DiACL – Diachronic Atlas of Comparative Linguistics Online. Description of Subsection Lexicology (2.0)

Authors: Gerd Carling, Sandra Cronhamn, Rob Farren, Rob Verhoeven (Lund University)

Date of publication: 2018-02-09

Description includes datasets:

Swadesh 100. URL: https://diacl.ht.lu.se/WordListCategory/Details/100

Swadesh 200. URL: https://diacl.ht.lu.se/WordListCategory/Details/200

Culture words for South America. URL: https://diacl.ht.lu.se/WordList/Index/

Culture words for Indo-European. URL: https://diacl.ht.lu.se/WordList/Index/

Culture words for Austronesia. URL: https://diacl.ht.lu.se/WordList/Index/

Culture words for Caucasus. URL: https://diacl.ht.lu.se/WordList/Index/

Culture words for Basque. URL: https://diacl.ht.lu.se/WordList/Index/

Culture words for Uralic. URL: https://diacl.ht.lu.se/WordList/Index/

Culture words for Middle-Eastern non-IE. URL: https://diacl.ht.lu.se/WordList/Index/

Culture words for Turkic. URL: https://diacl.ht.lu.se/WordList/Index/


URL: https://diacl.ht.lu.se/


URL: https://diacl.ht.lu.se/Lexeme/Index
§1. Subsection Lexicology: aim, demands and basic models

§1.1. Introduction

The basic aim of the lexicology subsection is to create a comparative lexical cognacy database, fulfilling the demands of phylogenetic, evolutionary, and lexicostatistical analysis but also accounting for information retrieved from comparative method, such as external/internal reconstruction, relative chronology, and semantic change. Since these methods are substantially different in the way they investigate lexical cognacy and change, the database hosts two types of datasets, which are basically different in the way they code cognacy. We label these two methods 1) Cognacy coding, and 2) Etymology coding, where the former represent a traditional lexical substitution dataset, as introduced by the lexicostatistics in the 1950s, and where the latter mirrors a comparative etymological model, where cognacy is based on etymological trees. The two types will be carefully described under §1.2.

The most commonly used datatype for phylogenetics and lexicostatistics is lexical datasets with basic vocabulary (Swadesh lists, Leipzig-Jakarta lists). These type of datasets are therefore also included in this subsection, as a separate Word List. The basis for lexicostatistics is the measuring of rate of substitution of cognates of a predefined set of lexical concepts (Dunn, 2014, pp. 193-194; Swadesh,
1955), a method that tabulates pairwise distances between languages, based on cognacy. An important criterion for inclusion of cognates in a list is that the semantic criteria match: if a categate changes its meaning, it is by necessity excluded from the list. Since our aim is to create datasets, which are prepared for phylogenetic analysis, it is a demand that they fulfil these criteria.

However, our aim is also to compile datasets that meet the demands of lexicography and comparative linguistics, to which lexicostatistical dataset represents a rough reduction of a very complex and varying reality. First, we intend to include, as far as possible, dictionary-type information about lexemes (transcription, script, IPA, polysemy, grammatical information, sources, see fig. 1), as well as etymological information, i.e., various types of cognacy relations to other lexemes within a language family. We also intend to meet the uncertainties and problems connected with the etymological method, in order to provide reliable and secure datasets, which are grounded in the most reliable etymological reference literature (see 2.5).

As mentioned before, the database contains basic vocabulary lists (Swadesh lists), but lexical data is also expanded beyond the domain of basic vocabulary, into other domains of the lexicon. This is particularly the case with lesser-researched languages, such as languages of South America, where we have compiled lexical data, by means of fieldwork, for languages that entirely lack dictionary resources.

At the centre of the Lexicology subsection of DiACL are lexical concepts or core concepts, a frequently occurring notion, used in comparative, contrastive and computational semantic research and data compilation (List & Cysouw, 2016). Concepts are typically organized by concept lists or concepticons, defined as ‘curated sets of concepts, minimally indexed via one or more words from a language, but perhaps, also more elaborately described using multiple languages’ (Poornima & Good, 2010). The concept list or concepticon model, as it is used by IDS or by Concepticon (List & Cysouw, 2016) has its roots in the model introduced by Buck (Buck, 1949), who only targeted one family, Indo-European. In our Indo-European dataset, the dictionary by Buck has been an important source. However, the aim of our database is mainly comparative and diachronic, and therefore we have selected to use a model of chunking lists by area and family, which is different from a lexical database such as Concepticon (List & Cysouw, 2016). We aim at highest reliability in etymological coding, following the principles laid out by, e.g., Hoffman and Tichy (Hoffman & Tichy, 1980), for securing reconstructions, avoiding as far as possible any paleo-linguistic speculation, substrate assumptions, or deep-family etymology. However, in fact, a vast number of etymologies boil down to an uncertain origin, where no reliable reconstruction is possible, but the apparent correspondence in sound structure and meaning cannot be overlooked or regarded as pure chance. Here, we open for multiple possible explanations, such as prehistoric migration words, loans, or possible substrate influence (Kroonen & Iversen, 2015).

A basic problem has been to solve this dilemma in the database construction. For that purpose, we have invented a unit Stub language, which we use at the very bottom of lexical etymologies. Stubs normally belong to a language family, and they indicate that lexemes are connected both by concept and by sound structure. Stubs normally lead to proper reconstructions – of which there may be several – in proto-languages. How this is solved technically, see §2.5, for solving of conflicting sources, see §4.
Figure 1. Screenshot of Lexeme beyki “beech” in Icelandic.

In order to meet the demands of lexicostatistics and comparative linguistics we have created two different instruments, which define relations between lexemes on the function and on the form side. By means of these instruments, we may measure rate of substitution (for phylogenetics, phylogeography, evolutionary linguistics etc.) and the history of individual etymologies can be traced (for the current status and advice of usages see §4).

These instruments are labelled Word Lists and Etymologies (see fig. 2).

**Word Lists** correspond to predefined lists of lexical concepts, such as a Swadesh list or a culture list from a specific area. These lists can be downloaded for lexicostatistic analysis.

**Word List Item** (e.g., OX, WHEEL, BLOOD) corresponds to lexical core concepts as defined in the literature (Dunn, 2014; Haspelmath & Tadmor, 2009), of which substitution is measured in lexicostatistics. These corresponds to Concepts in the Concepticon database (List & Cysouw, 2016). A Lexeme connected to a Word List Item typically targets the first/main meaning in the language, but if there are two or several meanings in a language with the same lexical concept, we include all. Connections between Word List Items therefore target a connection on the function side between two Lexemes (note however that there are differences between Lexical cognacy and Etymology coding, see §1.2.).

**Etymologies** connect lexical cognates on the form side and can account for all types of complex relations between lexical cognates, including borrowing, derivation, and semantic change. The correlation between Word List Items and Etymologies can be seen in figure 2, exemplified on a well-known etymology, the Indo-European word for MEAT and BLOOD. The difference between Cognacy and Etymology coding is described under §1.2. The organization of Word List and Etymology parts in the database will be described more in detail under §2.
Figure 2. Graph explaining the difference between the cognacy and etymology methods: in cognacy method blue circles and orange circles belong to two different cognacies, BLOOD versus MEAT, in the etymology method all circles belong to one tree, stemming either from a stub MEAT or BLOOD, both occurring as core meanings in branches of the tree.

Figure 3. Overview of tables and relations in the DiACL database
§1.2. Cognacy coding and Etymology coding

§1.2.1. Cognacy coding (Swadesh lists)

As mentioned in previous chapter, we use two types of coding, which are reflected by means of the structure of the etymological trees of lexical concepts. The first method, labelled cognacy coding, corresponds to the lexicostatistical method, the way in which it was designed by Swadesh and his followers in the 1950s (Swadesh, 1952, 1955). In the database, cognacy coding is used for Swadesh lists (100, 200) only. There is a rich literature on advantages and problems of the lexicostatistical method, and there are different views, e.g., on whether synonyms should be included, or if only one single lexeme per lexical concept is allowed in a language, how to treat semantic matches, and how to define cognacy precisely (Chang, Cathcart, Hall, & Garrett, 2015). We stick to a relatively traditional lexicostatistical method, which means that we keep cognacy within the semantic field of the lexical concept, we exclude loans, but we allow for more than just one lexeme per language, if they represent the targeted slot of the lexical concept. The coded cognacy is entirely flat: we do not build etymological trees with Swadesh vocabulary data. All lexemes of a cognacy tree are, on equal terms, drawn back to a node, which is either a reconstructed form of a proto-language (e.g., Proto-Indo-European), or a Stub, which we define as Stub Swadesh …, followed by the name of a language family, containing empty labels, e.g., egg-1, egg-1, eat-0, fingernail-1, with the Swadesh-term and the cognate number. These Stub Swadesh languages (which can be reached via the tab Language > Language tree > Stub languages), represent empty nodes at the bottom of Swadesh lists, where no reliable reconstruction is to be found in the literature.

Cognacy coded lists, e.g., Swadesh lists, are connected to their word lists only in attested languages, never at reconstructed states. For instructions how to download and use these lists, see §4.

§1.2.2. Etymology coding (culture lists)

For the so-called culture lists of our database, we are introducing a different coding system, which rather reflects a historical-comparative etymological than a lexicostatistical model. This model is different from the cognacy coding, described in previous chapter, in several aspects, and it is likely that any analysis using these datasets will yield a different outcome as compared to the cognacy coded sets. However, it always is possible, by means of filtering, to reduce the datasets with etymology coding to cognacy coding (see §4). The culture data sets make full use of the etymology controller tool of the dataset, which is described more carefully in §2.3. Basically, the etymology coding is based on core concepts in combination with etymological trees, which include all changes, including meaning change, lexical derivation etc, that occurs in etymological trees as long as the meanings mainly stay within the semantic domain of the targeted core concept. Like in etymological dictionaries, a reconstruction at the bottom of a tree is often a verbal root, but compared to dictionaries such as IEW (Pokorny, 1994), not all derivations of a root are part of an etymological tree. Included lexemes embrace only branches that pertain to core concepts (which may be several in a tree). This implies that if the core concepts is BULL, the etymological trees and branches attached to this core concept are those in which a substantial part of the lexemes of the tree have the core meaning BULL. If there is a meaning change in a language, which is not accompanied by a morphological derivation, the lexeme is still included, both to the etymological tree as well as to the core concept BULL (the meaning change is of course reflected in the Meaning field of the lexeme). On the other hand, if there is a morphological derivation of a lexeme (e.g., from another root or lexeme), but for which the meaning is BULL, the lexeme is also included. All types of occurring relations, derivation, borrowing, inheritance, or uncertain origin, can be mirrored through the etymological controller tool (§2.3.). The result of this coding model are conglomerates of etymological trees, clouing around core concepts, e.g., concepts targeting prototype meanings of high age and cultural salience (§4.1.), including smaller and larger semantic deviations as well as etymological and semantic links to other core concepts. We have selected this model for a purpose: a reduced lexicostatistical dataset can always be retrieved out of these conglomerates, but these conglomerates,
which more carefully reflect etymologies retrieved by comparative method, can never be filtered out of a lexicostatistical set.

§2. Tables and relations of the Lexicology subsection

§2.1. Lexemes: core of the subsection Lexicology

§2.1.1 What counts as a lexeme?

Lexemes constitute the core of the DiACL subsection Lexicology (see fig. 3). Lexemes are given for both attested (contemporary and historical) and reconstructed languages (see fig. 1). In the case of reconstructed languages, lexemes are given with an asterisk (*), as usual in comparative linguistic literature. By definition, a Lexeme equals to a *cognate* but in a specific language, meaning that lexemes may have different variant forms, such as variants in spelling, phonemic structure, or with allomorphs. If a lexeme differs in morphological derivation (but potentially has the same meaning), then it is a different lexeme.

§2.1.2 Organization of the Lexeme table

The core of the Lexeme table is the **Transcription** field, which gives the transcribed form of the lexeme in Latin script, adapting an orthographic policy which is described under §2.1.3. Next follows a field **Script**, which yields various native writing systems, such as Georgian or Cyrillic script. Further, there is a possibility to render the **IPA** transcription of a lexeme (this field is at current state not filled for any language). The **Meaning** field targets the full meaning of a lexeme, not just the connected lexical concept (Word List Item, e.g., *HEART, BULL*). In this field, synchronic colexification can be accounted for (diachronic meaning change is accounted for in a different way, see below). The following field **Meaning note** gives information connected to the meaning of the lexeme. Thereupon, a field for **Grammatical data** is given. This field typically gives information about inflection/conjugation of the lexeme, such as the gender of nouns. Finally, a field **Note** gives a possibility to add relevant data, which does not fit into any other field. This field may contain discussions both concerning the cognacy status of the lexeme, such as various etymologies, loan status (for instance if not fully implemented in the etymological tree hierarchy), or discussions on the form or use of the lexeme itself. Following an overarching principle of the entire database (also including the typological section), a lexeme has to be sourced, either in a literary source (dictionary, paper), or in a data set retrieved from a native speaker. These two types of sources are distinguished in the source section (Literary source vs. Informants).

§2.1.3. Policies for orthography, base form of lemmas, and hyphenation

Data of the lexicographic subsection of DiACL have been compiled from multiple sources, dictionaries, unpublished material, and new or earlier fieldwork. Our aim has been to use an orthography of the Transcription field, which meets an international scientific standard, is readable to native speakers, but which is still consistent both language-externally as well as, if possible, cross-linguistically. This is not a trivial task, in particular in cases where there are conflicting orthographies or in cases where there are no available consensus for a standard Latin transcription. This is the situation with lesser researched areas where there are native writing systems, and/or most scientific literature in non-Latin script (Cyrillic, Georgian), such as the Caucasian area. However, it is also a problem in areas where there are no previous standardized writing systems, such as in South America. Further, it is also a problem in areas where there is a rich scientific literature, such as for (Indo-European or other) reconstructed languages, or philological transcriptions and transliterations of doculects, such as Sanskrit or Avestan, but where the orthographic systems are conflicting or related to different scholarly disciplines.

An ultimate constraint to the selection of orthography has been the presence of non-combined Unicode characters. A database such as ours, which aims at making data available from an interface using any standard web browser, including downloading of data into formats such as JSON and XML for further migration into other programs, is entirely dependent on non-combined Unicode characters. The currently available set of Latin non-combined Unicode characters basically covers our demands, but we
have frequently been urged to make orthographic selections related to the availability of non-combined Unicode characters. For instance for reconstructed Indo-European, we have selected the system of using *w/*y instead of *u̯/*i̯ which are not available as superscript, non-composed characters. However, in a couple of instances, we have been forced to use a combination of characters to form characters with diacritic marks, which are not available as non-composed Unicode characters. In these cases, the principle has been, consistently, to use two characters, where the diacritic mark follows the character.

Another issue is the policy for rendering the base form of lemmata in the Transcription field. This policy is also connected to the policy of hyphenation. Beginning with nouns, the policy is to render the nominative singular form, or in the case of languages that lack a form for nominative singular, the morphological bare stem. In cases which are supposed to be (or exist only) in plural or collective, we use the plural/collective nominative. This is also the case for adjectives, where we, in case of a three-gender system, use the masculine nominative singular form. For verbs, we normally give the infinitive form, and, in the Meaning field, the translation is rendered as ‘to …’. Here, we mainly follow the standard of dictionaries of various languages.

For reconstructed languages, we use a different policy, which is related to standard of comparative-historical dictionaries. We give the stem form of nouns and adjectives with a hyphen, and the root or (when appropriate) the stem form of verbs, also with an hyphen.

As for hyphenization internally with respect to different languages, a very complex issue, there is no specific standard, rather, we have selected to follow the sources and adapted the data so that it is language-internally consistent.

The most important policy is, under all circumstances, that languages are internally consistent as concerns all these policies mentioned above, orthography, rendering of lemmata in the Transcription field, and hyphenization.

§2.2. Word Lists: functional hierarchies of lexical concepts
Lexical data of DiACL/ Lexicology is organized into semantic taxonomies defined by geography, labelled Word Lists, which can be described as a system of organizing lexical concepts into functional and environmental hierarchies (§3). The hierarchical system is not implemented in basic vocabulary (Swadesh lists), which are not distinguished by geography and have a flat hierarchy (§1.2.1).

The design of the database follows a basic model where linguistic features (lexical, typological) are organized functionally into hierarchies. Basically, the main levels are corresponding between the geographic areas and language families, whereas lower levels contain a higher degree of granularity and geographic adaptation. As with typology, this is also the case for vocabularies (see fig. 4). Languages are organized geographically, by Focus Area (macro-area), of which there are currently three, Eurasia, Pacific, and South America. The level below that is Word List, which targets a specific type of list of lexical concepts, which is adapted to a geographic area. Here, the geographic adaptation can be more fine-grained than the Focus Area definition, e.g., “Culture vocabulary lists” of Focus Area “Eurasia” can be divided into, e.g., “Culture words for Indo-European”, “Culture words for Caucasus”, and “Culture words for Basque”. This gives a possibility to control and make judgements about which type of Word List is suitable for definition. The geographically adapted lists, which we label “culture lists”, aim at capturing vocabularies which demonstrate a high age, which have a high functional stability and which still reflect the dynamics of geography, ecology, and subsistence system of languages and language families (see §3).
The hierarchical functional organization of Word Lists (fig. 4) is implemented into a chain of tables in the database, as follows (see fig. 3):

**Word List.** This level specifies the culture lists by area and/or family, which are defined by Focus area, Language area (each Language belongs to a specific Focus Area) or by language family. Here, we have, e.g., Culture words for the Caucasus, Culture words for South America, Culture words for Indo-European. The definition of these sets often aims at a specific geographic area or the occurrence of joint cognates and historical convergence, and can have different levels of detail. The current culture word lists are mainly aiming at subsistence vocabulary (see further §3).

**Word List Category.** This level specifies the over-arching semantic category of lexical generic meanings, such as Astronomy, Wild Animals. This level, though general, is adapted to geographic area and subsistence system, which makes it more fine-grained and varied compared to the Semantic field definition by Concepticon (List & Cysouw, 2016), which reflects the classification by Buck (Buck, 1949).

**Word List Item.** This level gives the lexical concepts, which are, in a Word List, selected by the characteristics of the area, the relevance of the subsistence system, cultural functionality and affordance, and occurrence in reconstructed vocabulary. Here, we find lexical concepts such as AXE, PLOUGH, SEW, OX, and so forth (Carling et al., 2016). A Word List Item corresponds to a Concept in the Concepticon database (List & Cysouw, 2016). At this level, all occurring lexemes in languages of the macro-area for a specific lexical concept (Word List Item) are displayed on a map and the data can be downloaded together with the spatial information.

**Lexeme.** This level gives the lexeme itself, connected to Language, independent of relation to Word List or Etymology (see §2.1.).
§2.3. The Etymology section

Lexical data is organized by etymologies, corresponding to cognates in other lexical cognacy databases (such as CoBL, http://www.shh.mpg.de/207610/cobldatabase). However, there are substantial differences, and the database currently incorporates two different types of cognacy coding 1) Lexical cognacy coding, and 2) Etymological coding. The Etymology section and its functionalities, which are specific to the DiACL Lexeme subsection, allows a higher degree of coding granularity than normally found in lexical cognacy databases. Here, a Lexeme can be linked to any other Lexeme within the database (also of other families); either as Ancestor Lexeme or as Descendant Lexeme (see fig. 5). Then, the nature of the connection can be specified by means of 7 different definitions, labelled Etymological Reliability: Unspecified/ Inherited/ Probably borrowed/ Certainly borrowed/ Uncertain origin/ Wanderwort/ Derivation.

Figure 5. Screenshot of the Etymology Details page, which defines relations between Lexemes.

The etymological connections for Lexemes are visualized as directed graphs on the web interface, where Lexemes function as the nodes of the graphs and the etymological links are the directed edges (that point from ancestor to descendant). The graph that is displayed for a given Lexeme will follow the etymological chain upwards towards the first ancestor(s) and downwards towards the last descendant(s). Sibling Lexemes are thus not included (fig. 6). All types of relations within lexical etymologies can be accounted for by means of this system: e.g., lexemes can be traced back to an unknown precursor (in the case of uncertain etymologies), loans between proto-states can be marked, attested loans can be coded, Wanderwörter (migration words) can be marked, lexical derivations can be coded. When lexical data is extracted as an XML file, all the etymological links are included (as links between a parent Lexeme and a child Lexeme).

§2.4. The Reliability definition

As mentioned under §2.3, lexemes within etymologies can be connected by relations of different character, labelled Reliability and specified as Unspecified/ Inherited/ Probably borrowed/ Certainly borrowed/ Uncertain origin/ Wanderwort/ Derivation.

1 Note that the etymological graph may contain cycles, because etymologies proposed by different sources may be proposed for a Lexeme. This may lead to a cycle for example in the case of two Lexemes of which one is a borrowing of the other, and two conflicting (sourced) etymologies are recorded that disagree on the direction of borrowing.
borrowed/ Uncertain origin/ Wanderwort/ Derivation (cf. §4 Advice for using the Lexicology subsection of DiACL: overview of current data status).

**Unspecified** means that the etymological relation has not yet been processed within the database, an indication that there is a cognacy relation (which is not loan or alike). It is the standard in Swadesh vocabularies, where loans are not included at the current state.

**Inherited** means that there is a secured cognacy relation between an ancestor and descendant lexeme, which has not been substantially altered by morphological derivation (e.g., compounding, suffixation).

**Probably borrowed** means that it is likely that the descendant lexeme is borrowed from its ancestor lexeme.

**Certainly borrowed** means that the descendant lexeme is borrowed from its ancestor lexeme.

**Uncertain origin** means that the lexeme has some correlation to an ancestor lexeme (i.e., the similarity on form and function is too close to be mere coincidence), but the exact relation is uncertain, and a number of alternatives might be possible, such as early loan, sound symbolic change, or analogical influence from other words.

**Wanderwort** means that the word is most likely borrowed from an ancestor lexeme, but the exact source and direction of borrowing cannot be defined (cf. description below).

**Derivation** means that the lexeme has a form which is marked by morphological derivation in relation to it’s another lexeme, which may be either an ancestor, or a base lexeme of the same language. Derivation includes all kinds of derivational morphology that change the form (and function) of a lexeme, e.g., pre-, in-, and suffixation, ablaut, or compounding. The inclusion of compounding means that a specific lexeme, if composed by two or more lexical roots, can be part of two or more etymological trees.

At the frontend, where etymologies are rendered graphically as trees (figs. 6, 7), the arrows of different type and colour represents these relations (table 1).
Figure 6. Screenshot of the Etymology for the Indo-European root for BEEP

Table 1. Arrow type and colour coding for relations (Reliability) of etymological trees (figs. 6, 7) at the frontend of the database

<table>
<thead>
<tr>
<th>Dashed line</th>
<th>Unbroken line</th>
<th>Dotted line</th>
<th>Dotted/dashed line</th>
</tr>
</thead>
<tbody>
<tr>
<td>Certainly borrowed: dark blue</td>
<td>Inherited: dark blue</td>
<td>Unspecified: red</td>
<td>Uncertain origin: red</td>
</tr>
<tr>
<td>Probably borrowed: medium blue</td>
<td>Derived: medium blue</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wanderwort: light blue</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

§2.5. Policy: how to deal with etymological reliability, Wanderwörter, and macro- etymologies

The discipline of etymology, though being capable of yielding precise information about a reconstructed past and therefore highly important to comparative linguistics, is nevertheless connected to difficulties and uncertainties (cf. Mailhammer, 2014). Its aim is fundamentally diachronic, but all results of reconstructions are potentially horizontally affected by conditions of a past synchrony. In the process of reconstruction, if assuming to many uncertain factors, such as far-gone semantic change,
occurrence of sporadic sound change, and/or prehistoric language contact, the discipline of etymology may be connected to uncertainties. Our policy is to use etymology in a strict sense, meaning that:

- We put meaning and function before form in judging etymological reliability;
- We are very hesitant about accepting macro-etymologies (i.e., etymologies using capitals such as N, V for indicating “vowel of some value”, “nasal of some value”);
- We are very careful about assuming prehistoric horizontal activities, such as loans from an unknown substrate language, or early borrowing between reconstructed states of languages.
- We do not include macro-family etymologies, such as Nostratic, Dené-Caucasian, or Ural-Altaic.

However, it is sometimes necessary to point out similarities in form between identical lexical concepts, shared within a family or an area, even though the details of transfer remain unclear. To illustrate some of the problems connected to etymology, we will describe how we deal with the complex issue of Wanderwörter.

By definition, Wanderwörter are “loans that are widely distributed throughout the languages of a region” (Epps 2015). It is typical for them to have obscure and complicated origins, which makes them a difficult case for visual representation.

Figure 7. Etymology for the Wanderwort for STAR in the South American region of Rondônia. For arrow type and colour coding, see table 1.
A good example of a Wanderwort is a lexical concept STAR, with a macro-etymology ($\sim wVrV(wVrV)$) that occurs in a group of unrelated, indigenous languages in and around the Brazilian state of Rondônia, which has been pointed out by (Crevels & Van der Voort, 2008). As the etymological relationships in our database are recorded as links with a direction (from ancestor to descendant), the difficulty of establishing an origin for this Wanderwort (both form- and affinitywise) could potentially be problematic. However, we have aimed at constructing a model as versatile as possible, which can be employed to represent a range of different situations. Some of the features that are relevant for this etymology include:

- **Arrow design.** In the picture above, two different arrow types are used: one that represents the relationship ‘Inherited’ (dark blue, unbroken line) and one that represents the relationship ‘Wanderwort’ (light blue, broken line). As described before, there are seven different arrows to choose from when making etymological connections, each representing a specific kind of relationship.
- **Stub languages.** Instead of attributing a formally and genetically obscure proto-form to an actual (proto-)language in the database, there is a possibility to create stubs of lexical concepts, which can be attributed to either a proto-language or a stub language (fig. 7). This design allows for connections and visual representation (albeit simplified) even in cases with limited amounts of information.
- **Lexeme placeholders.** Lexeme placeholders serve a purpose similar to stub languages, that is, they represent something that for some reason cannot be rendered in its accurate form. Lexeme placeholders typically occur in stub languages, but they may also occur in any reconstructed language, as long as there is no reliable reconstruction available.

To summarize these features into a general remark about the etymology in question, we have a situation in which we find several languages (most of which are unrelated to one another) sharing similar lexical forms for the meaning ‘star’. Some of these languages are taxonomic sisters and have been grouped together by inheritance from an obscured proto-form. The forms are (obviously) hypothesized to stem from somewhere, but in the absence of both an ancestor lexeme and genetic affinity for the language containing it, different kinds of placeholders are used to represent the hypothesis: a stub language *Stub Culture Amazonia*, unrelated to anything else in the database, holds a stub lexeme *star-0*, representing the proto-form of the Wanderwort.

Like all visual models, the actual reality is simplified. It is, for example, not likely that the attested forms derive directly from a single proto-form as in the picture. Rather, the lexeme has spread according to a complex network of contact between the languages. The model allows us to connect the lexemes etymologically despite the fact that the proto-form is not reconstructable by comparative method. It also offers a way to connect related lexemes without conflating with etymologies of greater certainty, thanks to the arrow design (i.e., Reliability) as well as various placeholders that serve the purpose of allowing for great complexity in the representation of various forms of etymological connections.

§2.5. Data sources

As a general and over-arching principle of the database, all data points need to be sourced. There are two types of sources: **Literary sources** or **Informants**. Since the database contains data from a wide range of languages, spanning from well-known, contemporary, to lesser-known, historical and reconstructed languages, the sources of the data in the lexical database is highly varying.

Indo-European data for contemporary languages has mainly been compiled from dictionaries. It has been a policy from the beginning already, not to judge literary sources by their reliability, but to use, for all languages, reliable dictionary sources only. This is also the case for etymological connections, where there is often margin for speculation. Sources for ancient or historical languages for the Indo-European
language family have basically been taken from etymological dictionaries. The series of etymological dictionaries by Brill (Lubotsky, 2010) have been important for tracing etymologies, but for data for individual languages, a wide range of dictionaries have been used. In the database, sources are listed in alphabetical order under Literary sources.

For lesser-known or undescribed languages, lexical data is often retrieved by fieldwork or populated by native language speaking collaborators, of which there are several in the project. This concerns in particular the language areas of Caucasus and Amazonia, for which many of the languages with lexical data in the database do not have reliable or any dictionary sources. The compilation and reliability control of this lexical data is done in collaboration with linguistic research institutes, which are listed on the database under Collaborators. Further, the sources of fieldwork data are listed under Informants.

§3. Culture vocabularies
§3.1. Theoretical background

The theoretical approach behind our model for organizing lexical concepts into hierarchical, semantic taxonomies, adapted to macro-areas, is founded in late 19th and early 20th century theories on the correlation between material culture, social structure and language, known as the Wörter und Sachen-theory. Here, cultural vocabularies play an important role in investigating (pre)historical contact and change (cf. Epps 2014, Carling 2016). A pre-condition for our systematization of cultural complexes into taxonomic schemes by subsistence is provided by the ethnographic classification of cultures, e.g., (Lomax et al., 1977; Murdock, 1969, 1981), also the foundation for resources such as e-HRAF and D-PLACE (Kirby et al., 2016). Another source is the cultural matrix model, evolving through systematic work on spread of language and culture in Amazonia (Eriksen, 2011; Hill & Hornborg, 2011). In our approach, focus is on stability, borrowability and productivity in lexicon in relation to functionality and affordance of cultural artifacts and practices of systems (Carling, 2013, 2016; Carling et al., 2016). We aim at compiling culture lists that reflect stability and change both language-internally and over distinct areas. Methodologically, research on borrowability by semantic category (Hasplmath & Tadmor, 2009) is an important prerequisite to the model, which recurs both in our selection of lexical concepts as well as in our organization of lexical data in the database. As for semantic taxonomy of culture vocabulary, we use a matrix of cultural main categories, which is organized into hierarchically organized features, by geographic area/language family. At the following macro-area-adapted level in the hierarchy, culture lists contain lexical generic meanings, which are selected according to: 1) geography and environment, 2) relevance to subsistence system, 3) cultural function or affordance, 4) occurrence in reconstructed vocabularies of language families (Campbell, 2013, pp. 346,ff.; Mallory & Adams, 2006). The aim is to provide a weighed selection, representative for quantitative analysis, also across language families and linguistic areas. In §3.2 and §3.3 we will look more specifically at the targeted areas.
§3.2. Word Lists Eurasia

From the Focus area Eurasia, there are several available Word Lists: Culture words for Indo-European, Culture words for the Caucasus, and Culture words for Basque. These lists contain the same lexical concepts, with a basic list of around 100 concepts judged to be of importance to sustainability from a historical perspective.

In our selection of cultural concepts, defined as lexical concepts, we focus on cultural concepts and artefacts, which are supposed to be of importance to subsistence, also from a historical perspective. Following the classification of cultures by subsistence by (Lomax et al., 1977; Murdock, 1969),
mentioned in previous chapter, the group in focus for the culture vocabularies of Eurasia are Plow Agriculturalists (PA), Eurasia (6) (Lomax et al., 1977, p. 665ff) (see table 2), which have several subgroups with languages in our database: PA-Indian Tribal (extensive agriculture, no dairy), PA-East Asian Irrigation (irrigation agriculture, no dairy), PA-Europe and PA-Middle East (intensive farming, dairy). Important crops are grain, wheat, and corn, animal husbandry includes pigs, sheep, and bovines (most, but not all groups have dairy consumption), and the most important agricultural tool is the plough. The subsistence classification is used as a matrix for selecting lexical concepts, also taken the reconstructed vocabulary into consideration, where we are well provided with data from the Indo-European family (Mallory & Adams, 1997; Schrader, 1917).

Table 3. List of thematic groups for which lexical concepts are selected in the culture vocabulary list for Eurasia, with OCM Subject category number (http://hraf.yale.edu/).

<table>
<thead>
<tr>
<th>Thematic groups</th>
<th>OCM Subject category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domestic animals</td>
<td>230 Animal husbandry</td>
</tr>
<tr>
<td>Wild animals: predators/scavengers</td>
<td>224 Hunting and trapping</td>
</tr>
<tr>
<td>Wild animals: game</td>
<td>224 Hunting and trapping</td>
</tr>
<tr>
<td>Cultivated produce, crops</td>
<td>240 Agriculture</td>
</tr>
<tr>
<td>Food products</td>
<td>250 Food processing</td>
</tr>
<tr>
<td>Agriculture activities</td>
<td>240 Agriculture</td>
</tr>
<tr>
<td>Agricultural tools</td>
<td>240 Agriculture</td>
</tr>
<tr>
<td>Food preparation</td>
<td>250 Food processing</td>
</tr>
<tr>
<td>Metals &amp; materials</td>
<td>320 Processing of basic materials</td>
</tr>
<tr>
<td>Trees</td>
<td>320 Processing of basic materials</td>
</tr>
<tr>
<td>Warfare implements</td>
<td>720 War</td>
</tr>
<tr>
<td>Warfare: activities &amp; implements</td>
<td>720 War</td>
</tr>
<tr>
<td>Religion</td>
<td>770 Religious beliefs</td>
</tr>
<tr>
<td>Social roles</td>
<td>560 Social stratification</td>
</tr>
<tr>
<td>Seasons</td>
<td>772 Cosmology/ 221 Annual cycle</td>
</tr>
<tr>
<td>Celestial bodies</td>
<td>772 Cosmology</td>
</tr>
</tbody>
</table>

§3.3. Word Lists South America

From the Focus area South America, culture vocabulary data have been compiled from a number of languages, including both members of greater and smaller families, and isolates. The type of subsistence system we target here is Incipient producers (I), South America, and the subgroups I-Highlands & Carribean (intensive agriculture), and I-Amazon (extensive agriculture) (Lomax et al., 1977, p. 665ff.) (see table 2). Both are characterized by agricultural production of roots, vegetables, and tree crops, (mainioc, sweet potato, maize), with game as main protein source.

Lexical data from indigenous languages of South America is scarce and most languages lack reliable sources. Further, there is little, or in some cases no etymological or cognacy judgements made on the lexical material. In the database, the lexical data is currently in a half-complete shape. Many language families are not complete in number of languages, cognacy judgements have not been completed for a number of lexemes and language families, and there are reconstructed forms available for a number of proto-languages and lexemes, for which there is no lexical data available for many languages. However, considering the shape, in general, of South American lexicology, the current dataset represents a considerable step forward in the direction of completing the lacuna of lexical data for South America.
Table 4. List of thematic groups for which lexical concepts are selected in the culture vocabulary list for South America, with OCM Subject category number (http://hraf.yale.edu/)

<table>
<thead>
<tr>
<th>Thematic groups</th>
<th>OCM Subject category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wild animals</td>
<td>224 Hunting and trapping</td>
</tr>
<tr>
<td>Wild plants</td>
<td>222 Collecting</td>
</tr>
<tr>
<td>Hunting and fishing</td>
<td>224 Hunting and trapping</td>
</tr>
<tr>
<td>Agricultural produce</td>
<td>240 Agriculture</td>
</tr>
<tr>
<td>Products</td>
<td>250 Food processing</td>
</tr>
<tr>
<td>Agriculture</td>
<td>240 Agriculture</td>
</tr>
<tr>
<td>Agricultural tools</td>
<td>240 Agriculture</td>
</tr>
<tr>
<td>Food preparation</td>
<td>250 Food processing</td>
</tr>
<tr>
<td>Metals</td>
<td>320 Processing of basic materials</td>
</tr>
<tr>
<td>Trees</td>
<td>320 Processing of basic materials</td>
</tr>
<tr>
<td>Warfare</td>
<td>720 War</td>
</tr>
<tr>
<td>Artefacts</td>
<td>252 Food preparation/ 300 Adornment</td>
</tr>
<tr>
<td>Settlement</td>
<td>360 Settlement</td>
</tr>
<tr>
<td>Social roles and positions</td>
<td>560 Social stratification</td>
</tr>
<tr>
<td>Religion</td>
<td>770 Religious beliefs</td>
</tr>
<tr>
<td>Time</td>
<td>772 Cosmology</td>
</tr>
<tr>
<td>Celestial bodies</td>
<td>772 Cosmology</td>
</tr>
</tbody>
</table>

§4. Advice for using the Lexicology subsection of DiACL: overview of current data status

At current state the Lexicology subsection of DiACL contain partly compete, partly incomplete datasets. The data of the subsection is supposed to be of controlled quality, fulfilling the criterion of sourcing datapoints and only use data from controlled fieldwork (see §2.4). This means that all individual data, for languages, etymologies, etc., can be used for quoting. As concerns the status of Word Lists that can be derived by using the XML downloading functions, the situation is somewhat different.

Beginning with Swadesh data, these families have complete, cognacy-coded sets:

- Chapacura – 200
- Indo-European – 100
- Kartvelian – 200
- Nambikwara – 200
- Romani chib – 200
- Tupí – 100

These sets are in a condition that etymological cognacy judgements are satisfactory and complete enough to use the data for phylogenetic analysis. Note, however, that the datasets have to be filtered out (by using the distinction Language Family) from the derived XML file, which is not distinguished by language family (note that you can get a list of all the languages in any subtree from the Language Tree page, in XML format). Lexical cognacies are not included in the derived XML file, rather, they have to be derived from the point of departure of lexical cognacies of every connected lexeme. Code for deriving complete sets from the database are found at the Zenodo account of DiACL.

For the other language families, data is available and complete, but cognacy judgements might be lacking or might not be enough curated.

In the future, cognacy judgements of lexical datasets will be improved. The Twitter feed on the database frontend will inform when:

- Datasets have been completed and can be used for phylogenetic analysis.
- New datasets have been added to the database.

As for culture vocabularies, only the Indo-European data set is currently in a complete condition by means of cognacy judgements. Here, a large body of etymological dictionaries are available, often suggesting alternative forms for reconstruction. Our policy has been to render reconstructions as close as possible to the sources of reliable dictionaries, but to conflate, as far as possible, redundancy in reconstruction provided by different orthographic standards (see §2.1.3). However, uncertainty or different standards of etymological reconstructions may result in redundant forms and unnecessary complexity in the etymological trees, which relate to different approaches (for instance on substrate/inheritance, version of the laryngeal theory, etc.) rather than orthographic conventions. We aim, in the future, to conflate this redundancy as much as possible – which is not a trivial task, since many of the cases of redundancy have their source in real uncertainties and problems of the reconstruction, for which various dictionaries offer alternative solutions.

References


Schrader, O. (1917). Reallexikon der indogermanischen Altertumskunde. Berlin: de Gruyter ::.
