

# **Linguistic Diversity: Empirical Perspectives**

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

Department for English Studies  
Friedrich Schiller University, Jena

Summer Semester 2020

# Contents

<b>1 Introduction</b>	<b>3</b>
Introduction to “Linguistic Diversity: Empirical Perspectives” . . . . .	3
Comparative Linguistics . . . . .	9
<b>2 Lexical Variation</b>	<b>19</b>
Lexical Variation (Historical Viewpoint) . . . . .	19
Lexical Variation (Typological Viewpoint) . . . . .	28
<b>3 Phonetic Variation</b>	<b>37</b>
Phonetic Variation (Historical Viewpoint) . . . . .	37
Phonetic Variation (Typological Viewpoint) . . . . .	46
<b>4 Structural Variation</b>	<b>55</b>
Structural Variation (Historical Viewpoint) . . . . .	55
Structural Variation (Typological Viewpoint) . . . . .	64

# 1 Introduction

The two introductory sessions set the topic for the seminar, focusing on practical aspects of the seminar, as well as introducing major perspectives on language comparison (historical, areal, typological).

# Introduction to “Linguistic Diversity: Empirical Perspectives”

## 1 Introductory remarks

## 2 How seminars will be structured

### 2.1 Main session

The main session will be streamed using Zoom. We will aim for about 45 minutes, during which the major topic will be presented, accompanied by a handout of approximately 8 pages with literature, which students can use to check on literature and also to deepen those points which cannot be completely touched during the seminar. After the 45 minutes, we plan for 15 minutes of questions. The remaining 30 minutes of a regular seminar in a regular university classroom will be dropped and pursued on an individual basis with the students. This will be done in form of personal talks about the seminar content or other topics with the students, which can be conducted either in form of video chats or in form of outside walks in save distance, during which the topics will be raised and notes will be taken. There will be time to make maximally two of these events per week (when calculating 30 minutes per event), so students can ask for individual appointments and – depending on the number of students – they can have up to two appointments in this form during the semester. Additionally, normal appointments to discuss term papers and other aspects are also possible.

### 2.2 Weekly reading exercises

Every week is accompanied by a text for reading which was already selected for 8 of the sessions (starting from Session 2). The students are encouraged to read the questions preceding each reading text and answer them. At the beginning of each seminar, the questions will be quickly discussed, and students can compare the answers with their own accounts. If they wish (but this is not in any way expected, just on a voluntary basis), students can send their solutions to me, before the end of the Monday before each next seminar session. To make it easier to recognize if students send emails with requests or simple with reading exercises, it is important that

- the emails with the solutions for the reading exercises do only contain the solution in form of a PDF document (any other formats won't be accepted),
- the email is addressed to my institutional email `list@shh.mpg.de`, and it contains as email subject the schematic formula [SOLUTION] NUMBER, where [SOLUTION] is the literal text (with angular brackets, and in capitals), and the NUMBER is the number of the exercise,
- the email does NOT contain any additional questions on different topics, but only a simple formulaic sentence mentioning that this email contains the solutions.

It is important to follow these rules strictly, since it would otherwise be impossible to filter emails directly in order to reduce the email load that I receive generally. If emails do not follow these requirements, they will be ignored.

## 2.3 Weekly essay exercises

To improve the writing skills of the seminar members, I propose weekly essay writing exercises, which usually take the reading text accompanying each seminar as the basis (but may also diverge). The general idea is to provide weekly exercises that can be answered in form of a short piece of writing, usually between half a page and one page (never more). Similarly to the reading exercises, students are encouraged to write down their solutions and send them to me before the end of the Monday before the seminar (to give me enough time to correct their essays). The tasks will vary and include a variety of styles. Focus will not be essentially on the grammar, but on the general line-of-thoughts and aspects of consistency, by which students answer the tasks. This will allow for a more direct feedback that students otherwise only receive in condensed form after handing in a term paper.

To make it easier to recognize if students send emails with proposed solutions for essay writing exercises, it is important that

- the emails with the essays do only contain the essay in form of a PDF document (any other formats won't be accepted),
- the email is addressed to my institutional email `list@shh.mpg.de`, and it contains as email subject the schematic formula [ESSAY] NUMBER, where [ESSAY] is the literal text (with angular brackets, and in capitals), and the NUMBER is the number of the exercise,
- the email does NOT contain any additional questions on different topics, but only a simple formulaic sentence mentioning that this email contains the essays.

It is again important to follow these rules strictly, since it would otherwise be impossible to filter the emails directly, and drastically increase my workload. Emails with questions should always be done separately.

After solutions have been received by the students, a sample solution, which I will write down myself, will be shared on the seminar website for each exercise. These are for orientation, they do not reflect the only solution, since we practice essay writing and scientific planning of article writing, and in this area, many solutions are often equally well.

## 3 Short overview on the seminar's content

### 3.1 General overview

The seminar is structured into five major parts. The first part consists of a more general introduction to the seminar topic, and we will try to arrive at a clear distinction between the three fields of historical, areal, and typological linguistics, and also try to arrive at a clearer understanding of "empirical studies" or "empirical approaches". As preparation, we will read the text by Aikhenvald (2007), which provides a short overview on the major research focus of areal linguistics.

The second block consists of two sessions devoted to *lexical variation*. We will approach this topic from the perspective of historical linguistics (starting from the notion of basic vocabulary and lexical change introduced by Swadesh 1952) and then look at areal and typological approaches, inspired specifically by the notion of *colexification*, first coined by François (2008), and how they can be implemented in larger data collections.

The third block consists of two sessions devoted to *phonetic variation*. Here again, we will approach the topic from the perspective of historical linguistics first, and discuss the

## 1 Introduction

phenomenon of regular sound change (which is roughly discussed in McMahon (1994)). When turning to areal and typological linguistics, we want to concentrate more closely on the notion of *sound patterns*, as it was used, among others, by Blevins (2004), and then discuss large-scale databases in which attempts are being made to collect larger amounts of data on phonetic diversity in the languages of the world.

The fourth block deals in two sessions with *structural variation*, which can be roughly compared with *grammatical variation*, although the term structure is a bit broader. Before we look into areal and typological aspects here, we make a quick excursus in the realm of *grammaticalization* as reflected, for example, in the work of Heine u. a. (2016). Having investigated the historical aspects, we then turn to areal and grammatical aspects, using Corbett (2004) as an example for a classical large-scale study on one specific aspect of structural variation, *number*. Based on these reflections, we will then look into larger data collections and general problems of establishing these collections for the languages of the world.

We will conclude in a single session in which we will either reflect about what we have learned so far, or by looking into topics inspired by the wishes of the seminar members.

### 3.2 Detailed seminar plan

#### 3.2.1 Session 1 from 07/05/2020

**topic** General Introduction to the Seminar

**content** introduction to the general plan of the seminar, mutual introduction of participants, etc.

#### 3.2.2 Session 2 from 14/05/2020

**topic** Comparative Linguistics

**content** introduction and comparison of historical, areal, and typological linguistics

**text** Aikhenvald (2007)

#### 3.2.3 Session 3 from 28/05/2020

**topic** Lexical Variation (Historical Viewpoint)

**content** looking at lexical variation from a historical linguistics viewpoint

**text** Swadesh (1952)

#### 3.2.4 Session 4 from 04/06/2020

**topic** Lexical Variation (Typological Viewpoint)

**content** looking at lexical variation from an areal and typological viewpoint

**text** François (2008)

### **3.2.5 Session 5 from 18/06/2020**

**topic** Phonetic Variation (Historical Viewpoint)

**content** looking at phonetic variation from a historical viewpoint

**text** McMahon (1994)

### **3.2.6 Session 6 from 25/06/2020**

**topic** Phonetic Variation (Typological Viewpoint)

**content** looking at phonetic variation from an areal and typological viewpoint

**text** Blevins (2004)

### **3.2.7 Session 7 from 02/07/2020**

**topic** Structural Variation (Historical Viewpoint)

**content** looking at structural variation from a historical viewpoint

**text** Heine u. a. (2016)

### **3.2.8 Session 8 from 09/07/2020**

**topic** Structural Variation (Typological Viewpoint)

**content** looking at structural variation from an areal and typological viewpoint

**text** Corbett (2004)

### **3.2.9 Session 9 from 16/07/2020**

**topic** Concluding Session

**content** pen discussion on free topics that can be selected by the participants

## **4 The seminar website**

The seminar website can be accessed and freely looked at at <https://digling.org/lidi/>. If you want to download the texts, you need username and password. These will be announced during our first meeting.

The website essentially offers the texts for download and provides the students with the seminar plan. It will also be used to share the handouts, the exercises, and the sample solutions.

### 5 Proving participation and grades for participation

For a normal prove of participation, 5 essay exercises are sufficient. If students experience difficulties in complying to this, I urge them to get in contact with me to discuss potential problems. For a graded prove of participation, students are asked to make an official appointment with me, either via video-chat, or by taking a walk outside in safe distance, in order to discuss the content of the term paper. Exams will as a default not be offered, but term paper content can be flexibly discussed.

### References

- Aikhenvald, A. Y. (2007). "Grammars in contact. A cross-linguistic perspective". In: *Grammars in contact*. Hrsg. von A. Y. Aikhenvald und R. M. W. Dixon. Oxford: Oxford University Press, 1–66.
- Blevins, J. (2004). *Evolutionary phonology. The emergence of sound patterns*. Cambridge: Cambridge University Press.
- Corbett, G. G. (2004). *Number*. Cambridge: Cambridge University Press.
- François, A. (2008). "Semantic maps and the typology of colexification: intertwining polysemous networks across languages". In: *From polysemy to semantic change*. Hrsg. von M. Vanhove. Amsterdam: Benjamins, 163–215.
- Heine, B., H. Narrog und H. Long (2016). "Constructional change vs. grammaticalization". *Studies in Language* 40.1, 137–175.
- McMahon, A. (1994). *Understanding Language Change*. Cambridge University Press.
- Swadesh, M. (1952). "Lexico-statistic dating of prehistoric ethnic contacts. With special reference to North American Indians and Eskimos". *Proceedings of the American Philosophical Society* 96.4, 452–463. JSTOR: 3143802.

# Comparative Linguistics

## 1 Preliminary Considerations

### 1.1 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 <sup>21</sup>	i <sup>55</sup>	xuei <sup>35</sup>	pej <sup>21</sup> fəŋ <sup>55</sup>	kən <sup>55</sup>	t <sup>h</sup> ai <sup>51</sup> ian <sup>11</sup>	t͡ʂəŋ <sup>55</sup>	tsai <sup>53</sup>	naə <sup>51</sup>	t͡ʂəŋ <sup>55</sup> luən <sup>51</sup>
Hakka Chinese	1	iu <sup>33</sup>	it <sup>55</sup>	pai <sup>33</sup> a <sup>11</sup>	pet <sup>33</sup> fuŋ <sup>33</sup>	t <sup>h</sup> uŋ <sup>11</sup>	nit <sup>11</sup> t <sup>h</sup> eu <sup>11</sup>	hək <sup>33</sup>	e <sup>53</sup>		au <sup>55</sup>
Shànghǎi Chinese	1	fi <sup>22</sup>		t <sup>h</sup> ɿ <sup>55</sup> tsɿ <sup>21</sup>	poʔ <sup>33</sup> foŋ <sup>44</sup>	taʔ <sup>5</sup>	t <sup>h</sup> a <sup>33</sup> fiä <sup>44</sup>	tsəŋ <sup>33</sup>	hɔ <sup>44</sup>		ləʔ <sup>1</sup> lə <sup>23</sup> tsa <sup>53</sup>
Běijīng Chinese	2	ʂei <sup>35</sup>			də <sup>55</sup>		pən <sup>35</sup> liŋ <sup>21</sup>	ta <sup>51</sup>			
Hakka Chinese	2	man <sup>33</sup>	jin <sup>11</sup>		k <sup>w</sup> ɔ <sup>55</sup>		vɔi <sup>53</sup>				
Shànghǎi Chinese	2	sa <sup>33</sup>	jin <sup>55</sup>		fiəʔ <sup>21</sup>		pən <sup>33</sup> zɿ <sup>44</sup>	du <sup>13</sup>			
Norwegian	1	nu:ravínˀŋ	ɔ	su:lŋ						kranlæt	ɔm
Swedish	1	nu:ɖanvɪndən	ɔ	su:lən		tyɪstadə	ən ɡɔŋ				ɔm
Danish	1	noʌʌnvenˀŋ	ʌ	so:lˀŋ	k <sup>h</sup> ʌm		enɡɔŋ	i sɖɛiðˀ			ʌmˀ
Norwegian	2	vem	a	dem	sŋ	vɑ:	ɖŋ	stæŋkæstə			
Swedish	2	vem	av	dəm	səm	va		starkast			
Danish	2	vemˀ	a	bŋ	ɖ	va	ɖŋ	sɖæʌɡəsɖə			

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

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

### 1.2 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* (Busmann 1996: 312). According to this model, languages are complex aggregates of different linguistic systems, which ‘coexist and influence each other’ (Coseriu 1973: 40).<sup>1</sup> 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

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

## 1 Introduction

linguistic varieties (dialects, sociolects) which are connected by such a standard constitute the “variety space” (*Varietätenraum*) of a language (Oesterreicher 2001), as shown in Figure 1.

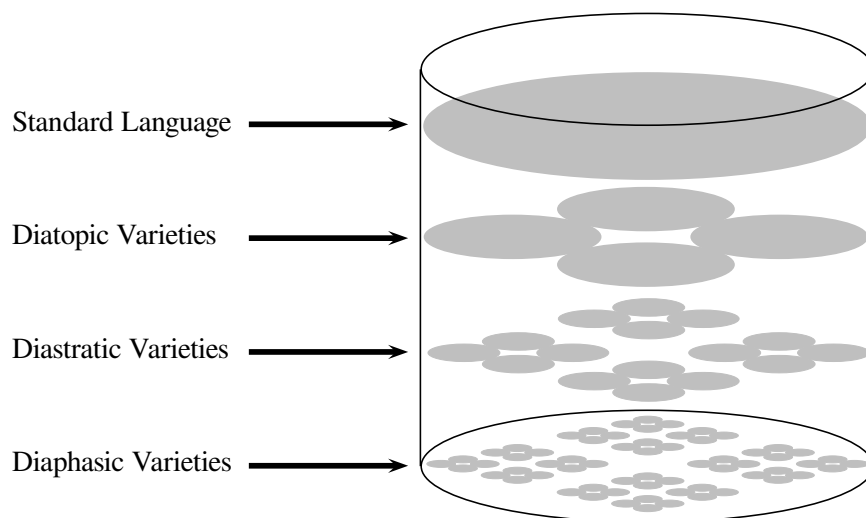


Abbildung 1: Language as a diasystem

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

### 1.3 What is a Linguistic Sign?

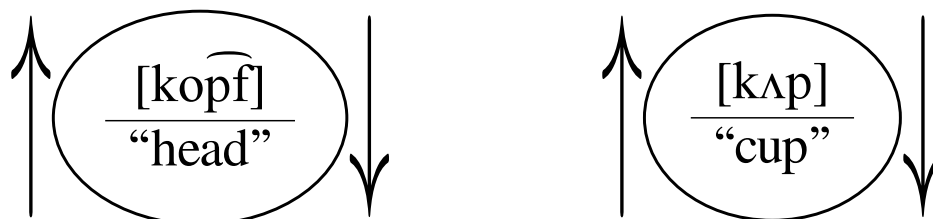
In historical linguistics, linguistic signs are usually treated in the context of the traditional sign model by Saussure (*Cours de linguistique générale*). As Roman Jakobson notes, we distinguish two sides: the form and the content:

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)

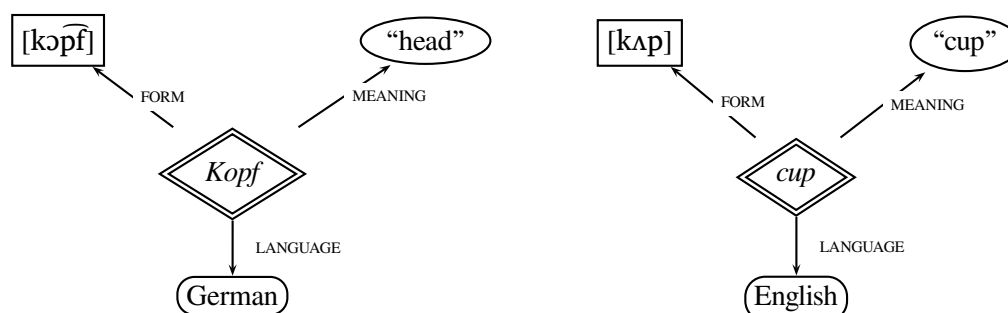
What does Jakobson mean with the words “material” and “intelligible”?

### 1.4 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	

## 1.5 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.

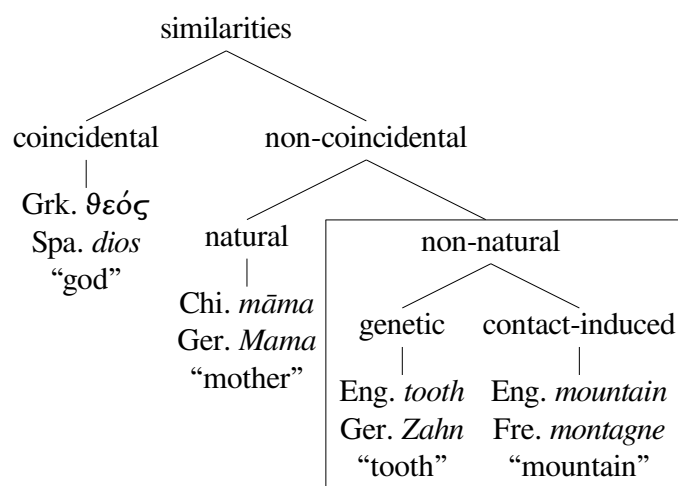
While this model of language as a bag of words may seem very simple, it is effectively the model that was underlying most of the quantitative comparative analyses that have been published so far. Additionally one should say, that even classical linguists who do not work in a quantitative framework tend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

## 1 Introduction

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

### 2.1 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?

## 2.2 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					Eng.	Grk.	Freq.
	<i>foot</i>	f	u	t	f	p	3 x
Ancient Greek	ποδ-	p	ɔ	d			
English	<i>father</i>	f	ɑ:	θ	f	p <sup>h</sup>	1 x
Ancient Greek	πατέρ-	p	a	t	ɹ	r	2 x
English	<i>fear</i>	f	ɪə	ɹ	θ	t	1 x
Ancient Greek	φοβέ-	p <sup>h</sup>	ɔ	b	t	d	1 x
English	<i>fire</i>	f	aɪə	ɹ			
Ancient Greek	πυρ-	p	y	r			

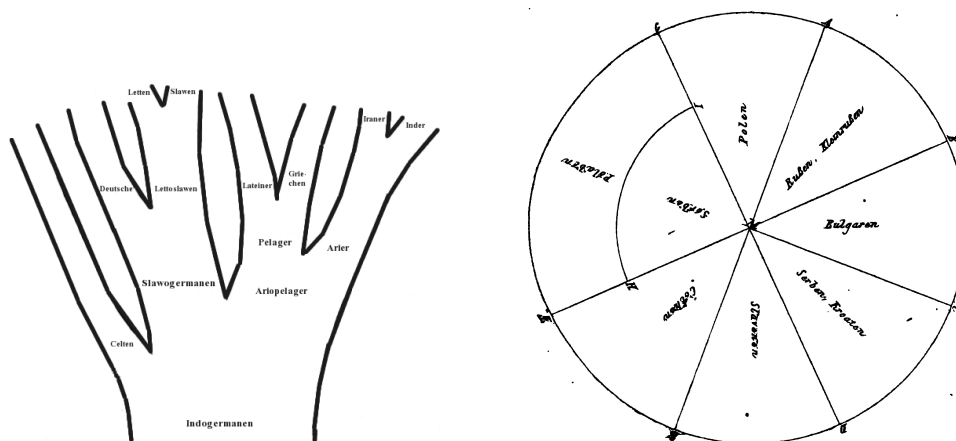
Detecting regular sound correspondences in classical historical language comparison.

## 2.3 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 u. a. 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.

## 1 Introduction

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

### 3.1 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?

### 3.2 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?

### 3.3 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 u. a. 2015), (3) population size (Bromham u. a. 2015), or (4) social structure (Lupyan und 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

### 4.1 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?

### 4.2 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?

### 4.3 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. und P. Stevenson (1998). *Variation im Deutschen. Soziolinguistische Perspektiven*. Berlin: de Gruyter.
- Bond, O. (2019). "Canonical Typology". In: *The Oxford handbook of morphological theory*. Hrsg. von J. Audring und F. Masini. Oxford: Oxford University Press, 409–431.
- Bromham, L., X. Hua, T. G. Fitzpatrick und 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., Hrsg. (1996). *Routledge dictionary of language and linguistics*. A. d. Deutschen übers. von G. Trauth und K. Kazzazi. London und New York: Routledge.
- Coseriu, E. (1973). *Sincronia, 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 und 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*. Hrsg. von D. V. Olmen, T. Mortelmans und F. Brisard. Berlin und New York: De Gruyter Mouton, 83–113.
- Jakobson, R. (1978). *Six lectures on sound and meaning*. A. d. Französischen übers. von J. Mephram. Mit einer Einl. von C. Lévi-Strauss. Cambridge und London: MIT Press.
- List, J.-M. (2014). *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.
- (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 und E. Baptiste (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. und 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*. Hrsg. von M. Haspelmath. Berlin und New York: Walter de Gruyter, 1554–1595.
- Saussure, F. de (1916). *Cours de linguistique générale*. Hrsg. von C. Bally. Lausanne: Payot; Deutsche Übersetzung: — (1967). *Grundfragen der allgemeinen Sprachwissenschaft*. A. d. Französischen übers. von H. Lommel. 2. Aufl. Berlin: Walter de Gruyter & Co., 1967.
- 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*. Bd. 1: *Kurzer Abriss einer Lautlehre der indogermanischen Ursprache*. Weimar: Böhlau.
- Schmidt, J. (1875). *Zur Geschichte des indogermanischen Vocalismus. Zweite Abteilung*. Weimar: Hermann Böhlau.



## **2 Lexical Variation**

The two sessions focus on lexical variation from a historical, areal, and typological perspective.

### Lexical Variation (Historical Viewpoint)

#### 1 Lexical Change

##### 1.1 Modeling Lexical Change

While we find many interesting challenges when trying to model external language history, the general picture becomes even more complicated when try to model internal language history. While many linguists probably see the greatest challenge in questions of grammaticalization, it is enough to look into those aspects of language that have been rather thoroughly investigated to find enough challenges to start with. One such aspect is *lexical change*. In a broad sense, lexical change refers to the way in which the lexicon of a human language evolves. In a narrower sense, which we will maintain here, it concentrates on the processes that affect the linguistic signs of a language during its history.

What are the major processes that can affect a linguistic sign?

##### 1.2 Three Dimensions of the Linguistic Sign

When concentrating on the words and how they are affected during language history, we need to identify the major processes that constitute the changes that affect them. Following Gévaudan (2007: 15-17), we can distinguish three different dimensions along which words can change, namely, the *semantic dimension* (a given word can change its meaning), the *morphological dimension* (new words are formed from old words by combining existing words or deriving new words with help of affixes), and the *stratic dimension* (languages may acquire words from their neighbors and thus contain strata of contact).

In the second session, we have discussed quickly a slightly extended model of the linguistic sign. To which degree does this model remind of the dimensions of lexical variation by Gévaudan?

##### 1.3 Lexical Change and Sound Change

The focus on three dimensions along which a word can change deliberately excludes sound change. Excluding sound change is justified by the fact that, in the majority of cases, the process proceeds independently from semantic change, morphological change, and borrowing, while the latter three process often There are, of course, cases where sound change may trigger the other three processes – for example, in cases where sound change leads to homophonous words in a language that express contrary meanings, which is usually resolved by using another word form for one of the concepts. An example for this process can be found in Chinese, where *shǒu* (in modern pronunciation) came to mean both “head” and “hand” (spelled as 首 and 手). Nowadays, *shǒu* remains only in expressions like *shǒudū* 首都 “capital”, while *tóu* 头 is the regular word for “head”. interact. Since the number of these processes where we have sufficient evidence to infer that sound change triggered other changes is rather small, we will do better to ignore it when trying to design initial models of lexical change.

People keep repeating that models do not necessarily need to be realistic. But if they are not realistic, what can we in the end gain from them?

## 1.4 Lexical Replacement

Important work on lexical change goes back at least to the 1950s, when Morris Swadesh (1909–1967) proposed his theory of *lexicostatistics* and *glottochronology* (Lees 1953, Swadesh 1952). What was important in this context was not the idea that one could compute the divergence time of languages, but the *data model* which Swadesh introduced. This data model is represented by a word-list in which a particular list of concepts is translated into a particular range of languages. While former work on semantic change had been mostly *onomasiological* – form-based, taking the word as the basic unit and asking how it would change its meaning over time – the new model used concepts as a *comparandum*, investigating how word forms *replaced each other* in expressing specific contexts over time. This *onomasiological* or concept-based perspective has the great advantage of drastically facilitating the sampling of language data from different languages. Swadesh's *concept-slot model* can be seen as some kind of a chest of drawers, in which each drawer represents a specific concept and the content of a drawer represents the words one can use to express that given concept. In such a model, lexical change proceeds by *replacement*: a word within a given concept drawer can be kicked out of the drawer in order to make place for another word. Unfortunately, we do not find many attempts to test the characteristics of this model in simulation studies. The only one known to me is a posthumously published letter from Sergey Starostin (1953-2005) to Murray Gell-Mann (Starostin 2007), in which he describes an attempt to account for his theory that a word's replacement range increases with the word's age ("Comparative-historical linguistics and lexicostatistics") in a computer simulation.

How can one explain what Starostin calls the "aging of words", i.e., the fact that the longer a word is part of a language, the more likely it is to be replaced?

## 1.5 Gain and loss

An alternative to Swadesh's concept-based model of lexical replacement is to treat a language as a bag of words in which – over time – certain words are added, and certain words are deleted. This model is very popular in evolutionary biology, where *gene families* correspond to the words in our bag of words, and evolution is modeled as a process of gene family gain or gene family loss (Cohen et al. 2008). The model is very easy to be applied to linguistics, where the gene family has a counterpart in the etymological *root* or the *word family*. Biologists have described the stochastic characteristics of different gain-loss models, and software packages that help to employ the models for inference of phylogenies are also available (Ronquist and Huelsenbeck 2003). While gain-loss models are frequently used by linguists to infer phylogenies (Gray and Jordan 2000, Sagart et al. 2019), they are less frequently used for plain simulation studies. Here, the only attempts that I know of are one study by Greenhill et al. (2009), where the authors used the TraitLab software (Nicholls et al. 2013) to simulate language change along with horizontal transfer events, and a study by Murawaki (2015), in which (if I understand the study correctly) a gain-loss model is used to model language contact.

What are the advantages and disadvantages of the gain-loss model in comparison with the concept slot model of lexical change?

### 1.6 Modeling lexical change with semantic shift

For the moment, no attempt to model morphological change as part of a model for lexical change is known to me (at least not from the perspective of g-linguistics). The problem of the gain-loss and the concept-slot models to account for semantic change, however, can be overcome by turning to *bipartite graph models of lexical change* (see Newman 2010: 32f for details on bipartite graphs). In such a model, the lexicon of a human language is represented by a bipartite graph consisting of *concepts* as one type of node and *word forms* (or forms) as another type of node. The association strength of a given word node and a given concept node (or its “reference potential”, see List 2014: 21f), i.e. the likelihood of a word being used by a speaker to denote a given concept, can be modeled with help of *weighted edges*. This model naturally accounts for *synonymy* (if a meaning can be expressed by multiple words) and *polysemy* (if a word can express multiple meanings). Lexical change in such a model would consist of the *re-arrangement* of the *weights* in the network. Word loss and word gain would occur if a new word node is introduced into the network or an existing node gets dissociated from all of the concepts. We can find this idea of bipartite modeling of a language’s lexicon in the early linguistic work of Sankoff (1969: 28-53), as reflected in the Figure 2 below.

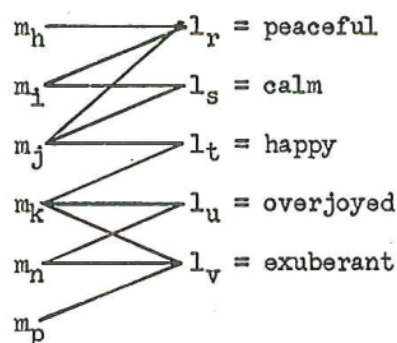


Figure 2: Bipartite graph model by Sankoff (1969: 36).

What are the advantages and what are the disadvantages of these models in comparison to the concept slot and the gain-loss models? And how would one model lateral transfer (borrowing)?

## 2 Morphological Change

### 2.1 Every Word has a Family

Linguists often emphasize that each word has its own history (“chaque mot a son histoire”), a statement, often attributed to Jules Gilliéron (1854-1926, see Campbell 1999: 189), but already Jacob Grimm (1785-1863) said in 1819 “jedes Wort hat seine Geschichte und lebt sein eigenes Leben” (Grimm 1819: xiv, compare also Koerner 1990: 13). Even more importantly, however, all human languages tend to build new words from existing words by various processes of word formation. As a result of word formation processes, we can say that certain words in the same language are *cognate*, i.e., that they form a *family*, since they go back to the same original word from which they were formed. Similar to a family, words can have different *degrees of relationship*, and if we try to resolve the relations among the members of a given word family, we tend to find *hierarchical relations* in which a given word gives rise to another word which itself gives rise to more words.

When talking of hierarchical relations in word formation, what does this mean in terms of a model?

## 2.2 Major Processes of Word Formation

If one tries to find information in typical text books of linguistics or historical linguistics, one will find a rich inventory of different word formation processes, but it is very difficult to understand how they are interrelated in concrete. The problem is – as we often find it in linguistics – that scholars use a terminology that mixes the description with the explanation. Instead of using a terminology which is exclusively descriptive, the terms used for certain processes are often also trying to explain why the phenomenon occurs. Thus, a term such as *reanalysis* not only describes the relation between a source form and a target form, but also tries to explain why the target form has the shape we observe. For this reason, we won't concentrate on any terminological examples, but rather look at the major processes based on inspecting the form alone. From this descriptive perspective alone, we can distinguish *contactenative*, *allomorphic*, and *subtractive* types of word formation (Schweikhard and List 2020).

Can you find an example for all three types?

## 2.3 Two Perspectives on Studying Word Formation

Word formation is predominantly studied from the perspective of language-specific processes, be it in a synchronic or in a diachronic perspective. This becomes already evident when looking at the term *productivity*, which is supposed to tell us how productive a given word formation device is for the formation of new words in a given language. However, there is another, much more interesting perspective of word formation that is rarely studied, but has tight connections to linguistic typology and language variation. When studying not which concrete affixes give rise to new words in a given language, but when instead trying to find out, which concepts are more often “recycled” to form new words, we can investigate cross-linguistics, language-family-specific or areal trends of *conceptual productivity* or *conceptual promiscuity* (List 2018). We will look at this in more detail in the session devoted to lexical variation from a typological perspective.

Why would one call this phenomenon “promiscuity”?

## 3 Semantic Change

### 3.1 It's the Meaning, Stupid!

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.

Why is the meaning part of the linguistic sign so much more difficult to handle than the form part?

### 3.2 Meaning is Different

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 20122)), 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. 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)<sup>1</sup>

### 3.3 Cumulation and Reduction

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

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

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

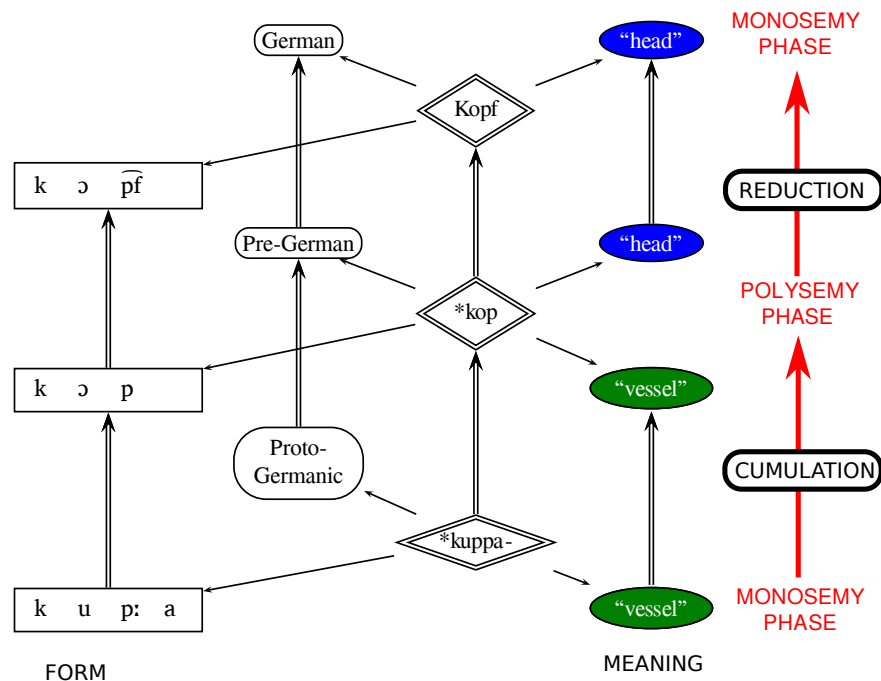
- “cup” > “head”, “see” > “think”

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

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

---

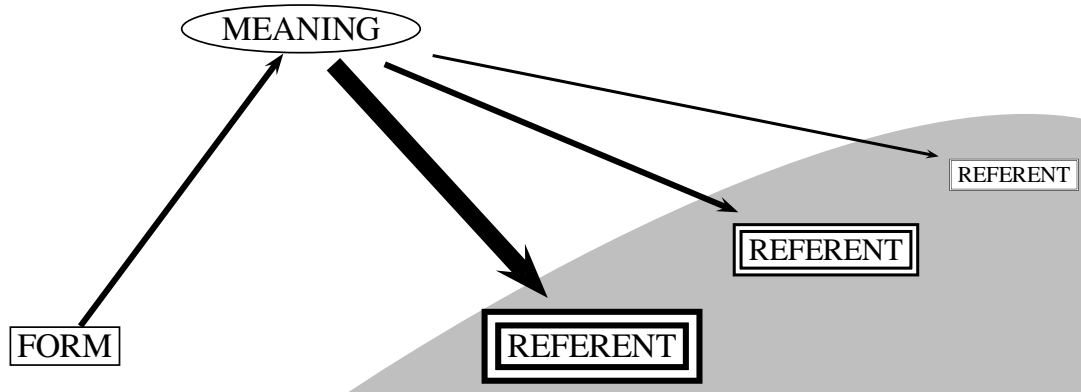
<sup>1</sup> My translation, original text: “Er besteht immer in einer erweiterung oder einer verengung des umfangs der bedeutung, denen eine verarmung oder bereicherung des inhalts entspricht. Erst durch die aufeinanderfolge von erweiterung und verengung kann eine von der ursprünglichen völlig verschiedene bedeutung sich bilden.”



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

### 3.4 Reference Potential

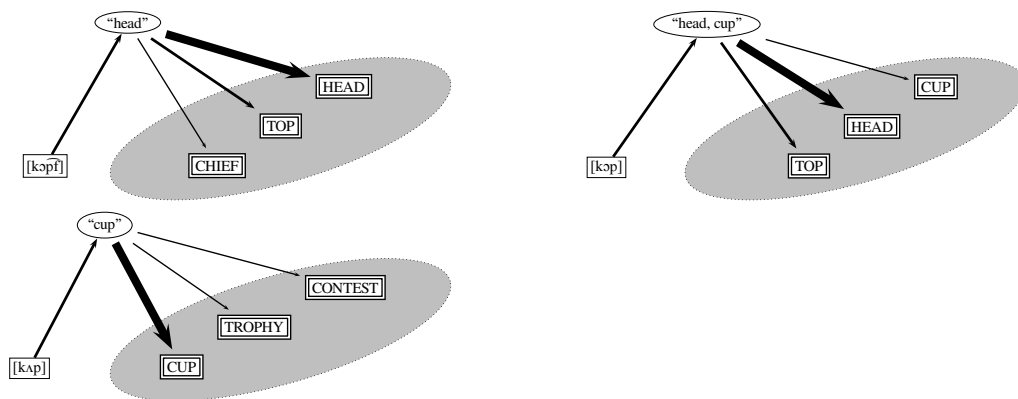
In contrast to the form part of the linguistic sign, it is difficult to find any kind of order in a sign's meaning part, because it lacks linearity, both because it is nothing concrete that we could grasp with our primary senses, and because it is not dependent of temporal succession. A further problem is that – because of the famous arbitrariness of the connection between sign form and sign meaning, “meaning is inherently fuzzy and non-systematic” (Hock and Joseph 1995 [2009]: 206). Unfortunately, there is no semantic theory that would be approved by a larger part of the linguistic community. While Saussure's model of the linguistic sign (Saussure 1916) is indifferent with respect to the question how a sign is used to refer to the real world. He simply emphasizes that the meaning part of the sign should not be confused with the object which it denotes (ibid.: 98). *Triadic sign models* try to close this gap by distinguishing the meaning of a word from its *reference* (Löbner 2003: 257). Since the reference of a linguistic sign can only be clear when the sign is used in a particular context, it is furthermore useful to make a distinction between the reference and the *reference potential* (Schwarz 1996: 175). The reference potential can be understood as the set of all possible referents that can be denoted by a given sign. Of course, the reference potential is dependent upon the meaning of the sign: if the meaning is very restricted, the amount of potential referents will also be restricted (Löbner 2003: 306).



If you compare words like German "Stein" and German "Ding", how do they differ with respect to their reference potential?

### 3.5 Semantic Change as Changing Reference Potential

The idea of the reference potential can help us to understand a bit better what happens during semantic change. If we only look for the most frequent referents of words like English *cup*, Dutch *cop* "head, cup", and German *Kopf*, we can find a continuum with respect to the referents of the words. It reflects not only former processes of semantic change, but may also give us hints on future processes. The change from German *Kopf* meaning "head" to the metaphoric meaning "chef", for example, is also attested in Chinese, as we have seen before.



If semantic change is something chaotic, and not easy to predict, what could nevertheless be the regularities that we might expect there, at least in some cases?

### 3.6 Pathways of Semantic Change

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.<sup>2</sup>

What is the major obstacle in constructing a database of attested semantic shifts?

## References

- Campbell, L. (1999). *Historical linguistics. An introduction*. 2nd ed. Edinburgh: Edinburgh Univ. Press.
- Cohen, O., N. D. Rubinstein, A. Stern, U. Gophna, and T. Pupko (2008). "A likelihood framework to analyse phyletic patterns". *Philosophical Transactions of the Royal Society B*.
- Fox, A. (1995). *Linguistic reconstruction. An introduction to theory and method*. Oxford: Oxford University Press.
- Gray, R. D. and F. M. Jordan (2000). "Language trees support the express-train sequences of Austronesian expansion". *Nature* 405, 1052–1055.
- Greenhill, S. J., T. E. Currie, and R. D. Gray (2009). "Does horizontal transmission invalidate cultural phylogenies?" *Proc. Biol. Sci.* 276.1665, 2299–2306.
- Grimm, J. (1819). *Deutsche Grammatik*. Vol. 1. Göttingen: Dieterichsche Buchhandlung. Internet Archive: bub\_gb\_fu0IAAAQAAJ.
- Gévaudan, P. (2007). *Typologie des lexikalischen Wandels. Bedeutungswandel, Wortbildung und Entlehnung am Beispiel der romanischen Sprachen*. Stauffenburg-Linguistik ; 45. Tübingen: Stauffenburg.
- Hock, H. H. and B. D. Joseph (2009). *Language history, language change and language relationship. An introduction to historical and comparative linguistics*. 2nd ed. Berlin and New York: Mouton de Gruyter.
- Koerner, K. (1990). "Jacob Grimm's position in the development of linguistics as a science". In: *The Grimm brothers and the Germanic past*. Ed. by E. H. Antonsen. Amsterdam and New York: John Benjamins, 7–23.
- Lees, R. B. (1953). "The basis of glottochronology". *Language* 29.2, 113–127. JSTOR: 410164.
- List, J.-M. (2014). *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.
- (2018). "Von Wortfamilien und promiskuitiven Wörtern [Of word families and promiscuous words]". *Von Wörtern und Bäumen* 2.10.
- Löbner, S. (2003). *Semantik. Eine Einführung*. Berlin: de Gruyter.
- Murawaki, Y. (2015). "Spatial structure of evolutionary models of dialects in Contact". *PLOS ONE* 10.7, e0134335.
- Newman, M. E. J. (2010). *Networks. An Introduction*. Oxford: Oxford University Press.
- Nicholls, G. K., R. J. Ryder, and D. Welch (2013). *TraitLab: A MatLab package for fitting and simulating binary tree-like data*.
- Paul, H. (1886). *Principien der Sprachgeschichte*. 2nd ed. Halle: Max Niemeyer. prinziendersp01paulgoog: ia.
- Ronquist, F. and J. P. Huelsenbeck (2003). "MrBayes 3: Bayesian phylogenetic inference under mixed models". *Bioinformatics* 19.12, 1572–1574.
- 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.
- Sankoff, D. (1969). "Historical linguistics as stochastic process". Dissertation. Montreal: McGill University.
- Saussure, F. de (1916). *Cours de linguistique générale*. Ed. by C. Bally. Lausanne: Payot; German translation: – (1967). *Grundfragen der allgemeinen Sprachwissenschaft*. Trans. from the French by H. Lommel. 2nd ed. Berlin: Walter de Gruyter & Co.
- Schwarz, M. (1996). *Einführung in die kognitive Linguistik*. Basel and Tübingen: Francke.
- Schweikhard, N. E. and J.-M. List (2020). *Developing an annotation framework for word formation processes in comparative linguistics*. Manuscript under Review.
- Sperber, H. (1923). *Einführung in die Bedeutungslehre*. Bonn and Leipzig: Kurt Schroeder.
- Starostin, S. A. (2007). "Computer-based simulation of the glottochronological process (Letter to M. Gell-Mann)". In: S. A. Starostin: *Trudy po yazykoznaniiyu [S. A. Starostin: Works in Linguistics]*. LRC Publishing House, 854–861.
- Swadesh, M. (1952). "Lexico-statistic dating of prehistoric ethnic contacts. With special reference to North American Indians and Eskimos". *Proceedings of the American Philosophical Society* 96.4, 452–463. JSTOR: 3143802.
- Traugott, E. C. (2012). "Pragmatics and language change". In: 249–565.
- 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.

<sup>2</sup>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>).

## Lexical Variation (Typological Viewpoint)

### 1 Introduction

While it seems pretty clear why one would study lexical variation from a historical viewpoint, the typological and areal questions that might be asked in this context may seem less clear at the first sight. The problem is again the definition of an angle for comparison. Thus, if we only look at the word for “apple” in various languages, this may sound very boring and not provide us with any interesting questions. However, if we look at apples and pears and bananas, we might be able to detect some interesting typological aspects of how fruits are denoted in the languages of the world. The major question is thus not the pronunciation of a given word itself, but the pronunciation of a given word in the context of other pronunciations, and the conclusions that can be drawn from this.

What conclusions can be drawn from similar pronunciations for different words in a given language?

### 2 Fixing meanings: the Concepticon

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 (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:

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, hundreds of different concept lists have been compiled for various purposes.

For what other purposes might scholars propose concept lists?

#### 2.1 Concept lists

Concept lists are collections of concepts which scholars decided to compile at some point. In an ideal concept list, concepts would be described by a *concept label* (*elicitation gloss*) 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. We find lists produced for historical language comparison (Swadesh 1952), subdivided lists of stable and less stable concepts (see Yakhontov's list mentioned in Starostin 1991), lists of the "most stable" concepts across all times and cultures (), classical questionnaires for linguistic field work (BDS), ranked lists (Starostin 2007), and many concept lists used in psycholinguistics, e.g., to study language acquisition (Ferguson 1964), to conduct naming tests (Ardila 2007), or to study specific semantic domains (Snoek 2013).<sup>1</sup>

What is meant by "naming tests" in this context?

## 2.2 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 (<https://concepticon.clld.org>, List et al. 2016) 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, and by now quite advanced workflows for curating and testing the data we have assembled so far. In the Concepticon project, all entries from different 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 (e.g. by including links to the Princeton WordNet University 2010).

Why could one not instead just start from Princeton WordNet as the source of definitions and senses? Why does the Concepticon need its own range of concept glosses?

## 3 Cross-Linguistic Data Formats

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

In what subfield of linguistics can data-free research be carried out?

<sup>1</sup> See List (2018) for details on the history of concept list compilation.

### 3.1 Data problems

The problem of data in linguistics is that it is all too often not FAIR in the sense of Wilkinson et al. (2016): **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable.

It is still very difficult to find particular datasets, since linguistic journals often do not have a policy on supplementary data and may lack resources for hosting data on their servers. It is also often difficult to access data, and many papers which are based on original data are still being published without the data<sup>1</sup> and having to request the data from the authors is sometimes a more serious obstacle than it should be. Due to idiosyncratic formats, linguistic datasets also often lack interoperability and are therefore not reusable. (Forkel et al. 2018: 2)

While it was less common to share one's data, or to even compile data directly, in the research of the nineties, and it was beyond doubt even difficult to find a good repository to share one's data up to the end of the first decade of the second millennium, it is disappointing to see to which degree modern linguistic research still fails to be based on FAIR data. While it is clear, that data sharing may be difficult for ethical reasons, there are still many people who think they own their data. While nobody should be required to share their data before a publication, it is clear, however, that a publication that does not offer the data is irreproducible and therefore scientifically questionable (see Berez-Kroeker et al. 2018 for the distinction between reproducible and replicable research).

The idea of reproducible research is nice, but how can one avoid to be scooped by colleagues who just grab the data and write papers on them?

### 3.2 Data standards

The Cross-Linguistic Data Formats initiative (CLDF, <https://cldf.clld.org>, Forkel et al. 2018) comes along with: (a) standardization efforts, (b) software APIs which help to test and use the data, and (c) working examples for best practice. (a) points to linguistic meta-data-bases like Glottolog (<https://glottolog.org>, Hammarström et al. 2018), Concepticon (List et al. 2016), and the Cross-Linguistic Transcription System initiative (CLTS, <https://clts.clld.org>, Anderson et al. 2018). These databases help scholars to make explicit what data (what *languages*, what *concepts*, what *sounds*) they are working with, and additionally aid them in merging different datasets into larger data collections. They aim, in brief, at increasing the *comparability* of linguistic data. (b) points to software (currently written in Python and R), which helps users to test how well their data conforms to the standards established by the CLDF initiative. The software contributes to the *transparency* of the data, as it requires data to be presented in both machine- and human-readable formats. (c) points to existing datasets which have been created by different scholars and try to illustrate how the standards can be used and implemented. These working examples (see, e.g., Sagart et al. 2019) increase both the *availability* of data, they also make them more *findable*, as they are shared on public repositories, with the necessary metadata that makes it easy to search for data in CLDF format, as well as contributing to *transparency* and *comparability*. At the moment, we are trying to lift the CLDF initiative to the next level, by working on new workflows that help for a more efficient creation and curation of cross-linguistic data. A first example for these efforts is the CLICS<sup>2</sup> database (List et al. 2018), which we will discuss in the next section.

What other possibilities apart from sharing examples of best practice and providing standards would we have to encourage and propagate data sharing in linguistics?

## 4 Cross-linguistic 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 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 than 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?

### 4.1 Polysemy, homophony, and colexification

Polysemy and homophony are two seemingly contrary concepts in linguistics. However, in the end they describe both the same phenomenon, namely that a word form in a given language can have multiple meanings. François (2008) therefore suggests to replace the two interpretative terms by the descriptive term colexification. Colexification in this context only means that an individual language “is said to colexify two functionally distinct senses if, and only if, it can associate them with the same lexical form” (ibid.: 171).

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

### 4.2 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.

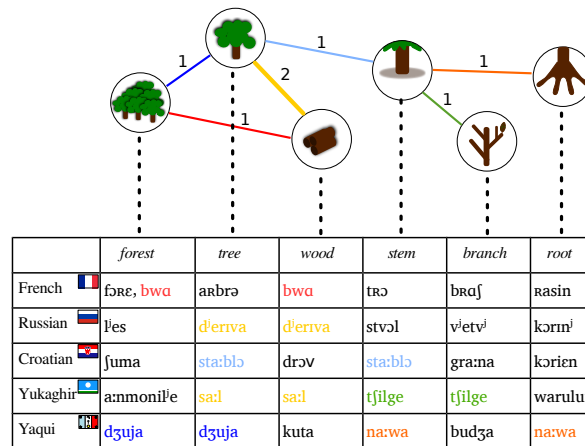
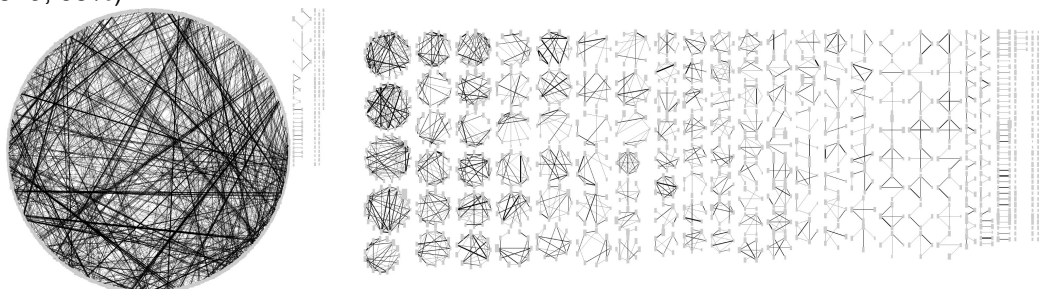


Figure 1: Reconstructing colexification networks from multi-lingual wordlists.

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

### 4.3 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 connections are dense but between which they are sparser” (Newman 2004: 4). In List et al. (2013), we used the algorithm by Girvan and Newman (2002) to analyse the network on the left. The result is given in the graphic on the right, where the originally almost completely connected network has been partitioned into 337 communities, with 104 being relatively big (5 and more nodes, covering a rather large parts of the 1289 concepts in our original database (879, 68%).



(a) complete networks

(b) analysed network

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?

## 4.4 Database of Cross-Linguistic Colexifications

CLICS<sup>2</sup> (<https://clics.clld.org>, List et al. 2018) is an online database of colexifications in currently 1220 language varieties of the world. CLICS<sup>2</sup> superseded the original Database of Cross-Linguistic Classifications, which established a computer-assisted framework for the interactive representation of cross-linguistic colexification patterns (Mayer et al. 2014). 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 the current version.<sup>2</sup>

## 4.5 Data curation and aggregation in CLICS<sup>2</sup>

The major advancement of CLICS<sup>2</sup> 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 carried out independently from each other. These stages include the *mapping of concepts* to Concepticon, the *referencing of sources* in the original data, the *linking of languages* to Glottolog, and the *cleaning of lexical entries* using a dedicated suite of Python scripts. 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 not requires to compare word forms *across* languages, but only *inside*, no further normalization (e.g., of the transcriptions) is needed.

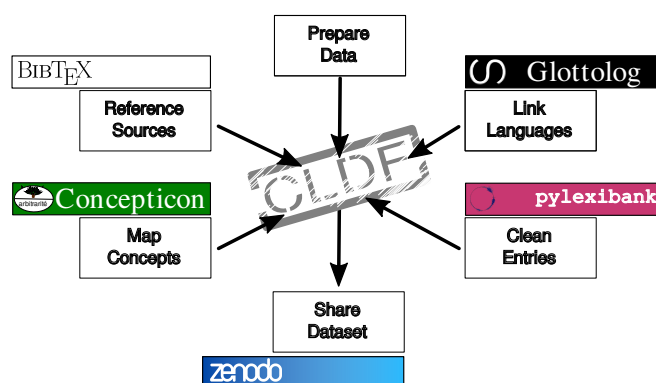


Figure 4: Workflow for data aggregation and curation in CLICS<sup>2</sup>.

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

## 4.6 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

<sup>2</sup>We are currently preparing an update that will further increase the coverage to more than 2000 language varieties.

## 2 Lexical Variation

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 illustrated in the screenshot in Figure 2, where the cluster around words for “tree” and “wood” is shown.

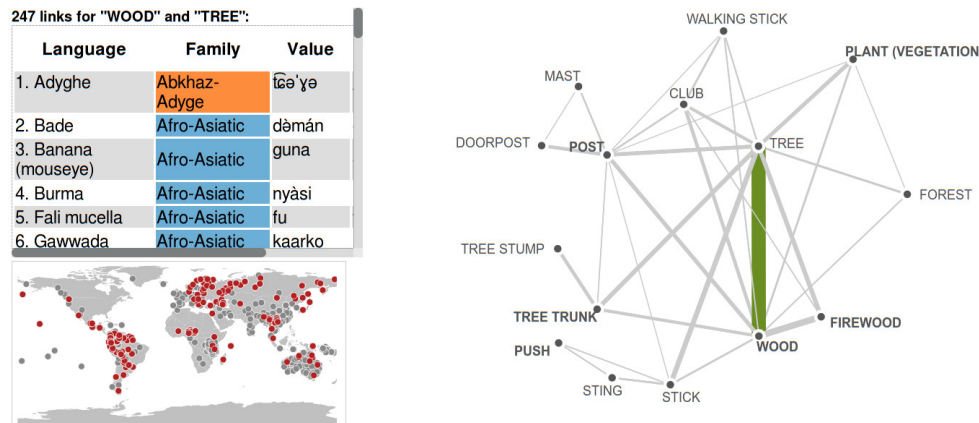


Figure 2: Screenshot from the CLICS<sup>2</sup> database (see infomap\_2\_WOOD).

What exactly does this visualization tell us?

## 5 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 CLICS<sup>2</sup> 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 help of polysemy patterns?

## 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.
- 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.
- Ardila, A. (2007). "Toward the development of a cross-linguistic naming test". *Archives of Clinical Neuropsychology* 22.3. Special Issue: Cultural Diversity, 297–307.

- Berez-Kroeker, A. L. et al. (2018). "Reproducible research in linguistics: A position statement on data citation and attribution in our field". *Linguistics* 56.1, 1–18.
- Bloomfield, L. (1973). *Language*. London: Allen & Unwin.
- Cysouw, M. (2010). "Semantic maps as metrics on meaning". *Linguistic Discovery* 8.1, 70–95.
- Ferguson, C. A. (1964). "Baby talk in six languages". *American Anthropologist* 66.6, 103–114.
- Forkel, R., J.-M. List, S. J. Greenhill, C. Rzymiski, 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.
- 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., R. Forkel, and M. Haspelmath (2018). *Glottolog*. Version 3.3. URL: <http://glottolog.org>.
- List, J.-M. (2018). "Towards a history of concept list compilation in historical linguistics". *History and Philosophy of the Language Sciences* 5.10, 1–14.
- 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., 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., S. J. Greenhill, C. Anderson, T. Mayer, T. Tresoldi, and R. Forkel (2018). "CLICS<sup>2</sup>. An improved database of cross-linguistic colexifications assembling lexical data with help of cross-linguistic data formats". *Linguistic Typology* 22.2, 277–306.
- Matisoff, J. A. (1978). *Variational semantics in Tibeto-Burman. The "organic" approach to linguistic comparison*. Institute for the Study of Human Issues.
- 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.
- 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.
- Snoek, C. (2013). "Using semantically restricted word-lists to investigate relationships among Athapaskan languages". In: *Approaches to Measuring Linguistic Differences*. Ed. by L. Borin and A. Saxena. Berlin: Mouton de Gruyter, 231–248.
- Starostin, S. A. (2007). "Computer-based simulation of the glottochronological process (Letter to M. Gell-Mann)". In: S. A. Starostin: *Trudy po yazykoznaniiu* [S. A. Starostin: *Works in Linguistics*]. LRC Publishing House, 854–861.
- Starostin, S. A. (1991). *Altajskaja problema i proischozhenije japonskogo jazyka* [The Altaic problem and the origin of the Japanese language]. Moscow: Nauka.
- Swadesh, M. (1950). "Salish internal relationships". *International Journal of American Linguistics* 16.4, 157–167. JSTOR: 1262898.
- (1952). "Lexico-statistic dating of prehistoric ethnic contacts. With special reference to North American Indians and Eskimos". *Proceedings of the American Philosophical Society* 96.4, 452–463. JSTOR: 3143802.
- (1955). "Towards greater accuracy in lexicostatistic dating". *International Journal of American Linguistics* 21.2, 121–137. JSTOR: 1263939.
- University, P. (2010). *WordNet. A lexical database for English*. Online Resource. Princeton.
- 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.
- 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.



## **3 Phonetic Variation**

The two sessions focus on phonetic variation from a historical, areal, and typological perspective.

## Phonetic Variation (Historical Viewpoint)

### 1 The detection of regular sound change

One of the most fundamental insights of historical linguistics, which has its origins in the very origins of the discipline, is the detection that sound change proceeds in what seems to be a mostly *regular* manner. What this exactly means has been subject of lengthy discussions. The notion of *regularity* also created a lot of confusion among those linguists and non-linguists without a detailed training in historical linguistics. This is in part reflected in computational approaches that seek to automate the classical approaches to linguistic reconstruction, which often quite naively ignore the fundamental aspect of regularity. Before we start to look into the techniques that linguists use in order to study sound change in detail, it is important to go back in history in order to review more closely how the concept of regularity evolved, and how it has been constantly challenged.

We will later spend more time on discussing what is actually meant by regularity, but judging from what you know about linguistics by now, what could regularity reflect in this context?

#### 1.1 Rask, Grimm, and the detection of sound shifts

The early detection that sound change may follow general tendencies in a given language family is usually attributed to Rasmus Rask (1787–1832), who pointed to what he thought were frequent transitions (of sounds) from Greek and Latin to Icelandic (Rask 1818: 169). When reading about these findings, Jacob Grimm (1785–1863) further investigated these systematic similarities between Greek, Latin, and Germanic languages (specifically Gothic), and expanded the second version of his *Deutsche Grammatik* considerably. Grimm identified regular correspondences between consonants in Greek, Gothic, and Old High German, as shown in Table 1 below.

gr.	goth.	alth.	gr.	goth.	alth.	gr.	goth.	alth.
P	F	B(V)	T	TH	D	K	..	G
B	P	F	D	T	Z	G	K	CH
F	B	P	TH	D	T	CH	G	K

**Table 1** Correspondences identified by Grimm (1822: 584).

What these formulas shown in the table meant that there were essentially many words of comparable meaning in the three languages in which the consonants formed patterns. If a word had a *p* in Greek (such as in ποδ- “foot”), it would reflect as *f* in Gothic (*fōtus*), and as *v* in Old High German (*vuoʒ*), yet not only in these three words, but in many more examples (see the detailed evidence in *ibid.*: 585). The crucial conclusion that Grimm drew from these observed patterns was that the *identity* of sounds (or letters) would not justify a comparison of words sharing a common origin. What would justify it instead was that the correspondence turned out to follow the rules he had detected while words with similar forms would result from coincidence or borrowing (*ibid.*: 588).

In the very same book, we also find him saying that the German word *Schrift* expressed “eight sounds in seven signs, since *f* stands for *p*” (*ibid.*: 3, my translation). What does this statement tell us about the historical context in which Grimm worked?

## 1.2 Verner and the dawn of regularity

Grimm's detection had pointed to an interesting tendency with respect to sound change, namely, that one could find patterns of corresponding sounds when comparing genetically related languages. The problem, however, was that these patterns did not seem to work in all cases. Grimm himself noted this, emphasizing that this consonant shift (he thought in terms of a change *from* Greek *to* Germanic) "proceeds in the majority but will never be pure in particular cases" (Grimm 1822: 590, my translation). Some reasons for certain exceptions had been mentioned already by Grimm himself. He noted, for example, that the patterns were influenced by the presence of liquids or sibilants (*ibid.*). Only later, however, linguists managed to find a proper explanation for these exceptions by refining the formulation of what they started to call *sound laws*. After exceptions of the basic correspondence pattern established by Grimm were listed systematically by Lottner (1862), Grassmann (1863) could explain the first class of exceptions by pointing to systematic assimilation processes in Greek and Sanskrit (Meier-Brügger 2002: L 348), by which of two aspirated sounds which follow each other, the first loses its aspiration (cf. Sanskrit *\*dhá-dhā-mi* > *dádhāmi*, "I put"). The second class of exceptions, finally, could be shown by (Verner 1877) to reflect a regular process in Germanic languages, during which the correspondence patterns varied in strict correlation with presumed stress patterns in the Proto-Germanic language (which themselves were still reflected in Vedic Sanskrit).

In Verner's original, he emphasizes that "Indo-European *k, t, p* first changed in all places to *h, þ, f*; these voiceless fricatives along with the voiceless fricative *s*, [...] became then voiced inside a word, when being in voiced neighborhood, but stayed voiceless when following after a stressed syllable" (*ibid.*). As an example, scholars often quote Gothic *broþar* in contrast to Vedic Sanskrit *bhrātar-* and Gothic *fadar* in contrast to Vedic Sanskrit *pitar-*. How should the stress be distributed in the Sanskrit words?

## 1.3 The Neogrammarian Manifesto

The fact that what formerly was thought to be a mere tendency could now, with help of refined rules that would allow to explain sound change as a process without exception lead to a great euphoria in the field and culminated in the so-called *Neogrammarian manifesto*:

All sound change, as long as it proceeds mechanically, follows exceptionless laws, i.e., the direction of the sound shift is the same with all members of a language community except from those cases in which the dialect split occurs, and all words in which the sound occurs in the same context are transformed without exception. (Osthoff and Brugmann 1878: XIII, my translation)

The principle assumption, that sound change proceeds without exceptions and that all apparent exceptions which one might observe can be regularly explained, be it by showing that the words under question are not cognate in the end, or that secondary processes have masked the former regularity, is still the working principle of classical historical linguistics and the first thing historical linguists learn during their training.

What are the two central aspects that can be found in the quote from the Neogrammarian manifesto?

### 1.4 Wang, Chen and the postulation of irregularity

Not all linguists would follow the opinion of the Neogrammarians. Especially dialectologists would often prefer to follow the famous slogan that “chaque mot a son histoire”, usually attributed to Jules Gilliéron (1854–1926, see Campbell 1999: 189). The doubts of the dialectologists were, however, not in direct contradiction to the Neogrammarian hypothesis of regularity, given that their theory did not state that *all words* in a given language change regularly, but rather emphasized that irregularities “could be accounted for [...] by certain less obvious mechanisms of borrowing and analogy” (Kiparsky 1988: 368).

In the 1960s, the situation changed drastically, when new research, which was almost exclusively based on the Chinese dialects, led to the postulation of a new mechanism of sound change which was in strict opposition to the hypothesis of the Neogrammarians.

Regarding the lexicon [they assumed] that a change always affects the whole lexicon, and can therefore be seen as an abrupt change. Regarding the sounds [they assumed] that the change proceeded step by step, and can therefore be seen as a gradual change. (Wang 2006: 109)<sup>1</sup>

The results of the analyses of the Chinese dialectologists, however, suggested that a certain mechanism of sound change, which they later called *lexical diffusion*, proceeds in the exact opposite way, namely, in “a manner that is phonetically abrupt but lexically gradual. As the change diffuses across the lexicon, it may not reach all the morphemes to which it is applicable. If there is another change competing for part of the lexicon, residue may result” (Wang 1969: 9). Examples were specifically drawn from cases where words with exactly the same pronunciation in Middle Chinese, the ancestor of most Chinese dialects, turned out to develop two different readings, which led the scholars conclude that “[when] a phonological innovation enters a language it begins as a minor rule, affecting a small number of words” which later “gradually spreads across the lexicon” (Chen 1972).

Character	Pīnyīn	Meaning	Middle Chinese	Shuāngfēng
步	bù	„to walk”	bo <sup>3</sup>	bu <sup>33</sup>
捕	bǔ	„to grasp”	bo <sup>3</sup>	p <sup>h</sup> u <sup>21</sup>
刨	páo	„to dig”	bæw <sup>1</sup>	bə <sup>33</sup>
跑	páo	„to scrape”	bæw <sup>1</sup>	p <sup>h</sup> ə <sup>21</sup>
盜	dào	„to rob”	daw <sup>3</sup>	də <sup>33</sup>
導	dǎo	„to lead”	daw <sup>3</sup>	t <sup>h</sup> ə <sup>35</sup>

**Table 2:** Examples for irregularities in the readings of Shuāngfēng (ZIHUI).

Some basic examples for the idea of Chen (1972) that homophonous readings in Middle Chinese developing different readings in a given dialect are given in Table 2. Do these examples provide an undisputable proof for the existence of lexical diffusion as an alternative mechanism of sound change?

### 1.5 Labov and the study of sound change in progress

The theory of lexical diffusion is not only contrary to the inherent model of sound change underlying, but also tackles its most important implication: If sound change is by and large regular, it means we can reconstruct ancient stages of languages not reflected in written

<sup>1</sup> My translation, original text: 作為詞彙,要變就都變,因而是一種突變。作為語音,變化是逐漸的,因而是一種漸變。

sources. But if considerable parts of our evidence turn out to reflect sound change processes that were not completely finished, this would make it much more difficult to carry out linguistic reconstruction. It seems, however, that the theory of lexical diffusion is not entirely correct. Firstly, Labov (1981) could show by investigating sound change in progress that there were two basic mechanisms of sound change, one mechanism that diffuses across the lexicon, and one in which a change captures all the words at the same time.

There is no basis for contending that lexical diffusion is somehow more fundamental than regular, phonetically motivated sound change. On the contrary, if we were to decide the issue by counting cases, there appear to be far more substantially documented cases of Neogrammarian sound change than of lexical diffusion. (Labov 1994: 471)

Secondly, what is even more important in this context, it would even open the door for speculations, if one would treat sound change as a process that could be in principle irregular, as it is to be expected that nobody would have tried to resolve Grimm's exceptions if one had thought that these were anyway impossible to be explained by means of regular "sound laws" (Hill 2016). Nevertheless, even when accepting the Neogrammarian idea of regularity, the question remains to which degree this regularity is persistent, given that we know well that processes like borrowing and analogy can mask it.

In this context, the term "mechanism" was used in order to distinguish lexical diffusion from Neogrammarian sound change. Would it not be possible to just use the term "process" instead of "mechanism"?

## 2 Techniques for the investigation of regular sound change

In order to accumulate the data needed to investigate how sound change proceeds, one needs specific techniques for inference. The most prominent method employed by scholars is traditionally called the *comparative method* (Meillet 1925 [1954]), which is essentially a bunch of techniques which are eclectically employed by linguists embarking on historical language comparison. If one asks different linguists, they will often differ with respect to what they think represents the comparative method best, and for this reason, this method is better treated as some kind of an *overarching framework* that scholars use in order to compare languages (Fox 1995, Jarceva 1990, Klimov 1990).

Why would linguists still talk of the comparative method, even if they know from their practice themselves that it is not a unified procedure?

### 2.1 Classical approaches in the framework of the 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

### 3 Phonetic Variation

which the sounds in the descendant languages (the reflex sounds) can be derived in the most convincing fashion.

Cognate List		Alignment			Correspondence List		
English	<i>foot</i>	f	u	t	Eng.	Grk.	Freq.
Ancient Greek	ποδ-	p	o	d	f	p	3 x
English	<i>father</i>	f	a:	θ ə ɪ	f	p <sup>h</sup>	1 x
Ancient Greek	πατέρ-	p	a	t ε r	ɹ	r	2 x
English	<i>fear</i>	f	ɪə	ɪ -	θ	t	1 x
Ancient Greek	φοβέ-	p <sup>h</sup>	o	b e	t	d	1 x
English	<i>fire</i>	f	aɪə	ɪ	<div>irregular match!</div>		
Ancient Greek	πυρ-	p	y	r			

**Table 3:** Detecting regular sound correspondences in classical historical language comparison.

Table 3 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?

## 2.2 Computer-assisted approaches

While traditional accounts on the inference of sound correspondences (and consecutively also accounts on the inference of sound change patterns) are still the predominant way in which linguists analyze the history of the world's languages, computational methods, specifically those that help linguists in their work rather than threatening to replace them, are constantly gaining ground. Among the most important techniques in this context are (1) techniques for *automated phonetic alignment*, which are needed as a basis for identifying corresponding sounds (Kondrak 2000, List 2014), (2) extended techniques for *automated cognate detection* (Arnaud et al. 2017, List et al. 2017), which make use of alignment techniques in order to search for the most likely candidates of related words across languages, and (3) relatively recent techniques for *automated correspondence pattern inference* (List 2019), which infer sound correspondences across multiple languages, offering a first starting point for phonological reconstruction.

The methods for phonetic alignments, cognate detection, and sound correspondence inference are quite advanced until now, and they start providing real help to linguists who investigate so far less thoroughly investigated language families (Chen 2019, Hill and List 2017, Kolipakam et al. 2018). With LingPy (<http://lingpy.org>, List et al. 2019), a stable software package offers basic algorithms for phonetic alignment analyses and cognate detection. Furthermore, the data processed with LingPy can be directly inspected with help of web-based tools, such as the Etymological Dictionary Editor (EDICTOR, <http://edictor.digling.org>, List et al. 2017), allowing linguists to quickly modify their data, correcting the errors made by the algorithm, or converting it to formats needed for the further analysis with help of phylogenetic software. Online tutorials (e.g., <https://calc.hypotheses.org>) along with print tutorials (List et al. 2018) run newcomers through the new techniques.

Another benefit of the methods that may be less evident from the first sight has been presented in a recent experiment on *word prediction*. Since scholars in fieldwork usually do not have time to elicit all words relevant for their study at ones, they can make use of the comparative method (either in a classical or a computer-assisted form) to predict how certain words would sound from the correspondence patterns they observe for the languages under investigation. This was in fact already mentioned by Grimm (1822: 589), who thought there would be a limited possibility to predict the consonantal shape of Germanic words if they were missing. In a recent experiment, we tested the usefulness of computer-assisted word

prediction techniques (Bodt and List 2019), the so far unpublished results indicate that the expert fieldworker was able to predict missing words in the data with an accuracy of about 75%. More studies and experiments will be needed to further test and enhance the suitability of the procedure which was laid out in this pilot study.

If we manage to predict words with an accuracy of 75% (by an expert who made use of computational pre-processing), what does this tell us with respect to the question of the regularity of language change? Is it now regular after all or not?

### 3 Describing Sound Change

#### 3.1 Types of Sound Change

In the following, I repeat some definitions of Trask (2000) on frequently described types of sound change:

- **assimilation** “Any **syntagmatic change** in which some segment becomes more similar in nature to another segment in the same sequence, usually within a single phonological word or phrase” (30).
- **dissimilation** “Any **syntagmatic change** in which one segment changes so as to become less similar to another segment in the same form” (95).
- **metathesis** “Any **syntagmatic change** in which the order of segments (or sometimes of other phonological elements) in a word is altered” (211).
- **tonogenesis** “Any process which leads to the introduction of tones into a language which formerly lacked them” (346).
- **sandhi** “Any of various phonological processes applying to sequences of segments either across morpheme boundaries (*internal sandhi*) or across word boundaries (*external sandhi*)” (296).
- **haplology** “A type of phonological change (of or phonological constraint) in which one of two adjacent syllables of identical or similar form is lost (or fails to appear in the first place)” (146).
- **elision (aphaeresis, syncope, apocope)** “Any of various processes in which phonological segments are lost from a word or a phrase. Specific varieties of elision are often given special names like **aphaeresis**, **syncope**, **apocope**, **synaeresis**, **synizesis**, **synaloepha**. Not infrequently this name is given to specific processes in particular languages” (102).
- **epenthesis** “Any phonological change which inserts a segment into a word or form in a position in which no segment was formerly present” (107).
- **prothesis** “The addition of a segment to the beginning of a word. [...] The opposite is **aphaeresis**” (266).
- **nasalization** “Any phonological process in which a segment acquires a nasal character which it formerly lacked” (224).

Try to find examples for each of the sound change types. What does Trask mean when talking about a ‘syntagmatic change’?

### 3.2 Alternative Sound Change Types

The typical types of sound change often repeated in the literature are no real types in the sense of a useful cross-linguistic classification, since they often are based on particular languages and particular examples. But we can propose a very rough classification that is independent of which language is spoken. This classification may seem trivial, but I consider it nevertheless as useful. We start from thinking of sound change as a function, which receives something as input and then outputs something. To classify sound change types now, we only need to set the input and the output into relation in order to qualify what happens. This leads us to the following five types:

1. continuation,
2. substitution,
3. insertion,
4. deletion, and
5. metathesis.

The following table provides examples for individual sound changes. Can you identify which sound change is happening in each of the examples?

Input	Output	Typ
Urslavisches * <i>žlty</i> 'gelb'	Czech <i>žlutý</i> [ʒluti:] 'gelb' (DERKSEN: 565)	
Althochdeutsch <i>angust</i> [aŋust]	Hochdeutsch <i>Angst</i> [aŋst]	
Althochdeutsch <i>hant</i> [hant]	Hochdeutsch <i>Hand</i> [hant]	
Althochdeutsch <i>ioman</i> [jo-man]	Hochdeutsch <i>jemand</i> [je-mant]	
Althochdeutsch <i>snēo</i> [sne:o]	Hochdeutsch <i>Schnee</i> [ʃne:]	

## References

- Arnaud, A. S., D. Beck, and G. Kondrak (2017). "Identifying cognate sets across dictionaries of related languages". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. (Copenhagen, 09/07–09/11/2017). Association for Computational Linguistics, 2509–2518.
- Běijīng Dàxué, ed. (1989). *Hànyǔ fāngyīn zìhuì* 漢語方音字彙 [Chinese dialect character pronunciation list]. 2nd ed. Běijīng 北京: Wénzì Gǎigé 文字改革; Electronic Edition: Wang, W. S.-Y. and C.-C. Cheng, eds. (1969). *DOC. Dictionary On Computer*. URL: <http://starling.rinet.ru/>.
- 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.
- Campbell, L. (1999). *Historical linguistics. An introduction*. 2nd ed. Edinburgh: Edinburgh Univ. Press.
- Chen, E. (2019). "Phonological reconstruction of Proto-Kampa consonants". *Berkeley Papers in Formal Linguistics* 2.1, 1–56.
- Chen, M. (1972). "The time dimension. Contribution toward a theory of sound change". *Foundations of Language* 8.4, 457–498. JSTOR: 25000618.
- Derksen, R., comp. (2008). *Etymological dictionary of the Slavic inherited lexicon*. Leiden Indo-European Etymological Dictionary Series 4. Leiden and Boston: Brill.
- Fox, A. (1995). *Linguistic reconstruction. An introduction to theory and method*. Oxford: Oxford University Press.
- Grassmann, H. (1863). "Ueber die aspiraten und ihr gleichzeitiges vorhandensein im an- und auslaute der wurzeln". *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete des Deutschen, Griechischen und Lateinischen* 12.2, 81–110.
- Grimm, J. (1822). *Deutsche Grammatik*. 2nd ed. Vol. 1. Göttingen: Dieterichsche Buchhandlung. Google Books: [MnsKAAAAIAAJ](https://books.google.com/books?id=MnsKAAAAIAAJ).
- Hill, N. (2016). "A refutation of Song's (2014) explanation of the 'stop coda problem' in Old Chinese". *International Journal of Chinese Linguistics* 2.2, 270–281.

- Hill, N. W. and J.-M. List (2017). "Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages". *Yearbook of the Poznań Linguistic Meeting* 3.1, 47–76.
- Jarceva, V. N., ed. (1990). *Lingvističeskij énciklopedičeskij slovar (Linguistical encyclopedical dictionary)*. Moscow: Sovetskaja Enciklopedija.
- Kiparsky, P. (1988). "Phonological change". In: *Linguistics. The Cambridge survey*. Vol. 1: *Linguistic theory. Foundations*. Ed. by F. J. Newmeyer. Cambridge et al.: Cambridge University Press, 363–415.
- Klimov, G. A. (1990). *Osnovy lingvističeskoj komparativistiki* [Foundations of comparative linguistics]. Moscow: Nauka.
- 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.
- Labov, W. (1994). *Principles of linguistic change*. Vol. 1: *Internal factors*. Malden, Oxford, and West Sussex: Wiley-Blackwell.
- Labov, W. (1981). "Resolving the Neogrammarian Controversy". *Language* 57.2, 267–308. JSTOR: 413692.
- List, J.-M. (2014). *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.
- (2019). "Automatic inference of sound correspondence patterns across multiple languages". *Computational Linguistics* 1.45, 137–161.
- 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., 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.
- List, J.-M., S. Greenhill, T. Tresoldi, and R. Forkel (2019). *LingPy. A Python library for quantitative tasks in historical linguistics*. Version 2.6.5. URL: <http://lingpy.org>.
- Lottner, C. F. (1862). "Ausnahmen der ersten lautverschiebung". *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete des Deutschen, Griechischen und Lateinischen* 11.3/4, 161–205. JSTOR: 40844790.
- Meier-Brügger, M. (2002). *Indogermanische Sprachwissenschaft*. In collab. with M. Fritz and M. Mayrhofer. 8th ed. Berlin and New York: de Gruyter.
- Meillet, A. (1954). *La méthode comparative en linguistique historique* [The comparative method in historical linguistics]. Repr. Paris: Honoré Champion.
- Osthoff, H. and K. Brugmann (1878). *Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen*. Vol. 1. Leipzig: Hirzel.
- Rask, R. K. (1818). *Undersøgelse om det gamle Nordiske eller Islandske sprogs oprindelse* [Investigation of the origin of the Old Norse or Icelandic language]. Copenhagen: Gyldendalske Boghandlings Forlag. GoogleBooks: cWgJAAAAQAAJ; English translation: – (1993). *Investigation of the origin of the Old Norse or Icelandic language*. Trans. by N. Ege. Travaux du Cercle Linguistique de Copenhague 26. Copenhagen: The Linguistic Circle of Copenhagen.
- Trask, R. L., comp. (2000). *The dictionary of historical and comparative linguistics*. Edinburgh: Edinburgh University Press.
- Verner, K. A. (1877). "Eine Ausnahme der ersten Lautverschiebung [An exception to the first sound shift]". *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete der Indogermanischen Sprachen* 23.2, 97–130.
- Wang, W. S.-Y. (1969). "Competing changes as a cause of residue". *Language* 45.1, 9–25. JSTOR: 411748.
- (2006). *Yǔyán, yǔyīn yǔ jìshù* 語言、語音與技術 [Language, phonology and technology]. Shànghǎi 上海: Xiānggǎng Chéngshì Dàxué.

## Phonetic Variation (Typological Viewpoint)

### 1 Open questions on sound change

In their very influential paper titled *Empirical foundations for a theory of language change*, Weinreich et al. (1968) proposed a set of problems for future work in historical linguistics. That Campbell (1999: 194f) repeats these actual problems, shows that there has not been much success in increasing our understanding with respect to these problems. Five major problems are summarized by Campbell: the problem of the (1) *constraints* of language change, the problem of the (2) *transition* of different change processes, the problem of the (2) *embedding* of change in linguistic and social relations, the problem of the (4) *evaluation* of change with respect to the speakers of a given language, and the problem of the (5) *actuation* of change, i.e., the question of why particular changes occur at particular times and places. Coseriu (1973: 65f) lists only three problems, namely “(a) el problema racional del cambio” (*why do languages change at all?*), “(b) el problema general de los cambios” (*under which circumstances do languages change?*), and “(c) el problema histórico de tal cambio determinado” (*why do particular changes take place*). In the following, we will discuss these three problems in more detail, specifically concentrating on sound change, and compare them with the ones mentioned by Campbell.

In this list of problems, actuation of sound change refers to the occurrence of particular changes in particular languages. Could one also understand the problem in a broader context?

#### 1.1 The rational problem of sound change

The question of why sound change happens after all is difficult to answer, since it is difficult to find a direct benefit resulting from the process for a given language system (Anttila 1976). In an evolutionary framework, we would thus say that there is no apparent selective pressure that would favor the modification of sounds. On the contrary, it is known that sound change may increase the amount of grammatical irregularities.<sup>1</sup> That not all changes in evolution need to yield a direct benefit, however, is nothing new for biologists, who have been investigating what they call phenomena of *drift* already for a long time. In order to explain why sound change happens, most scholars nowadays assume variation at the synchronic level as its starting point (Ohala 1989, Kümmel 2008: 22, Paul 1880 [1886]: 30). It is further assumed that language systems are *robust* enough to tolerate a certain amount of sound change (Hockett 1965: 203f). Robustness itself results from the redundancy of speech (*ibid.*),<sup>2</sup> which can be seen as an important feature of language, as it guarantees its functioning as a communication system. While these neutral theories of sound change seem to be obvious (and have been mentioned already quite early in the linguistic literature), it is less clear to which degree certain selective aspects could not also play a role in sound change. Blasi et al. (2019), for example, assume that the pronunciation of labiodentals was greatly facilitated along with changes in the diet of early humans. Everett et al. (2015) claim that tone languages evolve more frequently in humid climates. While neutral theories of evolution can in principle explain why sound change should be possible, we are still far away from being able to draw a conclusive picture of all the factors that may influence it.

<sup>1</sup> Anttila calls this *Sturtevant's paradox*, namely that *regular* sound change produces *irregularity* in language systems, while *irregular* analogy produces *regularity* in language systems.

<sup>2</sup> See Winter (2014) for a detailed discussion of robustness.

Labov (2001: 15) emphasizes that "the evolution of species and the evolution of language are identical in form, although the fundamental mechanism of the former is absent in the latter". What does he mean?

## 1.2 The general problem of sound change

If we look at the general patterns of sound change that can be observed for the languages of the world, we can distinguish two basic conditions of sound change, *phonetic conditions* and *systemic conditions*. Phonetic conditions can be further subdivided into *articulatory* and *acoustic conditions*. When trying to explain why certain sound changes can be observed more frequently across different languages of the world, many linguists tend to explain this by invoking phonetic factors. If the sound *p*, for example, turns into an *f*, this is not necessarily surprising given the strong similarity of the sounds. But similarity can be measured in two ways: one can compare the similarity with respect to the production of a sound by a speaker, and with respect to the perception of the sound by a listener. While production of sounds is traditionally seen as the more important factor contributing to sound change (Hock 1991: 11), there are clear examples for sound change due to misperception and re-interpretation by the listeners (Ohala 1989: 182). Some authors go as far as to claim that production-driven changes reflect regular *internal* language change (which happens gradually during acquisition, or – depending on the theory – also in later stages Bybee 2002), while perception-based changes rather reflect change happening in second language acquisition and language contact (Mowrey and Pagliuca 1995: 48).

While the interaction of production and perception has been discussed in some detail in the linguistic literature, the influence of systemic factors has so far only rarely been regarded. What I mean by this factor is the old structural idea that a language can be seen as a system, and that certain changes in the system may be explained exclusively as resulting from systemic constellations. As a straightforward example, consider the difference in *design space* for the production of consonants, vowels, and tones. In order to maintain pronunciability and comprehensibility, it is useful for the sound system of a given language, to fill in those spots in the design space that are maximally different from each other. The larger the design space and the smaller the inventory, the easier it is to guarantee its functionality. Since design spaces for vowels and tones are much smaller than for consonants, however, these sub-systems are more easily disturbed, which could be used to explain the presence of chain shifts of vowels, or *flip-flop* in tone systems (Wang 1967: 102). Systemic considerations play an increasingly important role in evolutionary theory, and, as shown in List et al. (2016), also be used as explanations for phenomena as strange as the phenomenon of Sapir's *drift* (Sapir 1953).

There is a lot of discussing in the linguistic literature with respect to the time when sound change occurs: should it occur during the life time of a human being, or should it rather occur only at the time of acquisition? What data would we expect for both scenarios?

## 1.3 The historical problem of sound change

The historical problems, i.e., the particular problems of sound changes in particular languages, are usually much better understood than the general or the rational problem, as presented above. As in all cases of historical language comparison, however, typological

(general) investigations and particular investigations should ideally guide each other. Unfortunately, general factors are rarely considered when discussing individual proposal for linguistic reconstruction. This was already criticized by Jakobson (1958), who criticized that linguists would rarely consider typological aspects when proposing their reconstructions for unattested languages, but the situation has not changed much in the meantime. The biggest problem in this context seems to be the general lack of cross-linguistic catalogs of attested or proposed sound change processes.

Why do linguists often defend to ignore typological evidence in reconstruction?

## 2 Open problems on sound patterns

While it is not entirely clear what *sound patterns* are, we can roughly follow the usage in Blevins (2004) in saying that sound patterns refer to certain characteristics of the sound systems of spoken languages. Apart from the *sound inventories* we find in spoken languages, we also find specific rules and restrictions on the combination of sounds to form words, so the *phonotactic restrictions* or the *syllable inventories* should also be addressed when talking about *sound patterns* and their differences from a typological perspective. While *sound change* refers to the diachronic dimension, *sound patterns* can be seen as its synchronic counterpart, yet it should never be forgotten that the two are intertwined, and that one cannot study one without the other. In the following, we will try to follow Coseriu (1973) again in assessing the specific problems of sound patterns.

Why would sound change and sound patterns be intertwined?

### 2.1 The rational problem of sound patterns

If we try transfer the idea of “el problema racional del cambio” to the synchronic notion of sound patterns, we have to start by asking concrete questions on sound patterns. While this is already done when studying language change, where the most mysterious question is probably *why* it happens after all, it is less clear what a “rational question” regarding sound patterns could look like. But if we carefully examine some major questions that are regularly being asked specifically in generative syntax, which in some sense deals more with rational than with general or historical questions, we can at least propose some rational problems for the study of sound patterns, which we can present in form of simple questions:

- Why are sound inventories of natural spoken languages always restricted?
- What complexity do phonotactic rules have in natural spoken languages?
- Are sound inventories similar to natural kinds or are they a product of history?
- If one cannot find a counterpart to sound inventories and the notion of a *phoneme* in signed languages (but this is so far unclear), what does that say about the importance of the double-articulation (Martinet 1984) for language in general?

Mielke (2008) makes an interesting analogy between two perspectives on phoneme inventories and phonological rules, one saying that they are like Starbucks, that is, a brand that has a clear-cut design, and one saying that they are like Ethiopian restaurants that are called "Blue Nile", that is, like an independently emerged structure of similar restaurants which are similar because of the cuisine they offer and the origin of the cuisine. How can we relate this to our "rational question" of sound patterns?

## 2.2 The general problem of sound patterns

When transferring the idea of a *general question of sound change* to the idea of sound patterns, this will result in specific questions regarding the structure of sound patterns of all spoken languages in the world. As Blevins (2004) shows nicely, there are many aspects of synchronic sound systems and phonotactic rules that are remarkably similar across the languages of the world, and analyzing these aspects systematically shows means addressing the general question of sound patterns.

One interesting general problem is the question of "symmetry" in sound change. We often find strikingly symmetric systems when looking at sound inventories across languages. The general question would be to which degree symmetry has an effect on concrete sound inventories which we can observe in the languages of the world. But could it not also be possible that symmetry is an artifact of our way to investigate languages, rather than a real phenomenon?

## 2.3 The historical problem of sound patterns

The historical problem of sound patterns is – similar to the historical problem of sound change – a problem that belongs to what Haspelmath (2019) calls *p-linguistics* as opposed to *g-linguistics*. P(articular) linguistics refers to those investigations that deal with a particular language or a particular language family, while g(eneral) linguistics refers to those investigations that try to find out something about language in general. Following Gabelentz (2016), we should add c(omparative) linguistics as a specific type of investigations in which more than one particular language is being compared, although one could say that c(omparative) linguistics is a broader subtype of p(articular) linguistics. In any case, when dealing with the historical problem of specific sound patterns, be it in a particular language family or a particular language, the questions that we need to ask differ quite drastically from the general questions we would ask otherwise.

What particular historical questions can we ask with respect to the problem of sound patterns?

## 3 Typological databases on sound patterns and sound change

Quite a few databases dealing with sound patterns have been published so far. Databases dealing with sound change, on the other hand, are extremely rare. What the databases offer differs substantially, with some only covering certain parts of the world, some covering

### 3 Phonetic Variation

syllable structures and phonological rules, while others restrict data to categorical variables, such as *large* or *small* phonological inventory size. In the following, we will quickly look at some examples for databases dealing with sound patterns and sound changes, and then present the recent effort to establish a reference catalog for transcription systems.

If one says a phoneme inventory of a given languages is large or small, what do you think large refers to and what do you think is meant with small?

#### 3.1 Databases of sound change patterns

The by far largest traditional study on the typology of sound change is Kümmel's (2008) book *Konsonantenwandel (consonant change)*, in which the author surveys sound change processes discussed in the literature on Indo-European and Semitic languages. As the title of the book suggests, it concentrates on the change of consonants, which are – probably due to the larger design space – also the class of sounds that shows stronger cross-linguistic tendencies. The book is based on the thorough inspection of the literature on consonant change in Indo-European and Semitic linguistics. The procedure, by which this collection was carried out, can be seen as the gold standard, which any future attempt of enlarging the given collection should be carried out. What is specifically important, and also very difficult to achieve, is the *harmonization* of the evidence, which is nicely reflected in Kümmel's introduction, where he mentions that one of the main problems was to determine what the scholars actually meant with respect to phonetics and phonology, when describing certain sound changes (Kümmel 2008: 35). The major drawback of the collection is that it is not (yet) available in digital form. Given the systematicity with which the data was collected, it should be generally possible to turn the collection into a database, and it is beyond doubt that this collection could offer interesting insights into certain tendencies of sound change.

Another collection of sound changes collected from the literature is the mysterious *Index Diachronica*, a collection of sound changes collected from various language families by an anonymous person who does not want to disclose her real name. Up to now, this collection even has a Searchable Index (<https://chridd.nfshost.com/diachronica/>) which allows scholars to click on a given sound and to see in which languages this sound is involved in some kind of sound change. What is a pity about the resource is that it is difficult to use, given that one does not really know where it actually comes from and how the information was in the end extracted from the sources. If the anonymous author would only decide to put it (albeit anonymously, or under pseudonym) on a public preprint server, such as, for example, Humanities Commons, <https://hcommons.org>, this would already be excellent, as it would allow those who are interested in pursuing the idea of collecting sound changes from the literature an excellent starting point to check the sources, and to further digitize the resource.

Right now, the resource seems to be mostly used by *conlangers*, i.e., people who create artificial languages as a hobby (or profession). Conlangers are often refreshingly pragmatic and may come up with very interesting and creative ideas on how to address certain data problems in linguistics, which “normal” linguists would refuse to do. There is a certain tendency in our field to ignore certain questions, either because scholars think it would be too tedious to collect the data to address a certain problem, or they consider it impossible to be done “correctly” from the start.

As a last and fascinating example, I have to mention the study by Yang and Xu (2019) in which the authors review studies on concrete examples of tone change in South-East Asian languages, trying to identify cross-linguistic tendencies. Before I read this study, I was not

aware that tone change had at all been studied in concrete, since most linguists consider the evidence for any kind of tendency far too shaky, and reconstruct tone exclusively as an abstract entity. The survey by Yang and Xu, however, shows clearly that there seem to be at least some tendencies, and that they can be identified by invoking a careful degree of abstraction when comparing tone change across different languages.

Are there any problems in the approaches to establishing a database of sound change patterns discussed so far?

### 3.2 Databases of phoneme inventories

There are quite a few different datasets of phoneme inventories. The tradition started most notably with the UPSID database which was established already in the 1980s (Maddieson 1984), listing sound inventories for some 400 of the world's languages. By now, the data is still used as legacy data in some projects, and it can still be retrieved (for example through an interface provided by the University of Frankfurt, <http://web.phonetik.uni-frankfurt.de/upsid.html>), but it is no longer actively maintained. When the World Atlas of Language Structures was published in 2005 (Haspelmath et al. 2005) and later published online, first in 2008, with substantial updates in the most recent version from (Dryer and Haspelmath 2013), it already contained part of the UPSID data, provided by specific chapters written, among others, by Maddieson (2013), expanded to more than 2000 languages. But while the UPSID database had offered the phoneme inventories in form of phonetic transcriptions, the information in WALS only gives approximate information, by dividing inventories in Small, Moderately small, Average, Moderately large, and Large.

In 2013, Maddieson et al. (2013) announced the LAPSyD database, the Lyon-Albuquerque Phonological Systems Database, which provides many enhancements with respect to UPSID, also in terms of coverage (with 692 varieties), and can also be searched online (<http://www.lapsyd.ddl.cnrs.fr/lapsyd/>). At about the same time, Moran et al. (2014) published the PHOIBLE database (ibid.), the *Phonetics Information Base and Lexicon*, which offers phoneme inventories for more than 2000 language varieties of the world and has recently been published in a new version (Moran and McCloy 2019), online available at <https://phoible.org>. Also in 2013, Donohue et al. (2013) published the *World phonotactics database*, which is by now, however, no longer publicly available, although a snapshot still exists on the Way Back Machine at <https://web.archive.org/web/20190608183108/http://phonotactics.anu.edu.au/>, but the first author has publicly called out linguists who had been using that older version without his permission, which renders the idea of free science useless, given that the database itself was derived from sources, which the authors used without explicitly asking for permission. Similar to the World Atlas of Language Structures, the World phonotactics database is based on categorical features instead of providing actual language data. Thus, as a result, we find features such as *CVC language*, *CV language*, etc., pointing to specific restrictions in the syllable inventories.

There are more sound inventory databases available, such as *PBase*, a database of phonological patterns, based on the original work by Mielke (2008), online available at <https://pbase.phon.chass.ncsu.edu/> Mielke (2015), which contains – apart from sound inventories – also phonological rules and information on the distribution of sound segments, or the *Database of Eurasian phonological inventories* (Nikolaev et al. 2015), which contains inventories for Eurasian languages, accessible at [eurasianphonology.info](http://eurasianphonology.info).

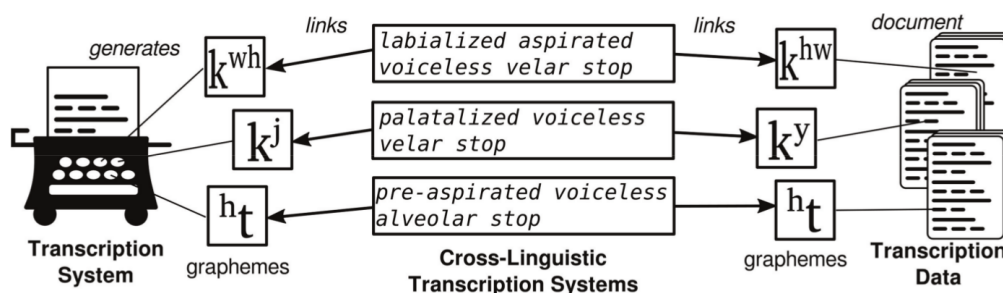
What is the disadvantage of collecting data in categorical form, as it was done for WALS and the World phonotactic database?

### 3.3 Cross-Linguistic Transcription Systems

Contrary to what non-practitioners might expect, the systems of phonetic notation used by linguists are highly idiosyncratic. Not only do various linguistic subfields disagree on the specific symbols they use to denote the speech sounds of languages, but also in large databases of sound inventories considerable variation can be found.

Inspired by recent efforts to link cross-linguistic data with help of *reference catalogues* (Glottolog, Hammarström et al. 2019, Concepticon, List et al. 2020) across different resources, the Cross-Linguistic Transcription Systems initiative presented initial efforts to link different phonetic notation systems to a catalogue of speech sounds (Anderson et al. 2018). This was achieved with the help of a data-base accompanied by a software framework that uses a limited but easily extendable set of non-binary feature values to allow for quick and convenient registration of different transcription systems, while at the same time linking to additional datasets with restricted inventories. Linking different transcription systems makes it possible to conveniently translate between different phonetic transcription systems, while linking sounds to databases allows users quick access to various kinds of metadata, including feature values, statistics on phoneme inventories, and information on prosody and sound classes. The current version of the CLTS database and software package links five different transcription systems and fifteen different transcription datasets (List et al. 2019). It can be accessed at <https://clts.clld.org> and <https://digling.org/calc/clts/> and an additional Python software package allows to use the data and code in Python applications.

The figure below describes the basic principles of the CLTS reference catalog. There seems to be a strict division between transcription systems and transcription data. What is the core of this distinction?



## References

- Anderson, C., T. Tresoldi, T. C. Chacon, A.-M. Fehn, M. Walworth, R. Forkel, and J.-M. List (2018). "A Cross-Linguistic Database of Phonetic Transcription Systems". *Yearbook of the Poznań Linguistic Meeting* 4.1, 21–53.
- Anttila, R. (1976). "The acceptance of sound change by linguistic structure". In: *Recent developments in historical phonology*. Ed. by J. Fisiak. The Hague, Paris, New York: de Gruyter, 43–56.
- 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.
- Blevins, J. (2004). *Evolutionary phonology. The emergence of sound patterns*. Cambridge: Cambridge University Press.
- Bybee, J. L. (2002). "Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change". *Language Variation and Change* 14, 261–290.
- Campbell, L. (1999). *Historical linguistics. An introduction*. 2nd ed. Edinburgh: Edinburgh Univ. Press.
- Coseriu, E. (1973). *Sincronía, diacronía e historia. El problema del cambio lingüístico* [Synchrony, diachrony, and history. The problem of linguistic change]. Madrid: Biblioteca Románica Hispánica.
- 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). *The World Atlas of Language Structures 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.
- Gabelentz, G. von der (2016). *Die Sprachwissenschaft: Ihre Aufgaben, Methoden und bisherigen Ergebnisse*. Ed. by M. Ringmacher and J. McElvenny. Berlin: Language Science Press.
- Hammarström, H., M. Haspelmath, and R. Forkel (2019). *Glottolog. Version 4.1*. Jena: Max Planck Institute for the Science of Human History.
- Haspelmath, M. (2019). "Confusing p-linguistics and g-linguistics: Philosopher Ludlow on «framework-free theory»". *Diversity Linguistics Comment* 7.6. eprint: <https://dlc.hypotheses.org/1801>.
- Haspelmath, M., M. Dryer, D. Gil, and B. Comrie (2005). *The world atlas of language structures*. Oxford: Oxford University Press.
- Hock, H. H. (1991). *Principles of historical linguistics*. 2nd ed. Berlin: Mouton de Gruyter.
- Hockett, C. F. (1965). "Sound change". *Language* 41.2, 185–204. eprint: 411873 (jstor).
- Jakobson, R. (1958). "Typological studies and their contribution to historical comparative linguistics". In: *Proceedings of the Eighth International Congress of Linguistics*. Oslo, 17–35; Reprint: – (1971). "Typological studies and their contribution to historical comparative linguistics". In: *Selected Writings*. Vol. 1: *Phonology*. The Hague: Mouton, 523–532.
- Kümmel, M. J. (2008). *Konsonantenwandel* [Consonant change]. Wiesbaden: Reichert.
- Labov, W. (2001). *Principles of linguistic change*. Vol. 2: *Social factors*. Malden, Oxford, and West Sussex: Wiley-Blackwell.
- List, J.-M., J. S. Pathmanathan, P. Lopez, and E. Baptiste (2016). "Unity and disunity in evolutionary sciences: process-based analogies open common research avenues for biology and linguistics". *Biology Direct* 11.39, 1–17.
- List, J.-M., C. Anderson, T. Tresoldi, C. Rzymiski, S. Greenhill, and R. Forkel (2019). *Cross-Linguistic Transcription Systems. Version 1.3.0*. Jena: Max Planck Institute for the Science of Human History.
- List, J. M., C. Rzymiski, S. Greenhill, N. Schweikhard, K. Panykh, A. Tjuka, M.-S. Wu, and R. Forkel (2020). *Concepticon. A resource for the linking of concept lists (Version 2.3.0)*. Version 2.3.0. URL: <https://concepticon.clld.org/>.
- Maddieson, I. (1984). *Patterns of sounds*. Cambridge and New York: Cambridge University Press.
- (2013). "Consonant Inventories". In: *The World Atlas of Language Structures Online*. Ed. by M. S. Dryer and M. Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology.
- Maddieson, I., S. Flavie, E. Marsico, C. Coupé, and F. Pellegrino. (2013). "LAPSyD: Lyon-Albuquerque Phonological Systems Database". In: *Proceedings of Interspeech*. (Lyon, 08/25–08/29/2013).
- Martinet, A. (1984). "Double articulation as a criterium for linguisticity". *Language Sciences* 6.1, 31–38.
- Mielke, J. (2008). *The emergence of distinctive features*. Oxford: Oxford University Press.
- (2015). *PBase. A database of phonological patterns*. Raleigh: North Carolina State University.
- Moran, S. and D. McCloy, eds. (2019). *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History.
- Moran, S., D. McCloy, and R. Wright, eds. (2014). *PHOIBLE Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.
- Mowrey, R. and W. Pagliuca (1995). "The reductive character of articulatory evolution". *Rivista di Linguistica* 7, 37–124.
- Nikolaev, D., A. Nikulin, and A. Kukhto (2015). *The database of Eurasian phonological inventories*. Moscow: RGGU.
- Ohala, J. J. (1989). "Sound change is drawn from a pool of synchronic variation". In: *Language Change: Contributions to the study of its causes*. Ed. by L. E. Breivik and E. H. Jahr. Berlin: Mouton de Gruyter, 173–198.
- Paul, H. (1886). *Prinzipien der Sprachgeschichte*. 2nd ed. Halle: Max Niemeyer. prinziendersp01paulgoog: ia.
- Sapir, E. (1953). "Language. An introduction to the study of speech".
- Wang, W. S.-Y. (1967). "Phonological features of tone". *International Journal of American Linguistics* 33.2, 93–105. JSTOR: 1263953.
- Weinreich, U., W. Labov, and M. I. Herzog (1968). "Empirical foundations for a theory of language change". In: *Directions for historical linguistics: A symposium*. Ed. by W. P. Lehmann and Y. Malkiel. Austin: University of Texas Press, 95–189 (–195?).
- Winter, B. (2014). "Spoken language achieves robustness and evolvability by exploiting degeneracy and neutrality". *BioEssays* 36, 960–967.
- Yang, C. and Y. Xu (2019). "A review of tone change studies in East and Southeast Asia". *Diachronica* 36.3, 417–459.



## **4 Structural Variation**

The two sessions focus on structural variation from a historical, areal, and typological perspective.

### Structural Variation (Historical Viewpoint)

#### 1 What is “Grammar”?

Before we start speculating too much about what a grammar is and what the term means, let's do what we should always do when trying to find a solution, and this is: look it up in a good dictionary of linguistics. Nowadays, one could also try and look it up in Wikipedia, but dictionaries of linguistics have the advantage that they can be attributed to one person or a group of persons who have compiled them, so we can quote them as any other source.

Originally, grammar designated the ancient study of the letters of the alphabet and in the middle ages of the entirety of Latin language, stylistics, and rhetoric. The term ‘grammar’ is presently used to refer to various areas of study. (Bussmann 1996: 482)

While this answer is surely not satisfying, Bussmann offers a more detailed account on the different notions of grammar, distinguishing at least four major notions:

- (1) Morphology and syntax of a language.
- (2) A system of rules.
- (3) A theory about how language works.
- (4) A description of the major regularities of a language.

Do you know at least one linguistic study devoted to each of the four notions of grammar mentioned in Bussmann?

##### 1.1 What is Structure?

##### 1.2 What is a Word?

If we follow the distinction of phonological rules (phonotactics), by which phonemes are manipulated in such a way that they form words, and morphosyntactical (or grammatical) rules (morphotactics), by which words are manipulated in such a way that they form sentences, it is obvious that the unit of the *word* needs a proper definition. When reading Dionysios Thrax (ca. 170-90 BC), who wrote one of the first attested grammars of Greek, the distinction between word and sentence looks very straightforward.

A word is the smallest part of an ordered sentence.<sup>1</sup>

A Sentence is a combination of words, either in prose or in verse, making complete sense.<sup>2</sup> (Technē grammatikē)

Defining the unit *word* cross-linguistically, however, is not as easy as it may seem. In German, for example, there is quite some debate with respect to certain constructions where it is unclear for many people to determine if they represent one or several words. Examples can be specifically found in a couple of verbal compounds, such as *radfahren* vs. *Rad fahren* “to cycle”, *leerkaufen* vs. *leer kaufen* “empty (a shop by buying all they offer)”, or *Biertrinken* vs. *Bier trinken* “drinking of beer”. What these cases have in common

<sup>1</sup>Original text: λέξις ἐστὶ μέρος ἐλάχιστον τοῦ κατὰ σύνταξιν λόγου.

<sup>2</sup>Original text: λόγος δέ ἐστι πεζῆς λέξεως σύνθεσις διάνοιαν αὐτοτελῇ δηλοῦσα.

is that they are a compound with a Verb as the second part, and a noun (adjective or substantive) as first part. Furthermore, in main clauses, the first part is usually placed after the verb part (*Ich fahre Rad, Ich trinke Bier, Ich kaufe den Laden leer*). While most German speakers probably agree that all three examples consist of two parts which all represent words themselves, there is considerable disagreement among speakers if they should be treated as two different words when being compounded or not.

The problem becomes even more complicated when looking at languages that are genetically even farer away from German and Greek, such as, for example, Chinese, the apparently clear picture begins to blur even more.

The 'word' is a clear and intuitive notion in English, because in the culture of English speakers the concept of the 'word' is particularly salient and robust [...]. This is what Chao called the sociological word (Chao 1968: 136-138): the unit that the society and culture takes to be the salient, critical subcomponent of an utterance [...]. In Chinese, however, the word is by no means a clear and intuitive notion. In Chinese language and culture, the clear and intuitive notion – the sociological word – is the *zì* 字. The term *zì* actually has two distinct meanings in popular usage: it can mean either a morpheme in the spoken language, or it can mean a written Chinese character. (Packard 2000: 14f)

One of the most crucial examples for the problems one has in determining *wordhood* in Chinese *chīfàn* 吃饭 "eat rice=to eat", *shuìjiào* 睡觉 "sleep a sleep=to sleep", or *tiàowǔ* 跳舞 "dance a dance=to dance". In all cases, the translational equivalents will be single words in German and English, and the meaning of the objects are completely bleached in Chinese, so that it seems straightforward to assume that they serve as mere placeholders, and that one should treat them semantically as one word when analyzing them, not two. But on the other hand, the objects behave as normal objects of Chinese transitive verbs, so it would syntactically make much more sense to treat them as two words.

If we accept that – at least for the time being – we cannot determine if the unit *word* has any cognitive reality in all speakers no matter what language they speak, it seems to be best to treat the unit *word* as a *comparative concept* in the sense of Haspelmath (2010), that is, some abstraction of which we hope that it will help us to make some general findings about human languages. If we do so, however, we always need to ask if a certain definition of *word* makes sense cross-linguistically by predicting properties of as many languages as possible, or – if we define a word for a particular language alone – whether the notion of a word helps us in the internal description of the language.

Packard distinguishes the "sociological", the "lexical", the "semantic", the "phonological", the "morphological", the "syntactic", and the "psycholinguistic" word. Try to imagine what he means with as many of these word types as you can.

### 1.3 What are Parts of Speech?

Not all words in a given language are the same with respect to the way in which they can be used to form sentences. This was already clear for the ancient Greeks, as we can see – again – in the grammar by Thrax, who assigned words to eight different classes.

There are eight parts of speech: Noun, Verb, Participle, Article, Pronoun, Preposition, Adverb, and Conjunction.<sup>3</sup>

<sup>3</sup> τοῦ δὲ λόγου μέρη ἐστὶν ὀκτώ· ὄνομα, ῥῆμα, μετοχή, ἄρθρον, ἀντωνυμία, πρόθεσις, ἐπίρρημα, σύνδεσμος. ἡ γὰρ προσηγορία ὡς εἶδος τῷ ὀνόματι ὑποβέβληται.

## 4 Structural Variation

When looking at the specific definitions of each of the eight parts of speech, we can see that Thrax made use of a mix of different criteria, including formal criteria (e.g., *declinability*, πτωτικόν as a criterion for nouns and adjectives), syntactic criteria (“A preposition is a word placed before any of the parts of speech [...]”, πρόθεσις ἐστὶ λέξις προτιθεμένη πάντων τῶν τοῦ λόγου μερῶν [...]), or semantic (“[...] signifying something either concrete or abstract [...]”, σῶμα ἢ πρᾶγμα σημαῖνον). Using various mixed criteria for the identification of the major parts of speech in particular languages is still the most common and most widespread technique (see e.g., the overview in Kempgen 1981 for the Russian tradition).

While the generally recognized parts of speech are more or less the same for most European languages, although there is a constantly ongoing debate on some marginal examples, it is again important to keep in mind that part of speech systems are mostly created for one language only and that languages may well differ quite drastically with respect to the classes into which they characterize their words. Chinese is – again – an interesting example, as we can see from Gabelentz (1881 [1953]: 112), who tells us that the classical Chinese classification of words divides these based on their syntactic function alone into two classes.

- a.) 實字 **šit-tsí**, full or essential words and 虛字 **hiũ-tsí**, empty, that is, immaterial or pure form words (particles)<sup>4</sup>
- b.) 活字 **huot-tsí**, living words are verbs in contrast to nouns, 死字 **ssí-tsí**, dead words. This distinction is important, since many words can be used as verbs and nouns at the same time. (ibid.)<sup>5</sup>

Below is a quote (translated from German) which praises the part of speech classification system developed by Thrax and other scholars in the context of European grammar writing. Do you think it is justified to compare these systems with the classification of plants or the identification of genes and their function in the human genome?

The big classification systems which were only developed in modern times in the natural sciences (or are still being developed by now, such as the description of the structure of the human genome) have – as far as language is concerned – already been established in antiquity. Weber (2002: 191)<sup>6</sup>

### 1.4 What are Grammatical Distinctions?

When talking of grammar, one is often tempted to talk of regularities that occur in a given language. However, scholars are – as far as these regularities are concerned – often quite eclectic when it comes to the regularities which they discuss in their studies. For example, the fact that Russian has a very regular structure by which a noun denoting an animal can be turned into its meat by adding *-ina* (compare *'kurica* “chicken” vs. *kur'jatina* “chicken meat”, *so'baka* “dog” vs. *sobachina* “dog meat”, see Hippius 1998: 1102).

<sup>4</sup>My Translation, original text: 實字 **šit-tsí**, volle oder Stoffwörter, und 虛字 **hiũ-tsí**, leere, d. i. immaterielle oder Formwörter (Partikeln).

<sup>5</sup>My translation, original text: 活字 **huot-tsí**, ‘lebende Wörter’ sind Verba im Gegensatz zu den Nominibus, 死字 **ssí-tsí**, ‘toden Wörtern’. Diese Unterscheidung ist wichtig, weil viele Wörter bald als Verba, bald als Nomina angewandt werden.

<sup>6</sup>My translation, original text: “Die großen Klassifikationssysteme, die in den Naturwissenschaften erst in der Neuzeit entwickelt wurden [...] oder noch werden (z. B. die Beschreibung der Struktur des menschlichen Genoms), sind für die Sprache bereits in der Antike geschaffen worden.”

However, nobody would include this pattern into a regular grammar of Russian, because it is generally considered that it is not describing the kind of regularity in terms of inflection that would be appropriate to be included in a grammar. In the words of Mel'čuk (1974: 98f), grammatical categories must present a "set of mutually exclusive (alternative) meanings" which is characterized by *obligatoricity*, *size*, *importance*, and *regularity of the meaning*. Even if the "meatative" in Russian would thus turn out to be extremely regular in its meaning, and if speakers would feel obliged to make a clear distinction between the animal and its meat, it could still be excluded from the investigation, since it would not fulfill the criterion of size. However, scholars are far away from arriving at a *communis opinio* in this debate, and especially also the last years have seen many debates and many attempts to define once and for ever what grammatical categories are and how one can test if a given language expresses them. Until now, however, one cannot see a big progress in the discussion.

In Chinese, there is a construction that is almost identical in its function with the *-ing* form in English. However, since almost no construction is really obligatory in Chinese, speakers may use it or may not use it. Given that Chinese is not the only language that shows this rather loose attitude towards obligatoricity, should we not better discard this as an criterion of whether a grammatical category is expressed in a language or not? Or is it possible that this will only increase our problems?

## 1.5 Universal Dependencies

In 2012, Petrov et al. proposed what they called a *Universal Part-of-Speech Tagset*. Working for Google, the authors were mainly driven by pragmatic motivations and less concerned with the typical linguistic problems of comparative concepts, particular languages, and the impossibility to find the correct part-of-speech system for a particular language. Their part-of-speech system consists of 12 "universal" part-of-speech tags, which they proposed to add to corpora which were already tagged for part-of-speech in order to render the general structures across different languages comparable. The GitHub project of the authors provides us with a list of their universal tags, which reads as follows:

- VERB - verbs (all tenses and modes)
- NOUN - nouns (common and proper)
- PRON - pronouns
- ADJ - adjectives
- ADV - adverbs
- ADP - adpositions (prepositions and postpositions)
- CONJ - conjunctions
- DET - determiners
- NUM - cardinal numbers
- PRT - particles or other function words

## 4 Structural Variation

- X - other: foreign words, typos, abbreviations
- . - punctuation

Using this set, they convert existing corpora (“treebanks”) for different languages into their uniform tagging system, as shown in the figure below.

sentence:	The	oboist	Heinz	Holliger	has	taken	a	hard	line	about	the	problems	.
original:	DT	NN	NNP	NNP	VBZ	VRN	DT	JJ	NN	IN	DT	NNS	.
universal:	DET	NOUN	NOUN	NOUN	VERB	VERB	DET	ADJ	NOUN	ADP	DET	NOUN	.

What would linguists say is the great danger of the Universal Dependencies approach?

## 2 How do Grammars Change?

Since we do not yet really know what we mean by grammar, we need to look at different aspects of grammatical change at the same time. We will do so by looking quickly at morphological change, syntactic change, and the large field of studies on grammaticalization.

Are there any other aspects of grammatical change which are not covered by the three aspects mentioned before?

### 2.1 Morphological Change

When talking of morphological change we usually think of derivational processes on the one hand, which would largely be have to be treated along with lexical change, since one would traditionally exclude these from the realm of grammar. On the other hand, morphological change deals with the change in inflectional paradigms, as reflected in analogical leveling (when paradigms get more regular than they were before) or in the development of irregularities of paradigms (when interfering with sound change), but also in the loss (syncretism) of distinctions. Koch (1996: 224) mentions the following categories of morphological change:

- morph replacement
- change in the formal realization of a morpheme (allomorphic change)
- change in the place of a boundary
- change in content/meaning/function
- change in morphosyntactic status
- reordering of morphemes
- morpheme doubling

He also discusses (ibid. 232) those factors that may influence which forms will be affected in the end.

1. Paradigm frequency: The variant occurring in the most forms in the paradigm prevails.

2. The variant occurring in the word form that expresses the semantically most basic category in the paradigm prevails. Basic forms express singular number, nominative case, third person, present tense, indicative mood, etc.
3. The variant occurring in the word form that occurs most frequently for the particular lexeme. [...]
4. The variant that most closely resembles invariant morphemes that occur in related paradigms.

Can you find an example for each of the categories mentioned by Koch?

## 2.2 Syntactic Change

While we have considerably good accounts on morphological change and even better accounts on sound change, syntactic change is difficult to investigate, since it is hard to reconstruct, and therefore there are generally much fewer examples of how syntactic change proceeds in different languages of the world. Roughly, we know that languages can change their basic word order and that this does not need to take too much time. We know this, or think we know this from the comparison of closely related languages in which different word order patterns can be found, such as, for example, in German, with its mixed word order of Subject-Object-Verb (SOV) in subordinate clauses and SVO word order in main clauses, while all other Germanic languages have exclusively SVO. But word order is not the only aspect of syntax that can change over time. Languages tend, for example, not to alter the order of determiner and determined (*he-dog* = determiner-determined), but we can find that even in the rather closely related Sinitic languages there are differences with respect to the order of determiner and determined (which surface in form of reconstructions like *he-dog* in Mandarin Chinese and *dog-he* in Cantonese).

What do you think are the major factors for syntactic change in language evolution?

## 2.3 Grammaticalization

Originally coined by Antoine Meillet “to indicate a process of linguistic change whereby an autonomous lexical unit gradually acquires the function of a dependent grammatical category” (Bussmann 1996: 488), grammaticalization nowadays mostly concentrate on semantic and pragmatic aspects. Bussmann (ibid.) mentions the following questions, which appear in the center of researcher's interest:

- (a) Is the change of meaning that is inherent to grammaticalization a process of desemanticization [...] or is it rather a case (at least in the early stages of grammaticalization) of a semantic and pragmatic concentration [...]?
- (b) What productive parts do metaphors and metonyms play in grammaticalization?
- (c) What role does pragmatics play in grammaticalization?

Classical examples for grammaticalization events include the development from verbs expressing possession (German *haben*) to markers of a resultative or perfect construction,

## 4 Structural Variation

which can also be observed in Cantonese from where it has already in parts spread to Mandarin Chinese (Heine and Kuteva 2002: 245). Even more frequent, however, is the development of verbs expressing possession to marks of constructions expressing obligation (*Ich habe zu tun...*, *ibid.* 242-245).

Rather strange examples include the change of *shì* 是 in Chinese from a demonstrative pronoun to a copula (although one can argue that Russian, a language that lost its copula, has a demonstrative pronoun appearing in a similar position, at times even being obligatory, such as *Зевс – это Юпитер*. “Zeus, that is Jupiter”, see Padučeva 1985: 165).

Studies on grammaticalization nowadays often focus on new patterns that can be observed in certain language families, or across certain areas, while scholars try, at the same time, to detect the basic rules underlying grammaticalization pathways.

How can one motivate the change from possession to obligation?

### 2.4 Grammaticalization Clines

It is very popular among linguists to postulate specific pathways of grammaticalization and to criticize those pathways which have been proposed by their colleagues. Koch (1996: 247), for example proposes the following two common pathways for inflection and derivation:

Lexeme → grammatical clitic → inflectional affix → part of lexical morph  
Lexeme → lexical component → derivational affix → part of lexical morph

Even more interesting are the increasing collections of grammaticalization processes as they have been collected by Heine and Kuteva (2002). What the authors offer here is a database-like collection of comparative concepts and pathways of change by which their transition can be described. Starting from an extensive definition of grammatical concepts (like *causative*, *case*, etc.) the authors then provide detailed evidence from the linguistic literature on grammaticalization processes observed in different languages of the world.

Looking at Heine and Kuteva's World Lexicon of Grammaticalization and having looked at different databases for other linguistic aspects in the past, how do you think could this collection be enhanced?

### 2.5 From Sharp Boundaries to Fuzzy Prototypes

What is interesting in this context is that we can also see that the discussion about grammatical constructions may be stated differently. If lexical items can turn into grammatical items (and potentially even the other way round, although there are only a few examples), this would also mean that any strict notion of a certain construction being “grammatical” has to deal with some kind of a continuum in the end.

To be sure, it was one of the main contributions of grammaticalization theory to show that the transition from lexical to grammatical categories is gradual rather than abrupt. Nevertheless, this work has also established that there is justification to distinguish prototypical lexical items, such as nouns and verbs, from prototypical grammatical items, such as markers for tense, aspect, case, (in)definiteness, number, gender, case, etc. (Heine et al. 2016: 161)

If we accept that there are prototypical nouns and verbs and prototypical grammatical items, how can we then distinguish the two semantically?

## 2.6 Excursus: Evolution of Directives

If we take directives in their simplest form, as commands, we find that many of the world's languages use similar techniques to express turn a phrase into a command, specifically those addressing the direct counterpart of an utterance (i.e., second person). While the design space is considerably large here, ranging from particles via specific lexical items up to the use of bare verbal stems (Aikhenvald 2010: 18), it seems possible to find even some tendencies, in so far as "Synthetic languages tend to mark imperatives with inflectional means. And isolating and highly analytic languages will employ particles (short independent function words) as command markers" (ibid.). Scholars have detected quite a few regularities with respect to directives. We often find similar strategies to avoid the imperative (e.g., by asking a question instead), as illustrated in depth by Aikhenvald (ibid.: 288), as well as we know that those constructions that are used to express the imperative may also quite often give rise to new functions (e.g., from imperative to conditional, as we can observe in Russian and German).

Givón (2005: 172) draws a continuum between prototypical imperatives (*Pass the salt!*) and prototypical interrogatives (*Was there any salt here?*). Can we use a similar way of reasoning to draw a continuum between imperatives and conditionals?

## References

- Aikhenvald, A. Y. (2010). *Imperatives and commands*. Oxford: Oxford University Press.
- 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.
- Chao, Y. (1968). *A grammar of spoken Chinese*. Berkeley, Los Angeles, and London: University of California Press.
- Gabelentz, G. v. d. (1953). *Chinesische Grammatik. Mit Ausschluss des niederen Stiles und der heutigen Umgangssprache*. Repr. Berlin: Deutscher Verlag der Wissenschaften. [Original edition: Leipzig: Weigel, 1881].
- Givón, T. (2005). *Context as other Minds: The pragmatics of sociality, cognition and communication*. Amsterdam and Philadelphia: John Benjamins.
- Haspelmath, M. (2010). "Comparative concepts and descriptive categories". *Language* 86.3, 663–687.
- Heine, B. and T. Kuteva (2002). *World lexicon of grammaticalization*. Cambridge: Cambridge University Press.
- Heine, B., H. Narrog, and H. Long (2016). "Constructional change vs. grammaticalization". *Studies in Language* 40.1, 137–175.
- Hippisley, A. (1998). "Indexed stems and Russian word formation: A network morphology account of Russian personal nouns". *Linguistics Faculty Publications* 36 (6), 1093–1124.
- Kempgen, S. (1981). *Wortarten als klassifikatorisches Problem der deskriptiven Grammatik: Historische und Systematische Untersuchungen am Beispiel des Russischen*. München: Otto Sagner.
- Koch, H. (1996). "Reconstruction in morphology". In: *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, 218–263.
- Mel'čuk, I. A. (1974). "Grammatical meanings in interlinguas for automatic translation and the concept of grammatical meaning". In: ed. by V. J. Rozencveig. Vol. 1. Frankfurt: Athenaeon.
- Packard, J. L. (2000). *The morphology of Chinese. A linguistic and cognitive approach*. Cambridge: Cambridge University Press.
- Padučeva, E. V. (1985). *Vyskazyvanie i ego sootnesennost' s deijstvitel' nost' ju*. Moskow: Nauka.
- Petrov, S., D. Das, and R. McDonald (2012). "A Universal Part-of-Speech Tagset". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation*. Istanbul: European Language Resources Association, 2089–2096.
- Téchnē grammatiké Téχνη γραμματική* [Art of grammar] (ca. 100 BC). By D. Thrāx (ca. 170–90 BC); Critical edition: Linke, K., ed. (1987). *Die Fragmente des Grammatikers Dionysios Thrax*. Sammlung griechischer und lateinischer Grammatiker 3. Berlin and New York: Walter de Gruyter, 1987; English translation: "The grammar of Dionysios Thrax" (1874). English. Trans. from the Greek by T. Davidson. *The Journal of Speculative Philosophy* 8.4 (1874), 326–339. JSTOR: 25665891.
- Weber, H. (2002). "Partizipien als Partizipien, Verben und Adjektive. Über Kontinuität und Fortschritt in der Geschichte der Sprachwissenschaft". In: *Sprache und Welt. Festgabe für Eugenio Coseriu zum 80. Geburtstag*. Ed. by A. Murguía. Tübingen: Narr, 191–214.

### Structural Variation (Typological Viewpoint)

#### 1 What is Structure?

Although as linguists we use the term *structure* a lot in our daily scientific life, it is not easy to give a clear-cut definition of what is usually meant by “structure”. The first thing that comes to mind is probably the term “structuralism”, which most people relate to the work of Ferdinand de Saussure (1857-1913), although it seems that Saussure himself was not even using the term *structure* in his work.

Even though de Saussure did not use the term ‘structure’ in his posthumously published *Cours de linguistique générale* (1916, based on lecture notes from the years 1906–11), but rather the terms *système* and *mécanisme*, he is none the less recognized as the ‘father’ and pioneer of structuralism, and his *Cours* is seen as a summary of the fundamental principles of structuralist linguistic description. De Saussure assumes that language is a relational system of formal, not substantial, elements, which can be precisely recorded and exactly represented. He sees research into the internal relations of language as the central task of linguistics and linguistics as an autonomous science that has no need to resort to psychology or the social sciences for aid in explanation. (Bussmann 1996: 1132)

As linguists, we meet the term *structure* in various contexts. For example, the famous work by Chomsky (2002) is called “syntactic structures”, when discussing words, we talk about their “phonological structure”, or we discuss their “semantic structure”. In all these cases, “structure” is used as a term to describe that we focus on the relations between a certain number of elements (words, phones, senses).

The paragraph mentions syntactic, phonological, and semantic structures. What alternative units of speech was not mentioned in this paragraph?

##### 1.1 Structure and Structuralism

Saussure is usually mentioned as the father of “structuralism” in linguistics, although it is all but clear what structuralism itself means when looking at the multiple ways in which the term is used in the field, as we can also easily see when looking up the term “structuralism” in a typical handbook of linguistics.

Collective term for a number of linguistic approaches in the first half of the twentieth century, all based on the work of F. de Saussure, but strongly divergent from one another. Depending on theoretical preconceptions, the term ‘structuralism’ is used in several ways. In its narrower sense, it refers to the pregenerative phase of linguistics before N. Chomsky’s Syntactic structures; in its broader sense, to all linguistic theories which focus on an isolated investigation of the language system, which would include generative transformational grammar. (Bussmann 1996: 1132)

However, given that it is quite clear from Saussure’s work that he was generally much more interested in the relations among linguistic objects than in the objects themselves we can still defend the view that “relation” is an important aspect of “structure” in linguistics.

What approaches are not representing “structuralism” in linguistics?

## 1.2 Broad Notion of Structure

If we say that structure in linguistics represents *relations* between *objects*, then there are many structures that we can find and investigate in linguistics. We can investigate the structure of sound systems, we can compare syntactic structures, we can look into semantic structures and lexical structures, and also discuss morphological and textual structures across different languages. If we assume, furthermore, that grammar (at least in one of the many notions of the term) is all about the relations between words and other units of speech, it is perhaps possible to see why this session is called *structural variation* and not *grammatical variation*.

What is the branch of linguistics that typically works with textual structures?

## 1.3 Assembling Structural Data

Since structure – in the broad notion introduced here – represents relations between objects, it can be treated as something opposed to phoneme inventory datasets as we have discussed them before. But this was not the only reason that the term *structural variation* was used for these sessions. The major reason was that there is a specific type of data which linguists have been collecting for quite some time in typology now, which has been called *structural dataset* and represents a certain kind of coding that is quite different from the datasets that were discussed so far (Forkel et al. 2018). Thus, in the introduction of the *World Atlas of Language Structures* (Haspelmath et al. 2005), Comrie et al. (2005) contrast their Atlas with the tradition of Dialect Atlases:

While dialect atlases show the geography of substantive linguistic features (such as particular cognate sounds, or particular words), WALS shows only structural features, i.e. abstract features of the language system that can be compared across unrelated languages.

This quote implies that the term “structure” is not intended to be very strict, leaving a lot of space for interpretation, when it comes to coding structure for different languages, and we will later see that this has a specific drawback both when it comes to analyzing structural data and when it comes to creating and curating datasets of structural features.

Judging from your knowledge of the WALS, how would you characterize the features used in this project?

## 2 How to Code Structure?

The majority of historical linguists compare words to reconstruct the history of different languages. However, in phylogenetic studies focusing on cognate sets reflecting shared homologs across the languages under investigation, there exists another data type that people have been trying to explore in the past. The nature of this data type is difficult to understand for non-linguists, given that it has a very abstract nature. In the past, it has led to a considerable amount of confusion both among linguists and among non-linguists who tried to use this data for quick (and often also dirty) phylogenetic approaches. For this reason, we need to look at this data type in more detail.

## 2.1 Examples

In order to illustrate the type of data we are dealing with here, let's have a look at a typical dataset, compiled by the famous linguist Jerry Norman to illustrate differences between Chinese dialects (Norman 2003). The table below shows a part of the data provided by Norman.

No.	Feature	Beijing	Suzhou	Meixian	Guangzhou
1	The third person pronoun is <i>tā</i> , or cognate to it	+	-	-	-
4	Velars palatalize before high-front vowels	+	+	-	-
7	The <i>qu</i> -tone lacks a register distinction	+	-	+	-
12	The word for "stand" is <i>zhàn</i> or cognate to it	+	-	-	-

In this example, the data is based on a questionnaire that provides specific questions; and for each of the languages in the sample, the dataset answers the question with either + or -. Many of these datasets are binary in their nature, but this is not a necessary condition, and questionnaires can also query categorical variables, such as, for example, the major type of word order might have three categories (subject-object-verb, subject-verb-object or other).

We can also see is that the questions can be very diverse. While we often use more or less standardized concept lists for lexical research (such as fixed lists of basic concepts, (List et al. 2016), this kind of dataset is much less standardized, due to the nature of the questionnaire: asking for the translation of a concept is more or less straightforward, and the number of possible concepts that are useful for historical research is quite constrained. Asking a question about the structure of a language, however, be it phonological, lexical, based on attested sound changes, or on syntax, provides an incredible number of different possibilities. As a result, it seems that it is close to impossible to standardize these questions across different datasets.

Although scholars often call the data based on these questionnaires "grammatical" (since many questions are directed towards grammatical features, such as word order, presence or absence of articles, etc.), most datasets show a structure in which questions of phonology, lexicon, and grammar are mixed. For this reason, it is misleading to talk of "grammatical datasets", but instead the term "structural data" seems more adequate, since this is what the datasets were originally designed for: to investigate differences in the structure of different languages, as reflected in the most famous, already mentioned *World Atlas of Language Structures* (Dryer and Haspelmath 2013).

If structural data is given in form of categorical variables or binary answers to questions in a questionnaire, can it then even be called "structural"?

## 2.2 Problems in Structure Coding

In addition to mixed features that can be observed without knowing the history of the languages under investigation, many datasets (including the one by Norman we saw above) also use explicit "historical" (diachronic in linguistic terminology) questions in their questionnaires. In his paper describing the dataset, Norman defends this practice, as he argues that the goal of his study is to establish an historical classification of the Chinese dialects. With this goal in mind, it seems defensible to make use of historical knowledge and to include observed phenomena of language change in general, and sound change in specific, when compiling a structural dataset for group of related language varieties.

The problem of the extremely diverse nature of questionnaire items in structural datasets, however, makes their interpretation extremely difficult. This becomes especially evident when using the data in combination with computational methods for phylogenetic reconstruction. This is problematic for two major reasons.

1. Since questions are by nature less restricted regarding their content, scholars can easily pick and choose the features in such a way that they confirm the theory they want them to confirm rather than testing it objectively. Since scholars can select suitable features from a virtually unlimited array of possibilities, it is extremely difficult to guarantee the objectivity of a given feature collection.
2. If features are mixed, phylogenetic methods that work on explicit statistical models (like gain and loss of character states, etc.) may often be inadequate to model the evolution of the characters, especially if the characters are historical. While a feature like “the language has an article” may be interpreted as a gain-loss process (at some point, the language has no article, then it gains the article, then it loses it, etc.), features showing the results of processes, like “the words that originally started in [k] followed by a front vowel are now pronounced as [tɕ]”, cannot be interpreted as a process, since the feature itself describes a process.

For these reasons, all phylogenetic studies that make use of structural data, in contrast to purely lexical datasets, should be taken with great care, not only because they tend to yield unreliable results, but more importantly because they are extremely difficult to compare across different language families, given that they have way too much freedom when compiling them. Feature collections provided in structural datasets are an interesting resource for diversity linguistics, but they should not be used to make primary claims about external language history or subgrouping.

This passage always talks about phylogenetic studies, but are these really the only cases in which the typical coding of structural data can be problematic?

## 2.3 Technical Recommendations for Structure Coding

As one step towards increasing the interoperability of cross-linguistic data, the Cross-Linguistic Data Formats (CLDF, <https://cldf.clld.org>) specification was published in 2018 Forkel et al. 2018. Having started with a series of workshops since 2014, in which linguists who make active use of cross-linguistic data in their research discussed the challenges of data standardization, the first version of the CLDF specification proposed standard formats along with evaluation tools for the most basic types of data encountered in cross-linguistics research, namely *word lists*, *structural datasets*, and *dictionaries*.

For structural datasets, we have already mentioned that the questions typically asked in the questionnaires are way too idiosyncratic to be comparable across different datasets. The language names, however, should typically be linked to Glottolog (Hammarström et al. 2019), and the specific characteristics of the features being investigated should be defined (Forkel and List 2020).

On the website <https://github.com/cldf-datasets> we have assembled examples of how datasets should be coded in CLDF to make them more easily comparable with other datasets. Structure datasets are considerably easy to code, and also easy to prepare. However, to avoid that certain errors in the coding procedure slip in during the

## 4 Structural Variation

data curation process, one should always use software to validate that no errors have been introduced into the data.

Why do datasets of structural features in linguistics run the risk of being idiosyncratic?

### 2.4 Modelling Recommendations for Structure Coding

When establishing a dataset of structural features, it is of great importance to be clear with respect to the question of what one wants to do with the data once they have been compiled. My recommendation is to avoid categorical data where possible and to prefer concrete data from which categorical data can be derived. Every step by which original, “raw”, data are converted to a higher level of abstractions involves an interpretation which can well be wrong or flawed, due to human error when coding the data. Instead of reading a grammar, counting the sounds in a phoneme chart in one’s head, and typing the count into a spreadsheet, one should better type off all sounds in a sheet, ideally directly making sure they are linked to reference catalogs such as CLTS (List et al. 2019). Even better, instead of typing off phonemes from a chart, one should type off words which illustrates how these phonemes can be used, so a lexicon or a wordlist should be preferred to a simple phoneme chart. It is clear that certain “questions” linguists like to ask are difficult to operationalize, but it should also be clear that what one sacrifices when asking “Does the language have a plural” instead of preparing diagnostic sentences in form of *interlinear-glossed text* or some “proof” in form of plural and singular noun forms for a specified number of words. But it should also be clear that modeling is quite difficult and may require at times years of thinking and planning and discussing.

What makes it so difficult to render some typical grammatical categories transparent in form of a questionnaire?

### 2.5 Structural Datasets

WALS (Dryer and Haspelmath 2013) was already mentioned above (<https://wals.info>). In addition, APLiCS Online (Michaelis et al. 2013), the *Atlas of Pidgin and Creole Language Structures Online* (<https://apics-online.info>) offers structural data for Pidgin and Creole languages. Carling et al. (2018) recently published the *Diachronic Atlas of Comparative Linguistics* (<https://diacsl.lu.se>), which offers lexical and structural data. Apart from these datasets which are accessible online, many scholars create structural data and present it only in form of tables in their papers, without sharing it in digital form, as the data by Norman (2003) mentioned before. A recent example is the collection of structural data on Mainland South-East Asian languages by Vittrant and Watkins (2019). What we can see from these examples (which are numerous) is that linguists often even do not directly think in analyzing their data with help of some computational tools, but rather investigating them merely qualitatively.

Where does the book by Corbett (2004) fit in this context?

## 3 How to Analyze Structure?

In the past, there have been many controversies about structural data. Given the misinterpretation of structural data as being “grammatical”, along with the unproven and misleading

claim by Nichols (2003) that certain grammatical features are more stable than lexical ones, one can often read about a controversy in linguistics: which aspects are more stable, and therefore more useful to study deep linguistic relationships, the lexicon or the grammar?

In this context, it is often ignored that we are not talking chiefly about the grammar when applying phylogenetic studies to structural datasets. It is also ignored that the original idea of the importance of “grammar” was pointing to homologies in complex and concrete morphological paradigms, as has been most prominently discussed by Meillet (1925 [1954]), later popularized by Nichols (1996), who also pointed to individual word forms, that is: predominantly lexical traits. “Grammar” never pointed to abstract similarities as they are captured in most structural datasets (see the excellent discussion by Dybo and Starostin 2008).

### 3.1 Examples

Leading scholars in historical linguistics have provided convincing arguments that genetic relationships among languages can only be demonstrated by illustrating regular sound correspondences in concrete form-meaning pairs across the languages under investigation (see especially the very good analysis by Campbell and Poser 2008). In spite of this, the rumor that “grammar” (i.e., structural datasets) might provide a shortcut to detect deep, so far unnoticed, relationships among the languages of the world is very persistent, as reflected in many different studies.

Among the examples, Dunn et al. 2008 claimed that language relationships for Papuan languages of Island Melanesia could be uncovered by means of phonological and grammatical (abstract) structural features; and Longobardi et al. 2015 used syntactic features to compare the development of European languages with the development of European populations. Zhang et al. (2018) used phonological inventories of more than 100 different Chinese dialects, coding the data for simple presence and absence of each of the more than 200 different sounds in the database, and analyzing the data with the STRUCTURE software (Pritchard et al. 2000), whose results tend to be notoriously misinterpreted.

What is important about these studies is that none of them (maybe with exception of the study by Dunn et al. 2008, but I am in no position to actually judge the findings) could make a convincing claim why the structural datasets would provide evidence of deeper relationships than could the lexicon. Even the study by Dunn et al., which tests the suitability of their small questionnaire of only 115 structural traits on Oceanic languages, has since then not led to any new insights into so far undetected language relationships, contrary to the hope expressed by the authors, “that structural phylogeny is an important new tool for exploring historical relationships between languages” (ibid. 734).

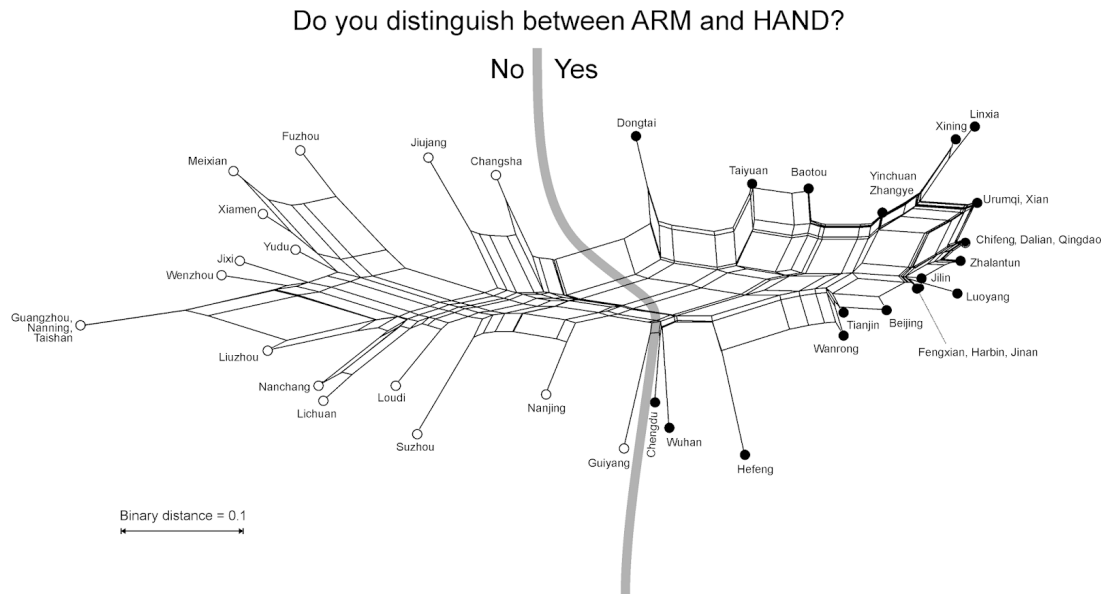
What features would you provide in order to show that the Indo-European languages are related?

### 3.2 Enhanced Network Analysis of Structural Data

In a study on structural data of Chinese dialects as provided by Szeto et al. (2018), we investigated the “quality” of the characters selected to yield sufficient genetic signal to help shed light on the classification of Chinese dialects (Grimm and List 2018). We first investigated all features provided by the authors thoroughly and classified them according to some basic categories (diachronic, semantic, lexical, etc.). We then check to which degree specific features were helpful to cluster the Chinese dialects in meaningful groups.

## 4 Structural Variation

The following image shows how the feature “Does the dialect distinguish ‘arm’ and ‘hand’?” in which we highlighted, where the feature splits the dialects into different groups.



What was interesting in this context is that the feature itself does not seem to be specifically meaningful when comparing the languages of the world. The reason is: the distinction of “arm” vs. “hand” does not seem to follow any specific pattern in the languages of the world, as we can easily see when consulting the CLICS database (List et al. 2019).



So what we can see from this example is that if one wants to create a structural dataset in order to investigate the history of a certain number of languages, one should be careful in choosing the features. If a feature cannot yield any evidence with respect to the history of the languages, there is no use in making a historical analysis with it. Accordingly, if a feature is typologically so randomly distributed as the distinction of “arm” and “hand”, it is questionable if it makes sense to investigate this feature for a particular area. Of course, one can investigate it, but one should not expect it to be very informative.

## References

- Busmann, H., ed. (1996). *Routledge dictionary of language and linguistics*. Trans. from the German by G. Trauth and K. Kazzazi. London and New York: Routledge.
- Campbell, L. and W. J. Poser (2008). *Language classification: History and method*. Cambridge: Cambridge University Press.
- Carling, G., F. Larsson, C. A. Cathcart, N. Johansson, A. Holmer, E. Round, and R. Verhoeven (2018). "Diachronic Atlas of Comparative Linguistics (DiACL). A database for ancient language typology." *PLOS ONE* e0205313, 1–20.
- Chomsky, N. (2002). "Syntactic structures."
- Comrie, B., M. S. Dryer, D. Gil, and M. Haspelmath (2005). "Introduction." In: Oxford: Oxford University Press, 1–8.
- 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.
- Dunn, M., S. C. Levinson, E. Lindstroem, G. Reesink, and A. Terrill (2008). "Structural phylogeny in historical linguistics: methodological explorations applied in island melanesia." *Language* 84.4, 710–759.
- Dybo, A. and G. S. Starostin (2008). "In defense of the comparative method, or the end of the Vovin controversy." In: *Aspekty komparativistiki* [Aspects of comparative linguistics]. Vol. 3: *Aspekty komparativistiki*. Ed. by I. S. Smirnov. Moscow: RGGU, 119–258.
- 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.
- Grimm, G. and J.-M. List (2018). "How languages lose body parts: once more about structural data in historical linguistics." *The Genealogical World of Phylogenetic Networks* 5.11.
- Hammarström, H., M. Haspelmath, and R. Forkel (2019). *Glottolog. Version 4.1*. Jena: Max Planck Institute for the Science of Human History.
- The world atlas of language structures* (2005). Oxford: Oxford University Press.
- List, J.-M., C. Anderson, T. Tresoldi, C. Rzymski, S. Greenhill, and R. Forkel (2019a). *Cross-Linguistic Transcription Systems. Version 1.3.0*. Jena: Max Planck Institute for the Science of Human History.
- 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, 05/23/2016–05/28/2016). 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., C. Rzymski, T. Tresoldi, S. Greenhill, and R. Forkel (2019b). *CLICS: Database of Cross-Linguistic Colexifications. Version 3.0*. URL: <http://clics.clld.org/>.
- Longobardi, G., S. Ghirotto, C. Guardiano, F. Tassi, A. Benazzo, A. Ceolin, and G. Barbujaan (2015). "Across language families: Genome diversity mirrors linguistic variation within Europe." *American Journal of Physical Anthropology* 157.4, 630–640.
- Meillet, A. (1954). *La méthode comparative en linguistique historique* [The comparative method in historical linguistics]. Repr. Paris: Honoré Champion.
- Michaelis, S. M., P. Maurer, M. Haspelmath, and M. Huber, eds. (2013). *APICS Online*. Jena: Max Planck Institute for the Science of Human History.
- Nichols, J. (1996). "The comparative method as heuristic." In: *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, 39–71.
- (2003). "Diversity and stability in language." In: *The handbook of historical linguistics*. Ed. by B. D. Joseph and R. D. Janda. Blackwell handbooks in linguistics. Malden, Mass.: Blackwell, 283–310.
- 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.
- Pritchard, J. K., M. Stephens, and P. Donnelly (2000). "Inference of population structure using multilocus genotype data." *Genetics* 155, 945–959.
- Szeto, P. Y., U. Ansaldi, and S. Matthews (2018). "Typological variation across Mandarin dialects: An areal perspective with a quantitative approach." *Linguistic Typology* 22.2, 233–275.
- Vittrant, A. and J. Watkins, eds. (2019). *The Mainland Southeast Asia Linguistic Area*. Berlin and Boston: Mouton de Gruyter.
- Zhang, M., W. Pan, S. Yan, and L. Jin (2018). "Phonemic evidence reveals interwoven evolution of Chinese dialects." *bioRxiv*.