EvoSem: A database of polysemous cognate sets

Mathieu Dehouck, Alexandre François, David Kletz, Siva Kalyan, Martial Pastor

*Evo*Sem: A database of polysemous cognate sets

Mathieu Dehouck
LATTICE, CNRS–ENS-PSL–USN
mathieu.dehouck@ens.psl.eu

Alexandre François
LATTICE, CNRS–ENS-PSL–USN
alexandre.francois@ens.fr

Siva Kalyan
University of Queensland
s.kalyan@uq.edu.au

Martial Pastor
Radboud University Nijmegen
LATTICE, CNRS–ENS-PSL–USN
martial.pastor@ru.nl

David Kletz
LATTICE, CNRS–ENS-PSL–USN
david.kletz@sorbonne-nouvelle.fr

Abstract

Polysemies, or "colexifications", are of great interest in cognitive and historical linguistics, since meanings that are frequently expressed by the same lexeme are likely to be conceptually similar, and lie along a common pathway of semantic change. We argue that these types of inferences can be more reliably drawn from polysemies of cognate sets (which we call "dialexifications") than from polysemies of lexemes. After giving a precise definition of dialexification, we introduce *Evo*Sem, a cross-linguistic database of etymologies scraped from several online sources. Based on this database (publicly available at http://tiny.cc/EvoSem), we measure for each pair of senses how many cognate sets include them both—i.e. how often this pair of senses is "dialexified". This allows us to construct a weighted dialexification graph for any set of senses, indicating the conceptual and historical closeness of each pair. We also present an online interface for browsing our database, including graphs and interactive tables. We then discuss potential applications to NLP tasks and to linguistic research.

1 Introduction

Colexification is the structural pattern whereby two meanings are expressed by the same word in a given language: e.g., Spanish *pueblo* colexifies the meanings PEOPLE and VILLAGE. While polysemy is defined semasiologically, as a property of a word, colexification is defined onomasiologically, as a property of a pair of meanings. These are two sides of the same coin: if a pair of meanings is colexified, then this means they are senses of the same polysemous word.

The concept of colexification was introduced by François (2008) in the context of lexical typology, with the aim of discovering universal patterns of conceptual structure, adapting the semantic-map approach that had already proven fruitful in typological studies of grammar (Anderson, 1974; Haspelmath, 1997). Since then, a number of works have been published on the topic of colexification. Some of these suggest additional sources of data (e.g. Östling, 2016), while others look for universals in lexical semantics (e.g. Georgakopoulos et al., 2022), or try to predict patterns of colexification from properties of the meanings themselves (e.g. Xu et al., 2020; Di Natale et al., 2021; Brochhagen and Boleda, 2022; Brochhagen et al., 2023). The growing body of research into colexification has also led to the creation of the CLICS database (Rzymski et al., 2020, available at https://clics.clld.org), now in its third edition: the empirical dataset that it provides makes it possible to test hypotheses about cross-linguistic patterns of colexification. This in turn has led to recent applications in the field of NLP, with presentations at major venues such as Bao et al. (2021) (GWC2021) who question the universality of common colexifications by comparing different colexification databases, Chen and Bjerva (2023) (SIGMORPHON 2023) who use colexification to create cross-lingual resources and Chen et al. (2023) (NoDaLiDa 2023) who infuse language embeddings with semantic typology using colexification information.

While it yields some insight into universal constraints on semantic change, "strict colexification" (François 2008, 171), defined in terms of synchronic properties of lexemes, misses the semantic links that are synchronically absent, yet can be revealed by studies of etymology. Incorporating semantic change into the study of lexical typology would contribute to a growing body of research on computational approaches in this domain (e.g. Kutuzov et al., 2018; Tahmasebi et al., 2021).

This issue is addressed by the new concept of *dialexification* (François and Kalyan, 2023, in prep.), the structural pattern whereby two meanings are expressed by members of the same cognate set. For example, knowledge of regular sound change in the Indo-European family shows that Norwe-

Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change, pages 66–75

gian *gård* 'land', Gothic *gards* 'house' and Polish *gród* 'city' are all cognate, since they all descend from the same Proto-Indo-European (p-IE) etymon *g^hórd^{h}os (Mallory and Adams, 1997, 199). The historical relations that link these three concepts cannot be captured by the notion of colexification, since none of these words has more than one of these meanings; but they can be described as instances of dialexification. More specifically, we can say that the semantic pairs {LAND–HOUSE}, {LAND–CITY}, and {HOUSE–CITY} are *dialexified* by (or under) the p-IE form *g^hórd^{h}os.

If two meanings *A* and *B* are dialexified, this means that either *A* evolved into *B*, *B* evolved into *A*, or both *A* and *B* evolved from a common source. In other words, dialexification is always indicative of a historical relation between two meanings—one that may not have been captured by earlier conceptual tools.

In this paper, we present *Evo*Sem, a database and a website (http://tiny.cc/EvoSem) dedicated to the study of dialexification. It consists of etymologies and definitions scraped from the English-language Wiktionary (https://en.wiktionary.org), itself a compilation of earlier scholarly work from various sources; as well as the *Austronesian Comparative Dictionary* or ACD (Blust and Trussel, 2013; Blust et al., 2023) and the *Sino-Tibetan Etymological Dictionary and Thesaurus* or STEDT (Matisoff, 2016). Among other features, *Evo*Sem allows us to measure how often any given pair of meanings is dialexified across the world's languages.

This paper is organized as follows. Section 2 will define the notion of dialexification mathematically, and contrast it with colexification. Section 3 will describe the process of data collection and post-processing. Section 4 will discuss the visualization of the data on our companion website. Finally, section 5 will discuss potential applications of *Evo*Sem to NLP tasks and to linguistic research.

2 Definitions

Let $\mathcal{L} = \{l_1, \ldots, l_n\}$ be a set of languages. For each language $l \in \mathcal{L}$, we have a vocabulary $\mathcal{V}_l = \{w_1, \ldots, w_{|\mathcal{V}_l|}\}$ of words and/or morphemes. Let $C(w)$ be the set of meanings (also called concepts or glosses) of the word $w \in \mathcal{V}_l$. Furthermore, let $e = a(w)$ be the earliest known ancestor (the "etymon") of w. We say that w is a *reflex* of e, and we call the set of all reflexes of e the *cognate set* to which w belongs.

Two concepts c_i and c_j are said to be "dialexified"—which we represent as "$\delta(c_i, c_j)$"—if there exist two words w_p and w_q such that w_p expresses c_i, w_q expresses c_j, and w_p and w_q are cognate (i.e., have the same etymon). Mathematically, dialexification is a symmetric and reflexive relation that can be formally defined as follows:

$$\forall c_i, c_j,\ \delta(c_i, c_j) \iff \exists w_p, w_q : a(w_p) = a(w_q) \wedge c_i \in C(w_p) \wedge c_j \in C(w_q).$$

As for colexification, it corresponds to the situation where w_p and w_q are the same; in this case, w_p and w_q are obviously cognate, since they necessarily descend from the same etymon. We write the colexification relation as $\kappa(c_i, c_j)$, and define it as follows:

$$\forall c_i, c_j,\ \kappa(c_i, c_j) \iff \exists w : c_i \in C(w) \wedge c_j \in C(w).$$

Like dialexification, this relation is symmetric and reflexive. Also, $\kappa(c_i, c_j) \Rightarrow \delta(c_i, c_j)$ for all c_i and c_j: in other words, any relation of colexification is also a relation of dialexification, though the converse is not true.

Note that the etymon of a given word is not always attested: it may be a proto-form reconstructed using the comparative method (Weiss, 2015). Its exact form may thus be uncertain; but this does not affect our ability to identify cases of dialexification, since all that matters for the definition is whether two words have the *same* etymon, i.e. belong to the same cognate set.

To put it another way, the domain of dialexification is not, strictly speaking, the etymon itself, but rather the cognate set that descends from the etymon. Thus, to say that a given etymon dialexifies concepts *A* and *B* is not a direct claim about the semantics of the original etymon: it is simply a statement about the meanings of its descendants. Strictly speaking, we could have defined dialexification in terms of cognate sets. But since it is more convenient to refer to a cognate set by its etymon than to list out all the cognate forms (and since there is a one-to-one correspondence between etyma and the cognate sets that descend from them), we prefer the etymon-based definition.

We make a distinction between a *root*, which is the minimal unit of historical reconstruction

(e.g. p-IE *g^herd^h- 'enclose'), and an *etymon*, i.e. a proto-form that is morphologically derived from a root: e.g., the nouns *$g^hórd^h$-*os* and *$g^hṛd^h$-*ós* (both glossed 'enclosure') are two distinct etyma derived from the root *g^herd^h-. Strictly defined, relationships of dialexification are always assessed at the level of the etymon rather than its root.[1]

Note that "cognate sets", as we define them in this paper, include not only direct descendants of etyma, but also borrowings. For example, the cognate set that descends from p-IE *$g^hórd^hos$ includes not only Russian *gorod* 'city', but also Yakut (Turkic) *kuorat* 'city', which is borrowed from the Russian word. This differs from the way cognate sets are usually defined in historical linguistics (i.e. excluding borrowings); however, we see no principled reason to distinguish between semantic changes that affect borrowed forms and those that affect inherited forms, and so this distinction is not relevant for defining dialexification. Regardless, we retain information about the borrowed status of lexemes in our database, to allow for future analyses that are sensitive to this distinction.

3 Data collection

We now describe how we went about assembling the *Ɛvo*Sem dataset.

3.1 Wiktionary

The bulk of our data comes from the English Wiktionary (https://en.wiktionary.org). Due to differences in the way different language families are organised on Wiktionary, we used slightly different procedures for extracting data for Indo-European; Semitic and Uralic; and all the remaining language families represented on Wiktionary, especially in terms of how we identified lemmas in the respective proto-languages. We describe these procedures in turn.[2]

3.1.1 Indo-European

Initially, we started with pages from the category "Proto-Indo-European roots" (653 entries on https://en.wiktionary.org/wiki/Category:Proto-Indo-European_roots). On each page, we looked for the section titled "Derived terms", and extracted every etymon derived from this root (e.g. the etyma *$g^hórd^h$-*os*, *$g^hṛd^h$-*yé-ti*, listed under the p-IE root *g^herd^h-).

We then proceeded to extract entire cognate sets, by listing every reflex of each etymon – e.g. Albanian *gardh* from *$g^hórd^h$-*os*, or Proto-Germanic (p-Gmc) **gurdijanǫ* from *$g^hṛd^h$-*yé-ti*. In the latter example, the reflex was itself a form from a proto-language (p-Gmc), the source of further reflexes. In such cases, the "Descendants" section of the relevant page was also scraped to yield further reflexes (e.g. Old English *gyrdan* and English *gird*, under p-Gmc **gurdijanǫ*). Descendants were crawled recursively until no more reflexes could be added. At every stage, relations of borrowing were noted (even though in our analyses, we do not treat borrowings separately from inherited forms).

While many p-IE lemmas in Wiktionary derive from a p-IE root, some are underived forms – i.e. morphologically simple rather than derived from a root (e.g. **oḱtṓw* 'eight'). We thus applied a similar scraping procedure to all underived p-IE lemmas to extract all of their reflexes.[3]

For each reflex of a given etymon, we then extracted all of its senses. The particular format of Wiktionary made it possible to design an approach based on the hyperlinks that usually appear in the (English-language) definitions of all entries. For example, Russian *grad* is defined as

> (*poetic*, *archaic*) town, city, used as a common city name suffix (Volgograd, Kaliningrad, Leningrad)

(where underlining indicates hyperlinks). We removed all parenthetical comments, and then extracted every hyperlinked word, with the idea that they would usually correspond to suitable English glosses;[4] in our example, this yielded a set of simple glosses {*town* | *city*}. Reducing the senses to

[1] In practice, the distinction between root and etymon only applies to Proto-Indo-European and Proto-Semitic, since for other proto-languages, the proto-forms listed in our sources are morphologically simple.

[2] All scraping of Wiktionary was done in R, using the `xml2` package (Wickham et al., 2023).

[3] In cases where the same reflex appeared under both a p-IE root and an underived p-IE lemma, only the entry with the root was kept.

[4] A limitation of this approach is that our use of English lemmas as glosses makes it hard to detect cases where a language distinguishes between two senses that are colexified in English: for example, German distinguishes between *kennen* 'to be acquainted with' and *wissen* 'to be aware of', but these are both glossed as *know* in Wiktionary. Ideally, we would be able to gloss the items in our database with WordNet synsets (Miller, 1995), for better granularity; but we are not aware of a reliable way to automate the matching of free-form definitions with synsets. We are grateful to an anonymous reviewer for highlighting this limitation, and acknowledge that it partially compromises the onomasiological perspective that motivates this work.

(mostly single-word) glosses would then make the meanings of different words easy to compare across languages—the very purpose of *Evo*Sem.

However, this hyperlink-based approach led to a couple of difficulties. Firstly, the words that are hyperlinked in a given Wiktionary definition often include not only the key words in the definition, but also auxiliary words such as *be* or *become*; this was addressed by excluding stopwords (using the list built in to the `stopwords` package in R, Benoit et al. 2021), unless the *only* hyperlinked words are stopwords (so as to not exclude words whose meaning is 'to be', etc.).

Secondly, while the definitions of non-English words tend to be succinct, and only contain hyperlinks to direct translations of the word being defined, the definitions of English words tend to be verbose, and contain hyperlinks to a wide variety of related concepts, running the risk of collecting noisy data. For example, the English word *gird* is defined as

1. (*transitive*) To bind with a flexible rope or cord.
2. (*transitive*) To encircle with, or as if with a belt.
3. (*transitive*, *reflexive*) To prepare (oneself) for an action.

Clearly, the hyperlinked words include both acceptable glosses (*bind*, *encircle*, *prepare*) and words that are only thematically related to the word being defined (e.g. *flexible*, *rope*, *belt*, *action*). In the absence of a reliable way to distinguish between the two types of links, we addressed the problem by forcing the gloss of every English word to be identical to the word itself (so that *gird* would only be glossed as *gird*). This meant erasing all polysemies in English; but we found this to be an acceptable alternative to an otherwise noisy dataset.[5]

Another problem we encountered is that many languages have homographs, i.e. lexemes with the same spelling but different etymologies: each of these homographs derives from a different etymon, and covers a different set of senses. For example, the Dutch word *vorst* means 'prince', 'frost', 'forest', and 'ridgepole'; but each of these meanings derives from a different etymon (p-IE **pṛh₂-is-*, **prustós*, **pṛkʷ-éw-s*, and **perst-*, respectively). When extracting the reflexes of a given etymon, there was no easy way to ensure that in cases like this, only the meanings corresponding to the correct etymon would be returned, and instead all meanings of the word were extracted, regardless of the etymology. (Thus, for example, *vorst* meaning 'forest' was initially listed under p-IE **pṛh₂-is-* as well as **pṛkʷ-éw-s*.) To remedy this, we ran a separate deduplication step, where for every word definition that appeared under multiple etyma (e.g. *vorst* meaning 'forest'), we searched the wikitext of the etymology for the `{{inh}}` ("inherited") and `{{der}}` ("derived") templates, to find the oldest mentioned ancestral form (in this case, p-West Germanic **furhiþi*), and then recursively searched for ancestors of this form until we arrived at a p-IE etymon (**pṛkʷ-éw-s*); this allowed us to eliminate cases where a definition of a word was listed under the wrong etymon.

In addition to extracting reflexes of p-IE roots and underived lemmas, we also extracted reflexes of proto-forms from each first-order descendant of p-IE (Proto-Germanic, Proto-Indo-Iranian, etc.), wherever these proto-forms are not themselves known to be descended from p-IE forms.

3.1.2 Semitic and Uralic

The same procedure that we applied to underived lemmas in p-IE was also applied to lemmas in Proto-Semitic and Proto-Uralic. We also deduplicated homographs in the same way.

For Semitic, we were able to have a domain expert (Chams Bernard) check the data manually, to correct errors in the extraction of glosses, and ensure that the etymologies reflect the state of the art in Semitic historical linguistics. We plan to also have the Indo-European and Uralic data manually checked by experts.

3.1.3 Other language families

To extract data from language families other than Indo-European, Semitic and Uralic, we first located all subcategories of "Lemmas by language" that have the form "Proto-[family] lemmas", exclud-

[5] In any case, English is just one of the 1,941 languages in our dataset, and accounts for only 2% of lemmas (though it is the most heavily-represented language in our dataset). An anonymous reviewer asks whether excluding English polysemies could lead to mis- or underidentification of cognate sets; this is not the case, as cognacy relations are determined purely by shared descent from a proto-form, which is not affected by our ability to accurately extract glosses from the definitions. Moreover, it does not introduce ambiguities into our results, beyond those that are inherent to the use of English lemmas as glosses. At some point, the glossing algorithm developed for non-Wiktionary sources (such as 3.2) could easily be applied to Wiktionary definitions as well, allowing us to recover a number of English polysemies; we plan to do this in future iterations.

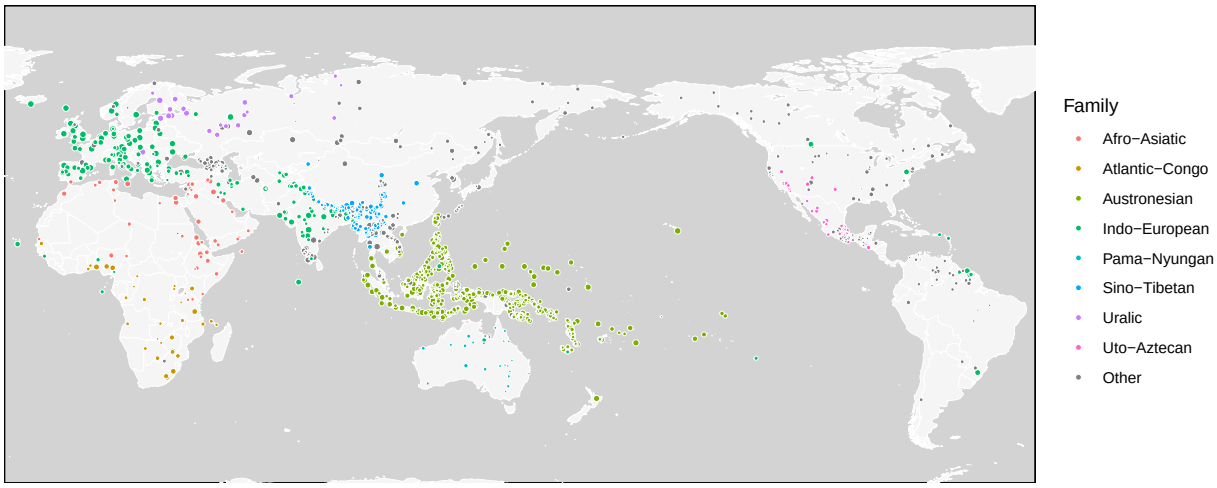

Figure 1: Geographic distribution of those languages covered by *ℰvo*Sem for which metadata is available from the Glottolog reference database of the world's languages (Nordhoff and Hammarström, 2012). Languages are color-coded by language family, and their size is proportional to $\log_{10}$ of the number of dialexifications that involve each language.

ing Proto-Indo-European, Proto-Semitic, Proto-Uralic, and all their descendant proto-languages. For each remaining proto-language, we then applied the same scraping procedure as for p-IE underived lemmas. Finally, whenever the same reflex was listed under multiple proto-languages at different hierarchical levels of the same language family (e.g. Proto-Austronesian and Proto-Malayo-Polynesian), only the entry under the highest-level proto-language was retained.

3.2 ACD (Austronesian) and STEDT (Sino-Tibetan)

We primarily drew on Wiktionary, since it is a rich and reliable resource for several language families – notably Indo-European – and generally provides references to published etymological research. Another reason for using it is ease of access, particularly since for many language families, the only other available etymological resources are printed publications. However, *ℰvo*Sem also incorporates data from other electronic sources when these are available. We thus added to our dataset two etymological resources we judged to be reliable: ACD and STEDT (see §1). (Other databases will be added to this list in the future, for other families.)

The ACD and STEDT databases were harvested using a web crawler that went through an index of their etyma. Each etymon was in turn associated with a cognate set, a list of reflexes for which our crawler collected all relevant details (language; family; form of the reflex; etymon; definitions).

Because these resources do not include hyperlinked words in their definitions like Wiktionary does, we developed a different parsing algorithm; in most cases, it proved able to convert wordy definitions into acceptable glosses. Our starting point was definitions such as the following description given by the ACD for the verb *a-kan* in the Kadazan Dusun language of Malaysia: {*to eat, consume, wear away; (in such games as chess) take a piece, destroy (as if by eating)*}.

Our first step is to identify key separators in the overall definition (e.g. ',' or ';'), so as to parse the text into separate glosses. Then all potential glosses are run through a regular expression that cleans out all non-essential lexicographic indications, such as content in parentheses, usage notes or special abbreviations. Likewise, we ignore certain stopwords at the beginning of a string, such as the article *a(n)* before nouns ('*a spoon*' → '*spoon*'), or the particle *to* before infinitives ('*to eat*' → '*eat*'). In the example of *a-kan* above, these first steps yield a set of separate strings, namely {*eat* | *consume* | *wear away* | *take a piece* | *destroy*}.

Next, a different parser attempts to isolate potential concepts for every clear gloss. In order to make sure that the glosses extracted from these databases are comparable to those extracted from Wiktionary, our parser matches every parsed string with the lemmas listed under `https://en.wiktionary.org/wiki/Category:English_lemmas` (which contains 729,370 entries). This matching operation recognizes '*eat*' and '*wear away*' as valid glosses,

but not '*take a piece*', which is not listed as an English lemma, and is thus eliminated from our results. This process allows us to filter many verbose definitions into a simple set of lemmatized glosses: e.g., the definition '*to open, as the fist or a book; to spread out, as a folded paper or mat*' is correctly parsed as {*open* | *spread out*}, leaving out the noise from other strings.

While this filtering script gave satisfying results, we noted that certain English words were not correctly identified, due to being inflected. For instance, conjugated verbs or plural forms like *children* would go unrecognized, as they do not correspond exactly to an English lemma in Wiktionary (unlike uninflected forms such as *child*, which do count as lemmas). Since we judged that such glosses ought to be retained rather than deleted altogether, we chose to accept them as well, as long as they belonged to a reference list of English non-lemmas (444,072 entries from https://en.wiktionary.org/wiki/Category:English_non-lemma_forms).[6]

Finally, some additional rules were made necessary by the different typological profile of certain language families. It is a well-known observation that parts of speech differ cross-linguistically (Croft, 2005); e.g. in various language families, *adjectives* tend to behave like a sub-class of verbs (Dixon, 2004; Van Lier, 2016). As a corollary, many dictionary authors choose to gloss property words as if they were verbs, with such definitions as '*be small*' (static reading), '*become small*' (dynamic reading), or even '*to be or become small*'. In order to make glosses compatible across language families, we decided to suppress these copulas: as a result, '*to be or become small*' (along with all possible variations thereof) is now correctly converted into a simple gloss {*small*}.

3.3 Summary of data

Table 1 summarizes the amount and diversity of data we were able to extract from each data source. Figure 1 shows the geographic distribution of languages, with dots colored according to language family and sized by how many dialexifications involve each language.

	Wiktion.	ACD	STEDT	combined
Languages	1,537	461	227	1,941
Families	55	6	5	58
Proto-lang.	91	9	19	115
Etyma	9,471	7,279	1,777	18,527
Reflexes	95,840	55,208	18,936	169,256
Meanings	26,822	13,569	3,327	31,143

Table 1: Summary statistics for each data source in current *Evo*Sem, as well as the combined dataset. Note that the statistics for the individual data sources do not always add up to the values in the "Combined" column, due to overlap in coverage between sources. The reason why ACD and STEDT cover more than one family each is that they both contain borrowings from Austronesian or Sino-Tibetan into other language families. The proto-languages covered by ACD include not only Proto-Austronesian, but also a number of proto-languages descended from it; likewise, the proto-languages covered by STEDT include not only Proto-Sino-Tibetan, but also a number of proto-languages descended from it, and Proto-Indo-Aryan.

4 Visualization

In this section, we present the tools provided on the *Evo*Sem website for exploring the database.

4.1 Dialexification graphs

From the collected data, we generate a weighted dialexification graph $G = (V, E)$ where $V = \{c \mid \exists c' : \delta(c, c')\}$ is the set of semantic concepts that participate in at least one dialexification, and $E \subset V \times V$ is the set of weighted dialexification relations, such that $(c_1, c_2) \in E \iff \delta(c_1, c_2)$.

The weight of an edge (c_1, c_2) is equal to the number of etyma (or cognate sets) that dialexify that pair of concepts:[7]

$$\begin{aligned} w(c_1, c_2) &= |\{e \mid \delta_e(c_1, c_2)\}| \\ &= |\{e \mid \exists w_1, w_2 : e = a(w_1) = a(w_2) \\ &\qquad \wedge c_1 \in C(w_1) \wedge c_2 \in C(w_2)\}|. \end{aligned}$$

The graph G represents all the dialexification relations between concepts in our database. Because it currently contains tens of thousands of concepts and more than a million edges, it is impossible to represent visually in its entirety. Instead, we propose to display subgraphs based on specific subsets of the concept set.

[6] A reviewer asked why we did not use a lemmatizer to address this issue. The main reason is that we did not want to lose the information conveyed by the inflections; since most dictionary definitions present the key defining words in an uninflected form, the use of an inflected form is likely to carry crucial information about the semantics of the word being defined, e.g. the fact that Italian *prole*, glossed as 'children', is in fact a collective noun (whose meanings also include 'offspring' and 'progeny').

[7] Given a pair of concepts, the weight of its edge is also called *dialexification score*, or *delta* score.

Given the definition of dialexification, a possible way to restrict the graph is to only select the concepts that are lexified by a given cognate set, as defined by an etymon e in language l_p: $G_e = (V_e, E_e)$ with $V_e = \{c \mid \exists w : c \in C(w) \wedge a(w) = e\}$, and $E_e = E \cap (V_e \times V_e)$. We call G_e the *etymograph* of e. Fig. 2 shows part of the etymograph of the etymon **deḱs(i)wós* in Proto-Indo-European.

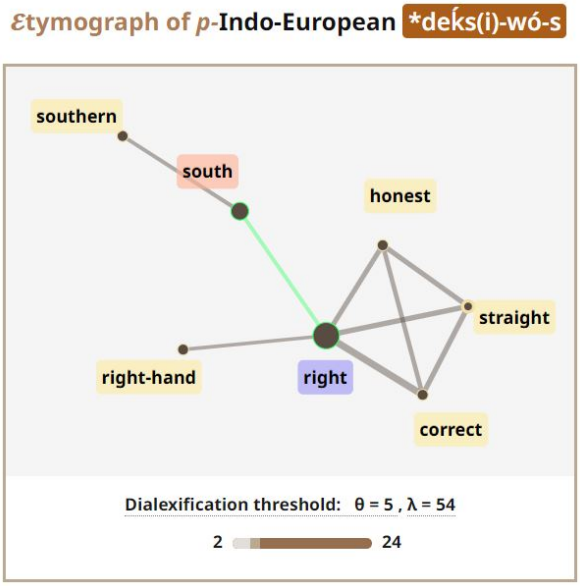

Figure 2: *Etymograph* of p-IE **deḱs(i)wós*, showing the concepts that are dialexified among its descendants. The highlighted edge indicates a pair of concepts {RIGHT–SOUTH} that is dialexified (according to *ℰvo*Sem) by 6 distinct etyma from four different families: its dialexification score is $\delta = 6$. The thickness of each edge reflects logarithmically the δ score of the concept pair for *ℰvo*Sem as a whole; for this etymon, the value of δ ranges from 2 to 24 (24 being the δ score of {RIGHT–CORRECT}. The current view has a threshold $\theta = 5$, i.e. it selects only those links whose dialexification score is $\delta \geq 5$; for this etymon, the number of dialex links displayed for $\theta = 5$ is $\lambda = 54$. As for the size of each vertex, it reflects the distribution of different concepts across the descendants of this etymon **deḱs(i)wós*: e.g. 5 of its reflexes have the sense RIGHT, but only one means HONEST.

Note that the weights of the edges are independent of the choice of etymon, since they are computed from the entire *ℰvo*Sem database.

Since there is one etymograph for each etymon, there are tens of thousands of etymographs in *ℰvo*Sem (see Table 1). However, each etymograph is of a limited size: the largest one (Proto-Austronesian **maCa*) has 292 nodes, and the median number of nodes in *ℰvo*Sem is 5. This makes etymographs much easier to view than the entire dialexification graph.

Because all pairs of concepts in an etymograph G_e are dialexified by e, an etymograph is, by definition, a clique. For this reason, we choose not to represent edges of weight $\delta = 1$. More generally, not displaying edges of weight 1 tends to reduce the noise resulting from faulty gloss extraction.

Even with the exclusion of edges of weight 1, some etymographs still have too many edges for easy visualization. To improve legibility, we offer the user the ability to set the weight threshold θ, so as to reduce the number of edges displayed. For example, Figure 2 shows the etymograph of the p-IE etymon **deḱs(i)wós* with $\theta = 5$. Decreasing the threshold brings more senses into view; increasing it reduces the number of nodes displayed.

From a technical standpoint, we store the information necessary to build each etymograph (concepts, reflexes, glosses, links to external sources) in a dedicated JSON file. When the user opens the etymon's dedicated page, an SVG representation of the etymograph is generated using the D3 Javascript library (Bostock, 2012) for computing the vertex layout. The user can then explore the graph, either by interacting with it directly, or by browsing the tables presenting the underlying data.

4.2 Data tables

The data related to an etymograph and its cognate set are presented in three tables: the *etymon-to-concepts* (E2C) *table*, the *concept-to-etyma* (C2E) *table*, and the *dialexification table*.

The *etymon-to-concepts table*, which appears directly alongside the etymograph, lists all the concepts lexified by at least one member of the cognate set. For each concept, a collapsible list of the relevant reflexes is provided (see Fig. 3).

Clicking on a concept cell, or clicking on the concept label directly on the graph, selects the given concept and opens the corresponding C2E table. The table ranks concepts by their frequency of attestation among reflexes; this is shown by the number in the last column, and by the node size in the graph (see Fig. 2). When a concept is selected, its row changes colors, and the concepts that are not dialexified with it at least θ times have their rows grayed out (see Fig. 3). On the graph, this is also reflected by a color change of the edges incident to the corresponding vertex.

The *concept-to-etyma table* lists all the etyma

▼ 19 meanings dialexified by *deḱs(i)-wó-s

MEANINGS	▶ REFLEXES		#
right	▼ language data		5
	Cornish	dyghow	
	Greek	δεξιός - dexiós	
	Irish	deas	
	Old High German	zeso	
	Old Irish	dess	
right-handed	▶ language data		2
south	▼ language data		2
	Irish	deas	
	Old Irish	dess	
able	▶ language data		1
close	▶ language data		1
convenient	▶ language data		1
correct	▶ language data		1
courteous	▶ language data		1
fortunate	▶ language data		1
honest	▶ language data		1

Figure 3: *Etymon-to-concepts table* for the p-IE etymon **deḱs(i)wós*, corresponding to the graph in Fig. 2. The table shows the first 10 of the 19 meanings dialexified by its descendants: RIGHT, SOUTH, CORRECT, etc. Clicking on the concept RIGHT has turned the row to blue (Concept_1). The rows in white show senses (e.g. CORRECT) that are dialexified with that Concept_1 at least θ times (here, $\theta = 5$); those that are dialexified fewer times appear grayed out. The sense SOUTH was selected as Concept_2, and thus appears in red. The collapsible lists for both selected senses are seen unfolded; they show that while Irish *deas* means both RIGHT and SOUTH, Greek *dexiós* only means RIGHT.

79 sources of concept *right*

FAMILY	ETYMON	▶ REFLEXES	
Indo-European	*deḱs(i)-wó-s	▸ language data	
Afroasiatic	*yamin-	▾ language data	
		Akkadian	𒅆𒈾 - imnum
		Arabic	يَمِين - yamīn
		Aramaic	יְמִינָא - yammīnā
		Hebrew	יָמִין - yamín, yāmîn
		Maltese	lemin
		Ugaritic	𐎘𐎌𐎍 - ymn
Algonquian	*keʔči-	▾ language data	
		Ojibwe	gichi-
Austronesian	*waNan	▸ language data	
Bantu	*-dɔ̀nga	▸ language data	
Germanic	*garihtiją	▾ language data	
		Old English	ġerihte
Tibeto-Burman	*g-(yr)a	▸ language data	
Dravidian	*wal	▾ language data	
		Telugu	వేల - vala
Indo-European	*deḱs(i)-teró-s	▸ language data	
Indo-European	*déḱs(i)-no-s	▾ language data	
		Bengali	দক্ষিণ - dokkhin
		Macedonian	дéсен - désen
		Old Church Slavonic	деснъ - desnŭ
		Pali	dakkhiṇa
		Sanskrit	दक्षिण - dákṣina
Indo-European	*h₃reǵ-tó-s	▸ language data	

Figure 4: *Concept-to-etyma table* for the concept RIGHT, opened by selecting that sense in the E2C table of p-IE **deḱs(i)wós* (blue row in Fig. 3). 79 etyma include reflexes that lexify the concept RIGHT, of which 11 are shown here. When unfolded, the reflex lists cite only those reflexes that have the target meaning.

that dialexify[8] the selected concept: e.g. the C2E table corresponding to RIGHT in Fig. 3 is given in Fig. 4. For each etymon, the C2E table provides the name of the language family to whose proto-language the etymon belongs; a link to the main source of data for that etymon; and a collapsible list of reflexes that lexify the concept of interest.

Clicking on a second (non-grayed out) concept has the effect of selecting the dialexification relation holding between Concept_1 (in blue) and Concept_2 (in red). Alternatively, the user can directly click on an edge of the graph. Selecting a dialexification edge replaces the C2E table with a new *dialexification table*: see Fig. 5.

The dialexification table lists all etyma that dialexify the pair of selected concepts. It works as if by combining together two C2E tables. The collapsible list of reflexes is now sorted and colored to reflect which concept is lexified by which reflex.

The top-most elements of the list (on a blue background) lexify only Concept_1, while the bottom-most elements (on a red background) lexify only Concept_2. The elements in the middle of the list, on a two-color striped background, are reflexes that colexify the two concepts.

It is always possible that some part of the list may be empty: e.g. in Fig. 5, no reflex of **deḱs(i)wós* means only SOUTH (red background). When a cognate set has no reflex that colexifies Concept_1 and Concept_2 together, one can speak of "pure dialexification". While the more common configuration is to find both dialex and colex in the same cognate set, cases of *pure dialexification* do occur.

5 Applications

*Evo*Sem allows us to observe historical connections between meanings that would be missed if we were to limit ourselves to looking at colexifications. For example, the meanings CHEST and

[8] Saying that an etymon dialexifies a concept implicitly means "with some other concept", since dialexification is a binary relation.

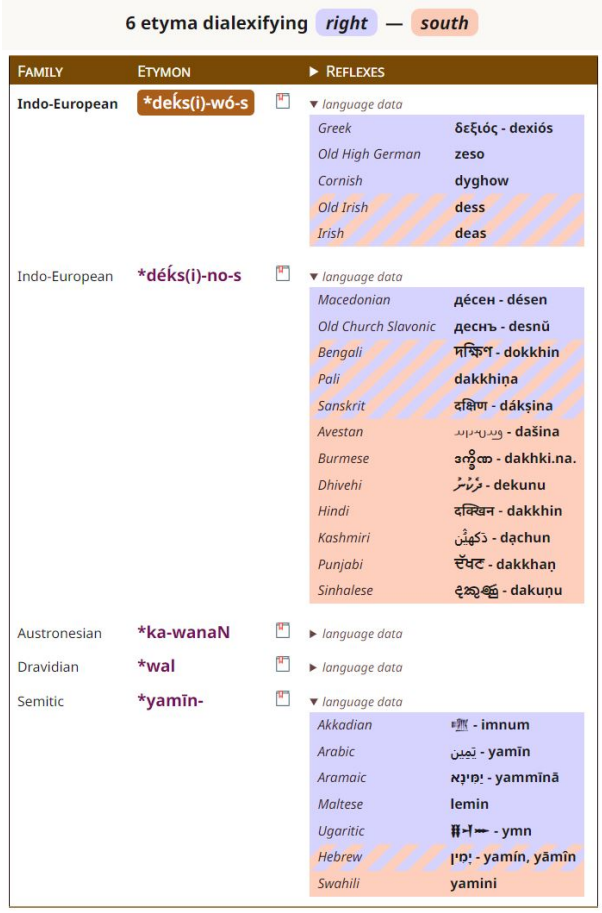

Figure 5: *Dialexification table* showing the six etyma that dialexify {RIGHT–SOUTH}. For each etymon, an unfolded list displays those reflexes that lexify only Concept_1 (blue background), or only Concept_2 (red background). When a reflex has both meanings at once, it is a case of colexification, made visible by the two-color stripe pattern. The clickable icon on each row (after the etymon) gives access to the online source.

STOMACH are dialexified 6 times in our data, but are not colexified even once.

Such instances of pure dialexification are useful to historical linguists, as they help to more accurately determine whether two forms with different meanings are potential cognates. They also provide insight into pathways of semantic change. Thus, while the pair {CHEST–STOMACH} is dialexified $\delta = 6$ times, {CHEST–HEART} has $\delta = 13$, and {HEART–STOMACH} has $\delta = 11$. From this, one can hypothesize that, if a form that once meant 'chest' later came to mean 'stomach', at some intermediate point it probably included 'heart' among its meanings.

Finally, dialexifications provide a way of measuring the similarities between concepts—much like colexifications, but in a manner that controls for shared descent. This opens up the possibility of using dialexifications to improve performance in similarity judgment tasks (as in Harvill et al., 2022), or to bootstrap the inference of semantic features in cross-lingual datasets (as in Chen and Bjerva, 2023).

References

Lloyd B. Anderson. 1974. Distinct sources of fuzzy data: ways of integrating relatively discrete and gradient aspects of language, and explaining grammar on the basis of semantic fields. In Roger W. Shuy and Charles-James N. Bailey, editors, *Towards Tomorrow's Linguistics*, pages 50–64. Georgetown University Press, Washington, D.C.

Hongchang Bao, Bradley Hauer, and Grzegorz Kondrak. 2021. On universal colexifications. In *Proceedings of the 11th Global Wordnet Conference*, pages 1–7, University of South Africa (UNISA). Global Wordnet Association.

Kenneth Benoit, David Muhr, and Kohei Watanabe. 2021. *stopwords: Multilingual Stopword Lists*. R package version 2.3.

Robert Blust and Stephen Trussel. 2013. The Austronesian comparative dictionary: A work in progress. *Oceanic Linguistics*, 52(2):493–523.

Robert Blust, Stephen Trussel, and Alexander D. Smith. 2023. CLDF dataset derived from Blust's "Austronesian Comparative Dictionary". Data set.

Mike Bostock. 2012. *D3.js - Data-Driven Documents*.

Thomas Brochhagen and Gemma Boleda. 2022. When do languages use the same word for different meanings? the goldilocks principle in colexification. *Cognition*, 226:105179.

Thomas Brochhagen, Gemma Boleda, Eleonora Gualdoni, and Yang Xu. 2023. From language development to language evolution: A unified view of human lexical creativity. *Science*, 381(6656):431–436.

Yiyi Chen, Russa Biswas, and Johannes Bjerva. 2023. Colex2Lang: Language embeddings from semantic typology. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 673–684, Tórshavn, Faroe Islands. University of Tartu Library.

Yiyi Chen and Johannes Bjerva. 2023. Colexifications for bootstrapping cross-lingual datasets: The case of phonology, concreteness, and affectiveness. In *Proceedings of the 20th* SIGMORPHON *workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 98–109, Toronto, Canada. Association for Computational Linguistics.

William Croft. 2005. Word classes, parts of speech, and syntactic argumentation. *Linguistic Typology*, 9(3):431–441.

Anna Di Natale, Max Pellert, and David Garcia. 2021. Colexification networks encode affective meaning. *Affective Science*, 2(2):99–111.

Robert MW Dixon. 2004. Adjective classes in typological perspective. *Adjective classes: A cross-linguistic typology*, pages 1–49.

Alexandre François. 2008. Semantic maps and the typology of colexification: Intertwining polysemous networks across languages. In Martine Vanhove, editor, *From Polysemy to Semantic Change: Towards a typology of lexical semantic associations*, Studies in Language Companion Series, pages 163–215. John Benjamins.

Alexandre François and Siva Kalyan. 2023. Dialexification: A tool for studying cross-linguistic patterns of semantic change. 16th International Cognitive Linguistics Conference.

Alexandre François and Siva Kalyan. in prep. Dialexification and the typology of lexical change. *Bulletin de la Société de linguistique*.

Thanasis Georgakopoulos, Eitan Grossman, Dmitry Nikolaev, and Stéphane Polis. 2022. Universal and macro-areal patterns in the lexicon: A case-study in the perception-cognition domain. *Linguistic Typology*, 26(2):439–487.

John Harvill, Roxana Girju, and Mark Hasegawa-Johnson. 2022. Syn2Vec: Synset colexification graphs for lexical semantic similarity. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5259–5270, Seattle, United States. Association for Computational Linguistics.

Martin Haspelmath. 1997. *Indefinite Pronouns*. Oxford University Press, Oxford.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *International Conference on Computational Linguistics*.

James P Mallory and Douglas Q Adams. 1997. *Encyclopedia of Indo-European Culture*. Fitzroy Dearborn Publishers, London, Chicago.

James A Matisoff. 2016. *STEDT: Sino-Tibetan etymological dictionary and thesaurus*. University of California, Berkeley.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Sebastian Nordhoff and Harald Hammarström. 2012. Cataloguing linguistic diversity: Glottolog/langdoc. Proceedings of Digital Humanities 2012.

Robert Östling. 2016. Studying colexification through massively parallel corpora. In Päivi Juvonen and Maria Koptjevskaja-Tamm, editors, *The lexical typology of semantic shifts*, pages 157–176. De Gruyter Mouton.

Christoph Rzymski, Tiago Tresoldi, Simon J. Greenhill, Mei-Shin Wu, Nathanael E. Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A. Bodt, Abbie Hantgan, Gereon A. Kaiping, Sophie Chang, Yunfan Lai, Natalia Morozova, Heini Arjava, Nataliia Hübler, Ezequiel Koile, Steve Pepper, Mariann Proos, Briana Van Epps, Ingrid Blanco, Carolin Hundt, Sergei Monakhov, Kristina Pianykh, Sallona Ramesh, Russell D. Gray, Robert Forkel, and Johann-Mattis List. 2020. The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data*, 7(1).

Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen. 2021. *Computational approaches to semantic change*. Language Science Press, Berlin.

Eva Van Lier. 2016. Lexical flexibility in Oceanic languages. *Linguistic Typology*, 20(2):197–232.

Michael Weiss. 2015. The comparative method. In Claire Bowern and Bethwyn Evans, editors, *The Routledge handbook of Historical linguistics*, pages 127–145. Routledge, London.

Hadley Wickham, Jim Hester, and Jeroen Ooms. 2023. *xml2: Parse XML*. R package version 1.3.5.

Yang Xu, Khang Duong, Barbara C. Malt, Serena Jiang, and Mahesh Srinivasan. 2020. Conceptual relations predict colexification across languages. *Cognition*, 201:104280.