinson et al. 2016), we started to curate collections of individual CLDF datasets that we have released with Zenodo. The most prominent collection here is *Lexibank* (`https://zenodo.org/communities/lexibank`), offering standardized word lists. Another collection (so far without a Zenodo community) is the collection *CLDF Datasets* (`https://github.com/cldf-datasets`) which offers various kinds of data that are not primarily lexical, including phoneme inventories like Phoible (Moran and McCloy 2019) or the World Atlas of Language Structures Online (Dryer and Haspelmath 2013). A larger collection of digital dictionaries in CLDF is offered by Dictionaria (`https://dictionaria.clld.org`), and a larger collection of wordlists with numeral systems from the worlds' languages is offered by Numeralbank (`https://github.com/numeralbank/`). Additionally, certain types of legacy data which are often no longer expanded, have been given their own CLDF collection, including the still slightly growing Intercontinental Dictionary Series (`https://github.com/intercontinental-dictionary-series`, (Key and Comrie 2016)), or the datasets discussed in List (2014), which are now accessible in individual CLDF datasets (`https://github.com/sequenceComparison/`).

Why did it take such a long time to publish the first version of the Lexibank database?

2 Lexibank

Background

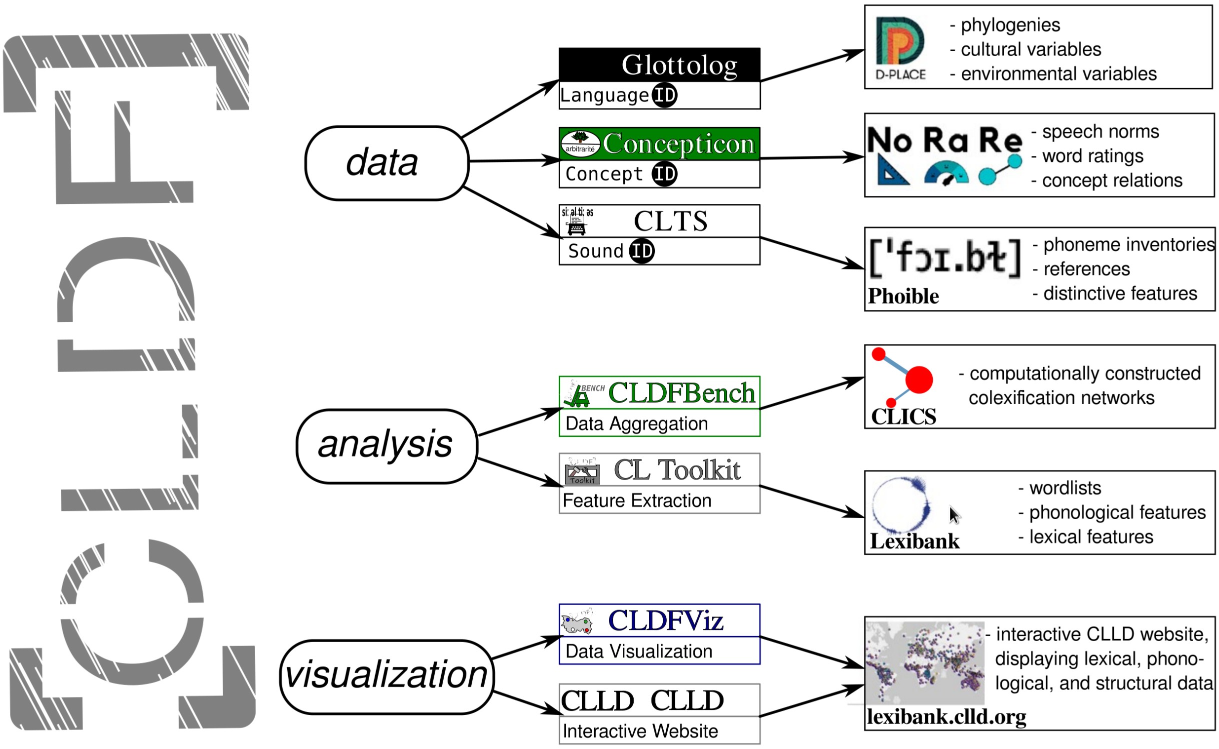
Lexibank is the largest collection of standardized wordlists in CLDF. Initiated in 2014, data collection reached its peak in 2018, after major components of the specific standards, like Concepticon (List et al. 2016) and CLTS (Anderson et al. 2018) had been published. Earliert test versions of Lexibank were published as part of the Database of Cross-Linguistic Colexifications (`https://clics.clld.org`), in 2018 (List et al. 2018) and in 2020 (Rzymski et al. 2020), which we will present in a separate session in more detail. In 2022, the first official version of Lexibank (Version 0.2) was published (List et al. 2022a). It consists of an aggregated CLDF dataset, in which data from 100 different CLDF datasets was aggregated and then consecutively analyzed, searching automatically for phonological and lexical features.

> Lexibank is a meta-collection of standardized wordlists compiled from various individual datasets. The standardized wordlists themselves are independently curated. Their curation fol- lows the data curation workflow of the Lexibank project, which uses the PyLexibank Python library (Forkel et al. 2021) to convert lexical data in custom formats into CLDF wordlists. The editorial board of the Lexibank project decides about the inclusion of individual datasets into the Lexibank wordlist collection. Datasets which are included in this collection need to be archived with Zenodo (`https://zenodo.org/`) and curated in a GIT repository (`https://github.com/`). Datasets included into the Lexibank wordlist collection are referenced with their Zenodo DOI and the URL of their GIT repository and classified for their level of standardization. (List et al. 2022a: 5/16)

What advantage has the automated search done in the context of Lexibank compared to a good traditional manual search in the grammatical literature?

Data Curation in Lexibank

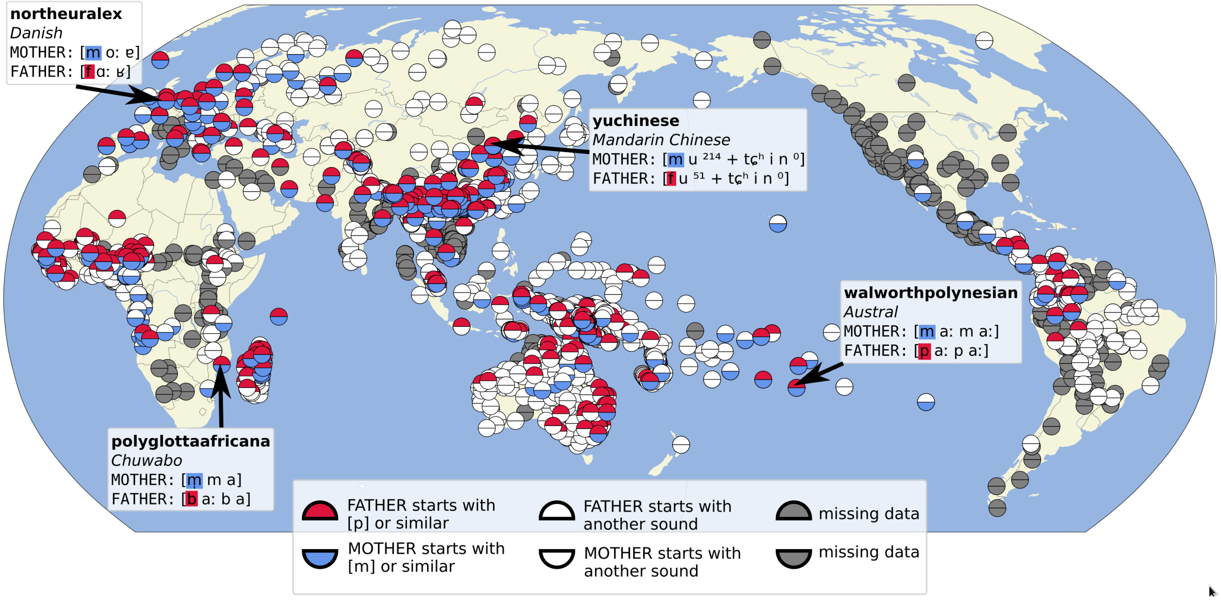
The curation process of Lexibank data makes use of the CLDFBench package (Forkel and List 2020) that was extended by the PyLexibank plugin (Forkel et al. 2021). The major idea is to standardize the process by which an individual dataset is converted to CLDF as well as possible. This means that we start from the raw data, which may be manually adjusted, and then try to parse the data in order to read it into tables. Having done this, we convert the tabular data to CLDF, providing additional information on the concepts (which we manually or semi-automatically link to Concepticon), on the languages (which we manually link to Glottolog) and the phonetic transcriptions, which we segment and normalize at the same time by creating an orthography profile (Moran and Cysouw 2018) and applying it to the original transcriptions. The transcriptions can themselves be preprocessed with the help of code for the handling of lexical entries provided by PyLexibank.



Reference catalogs have the advantage to allow us to access data from other databases that has been collected for different purposes, such as, for example, cultural data for language varieties or norm data for concepts. What research questions could one investigate with the extended access to cultural data and norm data from psychology?

Automated Feature Extraction

Once an aggregated word list is available, we can extract various phonological and lexical features automatically from the data. As an example, consider the feature observed by Jakobson (1960), who discussed the observation made by many people and linguists, that many languages have a word form “mother” that starts with *m-* and often sounds like *mama* and a word for “father” that often starts with *p-* or *f-* and sounds like *papa*. Since our transcriptions are provided in the form of CLTS (Anderson et al. 2018, List et al. 2021a) and our concepts are linked to Concepticon (List et al. 2022b), we can easily formulate a query, in which we state that we search for words that start with an m-like sound for the concept MOTHER and with a p-like sound for the concept FATHER. The resulting data can then be easily plotted on a map with the help of the CLDFViz package (Forkel 2021).



What are the reasons that languages frequently choose *mama* and *papa* as words for “mother” and “father”?

Future of Lexibank

We are currently working on extended functionalities of Lexibank. Our main objective as of now is to look into fast database queries that can be applied to the extended Lexibank data (which will have many more phonetically transcribed language varieties in its upcoming version). The idea is that we directly convert the Lexibank CLDF data into a SQLITE database (or any other database system) and then query the data in order to answer specific questions. Using databases rather than CSV files has the advantage of speed, and as a result, many different queries can be “asked” quickly in order to test and generate hypotheses. Queries could even be asked on a website, given that the current version of all Lexibank data as an SQLITE database does not exceed 250 MB. Queries can output data in various forms. One could create a word list in LingPy’s format (List et al. 2018) or the format required by EDICTOR (List 2017, List 2021). One can also create a CSV file with the values that would be sufficient to plot features of individual languages on a geographic map with the help of CLDFViz. All in all, we hope that queries in this form allow us to establish a Basic Linguistic Search Service (BLISS) that could play a similarly important role as the Basic Local Alignment Search Tool (BLAST) that revolutionized evolutionary biology (Altschul et al. 1990).

What kind of queries could we ask a Lexibank database?

3 CLDF Datasets

Background

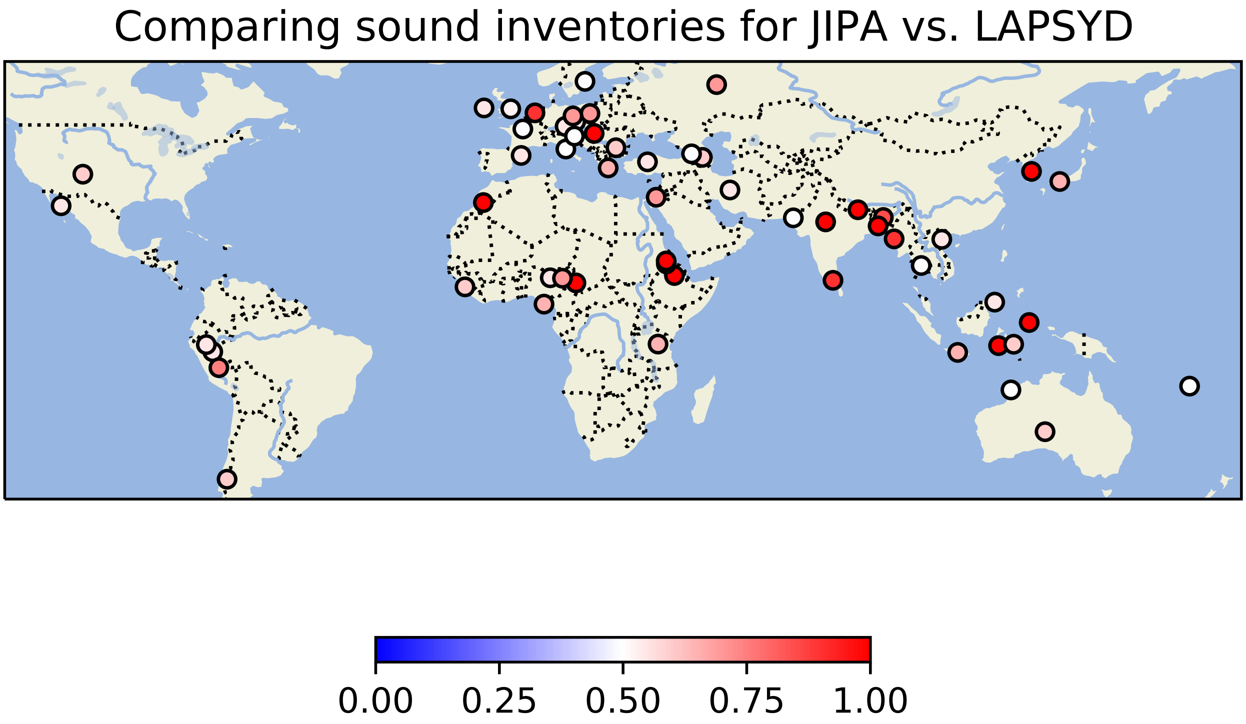
When starting to prepare lexical word lists in CLDF, we quickly realized that there are many other kinds of linguistic data that might also be worthwhile to be standardized. As a result, we began to prepare individual structural datasets in CLDF format, based on personal interests (List 2018). Later, it was decided to start collecting these datasets in a dedicated GitHub organization to not loose track of them. This organization, called CLDF Datasets (`https://github.com/cldf-datasets`) has still fewer repositories than the Lexibank organization, but it contains already 80 different public datasets as of today and is constantly growing. While adding datasets to the CLDF Datasets organization, we hope that we can identify specific sub types of data that might later also be either aggregated or compared.

What other kinds of cross-linguistic data apart from word lists could be similar enough to call for a consistent aggregation in a similar way in which this was done for Lexibank?

Phoneme Inventories

Based on our interest in different transcription systems, as reflected in the work on the Cross-Linguistic Transcription Systems reference catalog, we started to collect dedicated collections of phoneme inventories, trying to standardize them in a similar way in which we use the Concepticon to help to aggregate data from different word lists. While Concepticon offers identifiers for concepts, CLTS offers identifiers for speech sounds, and the idea was that we could use the identifiers in a similar way across datasets and in this way also point to differences in phonetic transcriptions across phoneme inventory collections.

Of the inventory datasets that are consistently linked to CLTS, we currently provide the LAPSyD database (Maddieson et al. 2013), the collection of phoneme inventories published in the Journal of the International Phonetic Association by Baird et al. (2021), and the Eurasian inventories by Nikolaev et al. (2017). Databases like Phoible (Moran and McCloy 2019) are available in CLDF and referenced by CLTS but do not yet include the explicit link to a given version of the CLTS reference catalog themselves. We estimate that there is a potential to link at least 10 more datasets in a similar form to CLTS. Thus, the currently still small collection could further grow in the future and allow for additional possibilities to build on initial studies that compare how well phoneme inventories for the same varieties correspond when collected by different authors (Anderson et al. 2021).

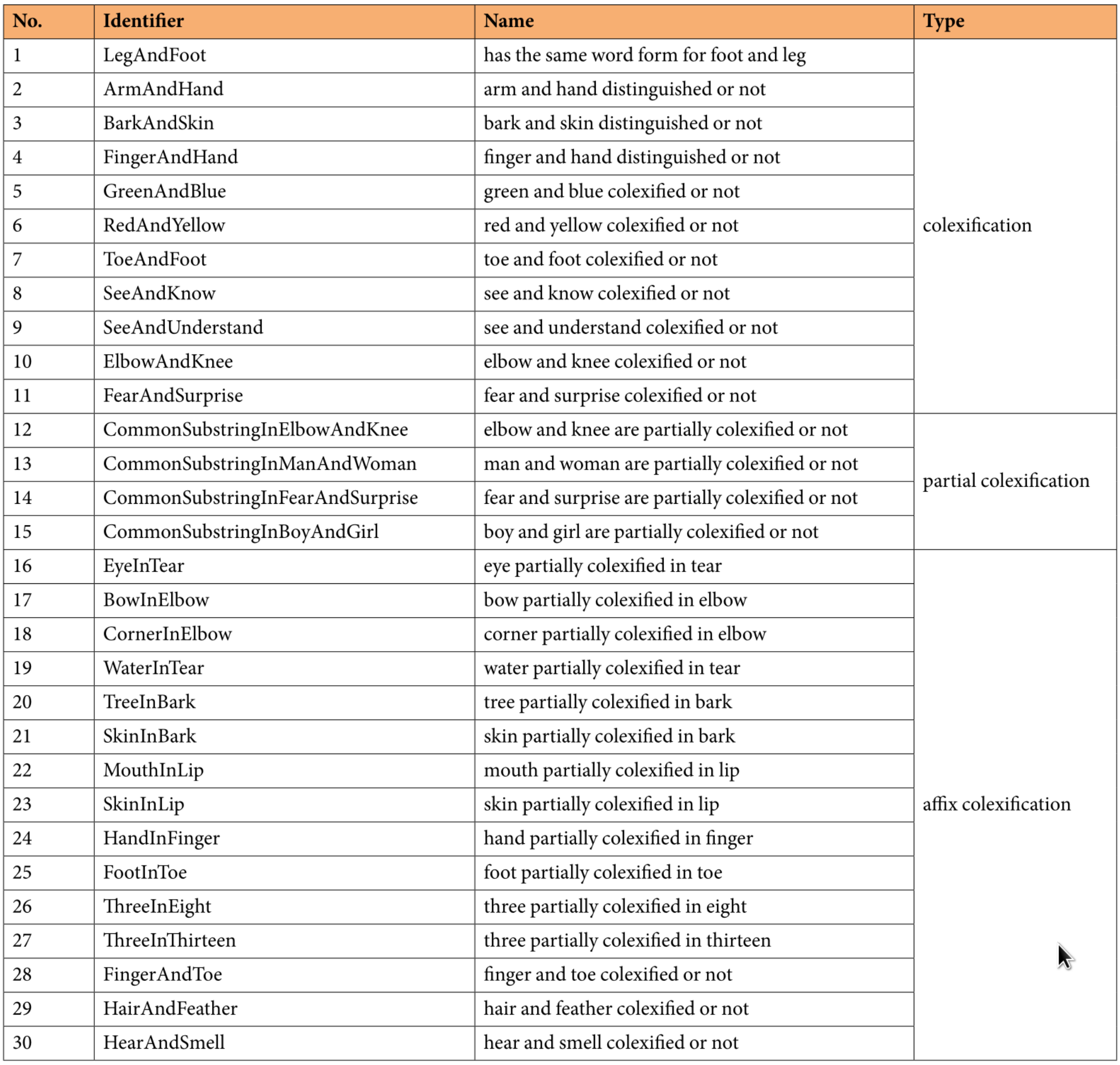


The map above (from Anderson et al. 2021) compares phoneme inventory sizes in the JIPA collection by Baird et al. (2021) and the LAPSyD database by Maddieson et al. (2013). Is there any pattern that can be detected with respect to the differences in the phoneme inventory sizes?

Structure Datasets

Structure datasets are very diverse by their nature, and we find collections of very specific features, such as "The third person pronoun is tā, or cognate to it." for Chinese dialect varieties (Norman 2003, `https://github.com/cldf-datasets/normansinitic`) or "Does the language have morphosyntactic plural markers?" (Tang and Her 2019, `https://github.com/cldf-datasets/tangclassifiers`). As a result, comparing individual datasets that have been collected so far is much more difficult if not impossible, than comparing word lists or phoneme inventory collections. In order to approach the problem, a metadata catalogue of structural properties of languages would be needed, and this catalogue would have to identify those features which frequently recur across the languages in the world.

While such an enterprise has not been undertaken yet, our work on the Lexibank project has initiated a first step into this direction. In automatically computing lexical and phonological features for the data in the Lexibank collection, we take direct inspiration from the features in the World Atlas of Language Structures Online (WALS, Dryer and Haspelmath 2013), and our feature computation workflow also notes similarities among the features we compute and their counterparts in the WALS database. Our idea was to further expand these collections of automatically computed features and to note more clearly and systematically which databases provide features that have been manually collected. In this way, a future reference catalog, albeit a small one to begin with, could well be prepared in the nearer future and also provide concrete information on the computability or the computability status of certain features that have been collected for the world's languages.



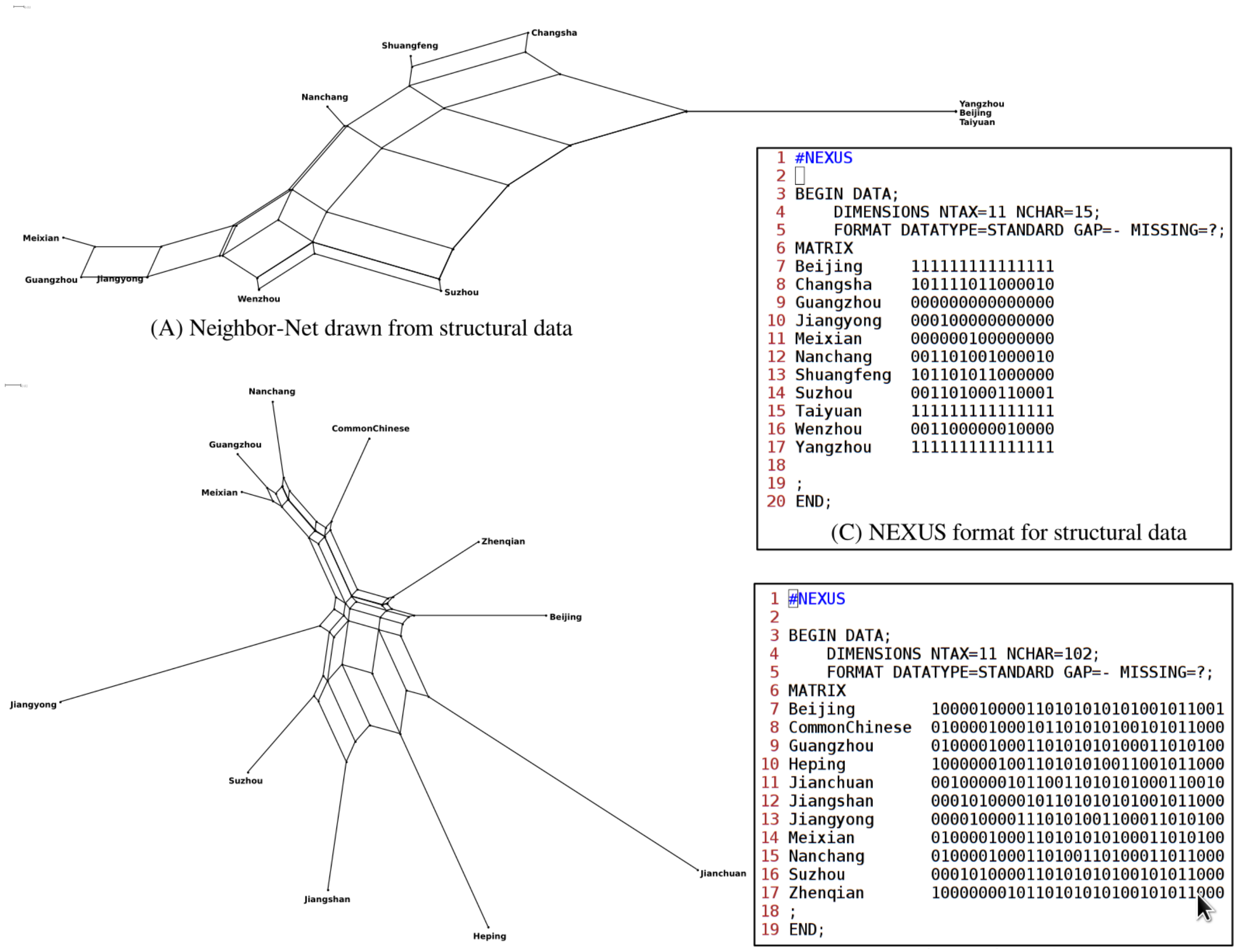
No.	Identifier	Name	Type
1	LegAndFoot	has the same word form for foot and leg	colexification
2	ArmAndHand	arm and hand distinguished or not	
3	BarkAndSkin	bark and skin distinguished or not	
4	FingerAndHand	finger and hand distinguished or not	
5	GreenAndBlue	green and blue colexified or not	
6	RedAndYellow	red and yellow colexified or not	
7	ToeAndFoot	toe and foot colexified or not	
8	SeeAndKnow	see and know colexified or not	
9	SeeAndUnderstand	see and understand colexified or not	
10	ElbowAndKnee	elbow and knee colexified or not	
11	FearAndSurprise	fear and surprise colexified or not	
12	CommonSubstringInElbowAndKnee	elbow and knee are partially colexified or not	partial colexification
13	CommonSubstringInManAndWoman	man and woman are partially colexified or not	
14	CommonSubstringInFearAndSurprise	fear and surprise are partially colexified or not	
15	CommonSubstringInBoyAndGirl	boy and girl are partially colexified or not	
16	EyeInTear	eye partially colexified in tear	affix colexification
17	BowInElbow	bow partially colexified in elbow	
18	CornerInElbow	corner partially colexified in elbow	
19	WaterInTear	water partially colexified in tear	
20	TreeInBark	tree partially colexified in bark	
21	SkinInBark	skin partially colexified in bark	
22	MouthInLip	mouth partially colexified in lip	
23	SkinInLip	skin partially colexified in lip	
24	HandInFinger	hand partially colexified in finger	
25	FootInToe	foot partially colexified in toe	
26	ThreeInEight	three partially colexified in eight	
27	ThreeInThirteen	three partially colexified in thirteen	
28	FingerAndToe	finger and toe colexified or not	
29	HairAndFeather	hair and feather colexified or not	
30	HearAndSmell	hear and smell colexified or not	

The table above shows lexical features computed from Lexibank data. What is meant with *partial colexification* and with *affix colexification*?

4 Combined Data

As a final example for our current efforts in working with CLDF and testing the limits of the format specification, we have started to create combined datasets in which we combine, e.g., features with words in a wordlist. Thus, we often find lexical word lists accompanied by phoneme inventories in the literature. When digitized, we can render both datasets in one combined CLDF datasets in which the language table is shared among both datasets.

As of now, we experiment with the combination of structure datasets with word lists, the combination of dictionaries with word lists (where a word list can be derived from a dictionary), and with the modeling of more complex features, such as, for example, colexifications (which we will discuss in an upcoming session). As a most complete way of combining data for a single resource, we hope to integrate interlinear-glossed text with additional resources, such as, for example, phoneme inventories, word lists, and even dictionaries in one single CLDF package. So far, however, we have not found data collections which would offer all these resources in combination (an initial example is discussed in List et al. 2021b). We will discuss how texts and corpus data can be handled in CLDF in an additional session.



(A) Neighbor-Net drawn from structural data

```
#NEXUS

BEGIN DATA;
    DIMENSIONS NTAX=11 NCHAR=15;
    FORMAT DATATYPE=STANDARD GAP=- MISSING=?;
MATRIX
Beijing     111111111111111
Changsha    101111011000010
Guangzhou   000000000000000
Jiangyong   000100000000000
Meixian     000000100000000
Nanchang    001101001000010
Shuangfeng  101101011000000
Suzhou      001101000110001
Taiyuan     111111111111111
Wenzhou     001100000010000
Yangzhou    111111111111111

;
END;
```

(C) NEXUS format for structural data

(B) Neighbor-Net drawn from lexical data

```
#NEXUS

BEGIN DATA;
    DIMENSIONS NTAX=11 NCHAR=102;
    FORMAT DATATYPE=STANDARD GAP=- MISSING=?;
MATRIX
Beijing         1000010000110101010101001011001
CommonChinese   0100001000101101010100101011000
Guangzhou       0100001000110101010100011010100
Heping          1000000100110101010011001011000
Jianchuan       0010000010110011010101000110010
Jiangshan       0001010000101101010101001011000
Jiangyong       0000100001110101001100011010100
Meixian         0100001000110101010100011010100
Nanchang        0100001000110100110100011011000
Suzhou          0001010000110101010100101011000
Zhenqian        1000000010110101010100101011000
;
END;
```

(D) NEXUS format for lexical data (excerpt)

Why are the Neighbor-net representations of the data by Norman (2003), as described in Forkel and List (2020) so different?

References

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman (1990). "Basic local alignment search tool." *Journal of Molecular Biology* 215.3, 403 –410.

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

Anderson, C., T. Tresolid, S. J. Greenhill, R. Forkel, R. Gray, and J.-M. List (2021). *Measuring variation in phoneme inventories*.

Baird, L., N. Evans, and S. J. Greenhill (2021). "Blowing in the wind: Using 'North Wind and the Sun' texts to sample phoneme inventories." *Journal of the International Phonetic Association* 0.0, 1–42.

Dryer, M. S. and M. Haspelmath, eds. (2013). *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Forkel, R. (2021). "CLDFViz. A python library providing tools to visualize data from CLDF datasets [Software Library, Version 0.5.0]."

Forkel, R., S. J. Greenhill, H.-J. Bibiko, C. Rzymski, T. Tresoldi, and J.-M. List (2021). *PyLexibank. The python curation library for lexibank [Software Library, Version 2.8.2]*. Geneva: Zenodo.

Forkel, R. and J.-M. List (2020). "CLDFBench. Give your Cross-Linguistic data a lift." In: *Proceedings of the Twelfth International Conference on Language Resources and Evaluation.* "LREC 2020" (Marseille). Luxembourg: European Language Resources Association (ELRA), 6997–7004.

Jakobson, R. (1960). "Why 'Mama´and 'Papa?." In: *Perspectives in psychological theory: Essays in honor of Heinz Werner*. Ed. by B. Kaplan and S. Wapner. New York: International Universities Press, 124–134.

Key, M. R. and B. Comrie (2016). *The Intercontinental Dictionary Series*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: https://ids.clld.org.

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

— (2017). "A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets." In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*. Valencia: Association for Computational Linguistics, 9–12.

— (09/03/2018). "Representing Structural Data in CLDF." *Computer-Assisted Language Comparison in Practice* 08.08.

— (2021). *EDICTOR. A web-based tool for creating, editing, and publishing etymological datasets. Version 2.0.0*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: https://digling.org/edictor.

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

List, J.-M., 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. Luxembourg: European Language Resources Association (ELRA), 2393–2400.

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

List, J.-M., N. A. Sims, and R. Forkel (2021b). "Towards a sustainable handling of interlinear-glossed text in language documentation." *ACM Transactions on Asian and Low-Resource Language Information Processing* 20.2, 1–15.

List, J.-M., A. Tjuka, C. Rzymski, S. J. Greenhill, and R. Forkel (2022b). *CLLD Concepticon [Dataset, Version 3.0.0]*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: https://concepticon.clld.org/.

List, J.-M., M. Walworth, S. J. Greenhill, T. Tresoldi, and R. Forkel (2018). "Sequence comparison in computational historical linguistics." *Journal of Language Evolution* 3.2, 130–144.

Maddieson, I., S. Flavier, E. Marsico, C. Coupé, and F. Pellegrino. (2013). "LAPSyD: Lyon-Albuquerque Phonological Systems Database." In: *Proceedings of Interspeech*. (Lyon, 08/25–08/29/2013).

Moran, S. and M. Cysouw (2018). *The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles*. Berlin: Language Science Press.

Moran, S. and D. McCloy, eds. (2019). *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History.

Nikolaev, D., A. Nikulin, and A. Kukhto (2017). *The database of Eurasian phonological inventories*. Moscow: RGGU.

Norman, J. (2003). "The Chinese dialects. Phonology." In: *The Sino-Tibetan languages*. Ed. by G. Thurgood and R. J. LaPolla. London and New York: Routledge, 72–83.

Rzymski, C. et al. (2020). "The Database of Cross-Linguistic Colexifications, reproducible analysis of cross- linguistic polysemies." *Scientific Data* 7.13, 1–12. URL: https://clics.clld.org.

Tang, M. and O.-S. Her (2019). "Insights on the Greenberg-Sanches-Slobin generalization: Quantitative typological data on classifiers and plural markers." *Folia Linguistica* 53.2, 297–331.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data* 3, 1–9.

Converting Data to CLDF

Johann-Mattis List (University of Passau)

1 Questions for the Practice Section

What kind of data do you have to deal with in your research?

In which format do you usually curate your data?

In which format do you usually share your data?

Where do you share your data?

When do you share your data?

2 Tasks for the Practice Section

Imagine you have to share your data in some kind of a tabular format, similar to CLDF or based on CLDF.

- Determine the main tables you will need for your data.
- Determine identifiers and foreign keys in your tables.
- Determine specific characteristics of the values in your cells (are they integers, are they some space-separated values, are they dates, are they words from a lexicon?).
- State which reference catalogs you would use to “enrich” your data or to outsource the curation of your data to other entities.
- State which reference catalogs you would likely have to initiate yourself.
- Write a first schematic overview of your new tabular table schema.

Computer-Assisted Language Comparison

Johann-Mattis List (University of Passau)

1 The quantitative turn in historical linguistics

Background

In the early 1950s, Morris Swadesh (1909–1967) presented a method to measure the genetic closeness between languages on the basis of a statistical formula that was ultimately based on counting the amount of shared cognates across standardized wordlists of different languages (Swadesh 1950). Although it seemed at first that the methods could revive the discipline of historical linguistics, which had past its prime after the structuralist turn in the begin of the 1920s , and had not seen any major methodological or analytical improvement since the begin of the 20th century.[1] Unfortunately, the original interest in the new ideas did not last long, and soon after it was first published, the new method was heavily criticized (Bergsland and Vogt 1962), and went out of vogue some 10 years later.

In the begin of the second millennium, Gray and Atkinson (2003) used similar data but different statistical methods to date the age of the Indo-European language family. They caused a similar stir as Swadesh had done almost half a century ago. But while Swadesh's method was filed away soon after it had been proposed, the method of Gray and Atkinson was part of a general *quantitative turn in historical linguistics*, which started at the begin of the second millennium. This quantitative turn is reflected in a large bunch of literature on such different topics as phonetic alignment (Kondrak 2000, Prokić et al. 2009), automated cognate detection (List 2014), and phylogenetic reconstruction (Atkinson and Gray 2006).

What may have been the reasons why Swadesh's approach was abandoned so quickly by historical linguists?

New studies on language evolution

We can distinguish four different aspects of research approaches in the course of the quantitative turn. As a first and most prominent aspect, we have research dealing with questions of *phylogenetic reconstruction* which usually involved *dating* as well. Language data are not only analyzed to yield a topology of the branching structure of the language family in question, but in addition, absolute branch lengths are often also inferred, which allow to estimate when a given language family has originated. The software and methods used for these studies are usually taken or inspired from approaches developed first in evolutionary biology. As of now, quite a few different language families have been analyzed in this way, including Indo-European (Chang et al. 2015, Gray and Atkinson 2003), Austronesian (Gray et al. 2009), Dravidian (Kolipakam et al. 2018), Bantu (Grollemund et al. 2015), Pama-Nyungan (Bowern et al. 2011), Japonic (Lee and Hasegawa 2011), and Sino-Tibetan (Sagart et al. 2019). In addition, scholars have also attempted to provide unified methods that could be applied in a completely automated fashion to all languages of the world (Holman et al. 2011).

Another strand of research deals with the computation of inference procedures which were traditionally only carried out manually. Most prominently, we find here various attempts to automate different aspects of the general workflow of the traditional *comparative method* for historical language comparison (Weiss 2015). Breaking down the workflow into some of its major parts, we thus find

[1] The last major improvement, the decipherment of Hittite, which also helped to proof that it was an Indo-European language dated back to Hrozný (1915).

(1) automated methods for the comparison of words, as reflected in methods for phonetic alignment (Kondrak 2000, Prokić et al. 2009) and automated cognate detection (Hauer and Kondrak 2011, List et al. 2016, Turchin et al. 2010), (2) automated approaches for the detection of borrowings (List 2015, Mennecier et al. 2016, Nelson-Sathi et al. 2011),[2] (3) automated approaches for linguistic reconstruction (Bouchard-Côté et al. 2013, Jäger 2019), and (4) automated approaches for the detection of sound correspondences (List 2019b).

While the second strand deals mostly with questions of inference, a third strand organizes inferred data in form of large-scale online databases that aggregate different kinds of information on the world's languages. The most prominent of these databases is beyond doubt the *World Atlas of Language Structures* (Dryer and Haspelmath 2013), but in addition we also find attempts to aggregate cross-linguistic information on phoneme inventories (Maddieson et al. 2013, Moran and McCloy 2019), polysemies (List et al. 2018), phonotactics (Donohue et al. 2013), borrowings (Haspelmath and Tadmor 2009), as well as datasets like D-Place, that compare cultural, environmental, and linguistic diversity (Kirby et al. 2016).

While the popular phylogenetic approaches deal with c-linguistics (or p-linguistics in a wider sense of the term), insofar as they deal with concrete languages in concrete times, trying to answer very specific (or *particular*) questions about their past, a fourth strand of research makes use of the new cross-linguistic databases along with results drawn from the phylogenetic approaches to investigate general aspects of language change, including questions like the rate of linguistic change and its correlates (Calude and Pagel 2011, Greenhill et al. 2017), the question to which degree environmental factors might have an impact on language evolution (Everett et al. 2015), or how language structures converge independent of contact or inheritance (Blasi et al. 2016).

Why is the aspect of dating, i.e., the inference of absolute phylogenies, so important for the new methods in historical linguistics?

Benefits of computational historical linguistics

Apart from the obvious benefit that the new quantitative methods have drastically revived the interest of scholars in historical linguistics, which also resulted in an increased amount of funding and a new generation of young scholars who are highly collaborative in their research and well trained in computational methods, the quantitative turn has also led to a considerable amount of rethinking in the field of historical linguistics, which offers new perspectives on the subject which have been ignored so far. First, we can see that the new methods shift the focus from *internal* to *external language* history, while at the same time turning away from the traditional focus on Indo-European alone.[3] We can also see that the new methods lead to the raise of new questions, specifically addressing *general* questions of language history.

This is also reflected in new research approaches, which are more explicitly *data-centered* nowadays and often based on statistical or stochastic modeling. While research in historical linguistics has always been data-centered, the new methods have shown that the classical approaches to deal with data – namely the individual collection of extensive personal notes from the literature, and the publication of new insights from these personal collections in form of extensive prose – are reaching their limits in times where the amount of data is constantly increasing. Although the attempts to automate the classical methods have so far not yet led to a situation where computers could beat the experts,[4] we

[2] See List (2019a) for an overview on these approaches.

[3] Compare classical handbooks such as the *Einführung in die vergleichende Sprachwissenschaft* by Szemerényi (1970), where the term *comparative linguistics* (which should be a general discipline) is seen as a synonym for *Indo-European linguistics*.

[4] This is also not to be expected shortly, given that the only areas in which machines outperform humans so far are restricted fields, such as chess, or the go-game (Silver et al. 2016), and not in problems that need to be solved in open worlds.

have won many important and new insights into the methods and the practice of historical language comparison, specifically also because the new methods challenged classical (traditional) linguists to revise the methods they use and to increase the degree of explicitness by which they apply them.

That languages interact with different factors is evident. What are the aspects that make it so difficult to study language change with help of computational frameworks?

Problems and criticisms

Not all linguists have enthusiastically welcomed the new methods. While the various critics range from justified criticism, via exaggerations, up to complete ignorance for the initial goals of the computational approaches, and at times rather reflect the insulted ego of those who consider themselves as indisputable experts, the new field faces a couple of serious problems that are worth being criticized and rigorously analyzed. Among the most important of these are (1) problems with the data that is used in quantitative analyses, (2) problems of applicability of the computational approaches, and (3) problems of transparency and (4) comparability with respect to the results and methods which scholars report, and (5) problems of the general accuracy of the computational methods in comparison with experts.

The data problems related to the way in which data are compiled and curated, and what judgments they are based upon. The general problem here is that most of the phylogenetic approaches still make use of human-annotated data, trusting the expertise of only a small amount of experts to be enough to annotated data for at times more than 100 different languages. The danger of this procedure (which is to some degree difficult to avoid) are potential problems of inter-annotator-agreement, which may themselves, of course, impact the results (Geisler and List 2010). The problem of applicability and transparency is reflected in large amounts of software solutions and datasets that are only discussed in the literature, but have not been openly shared (List et al. 2017). As a result, there are quite a few methods out there that could provide valid solutions, but which have only been tested on one dataset and never officially been published, which comes close to a crisis of irreproducibility as it has been noted in many branches of science since the beginning of this millennium (Nature 2013).[5]

The problem of comparability results from missing standards in our field, which make it difficult to compare results across datasets, since it is often very tedious to lift the data used by different scholars to a level where they could be easily compared. The problem of accuracy, finally, is probably the hardest problem to address, since the problems of historical linguistics are often quite hard to solve automatically, specifically also because – as a rule – data is sparse, while most computational methods have been built based on the assumption that data to test and train algorithms would be abundantly available.

What solutions can you think of to overcome the problems of transparency and comparability, which were mentioned above?

2 Towards a qualitative turn in diversity linguistics

Reconciling classical and computational research

The use of computer applications in historical linguistics is steadily increasing. With more and more data available, the classical methods reach their practical limits. At the same time, computer applications are not capable of replacing experts' experience and intuition, especially when data are sparse.

[5] Luckily, this picture is slowly changing, thanks to extensive efforts to propagate free data and free code. A our department, for example, we have now decided to refuse to review papers where we are not given code and data, if they are needed for replication, following the idea of referee's rights as expressed by the editorial board of the journal Nature in 2018.

If computers cannot replace experts and experts do not have enough time to analyse the massive amounts of data, a new framework is needed, neither completely computer-driven, nor ignorant of the assistance computers afford. Such computer-assisted frameworks are well-established in biology and translation. Current machine translation systems, for example, are efficient and consistent, but they are by no means accurate, and no one would use them in place of a trained expert. Trained experts, on the other hand, do not necessarily work consistently and efficiently. In order to enhance both the quality of machine translation and the efficiency and consistency of human translation, a new paradigm of computer-assisted translation has emerged (Barrachina et al. 2008: 3).

Do you have experience with computer-assisted translation? If not, what role do computers and computer tools play for your research?

Computer-assisted language comparison

Following the idea of computer-assisted frameworks in translation and biology, a framework for computer assisted language comparison (CALC) is the key to reconcile classical and computational approaches in historical linguistics. Computational approaches may still not be able to compete with human experts, but when used to pre-process the data with human experts systematically correcting the results, they can drastically increase the efficiency of the classical comparative method and make up for the insufficiencies of of current computational solutions. At the same time, bringing experts closer to computational and formal approaches will also help to increase the consistency or classical research, forcing experts to annotated their specific findings and corrections in due detail, without resorting to texts in prose and ad-hoc explanations.

Classical linguists working on etymological research often emphasize the importance of looking into all details of language history, invoking the slogan “chaque mot a son histoire”, which is, according to Campbell (1999: 189) traditionally attributed to Jules Gilliéron (1854-1926). Even if this was completely true, how can we still defend the recent attempts of computer-assisted and computer-based strategies in historical linguistics to work on a more formal and more quantitative handling of linguistic data?

Data, Software, and Interfaces

In the framework of computer-assisted language comparison, data are constantly passed back and forth between computational and classical linguists. Three different aspects are essential for this workflow: Specific *software* allows for the application of transparent methods which increase the accuracy and the application range of current methods in historical linguistics and linguistic typology. Interactive *interfaces* serve as a bridge between human and machine, allowing experts to correct errors and to inspect the automatically produced results in detail. To guarantee that software and interfaces can interact directly, *data* need to be available in human- and machine-readable form.

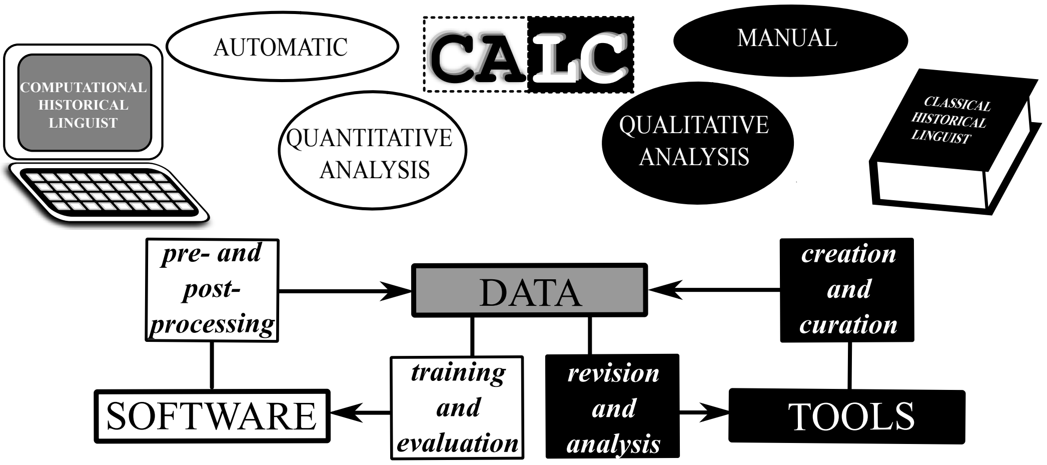


Fig. 1: Interplay of data, software, and interfaces in computer-assisted language comparison.

How exactly should one imagine data that are human- and machine-readable at the same time?

CALC project at the MPI-SHH in Jena

In the ERC-funded research project CALC (Computer-Assisted Language Comparison, List 2016), we try to establish a computer-assisted framework for historical linguistics. We pursue an interdisciplinary approach that adapts methods from computer science and bioinformatics for the use in historical linguistics. While purely computational approaches are common today, the project focuses on the communication between classical and computational linguists, developing interfaces that allow historical linguists to produce their data in machine readable formats while at the same time presenting the results of computational analyses in a transparent and human-readable way.

As a litmus test which proves the suitability of the new framework, the project attempts to create an etymological database of Sino-Tibetan languages (see Sagart et al. 2019 for initial attempts and results). The abundance of language contact and the peculiarity of complex processes of language change in which sporadic patterns of morphological change mask regular patterns of sound change make the Sino-Tibetan language family an ideal test case for a new overarching framework that combines the best of two worlds: the experience of experts and the consistency of computational models.

What may be the reason for choosing an interdisciplinary approach, and what are the most likely disciplines from which the project could take inspiration?

References

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

Barrachina, S. et al. (2008). “Statistical approaches to computer-assisted translation.” *Computational Linguistics* 35.1, 3–28.

Bergsland, K. and H. Vogt (1962). “On the validity of glottochronology.” *Current Anthropology* 3.2, 115–153. JSTOR: 2739527.

Blasi, D. E., S. Wichmann, H. Hammarström, P. Stadler, and M. H. Christiansen (2016). “Sound–meaning association biases evidenced across thousands of languages.” *Proceedings of the National Academy of Science of the United States of America* 113.39, 10818–10823.

Bouchard-Côté, A., D. Hall, T. L. Griffiths, and D. Klein (2013). “Automated reconstruction of ancient languages using probabilistic models of sound change.” *Proceedings of the National Academy of Sciences of the United States of America* 110.11, 4224–4229.

Bowern, C., P. Epps, R. Gray, J. Hill, K. Hunley, P. McConvell, and J. Zentz (2011). “Does Lateral Transmission Obscure Inheritance in Hunter-Gatherer Languages?” *PLoS ONE* 6.9, e25195.

Calude, A. S. and M. Pagel (2011). “How do we use language? Shared patterns in the frequency of word use across 17 world languages.” *Philosophical Transactions of the Royal Society B* 366, 1101–1107.

Campbell, L. (1999). *Historical linguistics. An introduction.* 2nd ed. Edinburgh: Edinburgh Univ. Press.

Chang, W., C. Cathcart, D. Hall, and A. Garret (2015). “Ancestry-constrained phylogenetic analysis ssupport the Indo-European steppe hypothesis.” *Language* 91.1, 194–244.

Donohue, M., R. Hetherington, J. McElvenny, and V. Dawson (2013). *World phonotactics database*. Canberra: Department of Linguistics. The Australian National University.

Dryer, M. S. and M. Haspelmath, eds. (2013). *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Everett, C., D. E. Blasi, and S. G. Roberts (2015). “Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots.” *Proceedings of the National Academy of Sciences of the United States of America* 112.5, 1322–1327.

Geisler, H. and J.-M. List (2010). “Beautiful trees on unstable ground. Notes on the data problem in lexicostatistics.” In: *Die Ausbreitung des Indogermanischen. Thesen aus Sprachwissenschaft, Archäologie und Genetik*. Ed. by H. Hettrich. Document has been submitted in 2010 and is still waiting for publication. Wiesbaden: Reichert.

Gray, R. D. and Q. D. Atkinson (2003). “Language-tree divergence times support the Anatolian theory of Indo-European origin.” *Nature* 426.6965, 435–439.

Gray, R. D., A. J. Drummond, and S. J. Greenhill (2009). “Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement.” *Science* 323.5913, 479–483.

Greenhill, S. J., C. H. Wu, X. Hua, M. Dunn, S. C. Levinson, and R. D. Gray (2017). “Evolutionary dynamics of language systems.” *Proceedings of the National Academy of Sciences of the United States of America* 114.42, E8822–E8829.

Grollemund, R., S. Branford, K. Bostoen, A. Meade, C. Venditti, and M. Pagel (2015). “Bantu expansion shows that habitat alters the route and pace of human dispersals.” *Proceedings of the National Academy of Sciences of the United States of America* 112.43, 13296–13301.

Haspelmath, M. and U. Tadmor, eds. (2009). Berlin and New York: de Gruyter.

Hauer, B. and G. Kondrak (2011). “Clustering semantically equivalent words into cognate sets in multilingual lists.” In: *Proceedings of the 5th International Joint Conference on Natural Language Processing*. (Chiang Mai, Thailand, 11/08–11/13/2011). AFNLP, 865–873.

Holman, E. W. et al. (2011). “Automated dating of the world’s language families based on lexical similarity.” *Current Anthropology* 52.6, 841–875. JSTOR: `10.1086/662127`.

Hrozný, B. (1915). “Die Lösung des hethitischen Problems [The solution of the Hittite problem].” *Mitteilungen der Deutschen Orient-Gesellschaft* 56, 17–50.

Jäger, G. (2019). “Computational historical linguistics.” *Theoretical Linguistics* 45.3-4, 151–182.

Kirby, K. R. et al. (2016). “D-PLACE: A Global Database of Cultural, Linguistic and Environmental Diversity.” *PLOS ONE* 11.7, 1–14.

Kolipakam, V., F. M. Jordan, M. Dunn, S. J. Greenhill, R. Bouckaert, R. D. Gray, and A. Verkerk (2018). “A Bayesian phylogenetic study of the Dravidian language family.” *Royal Society Open Science* 5.171504, 1–17.

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

Lee, S. and T. Hasegawa (2011). “Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages.” *Proc. Biol. Sci.* 278.1725, 3662–3669.

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

— (2015). “Network perspectives on Chinese dialect history.” *Bulletin of Chinese Linguistics* 8, 42–67.

— (2016). *Computer-Assisted Language Comparison: Reconciling Computational and Classical Approaches in Historical Linguistics*. Jena: Max Planck Institute for the Science of Human History.

— (2019a). “Automated methods for the investigation of language contact situations, with a focus on lexical borrowing.” *Language and Linguistics Compass* 13.e12355, 1–16.

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

List, J.-M., S. J. Greenhill, C. Anderson, T. Mayer, T. Tresoldi, and R. Forkel (2018). “CLICS^2. An improved database of cross-linguistic colexifications assembling lexical data with help of cross-linguistic data formats.” *Linguistic Typology* 22.2, 277–306.

List, J.-M., S. J. Greenhill, and R. D. Gray (2017). “The potential of automatic word comparison for historical linguistics.” *PLOS ONE* 12.1, 1–18.

List, J.-M., P. Lopez, and E. Bapteste (2016). “Using sequence similarity networks to identify partial cognates in multilingual wordlists.” In: *Proceedings of the Association of Computational Linguistics 2016 (Volume 2: Short Papers)*. Association of Computational Linguistics. Berlin, 599–605.

Maddieson, I., S. Flavier, E. Marsico, C. Coupé, and F. Pellegrino. (2013). “LAPSyD: Lyon-Albuquerque Phonological Systems Database.” In: *Proceedings of Interspeech*. (Lyon, 08/25–08/29/2013).

Mennecier, P., J. Nerbonne, E. Heyer, and F. Manni (2016). “A Central Asian language survey.” *Language Dynamics and Change* 6.1, 57–98.

Moran, S. and D. McCloy, eds. (2019). *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History.

Nature, E. B. (2013). “Reducing our irreproducibility.” *Nature* 496.4, 398.

— (2018). “Referees´rights.” *Nature* 560, 409.

Nelson-Sathi, S., J.-M. List, H. Geisler, H. Fangerau, R. D. Gray, W. Martin, and T. Dagan (2011). “Networks uncover hidden lexical borrowing in Indo-European language evolution.” *Proceedings of the Royal Society of London B: Biological Sciences* 278.1713, 1794–1803.

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

Sagart, L., G. Jacques, Y. Lai, R. Ryder, V. Thouzeau, S. J. Greenhill, and J.-M. List (2019). “Dated language phylogenies shed light on the ancestry of Sino-Tibetan.” *Proceedings of the National Academy of Science of the United States of America* 116 (21), 10317–10322.

Silver, D. et al. (2016). “Mastering the game of Go with deep neural networks and tree search.” *Nature* 529.7587, 484–489.

Swadesh, M. (1950). “Salish internal relationships.” *International Journal of American Linguistics* 16.4, 157–167. JSTOR: `1262898`.

Szemerényi, O. (1970). *Einführung in die vergleichende Sprachwissenschaft*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Turchin, P., I. Peiros, and M. Gell-Mann (2010). “Analyzing genetic connections between languages by matching consonant classes.” *Journal of Language Relationship* 3, 117–126.

Weiss, M. (2015). “The comparative method.” In: *The Routledge handbook of historical linguistics*. Ed. by C. Bowern and N. Evans. New York: Routledge, 127–145.

Sequence Comparison

Johann-Mattis List (University of Passau)

1 Background

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

Can recipes always be reduced to sequences of instructions?

Discreteness and Continuity

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

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

Defining a Sequence

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

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

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

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

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

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

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

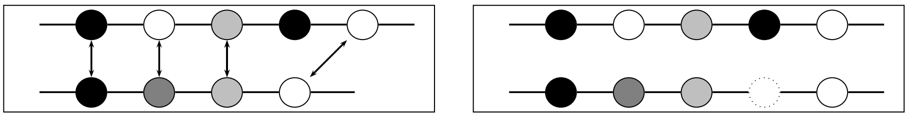
Why is it important to distinguish a subsequence of a substring, and what would be a general term for both suffix and prefix?

2 Phonetic Alignment

Alignment Analyses in General

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

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



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

Phonetic Alignment Analyses in Specific

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

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

(a) Globale Alinierung

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

(b) Lokale Alinierung

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

Types of Sound Change

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

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

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

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

Sound Classes

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

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

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

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

Morphemes and Secondary Structures

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

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

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

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

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

Alignability

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

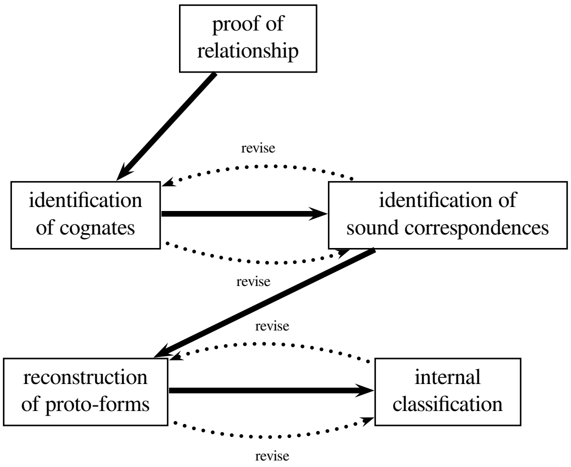
DOCULECT	SEGMENTS	ROOT	STEM	DERIVATION
French	sol←ej	*soh$_2$wl-	*soh$_2$wl + ?	RECTUS DIM
Spanish	sol	*soh$_2$wl-	*soh$_2$wl	RECTUS
German	zɔnɛ	*soh$_2$wl-	*sh$_2$en	OBLIQUUS
Swedish	suːl	*soh$_2$wl-	*soh$_2$wl	RECTUS

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

3 Cognate Detection

The Comparative Method

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


proof of relationship
revise
identification of cognates
identification of sound correspondences
revise
revise
reconstruction of proto-forms
internal classification
revise

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

Traditional Approaches to Cognate Detection

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

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

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

4 Automatic Cognate Detection

Quantifying Sound Correspondences

In bioinformatics, it is important to compute the probability of correspondences in DNA and protein alignment. This is done by comparing an *attested* with an *expected* distribution. Transferred to linguistics, this means that we compare a list of corresponding sounds with a distribution which we would expect if the languages were not genetically related. In order to substantiate this, linguists usually show long lists of potential cognates, as shown in the list below:

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

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

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

“square”	p j a ʦː a p l a s -
“feather”	p j u m a p l y m -
“flat”	p j a n o p l ɑ̃ - -

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

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

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

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

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

Clustering

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

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

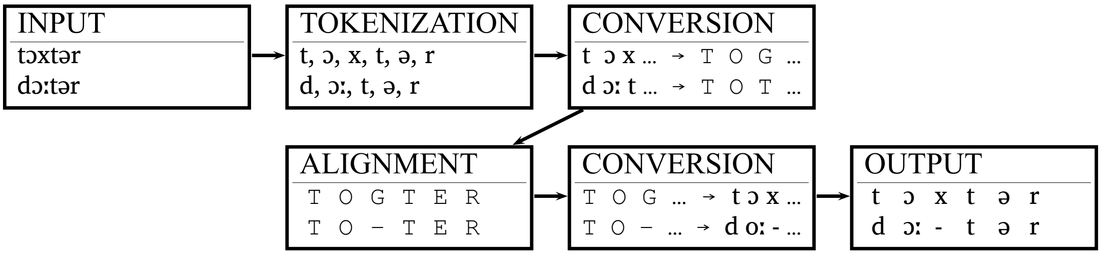
Automatic clustering has the advantage that the evidence which may be missing when comparing only one language pair, can be backed up by additional evidence. This nicely accounts for the use of *cumulative evidence* (Sturtevant 1920: 11), which is a fundamental aspect of the comparative methods for historical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

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

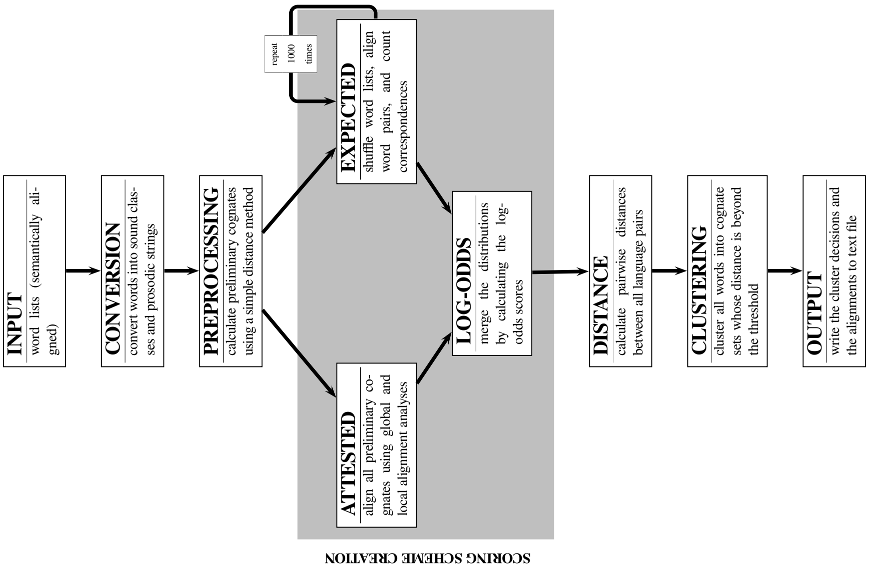
LexStat

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

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

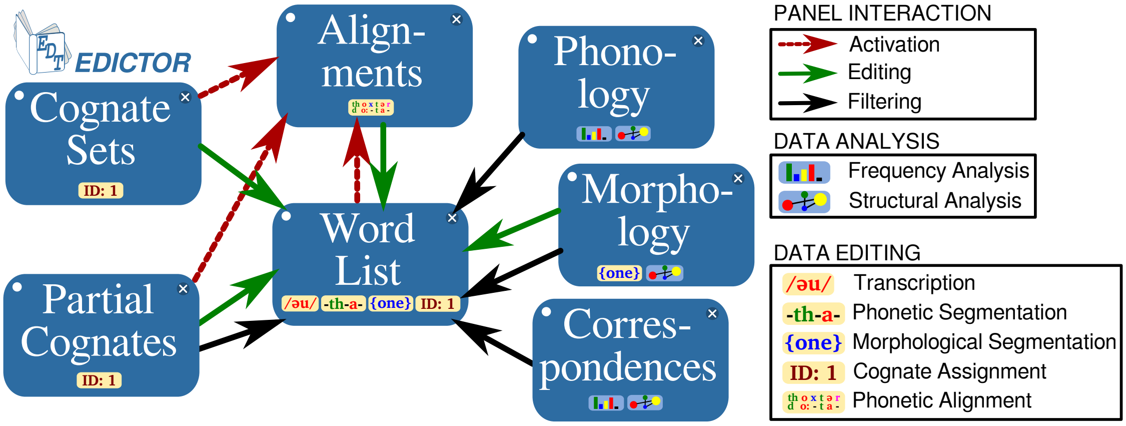


LexStat often has problems to distinguish true cognates from borrowings if borrowings are abundant. Why is that so?


INPUT
word lists (semantically aligned)
CONVERSION
convert words into sound classes and prosodic strings
PREPROCESSING
calculate preliminary cognates using a simple distance method
repeat 1000 times
EXPECTED
shuffle word lists, align word pairs, and count correspondences
ATTESTED
align all preliminary cognates using global and local alignment analyses
LOG-ODDS
merge the distributions by calculating the log-odds scores
SCORING SCHEME CREATION
DISTANCE
calculate pairwise distances between all language pairs
CLUSTERING
cluster all words into cognate sets whose distance is beyond the threshold
OUTPUT
write the cluster decisions and the alignments to text file

5 Cognate Annotation

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



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

References

Böckenbauer, H.-J. and D. Bongartz (2003). *Algorithmische Grundlagen der Bioinformatik.* German. Stuttgart, Leipzig, and Wiesbaden: Teubner.
Campbell, L. and M. Mixco (2007). *A glossary of historical linguistics.* Edinburgh: Edinburgh University Press.
Chao, Y. (1968). *A grammar of spoken Chinese.* Berkeley, Los Angeles, and London: University of California Press.
Derksen, R., comp. (2008). *Etymological dictionary of the Slavic inherited lexicon.* Leiden Indo-European Etymological Dictionary Series 4. Leiden and Boston: Brill.
Dolgopolsky, A. B. "Gipoteza drevnejšego rodstva jazykovych semej Severnoj Evrazii s verojatnostej točky zrenija [A probabilistic hypothesis concering the oldest relationships among the language families of Northern Eurasia]." *Voprosy Jazykoznanija* 2 (1964), 53–63; English translation: Dolgopolsky, A. B. "A probabilistic hypothesis concerning the oldest relationships among the language families of northern Eurasia." In: *Typology, relationship and time. A collection of papers on language change and relationship by Soviet linguists. Typology, Relationship and Time. A collection of papers on language change and relationship by Soviet linguists.* Ed. and trans. from the Russian by V. V. Shevoroshkin. Ann Arbor: Karoma Publisher, 1986, 27–50.
Gusfield, D. (1997). *Algorithms on strings, trees and sequences.* Cambridge: Cambridge University Press.
IPA, ed. (1999). *Handbook of the International Phonetic Association. A guide to the use of the international phonetic alphabet.* Cambridge: Cambridge University Press.
Kruskal, J. B. (1983). "An overview of sequence comparison. Time warps, string edits, and macromolecules." *SIAM Review* 25.2, 201–237. JSTOR: `2030214`.
List, J.-M. (2012). "LexStat. Automatic detection of cognates in multilingual wordlists." In: *Proceedings of the EACL 2012 Joint Workshop of Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources.* "LINGVIS & UNCLH 2012" (Avignon, 04/23–04/24/2012). Stroudsburg, 117–125.
— (2014). *Sequence comparison in historical linguistics.* Düsseldorf: Düsseldorf University Press.
— (2017). "A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets." In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations.* Valencia: Association for Computational Linguistics, 9–12.
List, J.-M. and R. Forkel (2016). *LingPy. A Python library for historical linguistics.* Version 2.5. URL: `http://lingpy.org`.
Sturtevant, E. H. (1920). *The pronunciation of Greek and Latin.* Chicago: University of Chicago Press. Internet Archive: `pronunciationgr00unkngoog`.
Trask, R. L., comp. (2000). *The dictionary of historical and comparative linguistics.* Edinburgh: Edinburgh University Press.

Semantic Networks

Johann-Mattis List (University of Passau)

1 Semantics and Semantic Change

It is well known and not surprising for practitioners of historical linguistics that semantics and semantic change are topics that are very difficult to handle systematically. The reason for this lies in what Sperber (1923: 1) calls the *psychological factors* of meaning, which are much more difficult to grasp and describe than it is to give logical definitions of certain concepts.

Apart from the general question where to allocate semantic change (in the domain of the lexicon or the domain of pragmatics, or as a transition between the two, see (Traugott 2012)), the reason for the problems one faces when dealing with semantic change can be found in the structural differences between sign form and sign meaning and the resulting processes by which both entities change. While the formal part of the linguistic sign is characterized by its sequential structure and sound change is characterized by the *alternation* of segments, the meaning part is better described as some kind of *conceptual network*, and semantic change is not based on an alternation but on the *accumulation* and *reduction* of potential referents,[1] for example by a reorganization of the sign's *reference potential* (List 2014: 36). Although change in meaning is traditionally considered to be notoriously irregular and unpredictable, with scholars emphasizing that "there is [...] little in semantic change which bears any relationship to regularity in phonological change" (Fox 1995: 111), it is also obvious that a large number of observed pathways of semantic change can be observed to occur independently in many different language families of the world. In some sense, we face the same problems we also found for the handling of regular sound change patterns. If we want to study pathways of semantic change cross-linguistically, we will need to find a way to make our data comparable. That this can be cumbersome and difficult could be observed for the Catalogue of Semantic Shifts (Zalizniak 2018, Zalizniak et al. 2012), which originally presented a larger collection of observed semantic change processes, but ultimately has problems to provide a rigorous specification of the different meanings that were tracked.[2]

How can we imagine this process of accumulation and reduction to take place, and what is meant by "reference potential"?

2 Multilingual Approaches to Semantic Change

We have repeatedly seen and discussed how notoriously difficult it is to study semantic change systematically, given that, once it comes to "meaning, one has as a guide only a certain probability based on common sense, on the personal evaluation of the linguist, and on the parallels that he can cite" (Wilkins 1996: 264). Interestingly, however, the often-invoked differences between semantic change and sound change become much less striking when we stop to think about sound change as something ultimately *regular*. In the last session, we have discussed the regularity of sound change a lot, and one of the important aspects was that the apparent regularity is nothing else than a change on a higher level, not at the level of the word alone, a change of the phoneme system, as emphasized

[1] This can already be found in the work of Herman Paul (1846–1921), who emphasizes that there is always an "extension or restriction of the extent of the meaning" and that "only the succession of extension and restriction allows the emergence of a new, from the original one completely different meaning" (Paul 1880 [1886]: 66, my translation).

[2] To my knowledge, the authors are currently working on a new version that will hopefully cope with the problems of the older version and also provide an increase in data (see `http://datsemshift.ru`).

early by Bloomfield (1933 [1973]: 351). If we look at the *substance* of sound change, at concrete patterns, and the incredible number of different sound segments which scholars propose to have found in certain languages (Anderson et al. 2018), however, sound change does not seem much more chaotic then semantic change. On the contrary: if it is possible to establish a first reference catalogue of phonetic transcriptions, and if we trust that the initial work done in the Concepticon project has been done thoroughly enough, and if we further keep in mind that diachronic patterns often can also be observed synchronically, we may be able to work on feasible solutions to at least approximately reconstruct basic semantic structure from cross-linguistic data.

How does semantic change surface in synchronic linguistic data?

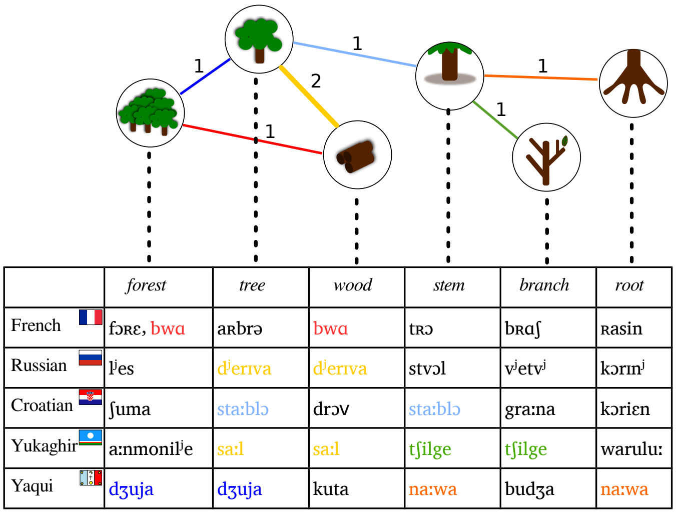
Polysemy, Homophony, and Colexification

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?

Colexification Networks

If one has enough data, it is considerably easy to construct *concept networks* from cross-linguistic colexifications (Cysouw 2010). The starting point are semantically aligned word lists for a large amount of different languages from different language families. By counting, in how many languages, or in how many language families a certain colexification recurs, we can further *weight* the edges of the network, as shown in Figure 1.



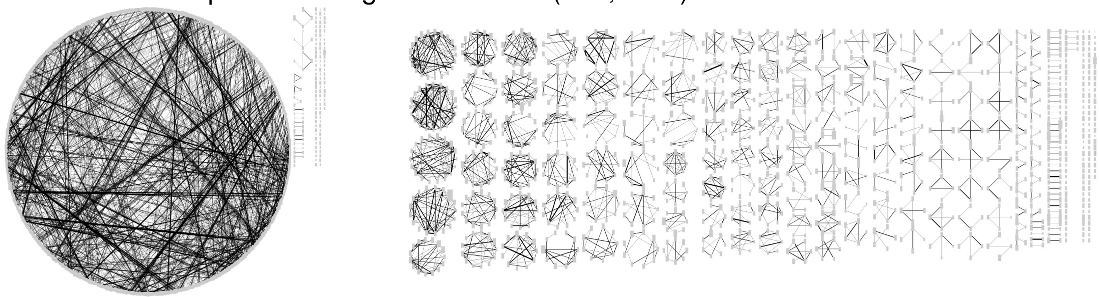
	forest	*tree*	*wood*	*stem*	*branch*	*root*
French	fɔʀɛ, bwɑ	aʀbʀə	bwɑ	tʀɔ	bʀɑʃ	ʀasin
Russian	lʲes	dʲerɪva	dʲerɪva	stvɔl	vʲetvʲ	kɔrɪnʲ
Croatian	ʃuma	staːblɔ	drɔv	staːblɔ	graːna	kɔriɛn
Yukaghir	aːnmonilʲe	saːl	saːl	tʃilge	tʃilge	waruluː
Yaqui	dʒuja	dʒuja	kuta	naːwa	budʒa	naːwa

Figure 1: Reconstructing colexification networks from multi-lingual wordlists.

Is there any straightforward way to derive directed graphs from weighted, undirected colexification networks?

Analyzing 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 connectionso 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**

Figure 3: Comparing clustered and unclustered colexification networks.

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?

Database of Cross-Linguistic Colexifications

CLICS^3 (`https://clics.clld.org`, Rzymski et al. 2020) is an online database of colexifications in about 2000 language varieties of the world. CLICS^3 is the third installment of the Database of Cross-Linguistic Classifications, following the second version published two years before (List et al. 2018), and an even earlier version from 2014 (List et al. 2014), which introduced the interactive representation of cross-linguistic colexification patterns (Mayer et al. 2014) which is still one of the major reasons why CLICS is so popular. While the original CLICS database was low in terms of cross-linguistic coverage and difficult to maintain, the strict adherence to the format specifications based on the CLDF initiative made it possible to grow the data drastically, from originally 221 language varieties in the original version up to 1220 varieties in second version (List et al. 2018), up to more than 2000 varieties in the third installment (Rzymski et al. 2020).[3]

2.1 Data Curation and Aggregation in CLICS^3

The major advancement of CLICS^3 was a new framework for data curation and aggregation, entirely built on the CLDF strategies. Essentially, this workflow consists of four major stages, which can be

[3] We have a new update for CLICS^4 in preparation, which will, however, no longer grow the number of languages covered, but rather concentrate on the quality of the data.

carried out independently from each other. These stages include the *mapping of concepts* to Concepticon (List et al. 2022b), the *referencing of sources* in the original data, the *linking of languages* to Glottolog (Hammarström et al. 2021), and the *cleaning of lexical entries* using a dedicated suite of Python scripts (later published as part of the Lexibank workflow List et al. 2022a). Once data are prepared in this form and rendered in PDF, aggregating data from different sources into a larger database is extremely straightforward. Since the investigation of colexification patterns furthermore does not require to compare word forms *across* languages, but only *inside*, no further normalization (e.g., of the transcriptions) is needed.[4]

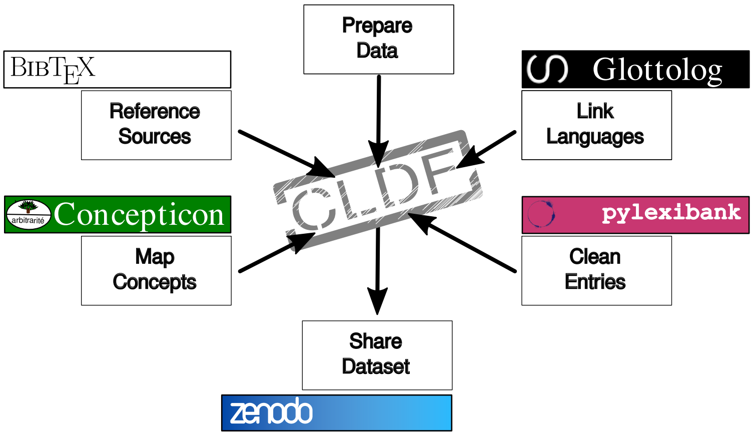


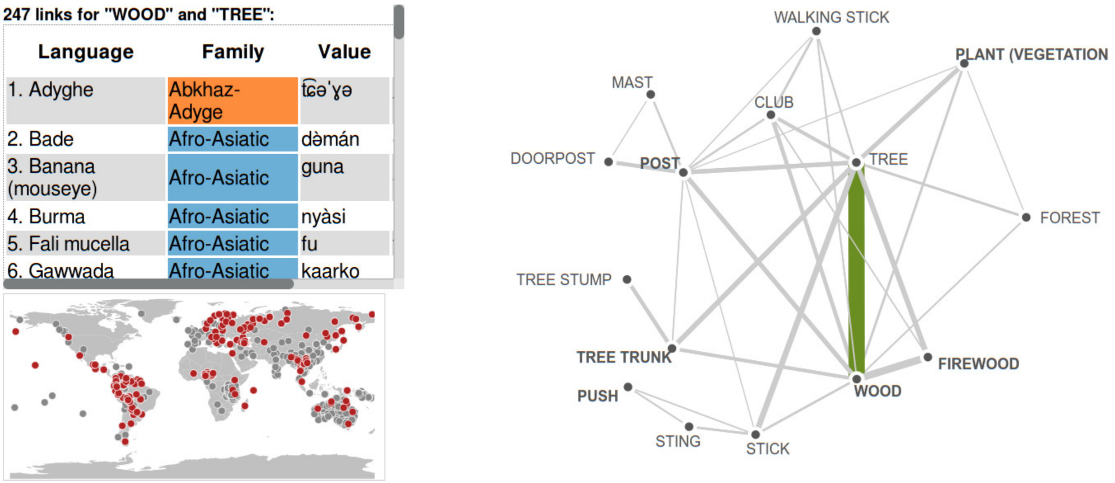
Figure 4: Workflow for data aggregation and curation in CLICS³.

What pitfalls should one avoid when trying to clean lexical entries?

2.2 Examples

The visualization framework used in CLICS is based on an interactive, force-directed, graph layout, written in JavaScript. The basic idea behind this visualization is to allow users to inspect both all the data underlying a given colexification (ideally up to allowing to trace the original datasets, the word forms, and the original elicitation glosses), while at the same time offering a bird's eye view on the global distribution of a given colexification pattern. This is illustrate in the screenshot in Figure 2, where the cluster around words for "tree" and "wood" is shown.

[4] The upcoming fourth installment of the CLICS database, however, will have fully transcribed word forms for a then slightly smaller amount of language varieties, since we decided that transcribed, unified transcriptions offer for more possibilities to analyze the data consistently.



247 links for "WOOD" and "TREE":

Language	Family	Value
1. Adyghe	Abkhaz-Adyge	t͡ɕəˈɣə
2. Bade	Afro-Asiatic	dəmán
3. Banana (mouseye)	Afro-Asiatic	guna
4. Burma	Afro-Asiatic	nyàsi
5. Fali mucella	Afro-Asiatic	fu
6. Gawwada	Afro-Asiatic	kaarko

Figure 2: Screenshot from the CLICS² database (see infomap_2_WOOD).

What exactly does this visualization tell us?

3 Beyond Colexification Networks

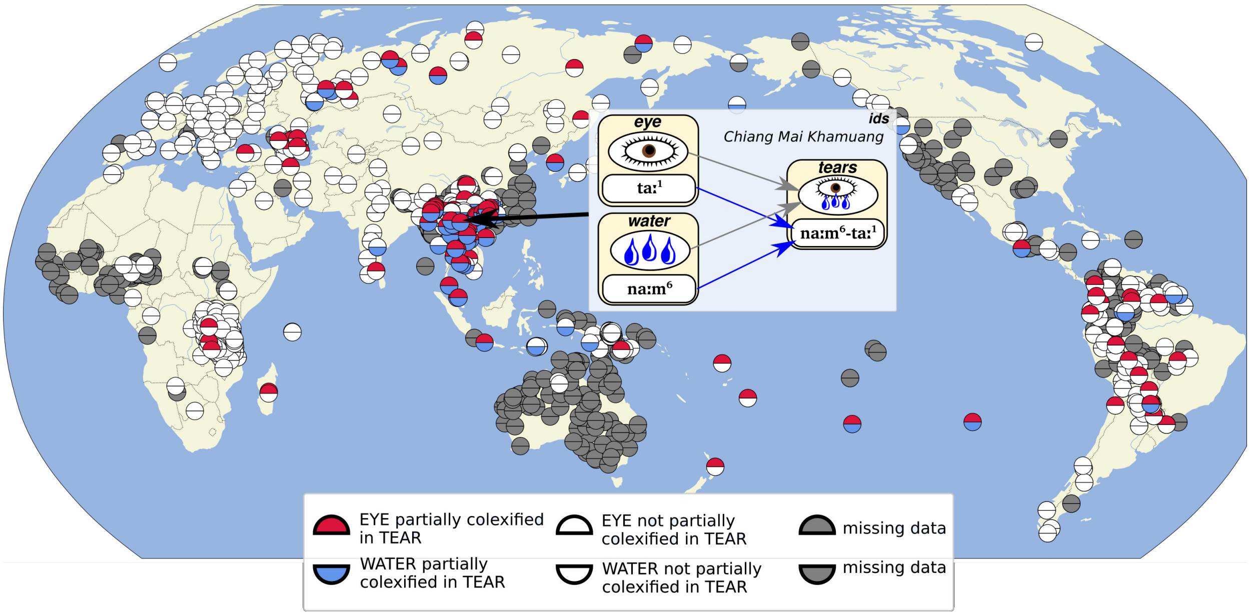
In contrast to the problem of sound change, the identification, the inference of cross-linguistically recurring polysemies can be rather straightforwardly done, by avoiding any distinction between polysemy and homophony in a first place, and then searching for those patterns which recur often enough in big colexification networks. Colexification networks as proposed in the CLICS³ database, however, do not solve all problems. First of all, they are a convenient way to present the data to linguists who are interested in the investigation of polysemy patterns due to their individual research. The colexification data as it was assembled with help of our improved CLDF data curation workflows, however, offer much more potential for future investigations. This is shown, for example, by Gast and Koptjevskaja-Tamm (2018) who study areal aspects of polysemy patterns, as well as by (Georgakopoulos and Polis 2018), who present new ideas to add a diachronic dimension. Additionally, there is a lot of potential for studies that *use* the colexification data in order to check linguistic, cognitive, and psychological theories and hypotheses.

What theories could, for example, be tested, with the help of polysemy patterns?

Lexibank and CL Toolkit

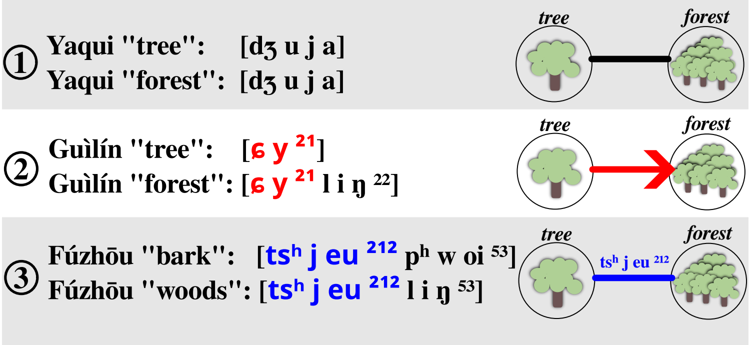
With the publication of the Lexibank database, we have shown how both phonological and lexical features can be automatically extracted from large aggregated collections of CLDF word lists (List et al. 2022a). For 30 exemplary lexical features, we also illustrate how they can be computed with the help of CL Toolkit (List and Forkel 2021), a package that facilitates the representation of features in code. All 30 lexical features defined in this form are based on *colexifications*, but not all features are based on *full colexifications*, but we also look for two types of partial colexifications, one based on the identification of common *substrings*, and one based on the identification of *part-of* relations (called *affix colexifications* in our study). This technique allows us to define individual colexification patterns and then search for them directly in the data in order to see how many languages show these patterns, and how many languages do not show them.

The figure shows the affix colexification for words for EYE being in an affix relation with words for TEAR and words for WATER being in an affix relation with words for TEAR in the Lexibank sample of languages. What do we find regarding the distribution of languages showing both patterns?

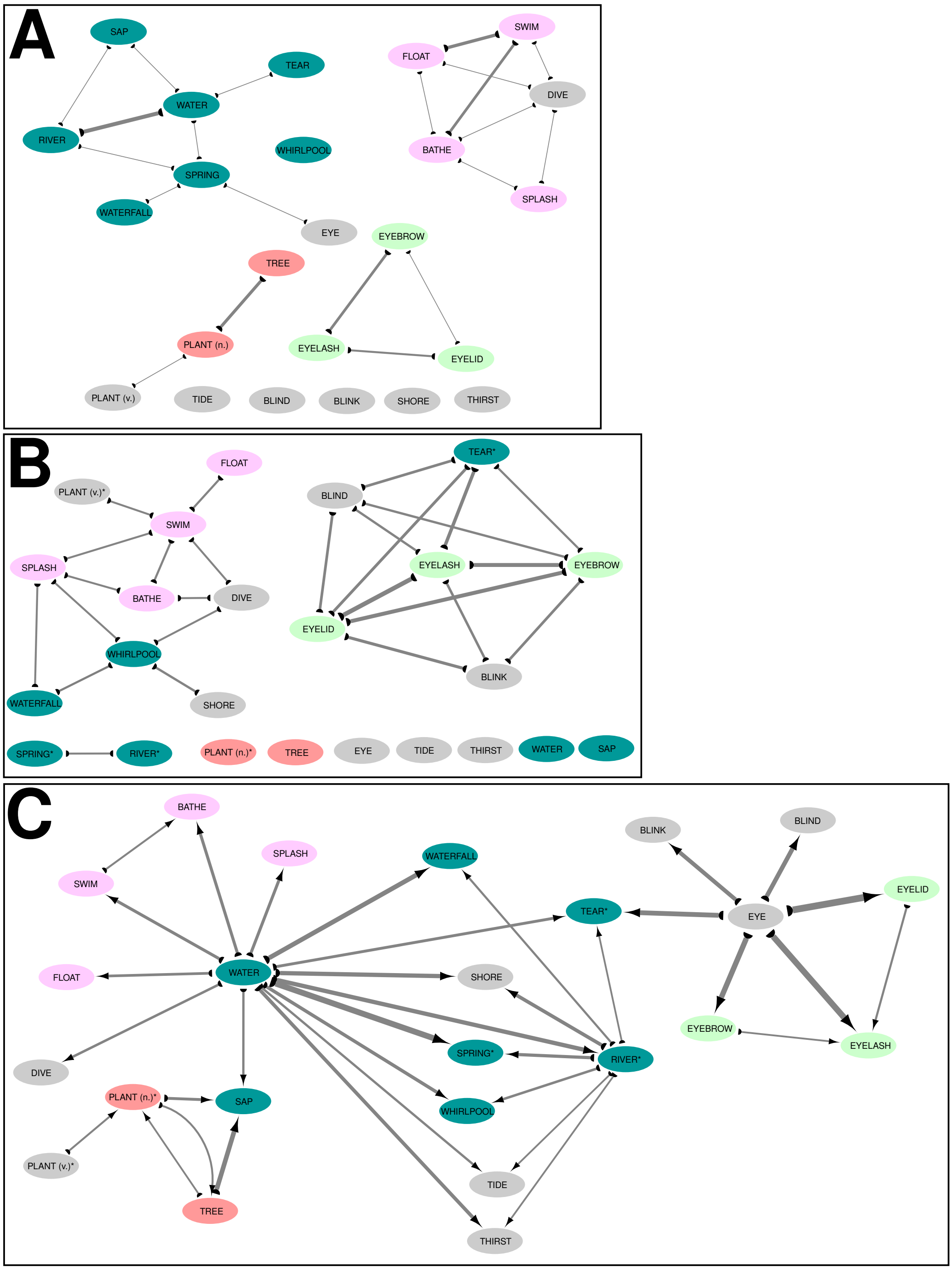

eye
ta:¹
water
na:m⁶
ids
Chiang Mai Khamuang
tears
na:m⁶-ta:¹
EYE partially colexified in TEAR
EYE not partially colexified in TEAR
missing data
WATER partially colexified in TEAR
WATER not partially colexified in TEAR
missing data

From CLICS to CLIPS

In a study under review (List 2023), we go one step further in trying to derive three kinds of colexification networks, including full, overlap, and affix colexifications from CLDF wordlists. While traditional colexification networks have been defined and used for a long time now (specifically as part of CLICS¹, CLICS² and CLICS³), the new pilot study defines two kinds of partial colexifications, following the earlier relations proposed in List et al. (2022a), by defining two specific kinds of partial colexifications, namely part-of relations and substring relations. While part-of relations should be modeled in directed networks, with the direction indicating what word is part of the other word, substring relations should be modeled in undirected networks, analogously to "full" colexification networks. To keep computation time at a reasonable level, the study introduces specific subtypes of part-of and substring relations: the affix relation (one word must be either a prefix or a suffix of the other word) and the overlap relation (two words can share a substring, but the substring must be either a prefix or a suffix in both strings).


① Yaqui "tree": [dʒ u j a]
Yaqui "forest": [dʒ u j a]
tree
forest
② Guìlín "tree": [ɕ y ²¹]
Guìlín "forest": [ɕ y ²¹ l i ŋ ²²]
tree
forest
③ Fúzhōu "bark": [tsʰ j eu ²¹² pʰ w oi ⁵³]
Fúzhōu "woods": [tsʰ j eu ²¹² l i ŋ ⁵³]
tree
tsʰ j eu ²¹²
forest

The results indicate that all three types of colexification networks are fundamentally different, while they are still semantically meaningful. Moreover, when modeling affix colexifications, we find that the weighted in-degree of these colexification networks correlates moderately (0.42, $r < 0.0001$) with the weighted degree of overlap colexification networks, while the weighted out-degree correlates moderately (0.50, $r < 0.0001$) with the weighted degree of full colexification networks (using Spearman rank correlations, Spearman 1904). These findings can be interpreted in such a way that they point to the tendency that concepts which are generally colexified very often are also frequently re-used as compounds or affixes in complex words. This shows that one could take the out-degree of affix colexifications as evidence for the phenomenon of *lexical root productivity* (the term is inspired by a discussion with Alexandre François, see List 2019a and List 2019b).



The correlation between the in-degree in affix colexification networks, that means, the tendency of words to re-appear in compounds, and the degree of overlap colexification networks, that means, the tendency of concepts to be expressed by a compound word, is not very surprising. It shows, however, that overlap colexifications can be used to compute the *compoundhood* of concepts, a property, that has only rarely been investigated for a larger number of languages.

Can we "see" the differences with respect to the in-degree and the out-degree of affix colexification networks in the figure (C) above when comparing them with full colexifications (A) and overlap colexifications (B)?

References

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

Bloomfield, L. (1933 [1973]). *Language.* London: Allen & Unwin.

Cysouw, M. (2010). "Semantic maps as metrics on meaning." *Linguistic Discovery* 8.1, 70–95.

Fox, A. (1995). *Linguistic reconstruction. An introduction to theory and method.* Oxford: Oxford University Press.

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.

Gast, V. and M. Koptjevskaja-Tamm (2018). "The areal factor in lexical typology. Some evidence from lexical databases." In: *Aspects of linguistic variation.* Ed. by D. Olmen, T. Mortelmans, and F. Brisard. Berlin and New York: de Gruyter, 43–81.

Georgakopoulos, T. and S. Polis (2018). "The semantic map model: State of the art and future avenues for linguistic research." *Language and Linguistics Compass* 12.2. e12270 LNCO-0727.R1, e12270–n/a.

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.

Hammarström, H., M. Haspelmath, R. Forkel, and S. Bank (2021). *Glottolog. Version 4.4.* Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: `https://glottolog.org`.

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

— (2019a). "Open problems in computational diversity linguistics: Conclusion and Outlook." *The Genealogical World of Phylogenetic Networks* 6.12.

— (2019b). "Typology of semantic promiscuity (Open problems in computational diversity linguistics 10)." *The Genealogical World of Phylogenetic Networks* 6.11.

— (2023). *Inference of Partial Colexifications from Multilingual Wordlists.*

List, J.-M. and R. Forkel (2021). *CL Toolkit. A Python Library for the Processing of Cross-Linguistic Data [Software Library, Version 0.1.1].* Geneva: Zenodo.

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

List, J.-M., S. J. Greenhill, C. Anderson, T. Mayer, T. Tresoldi, and R. Forkel (2018). "CLICS². An improved database of cross-linguistic colexifications assembling lexical data with help of cross-linguistic data formats." *Linguistic Typology* 22.2, 277–306.

List, J.-M., T. Mayer, A. Terhalle, and M. Urban (2014). *CLICS: Database of Cross-Linguistic Colexifications. Version 1.0.* Marburg: Forschungszentrum Deutscher Sprachatlas. URL: `https://lingpy.org/clics/`.

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.

List, J.-M., A. Tjuka, C. Rzymski, S. J. Greenhill, and R. Forkel (2022b). *CLLD Concepticon [Dataset, Version 3.0.0].* Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: `https://concepticon.clld.org/`.

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.

Newman, M. E. J. (2004). "Analysis of weighted networks." *Physical Review E* 70.5, 056131.

Paul, H. (1880 [1886]). *Principien der Sprachgeschichte.* 2nd ed. Halle: Max Niemeyer. prinzipiendersp01paulgoog: `ia`.

Rzymski, C. et al. (2020). "The Database of Cross-Linguistic Colexifications, reproducible analysis of cross- linguistic polysemies." *Scientific Data* 7.13, 1–12. URL: `https://clics.clld.org`.

Spearman, C. (1904). "The Proof and Measurement of Association between Two Things." *The American Journal of Psychology* 15.1, 72–101.

Sperber, H. (1923). *Einführung in die Bedeutungslehre.* Bonn and Leipzig: Kurt Schroeder.

Traugott, E. C. (2012). "Pragmatics and language change." In: 249–565.

Wilkins, D. P. (1996). "Natural tendencies of semantic change and the search for cognates." In: *The comparative method reviewed. Regularity and irregularity in language change. The comparative method reviewed. Regularity and irregularity in language change.* Ed. by M. Durie. With an intro. by M. D. Ross and M. Durie. New York: Oxford University Press, 264–304.

Zalizniak, A. A. (2018). "The Catalogue of Semantic Shifts: 20 years later." *Russian Journal of Linguistics* 22.4, 770–787.

Zalizniak, A. A., M. Bulakh, D. Ganenkov, I. Gruntov, T. Maisak, and M. Russo (2012). "The catalogue of semantic shifts as a database for lexical semantic typology." *Linguistics* 50.3, 633–669.

Workflow Development and Testing

Johann-Mattis List (University of Passau)

1 Questions for the Practice Session

Do sequences and sequence comparison play an important role in your research?

Do networks play an important role in your research?

Do the research objects with which you work *evolve* in any way?

Does your research involve unordered collections of distinct objects (sets)?

2 Tasks for the Practice Session

Try to construct sequential objects of the research objects you use in your work (even if this does not seem to make sense) and make a plan of comparing them using alignments or similar methods.

Try to construct network objects of the research objects you use in your work (even if this does not seem to make sense) and make a plan of identifying the dynamics underlying the research objects (how they interact, if one can transition into the other, etc.).

Make a concrete research plan for the comparison of various research objects relevant for your study with the help of computational methods, by which the differences between the objects can be displayed in a distance matrix or a similarity network.

Chinese Computational Linguistics

Johann-Mattis List (University of Passau)

1 Background

There are many myths regarding the Chinese language or the Chinese language**s**, and without having had proper insights into both the language structure and the writing system that is used to write the language, it may be difficult to assert whether something belongs to the realm of myths or the realm of known facts. There is not enough time to discuss all myths and facts in one session, but we will try to quickly look at the grammatical structure of Chinese and at the role that dialectal variation play in the context of the language.

What myths about Chinese do you know?

Grammatical Structure of Chinese

As a language, the major characteristics of Chinese are its *isolating* structure, reflected in the quote below by the Sergey Yakhontov:

> In Chinese, grammatical relations among words in a sentence are expressed by word order or by the use of specific function words, for example, prepositions, but not by modifying the word forms themselves. (Jachontov 1965: 12 [1])

What is meant by this structure can be easily seen when inspecting examples in interlinear-glossed text.

(1)

我	爸爸	不	在
wǒ	bàba	bú	zài
I	father	not	be present

„My father is not here."

(2)

我	会	告诉	他	的
wǒ	hùi	gàosu	tā	de
I	function verb	tell	he	*particle*

„I shall tell him."

(3)

我	以前	在	柏林	学习
wǒ	yǐqián	zài	Bólín	xuéxí
I	earlier times	be present	Berlin	study

„I used to study in Berlin."

What is remarkable with respect to parts of speech in the second example sentence above?

[1] My translation, original text: «В китайском языке грамматические отношения между словами в предложении выражаются порядком их расположения, а также специальными служебными словами, например предлогами, но не изменением формы слов.»

Linguistic Variation

According to the official definition of *pǔtōnghuà* 普通话, the variety often called *Mandarin Chinese* takes the variety of Běijīng as its phonetic basis (*yǐ Běijīng yǔyīn wéi biāozhǔn* 以北京语音为标准), while following the classical texts composed in *Báihuà* 白话 (*yǐ diǎnfàn de báihuàwén zhùzuò wéi yǔfǎ guīfàn* 以典范的白话文著作为语法规范) grammatically (Huáng and Liào 2002: 4). In practice, however, the Standard Chinese is spoken with many flavors by the multitude of people who speak different dialects of Chinese as their first language. In the last decades, we can see a rise of proficiency in Standard Chinese among younger people, accompanied by a drastic loss of dialectal variation. Nevertheless, it is problematic to speak of Chinese as one single language without knowing about the specific sociolinguistic situation in which this “language” is realized. In practice, however, linguists still tend to talk about the Chinese language as a certain kind of unity. The reason that justify to treat Chinese with all its Sinitic varieties as one unit is the sociolinguistic situation in which many distinct varieties share a common history, a common writing system, and a common “roof language” that is used to communicate across the individual dialectal varieties.

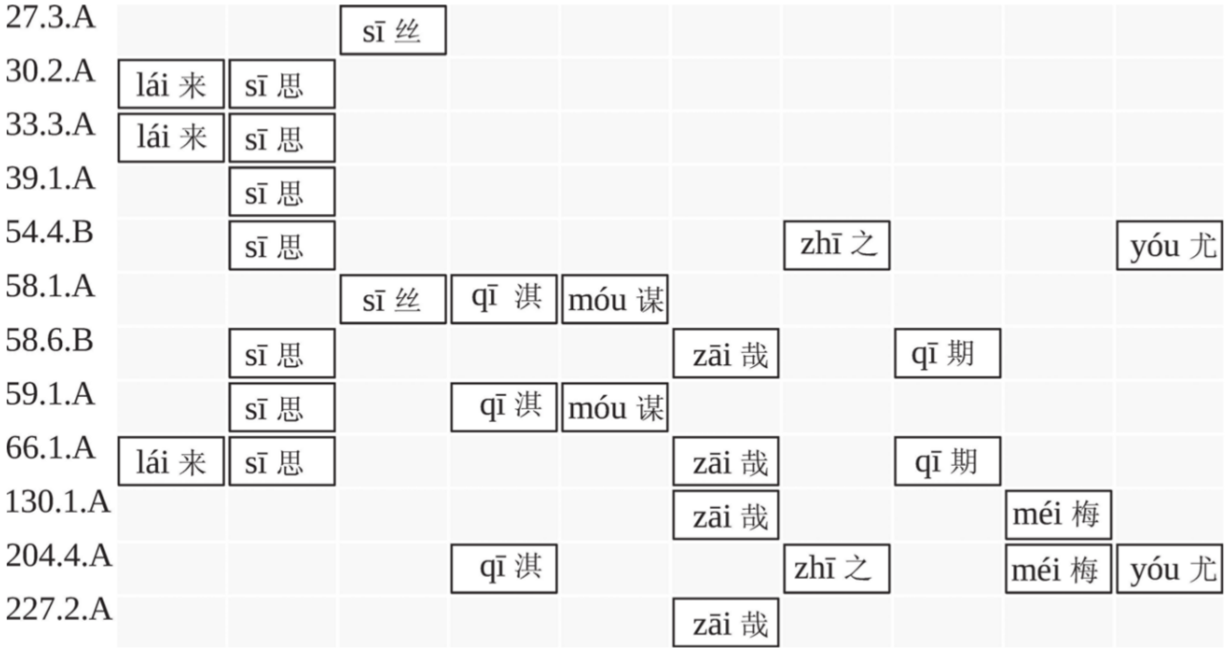
What differences and similarities can we find in the sociolinguistic context of language in China with the sociolinguistic context of language in Europe?

2 Rhyme Analysis

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 practiced by the ancestors (as advocated by Lù Démíng 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 (ibid.: 153-157). This is illustrated in Table 1.

Chinese Text	Translation	RW	Patterns	MCH	OCBS-Rhyme
燕燕於飛	The swallows go flying	*fēi* 飛	A	*pjɨj	*-ər
下上其音	falling and rising are their voices;	*yīn* 音	B	*ʔim	*-əm
之子於歸	This young lady goes to her new home,	*guī* 歸	A	*kjwɨj	*-əj
遠送於南	far I accompany her to the south.	*nán* 南	B	*nom	*-əm
瞻望弗及	I gaze after her, can no longer see her,	[*jí* 及]	–	[**gip*]	[**-əp*]
實勞我心	truly it grieves my heart	*xīn* 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 estab- lish 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’ (Gēng 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 com- pared 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 ide- ally do not show any more transitions among each other. This approach is essentially similar to the ‘linking method’ *xìlián 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.

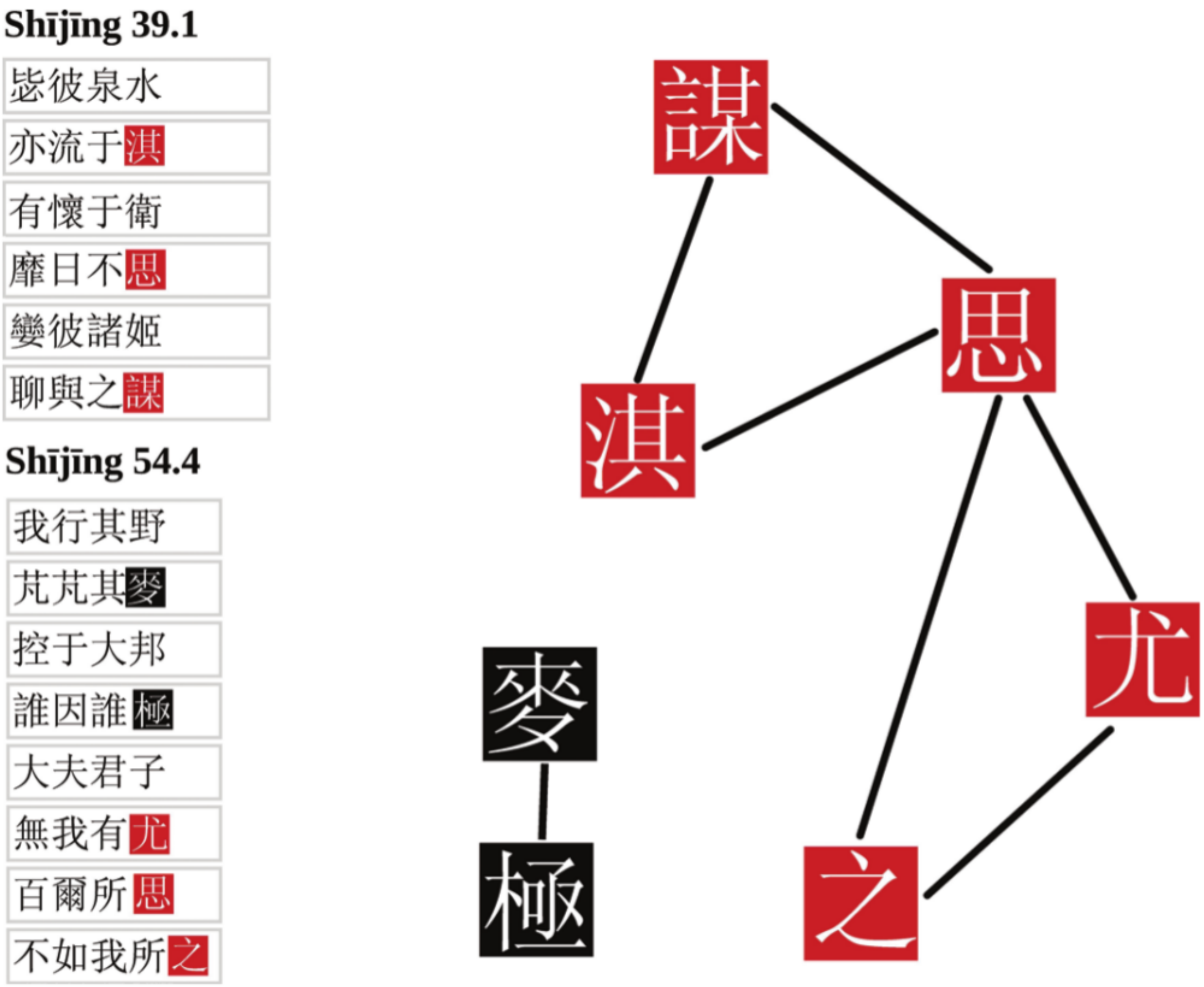


27.3.A			sī 丝							
30.2.A	lái 来	sī 思								
33.3.A	lái 来	sī 思								
39.1.A		sī 思								
54.4.B		sī 思					zhī 之			yóu 尤
58.1.A			sī 丝	qī 淇	móu 谋					
58.6.B		sī 思				zāi 哉		qī 期		
59.1.A		sī 思		qī 淇	móu 谋					
66.1.A	lái 来	sī 思				zāi 哉		qī 期		
130.1.A						zāi 哉			méi 梅	
204.4.A				qī 淇			zhī 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

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

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 Shījīng:



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. (2017) 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?

The Shījīng 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 Shījīng 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 2017). 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	mɯɯl	m^{ʕ}əj	muəi	mə̄j	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,伐其條肄	2.AB	sijH	ph-ljɯds	s-ləp-s	jiet	slhəps	B	质
既見君子,不我遐棄	2.CD	khjijH	khids	[k^{h}]i[t]-s	khiet	khijs	B	质
魴魚赬尾	3.A	mj+jX	mɯlʔ	[m]əjʔ	miuəi	məjʔ	C	微
王室如燬	3.B	xjweX	qh^{w}ralʔ	[m̥](r)ajʔ	xiuəi	h^{w}ejʔ	C	微
雖則如燬	3.C	xjweX	qh^{w}ralʔ	[m̥](r)ajʔ	xiuəi	h^{w}ejʔ	C	微
父母孔邇	3.D	nyeX	mljelʔ	n[ə][r]ʔ	njiei	n(h)ejʔ	C	脂

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

3 Character Analysis

The Chinese writing system, as we know it today, is famous for its structural properties reflected by a complicated interaction of phonetic and semantic elements.

Chinese characters can be divided into elements carrying phonetic function and in elements carrying semantic functions. As a result, scholars tend to call it a "semanto-phonetic writing system" (*yìyī wénzì* 意音文字, cf. Zhōu 1998: 60. But this characterization exaggerates the potential of character elements to *predict* a certain pronunciation or meaning.

> Most of the "phonetic" characteristics of the [Chinese writing system] are relics of the processes of character formation which, as they took place asynchronously, were always characterized by a complex interaction between the Chinese language spoken at different times of its history, the socio-cultural background of those people who created the characters, and general patterns of reasoning and conceptualization. (List et al. 2016: 49)

Thus, although people often *say* that the Chinese characters have phonetic and semantic characteristics, from the perspective of their potential, these aspects are very limited, since the predictive force is extremely limited. The reason lies in the fact that the Chinese writing system has been derived at different stages, similar to the way in which the lexicon of human languages shows different layers of transparency and opacity with respect to the motivation of individual lexemes. What can be said for sure is that – in the majority of all cases – one Chinese character expresses one morpheme of the Chinese language and that the morpheme is usually one syllable in length.

Below is a quote that defines *motivation* in linguistics. What are the reasons that motivation is lost in the lexicon of human languages, and what consequences would this have for the Chinese writing system?

> **Motivation**: Extent, to which the [complex word] can be understood as the result of its parts and their composition. (Glück 2000: s.v. ”Motivation”)[2]

Phonetic Elements in Chinese Writing

Chinese characters were developed over millennia and their formation (*zàozìfǎ* 造字法, Qiú 1988 [2007]) is best seen as a derivational process with striking similarities to word formation processes (Kunze 1937, List 2008).

> This derivational process applies specifically to the phonetic characteristics of the writing system, as reflected in the category of *xiéshēng* 諧聲 characters, which consist of one element that hints at the pronunciation of the word encoded by the character (the phonophoric determinative), and one element that hints at the word’s meaning (the semantic determinative) [...]. For example, the character 被, which writes the word *bjeX* ‘cover oneself with’ is composed of the phonophoric determinative 皮 , which as a character itself represents the word bje ‘skin’, and the semantic determinative 衤 , a contracted version of 衣 *’jɨj* ‘clothes’. (Hill and List 2019: 186)

Chinese characters thus show some degree of *recursion*: phonetic elements can themselves consist of complex characters and contain formerly transparent semantic and phonetic elements.

Is recursion the correct term to describe the derived character of phonetic elements in the Chinese writing system?

Network Analysis of Phonetic Elements

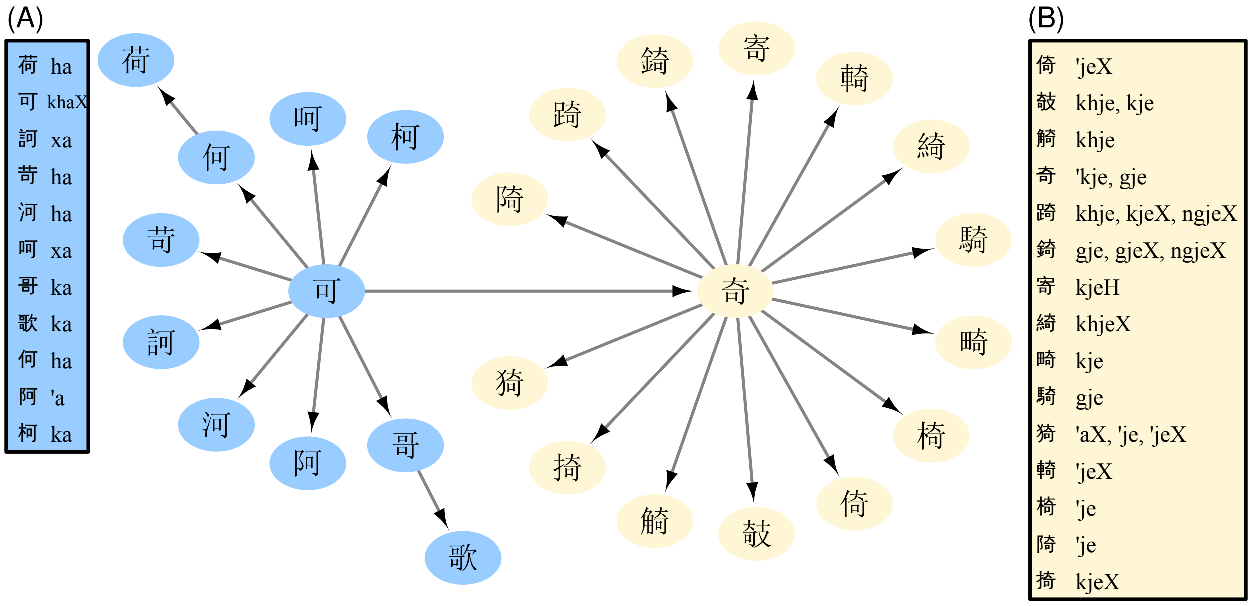
The parallel between word formation and Chinese writing can be used as a source of inspiration for the modeling of Chinese character formation processes.

> A crucial aspect of word formation (and also of character formation) is the hierarchical process by which words are derived from each other at different times. If we have a compound word, like German Krankheitsverlauf ‘disease progression’, we can recursively split the word into its respective components which usually were coined at different moments in history. (ibid.: 190)

These components can be modeled with the help of directed networks. The benefits of this approach is that the organization of phonetic elements in the Chinese writing system can be made much more transparent than it has been done in previous work, where scholars would assign individual characters to monolithic clusters (Karlgren 1957). In our pilot study (Hill and List 2019), we show the benefits of this approach by testing concrete hypotheses on the pronunciation of specific characters in Old Chinese.

How can the different phonetic realizations of the yellow characters in the table (B) be explained linguistically?

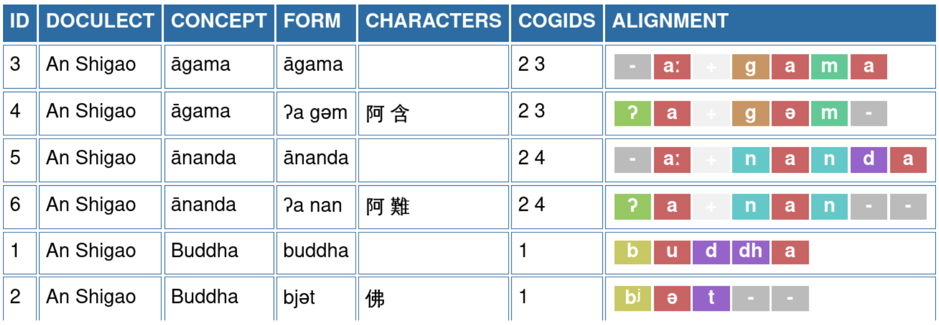
[2] My translation, original text: “**Motivation**: Ausmaß, in dem [das komplexe Wort] sich als Summe seiner Teile und der Weise ihrer Zusammenfügung verstehen lässt”.



4 Further Topics

The field of Chinese Historical Phonology is full of interesting topics that have not yet been sufficiently analyzed with the help of computational methods. While scholars have always applied some quantitative approaches, counting occurrences of examples in texts, or making quantitative analyses on sheets of paper, discussing them in prose, a rigorously formal treatment of many topics is still missing. In a forthcoming study (List forthcoming), I discuss three topics that could further advance the field of Computational Chinese Historical Phonology in the future, namely *corpus studies*, specifically studies that inspect the description of character pronunciations in Chinese texts, *additional network approaches*, specifically those that target the description of meanings in tags, as it can be frequently found in Chinese literature, and finally *alignment analyses* that could help to make the transcription of foreign names with the help of Chinese characters more transparent.

The figure below shows an example for the alignment of transcriptions of Buddhist terms in Chinese. So far, these transcriptions have only been analyzed in a manual fashion. Which chances and which challenges do alignment analyses bear for the investigation of transcriptions and transliterations and what alternative methods could be used?



ID	DOCULECT	CONCEPT	FORM	CHARACTERS	COGIDS	ALIGNMENT
3	An Shigao	āgama	āgama		2 3	- aː + g a m a
4	An Shigao	āgama	ʔa gəm	阿 含	2 3	ʔ a + g ə m -
5	An Shigao	ānanda	ānanda		2 4	- aː + n a n d a
6	An Shigao	ānanda	ʔa nan	阿 難	2 4	ʔ a + n a n - -
1	An Shigao	Buddha	buddha		1	b u d dh a
2	An Shigao	Buddha	bjət	佛	1	bʲ ə t - -

References

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.

Glück, H., ed. (2000). *Metzler-Lexikon Sprache*. 2nd ed. Stuttgart: Metzler.

Gēng, Z. 耿振生. (2004). *20 shìjì Hànyǔ yǔyīnxué fāngfǎ lùn* 20世纪汉语音韵学方法论 [20th centuryś methods in traditional Chinese phonology]. Běijīng 北京: Běijīng Dàxué 北京大學.

Hill, N. W. and J.-M. List (2019). “Using Chinese character formation graphs to test proposals in Chinese historical phonology.” *Bulletin of Chinese Linguistics* 12.2, 186–200.

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.

Huáng, B. and X. Liào (2002). *Xiàndài Hànyǔ* 现代汉语 [Modern Chinese]. 3rd ed. Vol. 1. 2 vols. Běijīng: Gāoděng Jiàoyù.

Jachontov, S. E. (1965). *Drevnekitajskij jazyk* [Old Chinese]. Moscow: Nauka.

Karlgren, B. (1957). “Grammata serica recensa.” *Bulletin of the Museum of Far Eastern Antiquities* 29, 1–332.

Kunze, R. (1937). *Bau und Anordnung der chinesischen Zeichen. Oder: Wie lernen wir leichter Zeichen lesen?* [Structure and assembly of Chinese characters. Or: How can we learn to read characters more easily?] Tokyo: Deutsche Gesellschaft für Natur und Völkerkunde Ostasiens.

List, J.-M. (2008). “Rekonstruktion der Aussprache des Mittel- und Altchinesischen. Vergleich der Rekonstruktionsmethoden der indogermanischen und der chinesischen Sprachwissenschaft [Reconstruction of the pronunciation of Middle and Old Chinese. Comparison of reconstruction methods in Indo-European and Chinese linguistics].” Magister thesis. Berlin: Freie Universität Berlin. PDF: `http://hprints.org/docs/00/74/25/52/PDF/list-2008-magisterarbeit.pdf`.

— (2017). *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.

— (forthcoming). “Chances and challenges for quantitative approaches in Chinese Historical Phonology.” *Bulletin of Chinese Linguistics* 0.0, 1–19.

List, J.-M., J. S. Pathmanathan, N. W. Hill, E. Bapteste, and P. Lopez (2017). “Vowel purity and rhyme evidence in Old Chinese reconstruction.” *Lingua Sinica* 3.1, 1–17.

List, J.-M., A. Terhalle, and D. Schulzek (2016). “Traces of embodiment in Chinese character formation. A frame approach to the interaction of writing, speaking, and meaning.” In: *Sensory-motor concepts. At the crossroad between language & cognition*. Ed. by L. Ströbel. Düsseldorf: Düsseldorf University Press, 45–62.

Liú, X. 刘晓南. (2006). *Hànyǔ yīnyùn yánjiū jiàochéng* 汉语音韵研究教程 [Reader in traditional Chinese phonology]. Běijīng 北京: Běijīng Dàxué 北京大學.

Lǚ, S. 吕胜男. (2009). “A brief study of the methodology of the study of ancient rhyme. And Concurrently on the study of the rhyme of “Jinwen Shangshu” 古韵研究方法论发微. 兼论今文《尚书》用韵研究 [History of ancient Chinese linguistics].” *Nányáng Shīfàn Dàxué Bào (Shèhuì Kēxué Bǎn)* 南阳师范学院学报(社会科学版) *[Journal of Nanyang Normal University (Social Sciences)]* 8.2, 57–61.

Pān, W. 潘悟云. (2000). *Hànyǔ lìshǐ yīnyùnxué* 汉语历史音韵学 [Chinese historical phonology]. Shànghǎi 上海: Shànghǎi Jiàoyù 上海教育.

Qiú, X. 裘錫圭. (1988 [2007]). *Wénzìxué gàiyào* 文字學概要 [Foundations of graphemics]. Běijīng: Shāngwù 商务.

Wáng, L. 王力. (1980 [2006]). *Hànyǔ shǐgǎo* 漢語史稿 [History of the Chinese language]. Repr. Běijīng 北京: Zhōnghuá Shūjú 中华书局.

Zhōu, Y. 周有光. (1998). *Bǐjiào Wénzìxué Chūtàn* 比较文字学初探 [Introductory investigations on the comparison of writing systems]. Yǔwén 語文.

Computer-Assisted Text Analysis

Johann-Mattis List (University of Passau)

1 Interlinear-Glossed Text

Linguists create linguistic resources with very specific purposes in mind. As a result, a resource can often only be used to address one specific question, although – if it had been carefully designed – it could be used for many additional analyses as well. We have seen enough examples of this lack of interest in the extensibility of resources, or the lack of *integration* as well as the efforts by the CLDF initiative (Forkel et al. 2018) to address these problems. This problem can also be found when dealing with *interlinear-glossed text*.

> Although annotation tools exist [...] their application is difficult due to a lack of cross-platform support [...], but also by a large degree of freedom offered by the respective software. Since the majority of IGT is still produced in research articles, and not in the form of standardized databases, errors in the glossing procedure are still rather common [...]. (List et al. 2021: 4/15)

We have tried to address these problems by providing a first framework that shows how interlinear-glossed text can be handled in standardized CLDF formats and how a consistent integration of interlinear glossed text resources can help us to retrieve many additional aspects of language data that the original interlinear-glossed text collection might not have been created for initially (ibid.).

```
Die               Katze     sitz-t        auf   den               Matratze-n.
ARTIC.NM.SGL.F    cat       sit-3.SG.IND  on    ARTIC.DT.PLR.F    mattress-PLR
The cat sits on the mattresses.
```

(1)

Word	Gloss
Die	ARTIC.NM.SGL.F
Katze	cat
sitz-t	sit-3.SGL
auf	on
den	ARTIC.DT.PLR.F
Matratze-n	mattress-PLR

(2)

Morpheme	Lexical Gloss	Grammatical Gloss
Die		ART.NOM.SG.F
Katze	cat	
sitz	sit	
t		3.SG
auf	on	
den		ART.DAT.PL.F
Matratze	mattress	
n		PL

(3a)

Lex. Concept	Concepticon
cat	1208 CAT
sit	1416 SIT
on	1741 ABOVE
matress	105 MATTRESS

(3b)

Gram. Concept	Leipzig Glossing Rules
ARTIC	ART
NM	NOM
SGL	SG
PLR	PL
...	...

(4)

Word	CLTS Transcription
Die	d iː
Katze	k a ts ə
sitz-t	s ɪ ts + t
auf	au f
den	d eː n
Matratze-n	m a t r a ts ə + n

(5)

Word	Cognacy
d iː	1
k a ts ə	2
s ɪ ts + t	3 4
au f	5
d eː n	1
m a t r a ts ə + n	6 7

Our workflow for the creation of integrated resources from interlinear-glossed text is shown above. Is it realistic to achieve all of the workflow steps in a completely automatic manner?

2 Rhyme Analysis

Having seen that it pays off, in general, to work on computer-assisted approaches in those cases where large amounts of data have to be handled, we might want to step back a bit from the very specific question of Chinese Historical Phonology and the rhyming practice (discussed in the previous session), and rather ask what we could do if we had a large database of poetry, and what questions we would like to ask. Once we have determined this sufficiently, we should decide what kind of data we want to have. In fact, most of the questions have already been discussed in the first section of this session. The question that now remains for us is how we can actually handle world-wide data on poetry in such a way that we can address these questions? In the following, we will first look at how it is actually being done, and then develop an alternative framework from the problems we observe in the current practice.

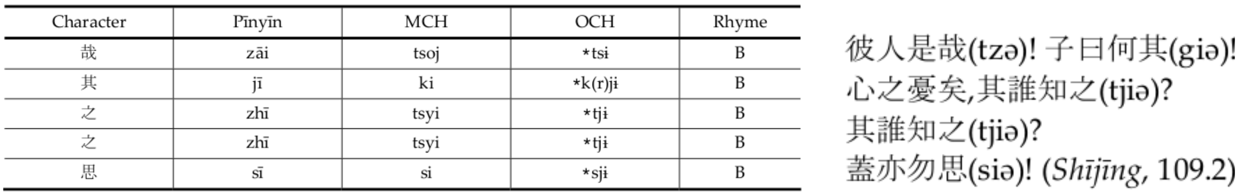
What specific questions would you like to ask about the evolution and typology of poetry?

Current Annotation Practice

When analyzing rhymes in poetry, one of the most crucial questions is what rhymes with what and where it rhymes. We can call such an analysis (which is a true analysis, since we may assume that experts commit errors in their assessment of either what the majority of language users think or what the author intended) a *rhyme judgment analysis*, similar to the term *cognate judgment*, which reflects the identification of potential cognate words by experts or algorithms. The ways in which scholars share their respective rhyme judgments in the literature is very diverse and makes a formal comparison of different rhyme analyses difficult. The problem here lies only to some degree in missing digital versions of important contributions, which would be merely a problem for pure computational approaches. A more significant problem is that many authors report their rhyme judgments in a form that is insufficiently explicit to infer the individual judgments made on individual poems and stanzas. Apart from scholars who presented only the results of their analyses, without providing the evidence, we also often find analyses that are extremely difficult to inspect, due to the way they present their judgments. In this sense, only a small amount of rhyme analyses is truly explicit. Among the few explicit rhyme analyses, we again face the problem that scholars differ widely in the formats they use for annotation, and also in the depth of annotation provided.

We have seen before that one can roughly distinguish between *inline* and *stand-off* annotation (Eckart 2012).[1] As an example illustrating the difference between the two annotation styles, consider the rhyme annotation employed by Baxter (Baxter 1992) as compared to the one by Wáng Lì Wáng 1980 [2006], for poem 109 (second part of stanza 2 in the Book of Odes). While Wáng Lì provides the rhyme judgements inline, Baxter (p. 625) basically uses a stand-off annotation by listing all relevant data in tabular form:

[1] While inline annotation manipulates the original data directly, for example, by adding tags, stand-off annotation only references the original data, without directly modifying it. Most annotation frameworks, however, typically use a mixture between the two types, although it is clear that stand-off annotation has the advantage of allowing for far more flexibility, especially if adding multiple layers of annotation to a given resource.



Character	Pīnyīn	MCH	OCH	Rhyme
哉	zāi	tsoj	*tsɨ	B
其	jī	ki	*k(r)jɨ	B
之	zhī	tsyi	*ʈjɨ	B
之	zhī	tsyi	*ʈjɨ	B
思	sī	si	*sjɨ	B

彼人是哉(tzə)! 子曰何其(giə)!
心之憂矣,其誰知之(tjiə)?
其誰知之(tjiə)?
蓋亦勿思(siə)! (*Shījīng*, 109.2)

In order to test their algorithm on automated rhyme detection, Haider and Kuhn (2018) uses a corpus in which poems are separated into stanzas, and stanzas are separated into lines, and rhyming is annotated by providing an attribute for each stanza, which reflects which line rhyme with which line, similar to the practice in school, using letters of the alphabet. What huge disadvantage has this system?

Preliminary Framework for Rhyme Annotation

Based on the discussions of the desiderata and past experiments which proved the particular insufficiency of certain annotation forms, the core annotation of a poem or a poem collection, as proposed in (List et al. 2017) now contains the following main components:

- ID: the identifier, which is a numerical ID.
- POEM: a name for the given poem.
- STANZA: the stanza of the poem (usually a numeric value, preceded by the name of the poem).
- LINE_IN_SOURCE: the line of the poem as we find it in the source from which the data is taken (especially containing original punctuation etc.).
- LINE: a double-segmented version of the line, in which words are separated with help of + as a separator, and spaces can be used to represent segments of phonetic values (similar to the format adopted by the LingPy software package to represent phonetic sequences and alignments).
- LINE_ORDER: A numerical value that provides the order of the lines of a poem in a given stanza.
- RHYMEIDS: A list of numerical identifiers, indicating which words in a the LINE rhyme by assigning the same ID to different words, using 0 to indicate that a given word does not rhyme.
- ALIGNMENT: A double-segmented version of the line that can, however, store aligned content, differing from the data in LINE, as well. This data comes in handy when trying to check questions of phonetic similarity of rhyme words, or of vowel purity, which would greatly facilitate automatic analyses as the one presented in List et al. (ibid.).

With these eight columns provided, poems can be annotated in a very straightforward way, regardless of the language in which they were written. One can, of course, add many more columns, depending on specific characteristics of the datasets, but for the general rhyme annotation, we think that these fields will be sufficient for most of the cases; it substantially exceeds rhyme annotation frameworks that have been proposed so far in terms of detail.

What is the obvious drawback of this annotation schema?

PoePy: Python Library for Quantitative Handling of Rhymes

We have developed a software API, called PoePy (`https://github.com/lingpy/poepy`), that allows one to parse, manipulate, and convert files following our new rhyme annotation schema in a convenient way, with help of the Python language. The framework builds heavily on LingPy, a Python library for quantitative tasks in historical linguistics (List and Forkel 2022), as well as SinoPy, a Python library for specialized tasks in Chinese historical lin- guistics (List 2018). The GitHub site of our API offers additional information for installing and using our software library. PoePy can read datasets in our general format mentioned above, it can also be used to align rhyme words, provided they are readily assigned to the data, and it can convert the data to different formats, that ease rhyme pattern inspection. Our stanza 2 from Ode 109 of the Shī- jīng, for example, can be rendered directly in the following tabular form, that greatly facilitates seeing the rhyme structure of the poem.

ID	STANZA	LINE	R:467	R:468
1733	109.2	園 有 **棘**	kiək	
1734	109.2	其 實 之 **食**	djiək	
1735	109.2	心 之 憂 矣		
1736	109.2	聊 以 行 **國**	kuək	
1737	109.2	不 我 知 者		
1738	109.2	謂 我 士 也 罔 **極**	qiək	
1739	109.2	彼 人 是 **哉**		tzə
1740	109.2	子 曰 何 **其**		giə
1742	109.2	其 誰 知 **之**		tjiə
1744	109.2	蓋 亦 勿 **思**		siə

How could the display be further enhanced?

Examples for Annotated Rhymes

As a first example, consider the first stanza of Bob Dylan's song "I want you" (from the album Blonde on Blonde, 1966). Here the rhyme patterns are more complex than in many other poems, but rhyming is in parts also more lax, with more imperfect rhymes, reflecting the typical style of Dylan's poetry.

ID	ST	LINE	R:1	R:2	R:3
1	1.1	The guilty undertaker *sighs*	`s - ai s`		
2	1.1	The lonesome organ grinder *cries*	`k r ai s`		
3	1.1	The silver saxophones *say*	`s - æi -`		
4	1.1	I should *refuse_you*		`r i f j uː s j uː`	
5	1.1	The cracked bells and washed-out *horns*			`h - ɔ r n s`
6	1.1	Blow into my face with *scorn,*			`s k ɔ r n -`
7	1.1	but it's not that way, I wasn't *born*			`b - ɔ r n -`
8	1.1	to *lose_you*		`- - - l uː s j uː`	

A further example is the song "Te doy una canción" by Silvio Rodriguez (from the album Mujeres, 1978), in which none of the three rhyme pairs which we have annotated in stanza 1.2 rhymes perfectly. One might thus assume that rhyming was generally not intended in this song, but we find a very similar pattern in stanza 1.4., and songs in which the words *tú* "you" and *luz* "light" co-occur in potential rhyming position are very frequent in Spanish songs. Our hope is, that with a growing body of datasets in this form, we may learn more about the difference between rhymes which are intended and rhymes which might occur simply by chance.

ID	ST	LINE	R:1	R:2	R:3
7	1.2	Te doy una canción si abro una *puerta*	puer ta		
8	1.2	Y de las sombras sales *tú*		tú	
9	1.2	Te doy una canción de *madrugada,*	madruga da		
10	1.2	Cuando más quiero tu *luz*		luz	
11	1.2	Te doy una canción cuando apareces			
12	1.2	El misterio del *amor*			a mor
13	1.2	Y si no apareces, no me importa:			
14	1.2	Yo te doy una *canción*			can ción

What complicates the problem of finding rhymes that occur by chance and rhymes that were intended by the authors?

RhyAnt: An Interactive Tool for Rhyme Annotation

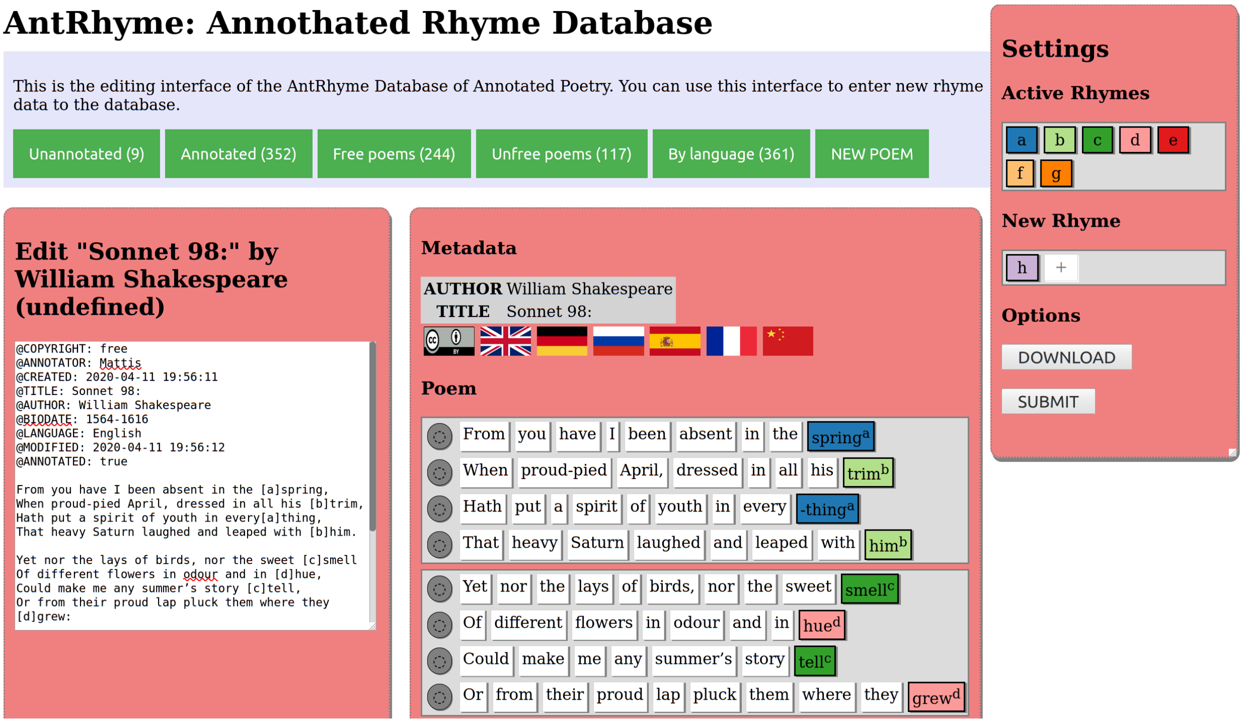
While I consider the inline-annotation format as rather complete by now (with all limitations resulting from inline-annotation), I realized, when trying to annotate poems by using the format, that it is no fun to edit text files in this way. I do not talk about small edits, like one stanza, or typing in some metadata, but annotating a whole rap song can become very tedious and even problematic, as one may easily forget which rhyme tags one already used, or oversee which words have been annotated as rhyming, or forget brackets and the like.

As a result I decided to write an interactive rhyme annotation tool which supports the inline-annotation format and can be edited both in the text and interactively at the same time, a bit similar to the text processing programs in blogging software which allow to write both in the HTML source and in a more convenient version that shows you what you will get.

This tool is now already online available (`https://digling.org/rhyant`, List 2020d). I call it RhyAnT, which is short for *Rhyme Annotation Tool*, and I have been using the tool in combination with a small server to populate a first database with rhymes in different languages that contains by now already more than 400 annotated poems (*AnTRhyme*, `https://digling.org/antrhyme`). This database can be accessed and inspected by everybody interested from its URL, but copyrighted texts from modern songs can – unfortunately – not be rendered by now (as I am not sure how many lines of them I would be allowed to share).

RhyAnt is freely available in the form of an interactive web application written in HTML/CSS and JavaScript. It can be used by opening the website `https://digling.org/calc/rhyant/`, or by downloading the code and opening the website offline with the help of a web browser. The tool is curated on GitHub, where it can also be downloaded (*https://github.com/digling/rhyant/*).

Is it a good idea to work with text that is not phonetically transcribed here?



AntRhyme: Annothated Rhyme Database

This is the editing interface of the AntRhyme Database of Annotated Poetry. You can use this interface to enter new rhyme data to the database.

Edit "Sonnet 98:" by William Shakespeare (undefined)

```
@COPYRIGHT: free
@ANNOTATOR: Mattis
@CREATED: 2020-04-11 19:56:11
@TITLE: Sonnet 98:
@AUTHOR: William Shakespeare
@BIODATE: 1564-1616
@LANGUAGE: English
@MODIFIED: 2020-04-11 19:56:12
@ANNOTATED: true

From you have I been absent in the [a]spring,
When proud-pied April, dressed in all his [b]trim,
Hath put a spirit of youth in every[a]thing,
That heavy Saturn laughed and leaped with [b]him.

Yet nor the lays of birds, nor the sweet [c]smell
Of different flowers in odour and in [d]hue,
Could make me any summer’s story [c]tell,
Or from their proud lap pluck them where they
[d]grew:
```

Metadata

AUTHOR William Shakespeare
TITLE Sonnet 98:

Poem

From you have I been absent in the spring[a]
When proud-pied April, dressed in all his trim[b]
Hath put a spirit of youth in every -thing[a]
That heavy Saturn laughed and leaped with him[b]

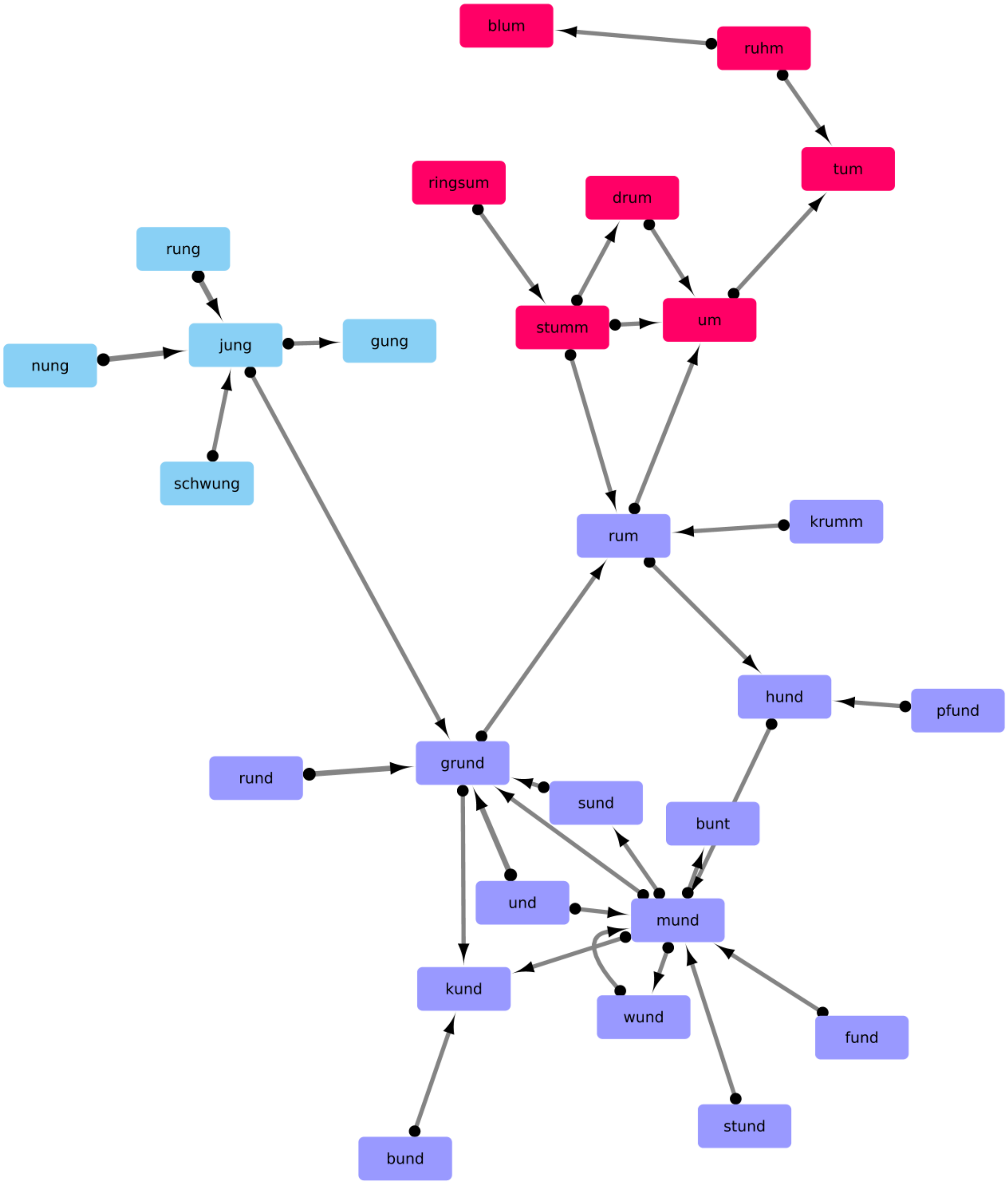
Yet nor the lays of birds, nor the sweet smell[c]
Of different flowers in odour and in hue[d]
Could make me any summer’s story tell[c]
Or from their proud lap pluck them where they grew[d]

An Initial Analysis of Rhymes in AntRhyme

While the analysis of rhyme networks in Chinese can be considered as quite advanced now, with quite a few examples having been published by now and active and in part ongoing discussions (see Wáng 2020 and List 2020c), network approaches to rhyming in languages other than Chinese are lacking. The reason can be found not only in the lack of data (rhymes are abundantly available for many languages in the world), but rather in the fact that the modeling of rhyme patterns needs to be advanced and that the inference of patterns cannot be done in the same straightforward way as it is done in Chinese, where one can naively assume that the last word of a row in a stanza always rhymes (BALEY 2022).

However, with the help of RhyAnT as an annotation tool for rhymes and with the extended more detailed schemas for rhyme annotation introduced along with RhyAnT and before (List et al. 2019), initial analyses of rhymes in German can be made. As an example and a proof-of-concept, I published a concrete example on rhyming in German, based on the small AntRhyme corpus, in which rhyme patterns are manually annotated (List 2020a, List 2020b).

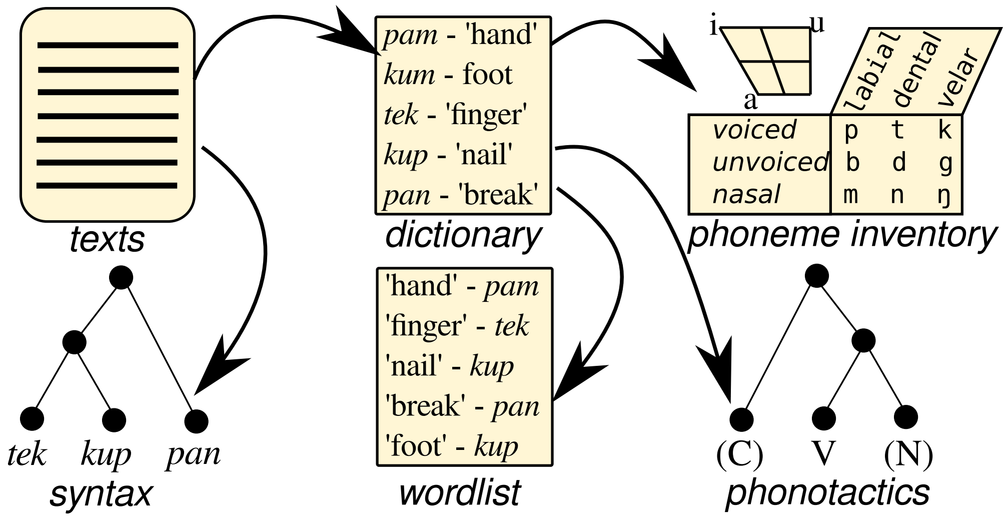
The rhyme network below is taken from the analysis of the AntRhyme corpus. What normalizations have been carried out in order to make the data comparable?



3 Outlook

There are many more topics that can be investigated in the future. For both the topic of interlinear-glossed text and the topic of general rhyme analysis, there is some hope that we can further advance them by increasing the availability of corpus data in standardized form. For interlinear-glossed text, concrete examples are currently being developed as part of CLDFViz (Forkel 2021, with new modules handling various forms of text in MarkDown, developed by R. Forkel, see `https://github.com/cldf/cldfviz`) and `pylingdocs` (Matter 2022, see `https://github.com/fmatter/pylingdo` Apart from this, I hope that we will manage to provide many more examples for *integrated* data, that is, data that does not only provide one type of information, but includes multiple types of information on various aspects of the same language variety which are interdependent and interlinked or even automatically derived from each other. In addition, we hope to be able to provide CLDF examples for all kinds of data used in rhyme analysis and in Chinese Historical Phonology.

Our integration goal for linguistic data is described below in the graphic. Is integration also important for your specific research topic?



References

BALEY, J. (2022). "Leveraging graph algorithms to speed up the annotation of large rhymed corpora." *Cahiers de Linguistique Asie Orientale* 51.1, 46 –80.

Baxter, W. H. (1992). *A handbook of Old Chinese phonology*. Berlin: de Gruyter.

Eckart, K. (2012). "Resource annotations." In: ed. by A. Clarin-D. Berlin: DWDS, 30–42.

Forkel, R. (2021). "CLDFViz. A python library providing tools to visualize data from CLDF datasets [Software Library, Version 0.5.0]."

Forkel, R., J.-M. List, S. J. Greenhill, C. Rzymski, S. Bank, M. Cysouw, H. Hammarström, M. Haspelmath, G. A. Kaiping, and R. D. Gray (2018). "Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics." *Scientific Data* 5.180205, 1–10.

Haider, T. and J. Kuhn (2018). "Supervised rhyme detection with Siamese recurrent networks." In: *Proceedings of Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. (Santa Fe, 08/25/2018), 81–86.

List, J.-M. (2018). *SinoPy: A Python library for quantitative tasks in Chinese historical linguistics*. Version 0.3.1. URL: `https://github.com/lingpy/sinopy`.

— (08/24/2020a). "Analyzing rhyme networks (From rhymes to networks 6)." *The Genealogical World of Phylogenetic Networks* 9.9.

— (08/24/2020b). "Constructing rhyme networks (From rhymes to networks 5)." *The Genealogical World of Phylogenetic Networks* 9.8.

— (2020c). "Improving data handling and analysis in the study of rhyme patterns." *Cahiers de Linguistique Asie Orientale* 49.1, 43–57.

— (2020d). *RhyAnT. A tool for interactive rhyme annotation*. Jena: Max Planck Institute for the Science of Human History.

List, J.-M. and R. Forkel (2022). *LingPy. A Python library for quantitative tasks in historical linguistics [Software Library, Version 2.6.9]*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

List, J.-M., S. J. Greenhill, and R. D. Gray (2017). "The potential of automatic word comparison for historical linguistics." *PLOS ONE* 12.1, 1–18.

List, J.-M., N. W. Hill, and C. J. Foster (2019). "Towards a standardized annotation of rhyme judgments in Chinese historical phonology (and beyond)." *Journal of Language Relationship* 17.1, 26–43.

List, J.-M., N. A. Sims, and R. Forkel (2021). "Towards a sustainable handling of interlinear-glossed text in language documentation." *ACM Transactions on Asian and Low-Resource Language Information Processing* 20.2, 1–15.

Matter, F. (10/2022). *pylingdocs [Version 0.0.10]*. Version 0.0.10.

Wáng, L. 王力. (1980 [2006]). *Hànyǔ shǐgǎo* 漢語史稿 [History of the Chinese language]. Repr. Běijīng 北京: Zhōnghuá Shūjú 中华书局.

Wáng, Z. (2020). "A linguistic study on rhyming in the Beijing dialect." *Cahiers Linguistiques Asie Orientale* 49.1, 21–42.

Final Discussion

Johann-Mattis List (University of Passau)

1 Final Questions

Does the integration of research data also play a role in your specific discipline?

Do you share the view reported here that it is worthwhile to try and advance the quality of our data in the humanities, rather than to advance the quality of methods to automatically handle data of bad quality?

What kind of textual data do you deal with in your research and do you think they would profit from being further standardized?

2 Final Tasks

Make a schema of the integration potential of data in your discipline.

Make a list of topics that could be automatically inferred from your data, provided they would be properly standardized and annotated.

Make a draft plan for a tool to ease the annotation of research data used in your discipline.