Developing an annotation framework for word formation processes in comparative linguistics

Nathanael E. Schweikhard, MPI-SHH, Jena
Johann-Mattis List, MPI-SHH, Jena

Word formation plays a central role in human language. Yet computational approaches to historical linguistics often pay little attention to it. This means that the detailed findings of classical historical linguistics are often only used in qualitative studies, yet not in quantitative studies. Based on human- and machine-readable formats suggested by the CLDF-initiative, we propose a framework for the annotation of cross-linguistic etymological relations that allows for the differentiation between etymologies that involve only regular sound change and those that involve linear and non-linear processes of word formation. This paper introduces this approach by means of sample datasets and a small Python library to facilitate annotation.

Keywords: language comparison, cognacy, morphology, word formation, computer-assisted approaches

1 Introduction

That larger levels of organization are formed as a result of the *composition* of lower levels is one of the key features of languages. Some scholars even assume that compositionality in the form of recursion is what differentiates human languages from communication systems of other species (Hauser et al. 2002). Whether one believes in recursion as an identifying criterion for human language or not (see Mukai 2019: 35), it is beyond question that we owe a large part of the productivity of human language to the fact that words are usually composed of other words (List et al. 2016a: 7f), as is reflected also in the numerous words in the lexicon of human languages.

While compositionality in the sphere of semantics (see for example Barsalou 2017) is still less well understood, compositionality at the level of the linguistic form is in most cases rather straightforward. Given that (as was early emphasized by de Saussure 1916: 103) the linguistic form is a function of time, the most straightforward way of combining two forms is to place them one after each other, as is usually done in word formation processes, such as *compounding* or *derivation* by prefixation or suffixation. Word formation processes are, of course, not limited to purely concatenative processes, as witnessed by well-observed phenomena such as *ablaut*, *umlaut*, or *template morphology* (Schwarzwald 2019), although from the perspective of their evolution, scholars often assume that nonlinear morphology has its origin in linear processes (Heine 2019: 7).

Considering the essential role that word formation plays not only for synchronic description but specifically also for diachronic investigation, it is surprising that scholars have not yet decided on a standardized way of representing the morphological relations between words inside and across related languages. Although the past has seen occasional attempts of formalization of etymological data (Crist 2005), the current practice of representing findings in historical linguistics is still in the typical form of etymological dictionaries, in which individual words are explained in prose with a minimal amount of formalization.

As an example for the current practice of etymological annotation, consider the entry for German *Frucht* ‘fruit’ in the online version (http://dwds.de) of the etymological dictionary of German by Pfeifer (1993), given in Figure 1A. Trained linguists can learn a lot from entries like this, specifically, that the form itself was borrowed from Latin *frūctus* ‘profit, fruit’ which itself was derived from *fruī* ‘to enjoy, to profit from’. By following a cross-reference to a different entry (Figure 1B) they can see that it is cognate with German *brauchen* ‘to use’, going back to Indo-European **bhrūg-* ‘to use’. For laypeople or scholars not familiar with the typical conventions of etymological prose, however, the two paragraphs are very hard to read and understand, specifically when comparing it with the illustration in Figure 1C where the major processes are displayed in form of a derivation graph.

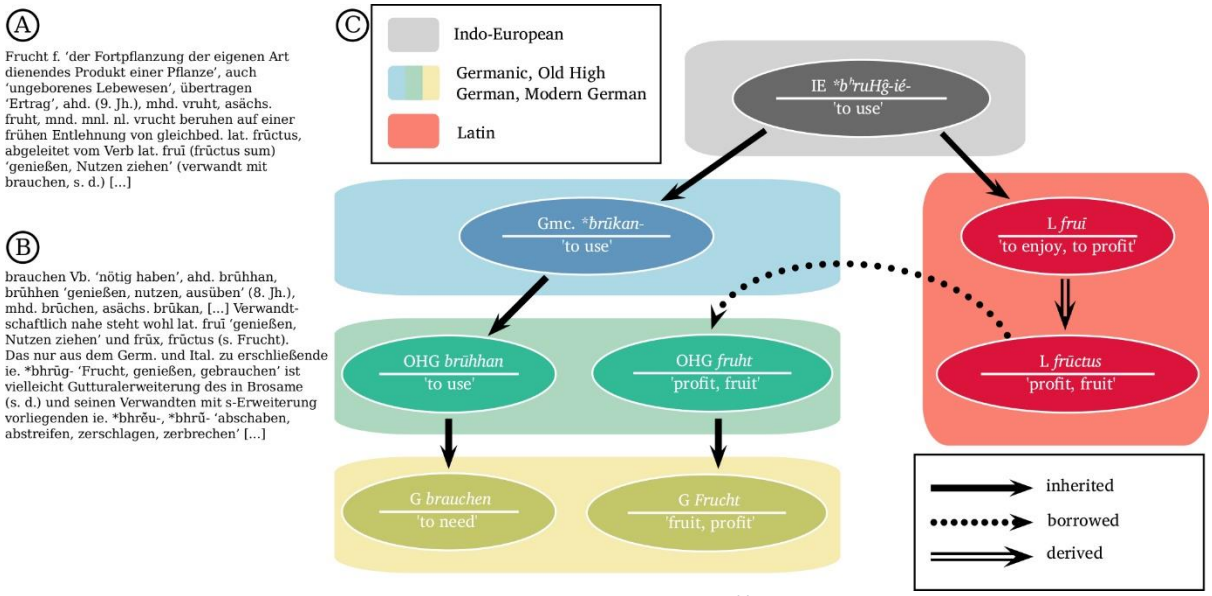


Figure 1: German *Frucht* and *brauchen* in Pfeifer (1993, also online at http://dwds.de) and in a derivation graph (inspired by a graphic on the same word family from Hans Geisler)

While a certain knowledge of specific practices of displaying information is required by all scientific disciplines, the current representation format of etymologies in historical linguistics has the serious disadvantage of limiting the application range of etymological dictionaries to purely qualitative studies. In order to draw a derivation graph of the words deriving from Indo-European **bhrūg* from the two entries in Pfeifer’s dictionary, one needs to attentively sift through the dictionary and collect the essential information from the text. Given that etymological dictionaries often differ in the way in which the information is shared with the readers, there is no automatic method that could parse the information consistently. This is a pity, given the wealth of knowledge underlying the large amount of etymological dictionaries which have been produced for many languages and language families of the world.

If it were possible to process this information consistently with the help of standard programming tools, we could harvest an abundant amount of information on attested and inferred patterns of word formation that could be used to test and improve morphological theory in general and assist scholars in producing etymologies for so far underinvestigated language families. If scholars adopted unified frameworks for the linguistic annotation of word formation processes and etymological relations, it would furthermore be much easier to check their individual proposals for overall consistency and plausibility.

In this paper, we present a new framework for the consistent annotation of word formation processes in etymological datasets in historical linguistics. We are thereby drawing from the wide-spread practice of *interlinear morphemic glossing* (Lehmann 2004). However, we shift the focus from the annotation of individual forms to the annotation of *etymological relations* between forms, while at the same time trying to guarantee that our annotations are both human- and machine-readable. Building on initial ideas for the annotation of morphological relations presented by Hill & List (2017), we expand their framework by (1) proposing more rigorous standards to distinguish grammatical from lexical morphemes, (2) allowing for a strict distinction between different etymological relations, and (3), as an outlook, introducing new ways to model word families in form of *derivation graphs*. Our framework comes along with annotation guidelines, usage examples presented in form of sample datasets, web-based tools assisting in data creation and curation, and a selection of scripts that assist users in checking their data for consistency. We hope this will support future cross-linguistic studies that utilize word list data or other forms of word annotations like interlinear glossing.

In the following, we will first discuss the role that word formation plays in historical language comparison (Chapter 2), and present some obvious problems of handling word formation consistently in historical linguistics (Chapter 3). We will then present our framework for a consistent handling of word formation in historical linguistics (Chapter 4) by introducing our basic ideas for data management in historical linguistics (§4.1), presenting how etymological word relations can be consistently annotated within our framework (§4.2), and showing how they can be checked for consistency with our Python library (§4.3). We conclude with pointing to open question which we could not resolve so far (§4.4) and presenting further applications of our framework in quantitative and computer-assisted frameworks (Chapter 5).

2 Word formation in historical language comparison

2.1 Historical relations between words

In order to handle morphological relations (be they still synchronically transparent or only detectable through linguistic reconstruction) with the help of a consistent framework for etymological annotation, it is important to be clear about the etymological relations which should be modeled by such a framework. Following Gévaudan (2007), and further elaborated in List (2016a), a straightforward model for etymological relations starts from the linguistic sign in the sense of de Saussure (1916) with *form* and *meaning* as its major constituents, which are realized in the system of a given *language*. With etymological relations being defined as those relations which reflect a shared history between two or more linguistic signs (List 2014: 56f), we can characterize individual etymological relations with respect to the different dimensions along which lexical change can proceed. Here, Gévaudan (2007) distinguishes the *morphological dimension*, affecting the *form* of a sign, the *semantic dimension*, affecting the *meaning* of a sign, and the *stratic dimension*, affecting the *language* in which a sign is being used. While the first two dimensions are straightforward and do not need further explanation, the third dimension was introduced by Gévaudan (2007) in order to allow for a proper handling of cases of lexical borrowing, a dimension usually excluded in the classical models of lexical change proposed in lexicostatistics (Swadesh 1952; Lees 1953).

Note that lexical change in this notion deliberately excludes questions of sound change (Gévaudan 2007: 14). Assuming that sound change is a regular process that usually does not have an impact on the abstract relations between the lexemes of a given language, this seems reasonable at first sight. However, as sound change impacts the phoneme system of a given language and because the lexemes themselves are built from phonemes, it can easily disrupt the lexical structure of a language, for example by forcing the replacement of a word in a specific meaning in order to avoid homophony. A prominent example where the impact of sound change on morphological structure is vividly discussed in historical linguistics is the development of Mandarin Chinese (and Sinitic languages in general), which apparently underwent a shift from a language with a rather complex syllable structure to a very simplified syllable model, accompanied by a rise in disyllabic compounds (Behr 2015; Sampson 2015).

In addition, we should also keep in mind that morphological processes can change the form of a word in a way that is quite different from regular sound change. Since these processes (such as *ablaut*, *umlaut*, *vowel harmony*, or *analogy* in its various forms) change the form of a sign in a fundamentally different way than regular sound change, we think it is worthwhile to include this information in a rigorous description of etymological relations.

We thus explicitly include both the information on regular sound change and on additional morphological processes that would change a given sign form more than it would have changed when only assuming sound change in a general model of etymological relations.

Summarizing the dimensions of lexical variation mentioned above, we thus find the *regularity dimension*, which deals with changes to a word's form that go beyond regular sound change, the *morphological dimension*, which deals with whether a sign and its cognate go back to the same word or to words formed from each other via a morphological process, the *semantic dimension*, which deals with the meaning of the sign, and the *stratic* dimension, which reflects whether a sign has been transferred from one "language stratum" to another.

All together we can combine types of variation along these dimensions in multiple ways. As shown in List (2016a), the typical terms for etymological relations, which at times also find direct counterparts in biology, result from controlling variation along one dimension. Since we add one more dimension in our review of etymological variations, there are 81 (3x3x3x3) possible combinations of the four dimensions, since we can control each dimension positively by requesting continuity or negatively by requesting change, or we can leave it uncontrolled. By adding the regularity dimension to our model of etymological relations, we can now also control for the continuous identity of word forms, which are thought to have only been affected by strictly regular sound change. List (2018b) proposes the term *regular cognates* for words showing continuity in this relation. However, we prefer the term *strict cognates* instead. Since any claims regarding the regularity of sound change processes depend on the analysis of the respective researchers, the term *strict cognacy* seems more appropriate, as it reflects that we are dealing with scholars' (potentially) individual assessments, as opposed to indisputable truths.

In Table 1, we present a revised schema of different *shades of cognacy*, following the representation proposed by List (2016a) along with our additional dimension. In contrast to the table by List, we add *strict cognacy* as an additional type of cognacy, and we also refuse to equate *orthology* with direct cognacy, as defined in List (2014), since it seems obvious that word formation as a linguistic process is far too specific to be fruitfully compared with any form of *homology* in biology.

Table 1: Revised table of etymological relations along with their counterparts in biology

Relation	Biological term	Regularity	Morphological continuity	Semantic	Stratic
traditional notion of cognacy	-	+/−	+/−	+/−	+
cognacy à la Swadesh	-	+/−	+/−	+	+
direct cognacy	~~orthology~~	+/−	+	+/−	+
oblique cognacy	-	+/−	−	+/−	+
etymological relation	homology	+/−	+/−	+/−	+/−
oblique etymological relation	xenology	+/−	+/−	+/−	−
strict cognacy	-	+	+/−	+/−	+/−

2.2 Patterns of word formation

With our multi-dimensional model of lexical variation, we can characterize etymological relations with a rather high degree of sophistication. Characterizing a set of etymologically related words by this model alone, however, won't solve the transparency problems of etymological dictionaries, which we have noted in the introduction, since it would still not allow us to annotate explicitly *where* words are cognate. While cognacy is often treated as a strictly binary concept, according to which two word forms in different languages are either cognate or not, we know well that word formation processes can easily alter the general shape of forms, thereby drastically reducing those parts in related words which actually share a common history.

As an example for the problem, consider word comparisons like Italian *sole* and French *soleil*, the former going back to Latin *sōl*, and the latter going back to Vulgar Latin **sōlĭculus* 'sunny, small sun' (Meyer-Lübke 1911, sec. 8067). While it is obvious that both words are related, given that **sōlĭculus* is a derivation of *sōl*, it is also clear that we cannot say that the word forms are *completely* cognate. The picture becomes even more complicated when adding words like German *Sonne* and Swedish *sol* to the comparison. While all four words go back to the Proto-Indo-European root **séh₂u̯el-* 'sun', the German word is a continuation of the oblique case of the root (**sh₂u̯én-*), which scholars consider to have been irregular already in Proto-Indo-European times. Given that it is rather the norm than the exception that etymologies show this degree of complexity in historical linguistics, it is evident that a clear-cut framework for a consistent annotation of etymological relations needs to be able to handle these cases as well. As a result, our framework should not only be capable of labeling etymological relations, but it should also allow for a transparent indication of the subtleties involving change along the formal and the morphological dimension of lexical variation.

In order to handle word formation consistently, it is useful to start from the patterns of word formation which are usually described in the literature. An overview can be found in Table 2. As a first example for a popular dichotomy, Haspelmath (2002) distinguishes *syntagmatic* and *paradigmatic* aspects of word formation (pp. 165–167). The syntagmatic perspective on word formation concentrates on *linear processes*, by which two or more morphemes are concatenated in order to form larger units. The most prominent types of

concatenative word formation are affixation (Trask 2000, s.v. *affixation*) and compounding. The paradigmatic perspective on word formation, on the other hand, concentrates on changes concerning the form of a whole word, including changes *within* morphemes, leading to allomorphs. The most prominent example for a word formation process that can be described not syntagmatically but paradigmatically is *ablaut* in Indo-European languages, reflected in vowel variation in the root of words, usually marking grammatical differences (Trask 2000: 2f), but other forms of morpheme alternations, such as, for example, *voicing alternation* in Sino-Tibetan languages (Hill 2014; Lai 2016), are also well-attested in the languages of the world.

Table 2: Types of word formation (terms and some examples from Trask (2000) and Haspelmath (2002))

Basic type	Process	Example
concatenative	compounding	*fish* + *tank* → *fish tank*
	affixation	*fish* + *er* → *fisher*
	full reduplication	Mandarin: *rén* ('person') → *rénrén* ('everyone')
	conversion	*fish* (noun) → *fish* (verb)
	...	
allomorphic	pattern-based	Sanskrit: *kulam* ('family') → *kaulam* ('belonging to a family')
	blending	*breakfast* + *lunch* → *brunch*
	infixation	Tagalog: *basag* ('to write') → *bumasag* ('wrote')
	reanalysis	*burglar* → *burgle*
	...	
subtractive	acronym	*radio detection and ranging* → *radar*
	clipping	*discoteque* → *disco*
	...	

It is clear that word formation processes are rarely strictly concatenative or allomorphic, especially because even a concatenative change directly alters the phonetic environment in which a morpheme occurs, which may then have an impact on the regular sound change processes by which the morpheme is further changed. Furthermore, there are cases in which it is difficult to distinguish concatenative from allomorphic processes. Consider the example of voicing alternation in Sino-Tibetan languages mentioned before. This process could either be seen as an allomorphic process by which the initial of a given morpheme is voiced or devoiced or as a concatenative process in which the initials are morphemes of their own which get prefixed to the remainder of the word. In many analyses by historical linguists, this alternation is interpreted historically in syntagmatic terms, by proposing some kind of prefix, whose form may be unknown, which either devoices (Mei 2012) or voices (Baxter & Sagart 1998) the initial of a given word as the result of a regular sound change process, but synchronically it seems more straightforward to describe it as a form of allomorphy. Another case is the suffix {-on} in Hebrew, which is used both on its own and in combination with pattern-based word formation processes. Yet also the derivations in which it seems to be used on its own could be analyzed as involving allomorphic processes, depending on whether one considers them derived from a specific other word form or from an abstract root (Schwarzwald 2019).

3 Problems of handling word formation in historical linguistics

Problems identified for the handling of word formation in historical linguistics can be characterized by assigning them to three different categories important for historical research, namely *modeling*, *inference*, and *analysis*. This triad, inspired by Dehmer et al. (2011, XVII) follows the general idea that scientific research in the historical disciplines usually starts from some kind of idea we have about our research object (the *modeling* stage), and based on which we then apply methods to infer the phenomena in our data (the *inference* stage). Having inferred enough examples for the phenomenon, we can then analyze it qualitatively or quantitatively (the *analysis* stage) and use this information to update our model. In the following, we will quickly discuss the major problems resulting from an insufficient handling of word formation in historical linguistics with respect to each of the three stages.

3.1 Problems of representing word formation

Problems of *modeling* word formation in historical linguistics are tightly connected to problems of *representing* word formation processes. The major problem here is, as we have already shown in the introduction, that scholars dispose over a very detailed knowledge of the complexity of word formation processes, but that they usually do not share this knowledge explicitly when proposing theories on cognacy. Word formation in this form is represented in linguistic prose describing the explanation for specific reconstruction proposals in detailed articles (for instance Cohen 2004; Mees 2014), or in form of summaries that do usually not have the ambition of being exhaustive, which are then published in larger collections such as etymological dictionaries.

The major problem of this way of handling word formation (detailed, but in prose, or by coarse annotation in etymological dictionaries), is a lack of *standardization* that decreases the comparability of etymological analyses. Furthermore, since word formation is a process that may counteract regular sound change, the failure to represent word formation consistently will also directly impact the way in which regular sound change is modeled in our analyses.

If we ignore the possibility of word formation and only consider words cognate that show fully regular sound correspondences, we will miss out on many potential cognate pairs. If we however, as is currently the norm, treat all cognate proposals the same in the way we represent them, independent of whether the words are strict cognates or not, we have a hard time assessing the overall regularity of a given analysis. While this may seem less important for those language families where scholars tend to know all sound laws including all disputed examples by heart, this is definitely not the case for less well studied language families where the number of experts is very small.

3.2 Problems of inferring word formation processes

Even more difficult than representing the etymological relations that hold for a set of etymologically related words is *inferring* them. This applies to classical, "qualitative" historical linguistics, but even more to computational approaches to historical language comparison. In computational tasks, like *automatic cognate detection*, for example, most available datasets for the testing and training of the algorithms do not provide the data in morphologically segmented form. As a result, algorithms which have been designed to identify cognates in multi-lingual

wordlists often fail when it comes to detecting deep etymological relations that are masked by word formation processes.

But this problem does not only apply to automatic approaches. In language families like for instance Sino-Tibetan, productive word formation processes which acted at different stages in the history of the language family have successively led to a situation where regular sound correspondences are extremely hard to infer.

Compounding is a major process of word formation in the Sino-Tibetan family (Matisoff 2003: 153f). If compounds are reduced due to contraction (Trask 2000, s.v. *contraction*; List 2016b), they obscure regular sound correspondences, and this may explain the large-scale inconsistencies in sound correspondences among Sino-Tibetan languages (Handel 2008: 425f).

Similar processes can be found in Indo-European languages as with the German word *Messer* (/mɛsɐ/, 'knife'), which goes back to the Old High German compound *mezzi-sahs*, literally 'food-knife', whose structure has become completely opaque due to regular sound changes that only applied to the compound form but not to the simplex words (Watkins 1990: 295). If the original compound and later forms of it would not be attested in historical documents, it would be very difficult to demonstrate this etymology.

3.3 Problems of analyzing etymological findings

Currently, etymological reconstructions tend to often be treated as the end goal of our endeavor as historical linguists. If they are utilized in follow-up studies, then most commonly in order to support or argue against another reconstruction. If they are used for other kinds of research questions, then those are typically interdisciplinary ones, e.g. using reconstructed words in order to reconstruct the culture and natural environment of the speakers, often in collaboration with anthropologists, biologists, and archaeologists. But they can lead to many more insights into language beyond that, also within linguistics proper.

For instance, developing statistics on the frequency of specific sound correspondences can help us determine how likely it is that a given sound turns into a given other sound, an important aspect of reconstruction that so far is based on the experience-based intuition of experts. The only existing large-scale project for aggregating sound changes (Index Diachronica, a version of it can be found under https://chridd.nfshost.com/diachronica/, last accessed on April 7, 2020) is undertaken by laypeople and makes use of non-scientific sources like Wikipedia because the scientific sources are less easily available.

Similarly, also studies on word family size, on the development of word formation patterns through time, or possibly even on semantic change could be undertaken easily and with much more detail and reliability provided accurately annotated data.

Such analyses could inform us about cross-linguistic typological tendencies of language change and possibly also point us to aspects of our model we need to further refine. But because of the way etymological reconstructions are presented thus far it is not easily possible to aggregate them for use in quantitative studies. We hope our framework will contribute to the solution of this issue.

4 Modeling and inferring word formation in historical linguistics

Our starting point are *wordlists* as they are now commonly used in computer-based and computer-assisted approaches to historical language comparison (List 2018c; List et al. 2020). While linguists tend to think of wordlists as tables in which concepts are listed in the first column, and translations of these concepts are then placed into the consecutive columns, reserving one column per language (see List 2014: 22–24), we make strict use of *long table formats* (Forkel et al. 2018), in which wordlists are represented by a table in which the first row contains a header, with an identifier in the first column, and each consecutive row represents one (and only one) *word form*, based on the content information provided in the header (List et al. 2018). We will discuss this format in more detail below.

4.1 Preliminary considerations

Before we provide a closer overview of our concrete suggestions for the handling of word formation, we need to discuss two important aspects of etymologically oriented investigation of word formation processes: *alignability* and *transparency*. Alignability is important for the annotation of regular sound correspondences with the help of alignments, while transparency is a more general requirement for annotation frameworks.

4.1.1 Alignability and strict cognacy

In the previous sections, we have tried to show that word formation is currently only insufficiently handled in etymological datasets, including etymological dictionaries (as the most prominent representative) but also etymological databases, or the now popular lexicostatistical wordlists, in which information on cognate words is coded in such a way that it can be analyzed with the help of software packages originally developed for applications in evolutionary biology. With our extended model of etymological word relations, in which we emphasize the importance of distinguishing between *strict* and *loose* cases of cognacy, with the former reflecting regular sound change processes and the latter reflecting those cases where morphological processes or sporadic sound change processes led to a further modification of the form part of the linguistics sign, we have introduced a first way to label different degrees of etymological relatedness. By distinguishing *concatenative*, *allomorphic*, and *subtractive* processes as the major processes of word formation, we can furthermore allow for a more fine-grained classification of these etymological relations which involve the formation of new words. What we need for our initial framework is a set of techniques by which we can annotate both (1) the specific *relations* among words, and (2) the *processes* by which words have been formed.

As a first and fundamental distinction for our annotation framework, we propose to distinguish *alignable* from *non-alignable* etymologically related word forms. This distinction accounts for the relation of *strict* as opposed to *loose* cognates and embraces the fact that only word forms which are strictly cognate can be aligned in a meaningful way. An alignment is hereby understood as one of the most general ways to compare sequences (with applications in many fields), in which sequences are compared by placing them in a matrix, one sequence per row, in such a way that corresponding segments appear in the same column, while those segments that do not have a counterpart in another sequence are represented by gap symbols in the other sequence (List 2014: 66–69; List et al. 2018). Strict cognates can be aligned with each

other in this manner by lining up those phonemes which correspond to each other across different sound sequences. Alignment analyses can also be carried out for partial cognates, provided that the partial cognacy itself is readily annotated (see List et al. 2016 for details on the representation of partial cognates in aligned form). An example for the use of alignments to annotate strict cognates in partial cognate sets can be seen in Figure 2.

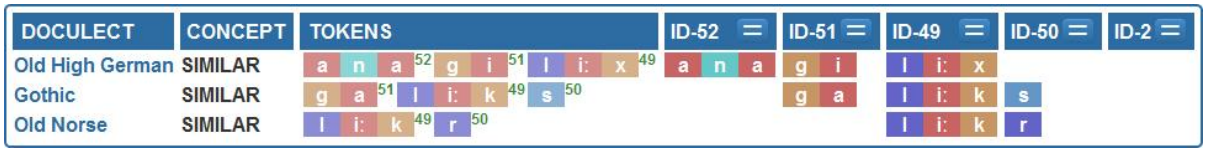


Figure 2: Partially cognate words for 'similar' in Old High German, Gothic and Old Norse. The cognate parts (e.g. *li:x* and *li:k*) are aligned using the EDICTOR tool (List 2017). Data from the Intercontinental Dictionary Series (Key & Comrie 2015)

With this approach, regular sound correspondences are presented in a transparent fashion. Since etymological analyses rely on the identification of regular sound changes, which are usually contrasted with the less systematic processes of word formation or analogy, it is essential for any etymological annotation framework to account for the regularity (or alignability) of etymologically related word forms. The advantage of this approach is that it allows us to both annotate linear word formation processes and to illustrate where we think that sound correspondences are regular. This would not be possible when using the typical cognate judgments in which cognate judgments relate to full words. We hereby follow a similar line of reasoning as proposed in Haspelmath (2002, 176f): morphemes may only exist as some kind of abstraction with respect to the relation between words, but they nevertheless are a helpful concept.

Yet if phonemes differ due to morphological processes, alignment analyses are useless since they can – per definition – only display which segments of related words correspond. But since correspondence may be severely hampered by processes beyond sound change, alignments are not apt to display word formation beyond the level of concatenation that does not leave traces on the pronunciation of the original morphemes of which the words were formed. Further details on how we suggest to handle those cases will be given in §4.2.3.

The problem of alignability can also occur in cases of transparent concatenation: if a regular sound change is involved in the relationship between two partially cognate forms, yet the cause for the sound change lies in the non-cognate part, the forms again cannot be aligned with each other. This can lead to the strange situation that – when comparing two cognate morphemes, one of which has been affected by a sound change whose source was in another morpheme – these morphemes are not strictly cognate, while they may be strictly cognate when comparing the whole words. An example for this can be found in Figure 3: The Gothic and Old High German words meaning 'poison' are strict cognates. However, in Old High German, the /b/ was geminated due to the /i/ (West Germanic /j/) following it (Braune 2004: 98), which constitutes the inflectional ending and thereby a different morpheme by our analysis. When comparing the whole words (on the top), these forms are fully alignable. However, when only comparing the stems (on the bottom), they are not alignable, as the cause for their difference lies in another part of the words, which can only be found in their combination. Assuming that these cases are rather rare, we will treat them in our annotation framework as examples of irregular sound change. Later research will have to show how we can consistently handle cases where alignability exceeds the level of the morphemes.

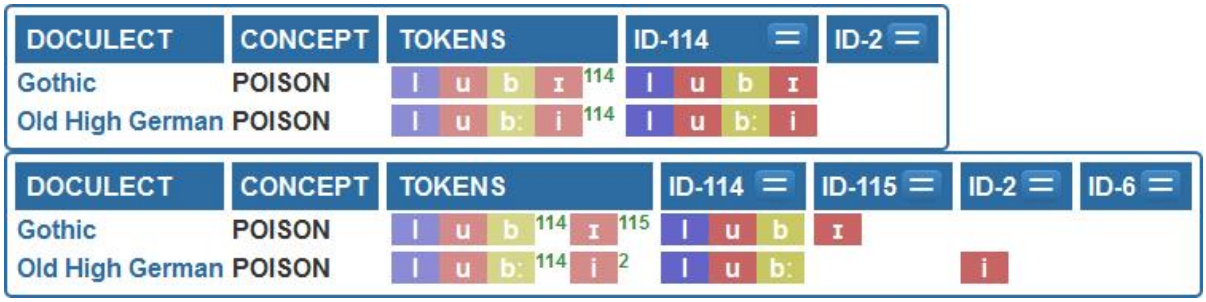


Figure 3: Cognate words for 'poison' in Gothic and Old High German, aligned using the EDICTOR tool (List 2017). Data from the Intercontinental Dictionary Series (Key & Comrie 2015)

As a second aspect of the framework we envision, we need to be able to represent the processes of word formation, including how words are composed of smaller parts. This relates to our concept of strict cognacy in so far as allomorphic and allophonic word formation disrupt alignability as a matter of principle. If the underlying phonology of the languages under investigation is well-know, problems of allomorphy can be handled by changing the original sound sequences that are used as the basis for a given comparison. Thus, instead of representing the German infinitive ending as a syllabic [n̩], when preceded by a consonant in the stem, one can choose a phonemic transcription, in which all infinitive endings are represented as /ən/ (e.g., /fiʃən/ instead of [fɪʃn̩] for *fischen* 'to fish'). Concatenative word formation on the other hand does not necessarily interfere with alignability and can therefore be annotated in a much easier way that we will present in §4.2.2.

4.1.2 Transparent and standardized annotation

In our annotation framework we handle word formation processes by breaking them down into several steps based on our inference methods. These methods are based on epistemological grounds. They represent what we assume we can know or deem probable at a given stage.

The main problem in this endeavor is that each step in reconstruction is based on assumptions gained from previous steps and it is not uncommon to change the result of an earlier step in a later step of the analysis. This *iterative* (not *circular*) procedure is nothing new for historical linguists working in the framework of the comparative method. The problem of this *iterative lifting of insights*, however, is, that it is hard to model it in a transparent way also accessible to machines. Although we know well that iterative reasoning is at the core of all comparative endeavors in historical linguistics, we try to reduce the number of times one has to go back and forth when annotating etymological word relations in our framework.

In our workflow, we first infer linear word formation processes by identifying morpheme boundaries based on the synchronic system of a given language. In the next step we infer regular sound correspondences by comparing morphemes of different languages. Here, we first only compare morphemes of words denoting the same concept ("Swadesh-style cognates") in order to get a reliable baseline for sound correspondences. Cognates across words denoting different concepts ("cross-semantic cognates") are only identified and annotated in a second step. With the knowledge on sound correspondences for strict cognate relations accumulated in the first stage (for which one may use recently proposed algorithms for sound correspondence pattern detection, e.g. List (2019)), it is easier to assess the regularity of cognates which differ in basic meaning. In the final step covered here, etymological relations further obscured by word formation get marked as such.

While the annotation framework we propose here may seem utterly simple, it is important to emphasize that current etymological frameworks usually do not distinguish the different levels of annotation we propose here. Cognacy is still largely treated as a binary concept: two words are either cognate, or not. That cognacy comes along in many different *shades* is what our framework tries to embrace. Although we know that we can barely address all complexities involving in the word formation processes that can be observed for the languages of the world, we are confident that the additional steps of annotation depths we propose here are important for future endeavors.

In the design of our framework we took inspiration from the standards proposed by the CLDF-initiative (Forkel et al. 2018). This means that we use a very straightforward data format based on comma- or tab-separated text files that can be edited with any text editor. The advantage of text formats is that they facilitate both the sharing and storing of data, while making it easy, on the other hand, to access the data with the help of common scripting languages. CLDF essentially gives two major recommendations with respect to data handling in historical and typological language comparison: on the one hand, CLDF propagates long table formats (discussed in the next section), on the other hand, CLDF recommends to *anticipate the need for more than one table*. We follow both suggestions by using long table formats throughout our whole annotation process, while at the same time using additional tables to represent those morphological relations we cannot represent in one table alone.

4.2 Annotation of word relations

4.2.1 Basic format

There are two basic types of annotation in common usage: Either the annotation is added into the data itself (inline annotation) or it is distinct from the data (stand-off annotation) (Eckart 2012: 31). We use a combination of both approaches, but mostly utilize stand-off annotation as this further facilitates keeping the data and our step-wise interpretation thereof distinct. This means that each of the columns mentioned will contain either the data itself, or an annotated version thereof, or pure stand-off annotation, depending on what we deemed most reasonable for representing our judgments.

The central part of our annotation framework consists of a table in which each row is reserved for a specific word (or word form), whereas different columns are used for different levels of annotation, leading to a very straightforward and flexible file structure. The header row specifies the content of each column. It thereby follows the standard input formats of LingPy (List et al. 2019a), a Python library for standard tasks in historical linguistics, and EDICTOR, an interface for analyzing and editing wordlists of cognate languages, which have been described in greater detail in the past (List 2017).

The most basic columns of our format comprise a *unique identifier* for each entry (an integer greater than zero, ID), the *name* of the language variety of the word form (usually in alphanumeric form, without brackets, commas, or other information, DOCULECT), and an elicitation gloss for the concept it denotes (CONCEPT). Including semantic information in a strictly onomasiological, or meaning-based, perspective, has two main advantages. First, it increases the comparability of data and results, since word forms denoting identical or similar concepts can be easily retrieved, specifically when the data is additionally linked to the Concepticon, a large collection of elicitation glosses and concept definitions which recur frequently in cross-linguistic datasets (List 2018c; List et al. 2020;

https://concepticon.clld.org), by adding a column specifying the Concepticon ID (CONCEPTICON_ID). Second, it provides practical help when working on less well investigated language families, since it is well known that initial search for cognates can be most reliably carried out for words denoting identical meanings. However, since the minimal requirement of our formats is only that both meaning and form are provided to form what Gévaudan calls a "lexical unit" (*lexikalische Einheit*; Gévaudan 2007: 28), the format can also be used for "traditional", semasiological (or form-based) annotation that centers around cognate word forms. An example for this basic format can be seen in Table 3. The annotated data underlying the figure is provided within the supplemental material and described in Appendix §2.3.

Table 3: Usage example of the basic format with Panoan data from the Intercontinental Dictionary Series (Key & Comrie 2015)

ID	DOCULECT	CONCEPT	FORM
193	Shipibo-Conibo	one	wɨstiora
194	Tacana	dust	epamo
195	Tacana	fire	ti

Additional columns contain the original data entry for the given word (VALUE), a corrected version of the original entry (FORM), in which specifically multiple variants of the same form are removed or other obvious errors are corrected, and a phoneme-segmented version of the word form (TOKENS), which ideally reflects a standardized transcription system, such as the B(road) IPA defined by the Cross-Linguistic Transcription Systems initiative (List et al. 2019; https://clts.clld.org; see Anderson et al. 2018 for details). While the choice of transcription system is not mandatory (and merely a suggestion to increase the general comparability of the data), our format standardizes the segmented version of word forms by requiring that those symbol sequences representing one sound unit in the transcription are separated by a space and by allowing (for the time being) only for one marker for morpheme boundaries (+) applied to distinguish morpheme boundaries at all levels (including phrases, compounds, clitics, or affixes). While it may seem useful to allow for a more fine-grained distinction of boundary markers (e.g., distinguishing word boundaries from morpheme boundaries) within the TOKENS column of our annotation format, practical annotation has shown that this increases the complexity for computational testing of consistency, while at the same time increasing the rate of errors introduced within the annotation process. If one wants to trace morphological information explicitly, we recommend to annotate it in an additional column, devoted only to this purpose.

Having assembled the data in this form, the core annotation of etymological relations can be done in different steps. To indicate partial cognate relations inside a given language and across multiple languages, we use integer identifiers for each part of a word form, which can be stored in two additional columns, one devoted to cross-semantic alignable cognate sets (called CROSSIDS, from *cross-semantic cognate identifiers*), and one devoted to non-alignable word forms (ROOTIDS). The cross-semantic alignable cognate sets are themselves linked to a column storing the phonetic alignment of each word part (ALIGNMENT), which is identical with the phoneme-segmented transcription (TOKENS), with the exception that gaps are introduced, represented by a dash (-) as gap symbol.

While the cognate set identifiers for alignable and non-alignable word parts along with the alignments allow already for a great deal of flexibility in etymological annotation that largely exceeds the formats that have been used in the past with respect to transparency and explicitness, allowing for an explicit annotation of both internal and external cognates, the identifiers and alignments alone do not provide any semantic or morphosyntactic information. In order to account for this, we use another column that stores morphological information in form of a morphological gloss building on the proposal by Hill & List (2017) (MORPHEMES). All columns of our annotation will be introduced in more detail in the following sections.

4.2.2 Annotation of alignable word relations

As mentioned in §4.1, a central notion of our annotation framework concerns the question whether a given morpheme is *alignable* with cognates in related languages, i.e. only differing from them via regular sound changes, or not. In our framework this would entail marking all morpheme boundaries in the data that have not been obscured by paradigmatic processes of word formation or by later sound changes. Determining the alignability of morphemes, however, depends on the detailed knowledge of sound correspondences, while the detailed knowledge of sound correspondences themselves requires to know the morpheme boundaries of the languages one investigates. When working with less well studied languages, the only way out of this circle is therefore to accept a certain degree of error in early stages of the analysis and to analyze the data in an iterative fashion in which both the annotation of cognates across and inside languages, as well as the analysis of sound correspondences by means of phonetic alignment analyses are repeated several times, reflecting the general iterative workflow of the comparative method (Ross & Durie 1996).

In practice, it has turned out useful to start by annotating all synchronically transparent morpheme boundaries. What counts as *transparent* in this context is of course difficult to determine. In the example annotations which we prepared for this study, we usually started from clear-cut examples of segments which occur both alone and in combination with other elements and thus give concrete hints on the semantics of a morpheme. A second class of transparent morphemes are those that occur in semantically similar words with a clear-cut semantic difference (e.g., *gender*, as in German *Schwiegermutter* 'mother-in-law' vs. *Schwiegervater* 'father-in-law'). In addition, cross-linguistic evidence can be consulted, for example, when encountering regular sound correspondences across parts of words, as those then point to cases of partial cognacy. Consider the following case in Tucanoan languages as reflected in Huber & Reed (1992). In Carijona, the word for 'seed' is *eheru*, but from our sample data of Carijona it is not clear whether this word is morphologically complex. Yet our dataset does contain the Macuna word for 'seed' as well, which is *ahe*, from which with certain confidence it can be assumed that the first part of *eheru* is cognate with *ahe* and that it is therefore segmentable as *ehe* + *ru*. Table 4 shows how these examples are reflected in our annotation format, see Appendix §2.6 for information on the whole dataset with examples on Tucanoan languages.

Table 4: Annotation of Tucanoan data from Huber & Reed (1992)

ID	DOCULECT	CONCEPT	FORM	TOKENS	CROSSIDS
3467	Carijona	seed	Eheru	e h e + r u	154 159
13571	Macuna	seed	Ahu	a h u	154

In a similar fashion, the knowledge of patterns of *semantic motivation* which are transparent in one language can be used to search for similar, less transparent patterns in another language. In order to maximize the amount of morphemes that can be found in this manner, we recommend that the data be provided in a phonemic transcription, not in a phonetic one, in order to avoid purely phonologically conditioned allomorphy. This same distinction is recommended by Lehmann (2004: 7) for morphophonemic representation of data.

As mentioned above, we annotate morpheme boundaries by adding a plus symbol (+) surrounded by spaces between the segmented representation of our word forms (as given in the column we call TOKENS). This procedure can be facilitated by using computer-assisted methods. If frequent morphemes are known to the researcher, a very simple (but in our experience also efficient) approach for marking at least a larger part of the morphemes semi-automatically is to use search and replace functionalities (in combination with regular expressions if needed). In this way, all instances of, for example, the Gothic infinitive suffix {-n} can be easily annotated by searching for the string n$ (i.e., the n occurring in the end of all words) and replacing it with + n. Our rudimentary collection of scripts for the curation and analysis of morphologically annotated wordlist data contains a script that automatizes this task (which should, of course, be double-checked manually in a second step). See Appendix §1.2.2 for details on this.

Once the morphological segmentation has been done for a considerable amount of words in a given language, *morphological glosses* can be added (in the column MORPHEMES). Here, we follow an idea proposed by Hill & List (2017), which allows for a quick but straightforward annotation of language-internal etymological word relations. In the original proposal, each morpheme in a morphologically segmented word-form was glossed by a short gloss representing either the basic meaning of the morpheme or its grammatical function. In practice, this is done by writing the morpheme glosses in free form, using a space as the character for segmentation. As a result, spaces are not allowed inside a morpheme gloss and need to be represented by dashes or underscores or other techniques. This results in "a language-internal word family analysis, as it allows us to identify cognates within the same language" (Hill & List 2017: 63, emphasis removed).

To ease the annotation procedure, it turned out to be straightforward to automatically generate the glosses from the elicitation glosses used for the concepts in a given wordlist in a first instance, and then to manually correct the cases where this very simple procedure fails (see Appendix §1.2.3 for more information). Additionally, one may want to distinguish between *content morphemes* and *grammatical morphemes*. We annotate the latter by adding an underscore at the beginning of their gloss. This is especially helpful if one wants to exclude, for example, infinitive suffixes from word family analyses, since they would otherwise strongly distort the results. In each row, there should be the same amount of morphemes in the segmented word form (column TOKENS) as morpheme glosses (in the column MORPHEMES). Our collection of scripts for data curation provides a small script that performs sanity tests to check for consistency in this regard (see Appendix §1.3 for details).

The glossing style is left to the annotator. There are two styles we can recommend: Either one uses the morphemes' forms themselves as their own glosses, or one glosses content morphemes with their basic meaning, and grammatical morphemes with their general function or with a combination of their form and their function. These glosses need to be unique only within the same language variety.

While the morpheme glosses serve only to annotate language-internal relations, cross-linguistic cognacy that crosses semantic boundaries needs to be annotated with the help of numeric cognate identifiers (column CROSSIDS). While tools for cognate and alignment annotation, such as EDICTOR, provide help to carry out this part of the analysis, it may be useful to pre-process the data automatically, using state-of-the-art software for phonetic alignments, sound correspondence pattern detection, and cross-semantic cognate detection (as offered, e.g., in the LingRex package, see List 2018a, 2019).

Table 5 provides a usage example illustrating how morpheme boundaries, glosses, and cognate identifiers can be added in our annotation framework (see Appendix §2.2 for information on the whole dataset with examples on Germanic languages). More examples are provided in the supplementary material, which is described in detail in Chapter 2 of the appendix accompanying this paper.

Table 5: Usage example of adding morpheme boundaries and glosses in our annotation framework with Germanic data from the World Loanword Database (Haspelmath & Tadmor 2009) and the Intercontinental Dictionary Series (Key & Comrie 2015)

ID	CONCEPT	DOCULECT	FORM	TOKENS	MORPHEMES	CROSSIDS
36	BOW	Old High German	bogo	b o g + o	bow _o-nom	51 37
40	ELBOW	Old High German	elinbogo	e l i n + b o g + o	ell bow _o-nom	53 51 37
41	RAINBOW	Old High German	reganbogo	r e g a n + b o g + o	rain bow _o-nom	54 51 37
44	ELBOW	Old Norse	ǫlnbogi	ɔ l n + b o g + i	ell bow _i-nom	53 51 37
45	BOW	Old Norse	bogi	b o g + i	bow _i-nom	51 37

Morpheme boundaries are an abstraction based on a specific language in time and may change through time due to becoming opaque or being created anew by reanalysis. Therefore, at later stages in annotation, morpheme boundaries might be found by language comparison or other evidence that were not transparent for the researcher before, and suspected morpheme boundaries might turn out to be the result of reanalysis. To avoid getting lost in annotation, it is important to make clear to oneself that any analysis is preliminary, and that – ideally – an analysis should always have a clear-cut and transparent reference point. When deciding to compare Germanic languages like Old High German and Gothic, for example, it would not make sense to annotate morpheme boundaries which were not perceived as such by the speakers of the common ancestor language Germanic.

4.2.3 Annotation of non-alignable word relations

While one could stop with the consistent annotation of alignable etymological relations between words, a typical etymological analysis has the ambition of listing all etymological relations that can be inferred, including those where etymological relationship has been obscured by paradigmatic processes of word formation or by sound changes which were triggered by conditioning contexts that cross morpheme boundaries (for the latter, compare the case of German *Messer*, discussed in §3.2).

In practice, it may be hard to distinguish the two processes. In Middle High German, for example, an *i* triggered the fronting of a back vowel in the syllable before, across morpheme boundaries, including the *i* in the diminutive-suffix {-lîn} (Modern German {-lein}). In Modern German, {-lein} still triggers basically the same *phonetic change*, but today this happens purely as the result of a productive *morphological pattern*. As a result, when dealing with Modern

German diminutives whose roots contain an original back vowel that underwent the process of *umlaut*, the vowel change may either be attributed to the synchronic morphological pattern or due to the diminutive having been lexicalized already in Middle High German (and simply retained its phonologically caused vowel alternation). Given these difficulties, our annotation does not strictly distinguish between these processes at this stage.

Concretely, the annotation targets again all morphemes in the segmented representation of the word forms, but this time, we introduce a deeper level of cognate identifiers (which we place in a column called ROOTIDS, i.e., *root cognate identifiers*), in which we first annotate all cases of cognates (language-internally and language-externally) which have been ignored in the previous step, since they turned out to be not alignable.

By combining the deeper etymological annotations with the shallower ones provided in form of cross-semantic, strict cognates, we can automatically create a multi-layered, directed word family graph, which starts from a given root identifier as the source and links to the strict cognate sets, which themselves link to the extant word forms. While the graph in this form lacks a hierarchy among the non-alignable cognates of a word family, this information can (if it is known to the researcher) be annotated with the help of an additional table that represents the etymological data in form of a directed derivation graph, with source and target nodes. We provide a script that creates the shallow network from a given dataset by listing all words sharing at least one identifier in a specified column in tabular form (see Appendix 1.4.1 for a usage example describing how to use this script). In Schweikhard & List (forthcoming) we describe a more exhaustive approach to representing both alignable and non-alignable etymological relations in such word family graphs.

Figure 4 provides an example of our tabular annotation of non-alignable word relations (A), the corresponding word family graph (B), the tabular annotation of a given hierarchy (C), and the corresponding word family graph derived from this hierarchy (D). The annotated data underlying the figure is provided within the supplemental material and described in Appendix 2.2.

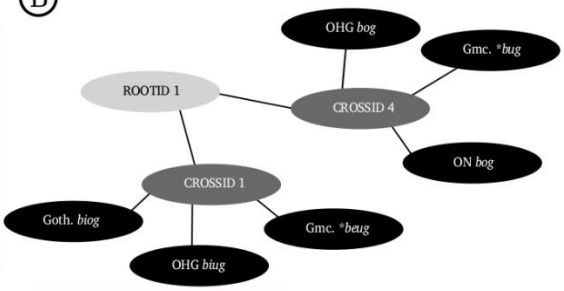
4.3 Annotation examples and code

The supplementary material accompanying this paper offers extended annotation examples for six different language families: Burmish languages (Sino-Tibetan family, based on Hill & List 2017, Appendix 2.1), Germanic languages (Indo-European family, based on Key & Comrie 2015, Appendix 2.2), Panoan languages (Pano-Tacanan family, based on Key & Comrie 2015, Appendix 2.3), Polynesian languages (Austronesian family, based on Walworth 2018, Appendix 2.4), Sanzhi Dargwa (Nakh-Daghestanian family, based on Forker 2019), and Tucanoan languages (Tucanoan family, based on Huber & Reed 1992). The examples were selected in such a way that they illustrate the general applicability of our annotation framework and the advantages of trying to follow a given set of guidelines consistently.

Ⓐ

ID	CONCEPT	DOCULECT	TOKENS	CROSSIDS	ROOTIDS
1	BEND	Old High German	b i o g + a + n	1 2 3	1 2 3
2	BOW	Old High German	b o g + o	4 5	1 5
3	BEND	Gothic	b i u g + a + n	1 2 3	1 2 3
4	BOW	Old Norse	b o g + i	4 5	1 5
5	BEND	Germanic	b e u g + a + n	1 2 3	1 2 3
6	BOW	Germanic	b u g + a n	4 5	1 5

Ⓑ



Ⓒ

Source	Source Form	Target	Target Form	Process
5	*beugan	6	*bugan	change to zero ablaut grade and exchange of ending
5	*beugan	1	biogan	regular sound change
5	*beugan	3	biugan	regular sound change
6	*bugan	2	bogo	regular sound change
6	*bugan	4	bogi	regular sound change

Ⓓ

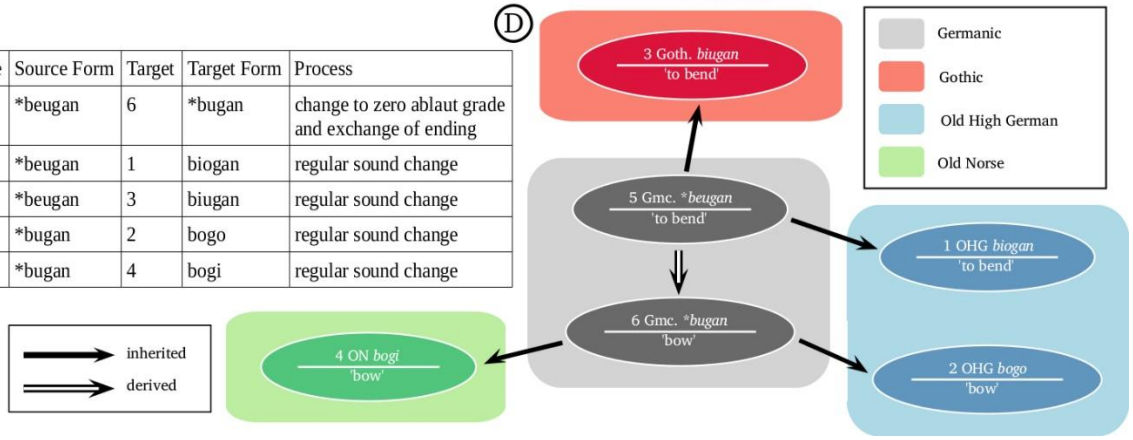


Figure 4: Annotating cognacy between full words. Data from the World Loanword Database (Haspelmath & Tadmor 2009) and the Intercontinental Dictionary Series (Key & Comrie 2015), reconstruction based on Kroonen & Lubotsky (2013, s.vv. **beugan-* and **bugan-)*

In addition, we provide a set of scripts, which are distributed in form of a small Python package, called worforpy, which can be used to ease the task of annotation by offering methods to (a) preprocess the data before starting with a given annotation, as well as methods to (b) validate the data, in order to make sure that the annotation has been done correctly. The usage of these scripts and routines, which were also mentioned throughout this paper, are described in detail in Chapter 1 of the Appendix in the supplementary information accompanying this paper.

Parts of the suggestions described here have already been implemented in other software packages, specifically LingPy (List et al. 2019a) and EDICTOR (List et al. 2017), as mentioned before. In the future, we hope that the new ideas which we have introduced here can find a broader support.

4.4 Caveats

While we are confident that our annotation framework is capable of handling a large section of etymological relations, there are some scenarios which are beyond our current scope. For one, we cannot easily describe in which way, by which exact combination of processes, the morphemes are related to each other which differ due to non-concatenative word formation processes. Additionally, in those cases in which morpheme boundaries have disappeared or were created anew, and also in the case of analogy where the morphological pattern applied can be traced back to a specific word (or small group of words) serving as a role-model, it is not possible to provide this information in this manner.

Furthermore, if no morpheme boundary is annotated between two originally distinct morphemes since they have merged by processes of word formation or sound change or since

the morpheme boundary has become opaque by other means, cognates of either morpheme would need to be annotated as cognate by receiving the same ID in the ROOTIDS-column. For these cases, we recommend linking the result of the merged morphemes to only one of the cognate sets its morphemes are related to. For example, both Old High German *būr* ('dwelling', 'cage') and *hār* ('hair') contain a morpheme that goes back to the Indo-European suffix {-ro} (Kroonen & Lubotsky 2013, s.vv. *būra-* and *hēra-*) but which cannot be considered synchronically transparent. It would be possible to give both of them the same ROOTID in order to mark their partial cognacy, yet this would be misleading as it would also give Old Norse *hār* the same ROOTID as *būi* ('dweller') since that word shares the same root with *būr* (Kroonen & Lubotsky 2013, s.v. *bōan ~ būan*). More generally, it seems advisable to focus on linking content morphemes and neglect grammatical morphemes, especially when the latter include a phonetic merger, but it is difficult to define a boundary what to include in the annotation and what not.

One situation in which the annotation of cognacy is possible but can be problematic affects inflectional grammatical morphemes as those may be easily replaced by a word switching into a different inflectional class. For instance, consider the Old Norse infinitive ending *-a*, which may go back to several different Germanic suffixes, depending on the verb in question (Ringe 2006: 235f). In order to determine with at least some level of certainty which of the stem-forming suffixes of other Germanic languages the infinitive suffix of a given Old Norse verb is cognate with, one needs to know the reflex of that verb in a Germanic language that has retained the differentiation between the suffixes, e.g. Gothic, where it however may not be attested, and adopt the assumption that the verb did not switch to a different inflection class in either of the languages involved. If the verb in question was only formed in Old Norse, then it is almost a philosophical question which Germanic suffix its infinitive is cognate with – it would be all and none simultaneously. We recommend to annotate it as being cognate to which seems most reasonable given the data, but to feel free to not posit cognacy with any in cases of doubt. In such cases, the ID 0 can be given.

The opposite situation can be seen in the reflexes of the thematic vowel of verbs in Old Norse in comparison to Gothic, in infinitives and past participles. In Old Norse, it is attested as *a* in the aforementioned infinitive ending (e.g. *drikka*), but as *i* in the past participle (e.g. *drukkinn*). In Gothic on the other hand, we find an *a* in both instances, in *drigkan* and in *drugkans*. The most likely explanation seems to be a context-dependent sound change, but some analogical process cannot easily be excluded. Since we assigned the same ID to all instances of the verbal thematic vowel in Gothic, but different ones to the different reflexes in Old Norse, we would annotate only one of the Old Norse reflexes as cognate with the Gothic reflexes – yet even this comparatively simple scenario begs the question of how we could decide which one to choose. An arbitrary decision is necessary here.

Similarly, reanalysis can lead to a shift in the morpheme boundary. Consider the English word *alone*. From internal evidence like the word *lonely*, among others, we can posit a synchronic morpheme boundary {a}{lone}. Cross-linguistic evidence on the other hand leads to the conclusion that the word is cognate with German *allein*, in which case the morpheme boundary is less obscured than in the English cognate, leading to a historical morpheme boundary of {al}{one} and partial cognacy with *all* and *one* (Pfeifer 1993, s.v. *allein*).

All this hopefully makes clear why we consider this linear form of annotation merely a helpful tool for finding regularities in sound correspondences between languages and a useful workflow for determining the most reliable cases of cognacy, but not a detailed way of

presenting all relationships between words. We are still working on a more exhaustive framework to fully annotate etymological relations which will also allow us to handle non-linear word formation processes, as we have hinted at at the end of §4.2.

5 Conclusion

In this paper we have proposed an annotation framework that is supposed to ease the investigation of word formation processes from a cross-linguistic perspective. Although we are aware of the complexity of the task, we see multiple use cases for this framework. For scholars working on etymologies, the framework along with the tools we propose and describe can be very helpful to increase the explicitness of their research, by allowing them to define concretely where and how they suggest words to be related. For scholars working in the field of semantic or lexical typology (Koptjevskaja-Tamm & Liljegren 2017), the framework can provide great help in the collection of examples that can be easily compared cross-linguistically. For scholars working on computational approaches in historical linguistics and linguistic typology, the framework can serve as the basis in which the software they create should read and write its findings. Scholars creating dictionaries and working in language description, furthermore, could annotate at least parts of their data more explicitly, using morpheme glosses, as described here, in order to make it easier for colleagues to inspect and digest original data they might want to use in their research.

We are well aware of potential limitations of the framework proposed here, and emphasize that it is best treated as work in progress. Nevertheless, we feel the importance to share the framework already in this initial stage, as we hope that more people could test it and thereby help to improve upon it. That there is a definite need for more standardization and more transparency in the field of diversity linguistics seems to be out of question. But how it can be satisfied is, of course, another question, for which we have tried to provide an initial answer with our framework.

Acknowledgments

This research was funded by the ERC Starting Grant 715618 “Computer-Assisted Language Comparison” (CALC, https://digling.org/calc). We thank the anonymous reviewer as well as Nathan W. Hill and Alexander Vertegaal for feedback on the draft version of this paper. We also thank Martin Haspelmath, Gereon Kaiping, Timotheus A. Bodt, Thiago C. Chacon, Mei-Shin Wu, Tiago Tresoldi, Roberto Zariquiey, and Yunfan Lai for discussing and/or testing our annotation framework with us.

Appendix and supplementary material

The appendix submitted along with this paper contains detailed instructions to apply the code and further information on our sample annotations. Together with the annotation examples and the code, it has been curated on GitHub (https://github.com/digling/word-formation-paper) and archived on Zenodo (https://zenodo.org/record/3889970).

References

Anderson, Cormac & Tresoldi, Tiago & Chacon, Thiago C. & Fehn, Anne-Maria & Walworth, Mary & Forkel, Robert & List, Johann-Mattis. 2018. A cross-linguistic database of phonetic transcription systems. *Yearbook of the Poznań Linguistic Meeting* 4(1). 21–53. (doi: 10.2478/yplm-2018-0002).

Barsalou, Lawrence W. 2017. Cognitively plausible theories of concept composition. In Hampton, James A. & Winter, Yoad (eds.), *Compositionality and concepts in linguistics and psychology*, 9–30. Cham: Springer International Publishing. (doi: 10.1007/978-3-319-45977-6_2).

Baxter, William H. & Sagart, Laurent. 1998. Word formation in Old Chinese. In Packard, Jerome L. (ed.), *New approaches to Chinese word formation: Morphology, phonology and the lexicon in Modern and Ancient Chinese*, 35–76. Berlin: de Gruyter.

Behr, Wolfgang. 2015. G. Sampson, "A Chinese phonological enigma": Four comments. *Journal of Chinese Linguistics* 43(2), 719–732.

Braune, Wilhelm. 2004. *Althochdeutsche Grammatik I: Laut- und Formenlehre*. (Ed. Reiffenstein, Ingo, 15th ed.). Tübingen: Max Niemeyer Verlag.

Cohen, Paul S. 2004. A new etymology for Latin *aquila*. In Clackson, James & Olsen, Birgit A. (eds.), *Indo-European word formation: Proceedings of the conference held at the University of Copenhagen October 20th – 22nd 2000*, 25–35. Copenhagen: Museum Tusculanum Press.

Crist, Sean. 2005. Toward a formal markup standard for etymological data. (http://www.sean-crist.com/professional/publications/crist_etym_markup.pdf) (Accessed 2019-07-29).

Dehmer, Matthias & Emmert-Streib, Frank & Graber, Armin, & Salvador, Armindo (eds.). 2011. *Applied statistics for network biology: Methods in systems biology*. Weinheim: Wiley-Blackwell.

de Saussure, Ferdinand. 1916. *Cours de linguistique générale*. Lausanne: Payot.

Eckart, Kerstin. 2012. Resource annotations. In *CLARIN-D user guide*, 30–42. (http://media.dwds.de/clarin/userguide/userguide-1.0.1.pdf) (Accessed 2018-11-05).

Forkel, Robert & List, Johann-Mattis & Greenhill, Simon J. & Rzymski, Christoph & Bank, Sebastian & Cysouw, Michael & Hammarström, Harald & Haspelmath, Martin & Kaiping, Gereon A. & Gray, Russell D. 2018. Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Nature Scientific Data* 5(180205). (doi: 10.1038/sdata.2018.205).

Forker, Diana. 2019. Sanzhi Dargwa dictionary. *Dictionaria* 5, 1–5533. (online version: https://dictionaria.clld.org/contributions/sanzhi).

Gévaudan, Paul. 2007. *Typologie des lexikalischen Wandels: Bedeutungswandel, Wortbildung und Entlehnung am Beispiel der romanischen Sprachen*. Tübingen: Stauffenburg.

Handel, Zev. 2008. What is Sino-Tibetan? Snapshot of a field and a language family in flux. *Language and Linguistics Compass* 2/3, 422–441.

Haspelmath, Martin. 2002. *Understanding morphology*. London: Arnold.

Haspelmath, Martin & Tadmor, Uri (eds.). 2009. *WOLD*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (online version: https://wold.clld.org/).

Hauser, Marc D. & Chomsky, Noam & Fitch, W. Tecumseh. 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science* 298, 1569–1579.

Heine, Bernd. 2019. On the grammaticalization of some processes of word formation in Africa. *SKASE Journal of Theoretical Linguistics*, 16(1), 2–18.

Hill, Nathan W. 2014. A note on voicing alternation in the Tibetan verbal system. *Transactions of the American Philosophical Society*, 112(1), 1–4.

Hill, Nathan W. & List, Johann-Mattis. 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.

Huber, Randall Q. & Reed, Robert B. (1992). *Vocabulario comparativo: Palabras selectas de lenguas indígenas de colombia [Comparative vocabulary. Selected words from the indigeneous languages of Columbia]*. Santafé de Bogota: Asociatión Instituto Lingüístico de Verano.

Key, Mary R. & Comrie, Bernard (eds.). 2015. IDS. Leipzig: Max Planck Institute for Evolutionary Anthropology. (online version: https://ids.clld.org/)

Koptjevskaja-Tamm, Maria & Liljegren, Henrik. 2017. Semantic patterns from an areal perspective. In Hickey, Raymond (ed.), *The Cambridge handbook of areal linguistics*, 204–236. Cambridge: Cambridge University Press.

Kroonen, Guus & Lubotsky, Alexander (eds.). 2013. *Etymological dictionary of Proto-Germanic*. Leiden: Brill.

Lai, Yunfan. 2016. Causativisation in Wobzi and other Khroskyabs dialects. *Cahiers de Linguistique Asie Orientale* 45(2), 148–175.

Lees, Robert B. 1953. The basis of glottochronology. *Language* 29(2), 113–127.

Lehmann, Christian. 2004. Interlinear morphemic glossing. In Booij, Geert & Lehmann, Christian & Mugdan, Joachim & Skopeteas, Stavros (eds.), *Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung* (Vol. 2.2), 1834–1857. Berlin: Walter De Gruyter.

List, Johann-Mattis. 2014. *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.

List, Johann-Mattis. 2016a. Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction. *Journal of Language Evolution* 1(2), 119–136. (doi: 10.1093/jole/lzw006).

List, Johann-Mattis. 2016b. Contraction. In Sybesma, Rint (ed.), *Encyclopedia of Chinese language and linguistics*. Leiden: Brill.

List, Johann-Mattis. 2017. A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets. In *Proceedings of the 15th conference of the European chapter of the Association for Computational Linguistics: Software demonstrations*, 9–12. Valencia: Association for Computational Linguistics. (online version: https://digling.org/edictor/)

List, Johann-Mattis. 2018a. *LingRex: Linguistic reconstruction with Lingpy. Version 0.1.3*. Jena: Max Planck Institute for the Science of Human History. (doi: 10.5281/zenodo.1544943).

List, Johann-Mattis. 2018b. Regular cognates: A new term for homology relations in linguistics. The Genealogical World of Phylogenetic Networks. 5(8) (url: https://phylonetworks.blogspot.com/2018/08/regular-cognates-new-term-for-homology.html).

List, Johann-Mattis. 2018c. Towards a history of concept list compilation in historical linguistics. *History and Philosophy of the Language Sciences* 5(10), 1–14.

List, Johann-Mattis. 2019. Automatic inference of sound correspondence patterns across multiple languages. *Computational Linguistics* 1(45), 137–161. (doi: 10.1162/coli_a_00344).

List, Johann-Mattis & Anderson, Cormac & Tresoldi, Tiago & Rzymski, Christoph & Greenhill, Simon J. & Forkel, Robert. 2019. *Cross-linguistic transcription systems*. *Version 1.2.0*. Jena: Max Planck Institute for the Science of Human History. (online version: https://clts.clld.org, doi: 10.5281/zenodo.2633838)

List, Johann-Mattis & Greenhill, Simon J. & Tresoldi, Tiago & Forkel, Robert. 2019a. *LingPy. A Python library for quantitative tasks in historical linguistics. Version 2.6.6*. Jena: Max Planck Institute for the Science of Human History. (url: http://lingpy.org).

List, Johann-Mattis & Lopez, Philippe & Bapteste, Eric. 2016. Using sequence similarity networks to identify partial cognates in multilingual wordlists. In *Proceedings of the Association of Computational Linguistics 2016: Short papers*, 599–605. Berlin: Association of Computational Linguistics.

List, Johann-Mattis & Pathmanathan, Jananan S. & Lopez, Philippe & Bapteste, Eric. 2016a. Unity and disunity in evolutionary sciences: Process-based analogies open common research avenues for biology and linguistics. *Biology Direct* 11(39), 1–17.

List, Johann Mattis & Rzymski, Christoph & Greenhill, Simon & Schweikhard, Nathanael & Pianykh, Kristina & Tjuka, Annika & Wu, Mei-Shin & Forkel, Robert. 2020. *Concepticon. A resource for the linking of concept lists. Version 2.3.0*. Jena: Max Planck Institute for the Science of Human History. (doi: 10.5281/zenodo.3706687).

List, Johann-Mattis & Walworth, Mary & Greenhill, Simon J. & Tresoldi, Tiago & Forkel, Robert. 2018. Sequence comparison in computational historical linguistics. *Journal of Language Evolution 3(2)*, 130–144. (doi: 10.1093/jole/lzy006).

Matisoff, James A. 2003. *Handbook of Proto-Tibeto-Burman: System and philosophy of Sino-Tibetan reconstruction*. Berkeley: University of California Press.

Mees, Bernard. (2014). The etymology of *rune*. *Beiträge zur Geschichte der deutschen Sprache und Literatur* 136(4), 527–537. (doi:10.1515/bgsl-2014-0046).

Mei, Tsu-Lin. 2012. The causative *s- and nominalizing *-s in Old Chinese and related matters in Proto-Sino-Tibetan. *Language and Linguistics* 13(1), 1–28.

Meyer-Lübke, Wilhelm (ed.). 1911. *Romanisches etymologisches Wörterbuch*. Heidelberg: Winter.

Mukai, Makiko. 2019. Productivity of recursive compounds. *SKASE Journal of Theoretical Linguistics* 16(1), 35–48.

Pfeifer, Wolfgang (ed.). 1993. *Etymologisches Wörterbuch des Deutschen*. Berlin: Akademie. (Retrieved from https://www.dwds.de/).

Ringe, Don. 2006. *From Proto-Indo-European to Proto-Germanic*. Oxford: Oxford University Press.

Ross, Malcom & Durie, Mark. 1996. Introduction. In Durie, Mark (ed.), *The comparative method reviewed. Regularity and irregularity in language change*, 3–38. New York: Oxford University Press.

Sampson, Geoffrey. 2015. A Chinese phonological enigma. *Journal of Chinese Linguistics* 43(2), 679–691.

Schwarzwald, Ora. 2019. Linear and nonlinear word formation in Hebrew – words which end with *-on*. *SKASE Journal of Theoretical Linguistics* 16, 109–120.

Swadesh, Morris. 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.

Schweikhard, Nathanael E. & List, Johann-Mattis. Forthcoming. Modeling word trees in historical linguistics. Preliminary ideas for the reconciliation of word trees and language trees. In Brogyanyi, Bela & Lipp, Reiner (eds.), *Sprach(en)forschung: Disziplinen und Interdisziplinarität. Akten der 27. Fachtagung der Gesellschaft für Sprache und Sprachen GeSuS e.V. in Warschau, 30. Mai-1. Juni 2019*. Hamburg: Verlag Dr. Kovač. (preprint: https://doi.org/10.17613/8h49-rp11)

Trask, Robert L. 2000. *The dictionary of historical and comparative linguistics*. Edinburgh: Edinburgh University Press.

Walworth, Mary. 2018. *Polynesian segmented data (version 1)*. Jena: Max Planck Institute for the Science of Human History. (doi: 10.5281/zenodo.1689909).

Watkins, Calvert. 1990. Etymologies, equations, and comparanda: Types and values, and criteria for judgment. In Baldi, Philip (ed.), *Linguistic change and reconstruction methodology* (Vol. 1), 289–303. Berlin: de Gruyter.

Nathanael E. Schweikhard
Department of Cultural and Linguistic Evolution
Max Planck Institute for the Science of Human History
Kahlaische Straße 10
07745 Jena, Germany
schweikhard@shh.mpg.de

Johann-Mattis List
Department of Cultural and Linguistic Evolution
Max Planck Institute for the Science of Human History
Kahlaische Straße 10
07745 Jena, Germany
list@shh.mpg.de

In SKASE Journal of Theoretical Linguistics [online]. 2020, vol. 17, no. 1 [cit. 2020-06-03].Available on web page http://www.skase.sk/Volumes/JTL43/pdf_doc/01.pdf. ISSN 1336-782X