# Lexical typology

## Comparing patterns of lexification

Alexandre FRANÇOIS

## 1 Introduction: The tenets of lexical typology

Scholars have discussed the nature of lexical meaning ever since the foundational works by philologists such as Paul (1880) or Bréal (1897) – the very linguist who coined the term *semantics*.[1] These authors outlined general principles of semantic structure and change (e.g. synonymy, polysemy, semantic broadening). Erudites before them (e.g. Du Marsais 1730; Fontanier 1825), inspired by a rich literature in philology and rhetoric anchored in Antiquity and Renaissance, had already started defining tropes such as metaphor, metonymy, synecdoche…, all relevant to the study of semantic extensions (cf. Blank 1999). Later, structuralist approaches – whether synchronic (Greimas 1966) or diachronic (Coseriu 1964) – showed how semantic structures follow patterns that can be studied for their own sake, independent of individual word histories. Then in the late 20th century, cognitive semantics (e.g. Talmy 1983; Langacker 1987; Sweetser 1990; Traugott & Dasher 2001) demonstrated how meaning can evolve along cognitive paths, and is structured in recurrent schemas, embodied frames, and mental analogies.

These foundational works were all essential in building a science of meaning, as they helped define fundamental concepts regarding semantic structure and change, and provide linguists with a shared terminology. Yet like most early research in linguistics, they usually rested on the observation of well-known European languages, whether ancient or modern. Among the examples proposed to illustrate semantic patterns, some may indeed be widespread in the world; but quite a few may have been, in fact, specific to just certain languages of Europe. One new question, which arose at the turn of this century, consists in assessing, among the principles of semantic structure and change, which patterns are indeed universal, and which show cross-linguistic variation. Such a scientific agenda had already begun to inform research on the diversity of grammatical structures; but it really started being applied to the lexicon in the early 21st century, giving rise to the new domain of *lexical typology*.

While all languages exhibit ambiguity and polysemy, and all go through semantic change over time, it would be a legitimate endeavour to compare specific patterns, assess their world-wide frequency, and study their distribution across language families. To take one example, the Proto-Indo-European (PIE) root **weyd*- encapsulates a link between vision and cognition, being both the ancestor of verbs meaning 'to see' (AGr. ἰδεῖν, Lat. *vidēre*, Rus. видеть) and of verbs meaning 'to know' (AGr. οἶδα, Old Norse *vita*, Grm *wissen*); one question then is whether that link {SEE–KNOW} is common across the globe, or is merely attested in Indo-European yet uncommon elsewhere (cf. Evans & Wilkins 2000). As for those semantic patterns that aren't evenly distributed across the globe, are they scattered haphazardly, or do they form areal clusters? Do these clusters reflect a history of human migrations and language contact? Can some of them be explained by cultural motivations?

To appear (2027?) in: A. Aikhenvald & RMW Dixon (eds.).
*The Cambridge Handbook of Linguistic Typology*, 2nd edition.

Another issue is how lexical polysemy and change vary, not just among language families, but across semantic domains. While cognitive semantics highlighted general principles of semantic change, it is safe to assume that certain domains will also be shaped by pragmatic or social factors, where human histories, cultural ideologies, local ecologies, play key roles – parameters to which other semantic fields may be less permeable. The dynamics of polysemy among, say, verbs of physical impact, must be quite different from what happens in kinship systems, or in the lexicon of emotion. Some subdomains may be prone to lexical renewal, while others are more stable. Some may tend to align their polysemies across languages, while other domains may accommodate more variation. Aspects of lexical semantics may correlate with the frequency of words in discourse, with their degree of abstraction, with their productiveness in constructions.

Such research questions have been central to the development of lexical typology. As we address them, the key is to avoid making assumptions based on introspection, and instead come up with robust methodologies that can help linguists define their scientific problems in a testable and falsifiable way, collect and analyse empirical data, and identify recurrent patterns.

Among the now abundant literature in lexical typology, the present chapter will focus on current research in *lexification patterns* – i.e. the packaging of meaning into lexemes, and the typology of polysemy. A substantial portion of that research revolves around the concept of *colexification* – i.e. the pattern whereby two senses can be encoded, in certain languages, by the same word (François 2008). After recalling the definition of this concept [Section 2], I will examine some of the research it has inspired in recent years [Section 3].

But while colexification is a synchronic notion, scholars have also been tackling the diachronic dimension, that of semantic change in the lexicon [Section 4]. The typology of semantic change informs new avenues of research: within the framework of lexification patterns, I will propose a new concept, that of *dialexification*, or semantic cognacy: two meanings are dialexified if they are expressed by descendants of the same etymon – like {SEE–KNOW} in the Indo-European family. This new approach, we shall see, can help us detect large networks of semantic connections, through the traces they have left in modern languages [Section 5].

This chapter will outline the latest research in the typology of lexification patterns – whether they are observed synchronically (colexification) or through the traces left by diachronic processes of change (dialexification). As I present the main lines of scholarship in these domains, I will also pay attention to the online resources that are currently being developed. Along the way, I will present three databases: CLICS, a catalogue of colexification [§2.3]; DatSemShift, a database of semantic shifts [§4.1]; and $\mathcal{E}\mathcal{v}\sigma$Sem, a catalogue of etymological links [§5.3]. These resources are proving useful in consolidating the empirical basis of research on lexical semantics.

## 2 Lexification studies: the onomasiological turn

### 2.1 The foundations of lexical typology

Several early publications pertain to lexical typology, even though they did not use this term. A classic reference in this respect would be the study of basic colour terms by Berlin & Kay (1969) – an exploration of semantic typology *avant la lettre*.

The notion of lexical typology began taking shape at the turn of this century, through a number of foundational publications. Three prominent researchers in this domain were Maria Koptjevskaja-Tamm, Martine Vanhove and Peter Koch; their paper "Typological approaches to lexical semantics" (Koptjevskaja-Tamm et al. 2007) may be seen as the birth certificate of the

new discipline. The early 2000s were already the period when a Moscow-based group led by Anna Zalizniak (Institute of Linguistics, Russian Academy of Sciences) began creating the *Catalogue of Semantic Shifts*, a long-running project that is still active today [see §4.1]. In parallel, the Paris-based Fédération *Typologie et universaux du langage* (TUL–CNRS) featured a research group led by Martine Vanhove, and titled "The typology of semantic associations"; the result of their debates took the form of a collective volume *From Polysemy to Semantic Change* (Vanhove 2008), which set the scene for some aspects of lexical typology – whether from a synchronic or diachronic viewpoint.

Because the present chapter proposes to focus on *patterns of lexification*, I will now turn to one particular chapter within that collective volume of 2008. François (2008)[2] outlined the concept of colexification, and proposed a roadmap for comparing semantic structures in the languages of the world. This approach would later be adopted by numerous scholars, who have been forming an active community in the field of lexical typology.

## 2.2 From polysemy to colexification

As we study the relationship between meaning and form, a key concept is that of *polysemy*. This term, coined by Bréal (1897:154–172), describes the case when a single word is endowed with multiple meanings or senses.[3] For example, French *bureau* can mean a 'desk', an 'office', or a 'bureau'; English *office* may refer to a liturgical ceremony, to a type of workplace, or a type of institution. A common observation is that the array of meanings that can be identified in a polysemous word in one language will hardly ever match that of a word in a different language. The polysemy of Eng. *office* and that of Fr. *bureau* do show some overlap, but they can hardly be mapped exactly onto each other.

For the typologist wishing to compare polysemies across languages, a key issue is whether they can be made comparable in the first place. When Louis-Ferdinand de Saussure defined the sign as the arbitrary pairing of a form and a "concept", he noted that each concept can only be defined negatively, insofar as it contrasts with other words within the system of that particular language:

> "Concepts (...) are purely differential; they are defined not positively by their contents, but negatively by their relationship to the other elements of the system." (Saussure 1916 [1972]:162)[4]

A corollary of this stance is that the semantic profile of a word is, in principle, unique to its language: it should not be expected to be found identical across different languages. In Saussure's formulation, the "concept" of Fr. *bureau* is supposed to encompass all its senses ('desk', 'office'…), and will thus be unique to French. In such a framework, polysemous words do not readily lend themselves to cross-linguistic comparison.

This reasoning led François (2008) to propose a new approach, around the notion of *colexification*. The following is a possible way to define that term:

(1) Colexification is the pattern whereby two different meanings can be encoded, in a given language, by the same lexical form.

On the face of it, *colexification* is just an equivalent of *polysemy*. However, the two notions differ in a crucial way. The perspective of polysemy starts from a word, and describes its different senses. Colexification flips the perspective: it starts from a pair of senses, and observes whether they can be expressed identically in a given language. While polysemy is semasiological – going from form to meaning – colexification adopts an onomasiological perspective, from meaning to

form. Indeed, the fact that two senses ($s_1$, $s_2$) can be expressed by the same word in a given language is, admittedly, a property of that word – which is traditionally captured by the term *polysemy*; but the same configuration, crucially, can also be seen as a property of that pair of senses {$s_1$–$s_2$}: it is an indication of their conceptual closeness. The power of the colexification approach, which may explain its later success in the scientific literature, arguably resides in this "onomasiological turn" in polysemy studies. This was the pivotal change needed to make cross-linguistic comparison possible.

The "meanings" or "senses" that constitute the starting point of these observations can be defined empirically – rather than assumed aprioristically – as any meaning that is "attested" in the world's languages (François 2008:168–170, 2022a:93–95):

(2) For the purpose of cross-linguistic comparison, a sense is considered *attested* if it defines the semantic contrast between two different forms, in at least one language.

Thus, the fact that certain languages have a special word to refer to FIREWOOD as distinct from WOOD (used as material) – e.g. Spanish, with respectively *leña* vs. *madera* – is enough to posit these two senses as distinct, even if many languages colexify them. Far from being unique to a specific language – like Saussure's "concept" – the senses identified under this approach are understood to be universally testable: the idea is that any language in the world should have a way to express the senses FIREWOOD and WOOD (MATERIAL) – whether that is done using the same form, or different ones. These individual senses constitute the *tertium comparationis* that was needed in order to compare polysemies across the world's languages: the "etic grid" they form can then serve as a background against which emic categories of particular languages can be assessed (see Evans 2010).

An important corollary is that colexification can be observed independently of the vexing problem of distinguishing *polysemy* from *monosemy* or *vagueness* (Cruse 2001:80). For a word like *hoyo* in Mwotlap (Oceanic, Vanuatu), which translates as 'long' (e.g. for a string) or as 'tall' (for a person), there could be debate whether this is a case of genuine polysemy, or simply different contextual readings of a single, more abstract meaning. In principle, solving this conundrum would be a prerequisite before deciding whether this word should be featured in a typological study of "polysemy". But with colexification, this question can be sidestepped. The mere fact that LONG and TALL are distinguished in at least some languages is enough to consider them two separate senses – and then observe which languages can encode them identically.

In principle, accidental homophony and homography are excluded from colexification. Indeed, the term "lexical form" used in the definition (1) refers to a lexical unit – whether a lexeme, a lexicalised phrase or a construction – whose senses are connected by a shared history, as established by a linguist or lexicographer. The fact that the surface form *seal* may refer, in English, to a marine animal and to an official stamp, is an accident of sound change, and does not result from a history of semantic associations. The sea animal derives from Proto-Germanic **selhaz* (same meaning), whereas the official *seal* is a Romance borrowing, ultimately from Lat. *sigillum* 'figurine, small sign'. While they share the same surface form, *seal*$_1$ and *seal*$_2$ constitute two different "lexical forms": therefore, they do not fall under the definition of colexification.

Certain authors, though, have understood colexification to include homophony (List et al. 2026).[5] This choice is understandable from the point of view of computational approaches involving the automatic matching of character strings, by an algorithm that does not take semantics into account; this must have been a useful step when parsing wordlists for the creation of tools like CLICS [§2.3]. That said, when discussed in the context of lexical typology proper, the general understanding of colexification is that it is the formal manifestation of a

lexical connection that is ultimately rooted in semantics – and therefore does not include accidental homophony.

The opposite of *colexification* is *dislexification* (François 2022a:95):[6]

(3) Dislexification is the pattern whereby two meanings, in a given language, are usually encoded using two separate lexical forms.

For example, the two meanings HEAR and FEEL are colexified in Italian, which can use the verb *sentire* when referring to auditory perception (*Non ti avevo sentito* 'I hadn't heard you') as well as sensations (*Sentivo dolore* 'I was feeling pain'). By contrast, a language like French encodes these two meanings using separate verbs, *entendre* and *ressentir*: the same two sentences would be rendered respectively as *Je ne t'avais pas entendue* (**ressentie*), and *Je ressentais* (**entendais*) *de la douleur*. This situation can be captured by saying that the two meanings HEAR and FEEL are colexified in Italian, but dislexified in French. Note that the two terms are purely descriptive, with no entailment as to which configuration should be regarded as expected or normal – thereby avoiding the risk of ethnocentrism.

I propose a notational convention to represent these patterns, using small capitals for senses, between braces. Colexification can be represented using the equal sign '=', while its opposite symbol '≠' would encode dislexification – yielding {HEAR = FEEL} vs. {HEAR ≠ FEEL}. Together, the two notions – co- vs. dislexification – can be subsumed under an umbrella term, namely *patterns of lexification* (François 2022a:92):

(4) *Patterns of lexification* are the patterns of distribution between senses and forms.

The typological study of these patterns can be labelled *lexification studies* – a central component of lexical typology.

Haspelmath (2023) assigns two readings to the notion "lexification patterns". One approach, which he calls *coexpression*, refers to "the availability of two meanings for a minimal form in different contexts"; this encompasses *colexification* – as defined above – and what he calls *cogrammification*, or grammatical polysemy. The author also proposes a second approach, which he labels *synexpression*, namely "the simultaneous presence of two meanings in a minimal form". Synexpression is a cover term for *syllexification* and *syngrammification*: for example, Eng. *kitten* syllexifies [cat + young]; and portmanteau inflections syngrammify more than one category at a time – such as Lat. *-erunt* (in *vid-erunt* 'they saw') which encodes [perfect + 3rd person + plural]. This domain of synexpression would warrant methodological approaches of their own, which would take us beyond our scope here. The present chapter will thus focus on patterns of coexpression (co- and dis-lexification) in the world's lexicons.

## 2.3 Semantic maps and online databases

Besides defining the concept of colexification, another idea central to François (2008) was the proposal to display observations using *semantic maps*. The idea was to build on a method that had proven effective in the field of grammatical typology (Haspelmath 2003; see Georgakopoulos 2019), and extend it to the lexical domain. The paper illustrated a potential workflow by discussing cross-linguistic trends in semantic extensions around the sense BREATHE (→ *rest; soul; spirit; ghost*…), and by handdrawing the sort of semantic maps that could represent them.

In 2014, these suggestions inspired Johann-Mattis List, with a team of linguists and engineers, to create an online resource[7] called "CLICS" – the Database of Cross-linguistic Colexifications (see Rzymski et al. 2020). CLICS imports a number of publicly available lexicons and wordlists (e.g. the Intercontinental Dictionary Series, among many) and analyses these datasets

so as to automatically detect cases of colexification. Using the community detection algorithm *Infomap*, CLICS then visualises its results in the form of force-directed graphs to reveal semantic networks. Figure 1 illustrates such a semantic cluster around sensory (HEAR, SEE…) and cognitive (KNOW, UNDERSTAND…) meanings.

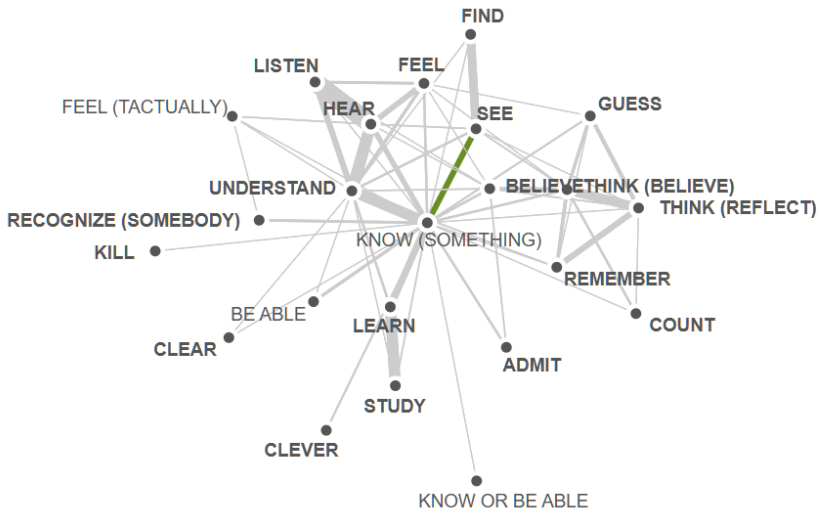


Figure 1 – A semantic subgraph around the notion HEAR, from CLICS3. The thickness of every edge reflects the relative frequency of colexification for a given pair of senses.

Nodes in the graph represent meanings; an edge between two nodes indicates that these two senses are colexified in some languages, with the thickness of each edge reflecting the relative frequency of colexification. Thus in their dataset, the pair of senses {HEAR=UNDERSTAND} is colexified in 50 languages, which is more than {HEAR=FEEL} (20) or {UNDERSTAND=BE ABLE} (4). Hovering over an edge brings up a table with a list of relevant languages and forms.

The CLICS database evolved over the years, and is now at CLICS4 (Tjuka et al. 2026). It has shown steady progress in its coverage (from 221 languages in 2014 to 3,447 in 2026) and in the number of its concepts (from 1,280 to 4,033). CLICS resorts to the "Concepticon" reference (List, Cysouw & Forkel 2016) in order to normalise or disambiguate its glosses: this is how Figure 1 distinguishes two senses of the English verb *think*, namely THINK (BELIEVE) and THINK (REFLECT). Under CLICS4, a decision was made to retain only languages with ≥180 concepts, and reduce its list of concepts to the 1,800 most frequent ones; while this aims at reducing sparsity, it also has the effect of considerably limiting the semantic breadth of the database, which now documents only 1,800 concepts.

Colexification has inspired other empirical datasets. Pericliev (2015) reports on a similar yet distinct dataset, derived from the ASJP database, and dedicated to basic vocabulary. Östling (2016) introduces a computational method to study colexification using parallel corpora – in this case, 1,142 translations of the New Testament into 1,001 languages; he shows how lexification patterns could be automatically detected with relative success. Finally, section 5.3 will report on *EvoSem*, a new resource dedicated to lexification patterns, targeting both synchrony and diachrony, and partly inspired by the CLICS project.

In sum, François' 2008 chapter planted the seeds for several projects – that of studying colexification in a systematic way, of visualising results using semantic maps, or of collecting large amounts of empirical data to allow for typological generalisations. Other authors later managed to transform these initial proposals into actual datasets that can be tested, completed, mined for information, and exploited to propose scientific claims and observations. In doing so,

they allow scholars to explore the full potential of lexification studies, bring about results, and enhance their reproducibility – a valuable objective in linguistics (cf. Berez-Kroeker et al. 2018). The next sections will summarise some of these findings.

# 3 Some recent research in lexification studies

## 3.1 Studies of different semantic domains

The last decades have seen a great number of studies on (co)lexification, whether they cited that term or adopted a similar approach. At the time of writing, the literature includes more than a thousand publications revolving around colexification, showing how active this research strand has become. For reasons of space, I will focus here on a sample of these publications, enough to illustrate the diverse types of scientific questions addressed by current scholarship in lexical typology.

Due to the sheer size of the lexicon, a reasonable approach often consists in selecting a specific semantic domain, and examining how it is structured across different languages. Quite a few semantic fields have been the object of scrutiny – whether through individual articles or entire volumes: events of cutting and breaking (Majid et al. 2008), of eating and drinking (Newman 2009; Stathi 2023), of putting and taking (Kopecka & Narasimhan 2012), of motion (Talmy 1985; Wälchli & Cysouw 2012), of falling (Ryzhova et al. 2024); physical qualities (Rakhilina et al. 2022; Koptjevskaja-Tamm 2015, 2022); sensory perception (San Roque et al. 2018; Jędrzejowski & Staniewski 2021; Norcliffe & Majid 2024); emotion and affection (Majid 2012; Jackson et al. 2019; Di Natale et al. 2021; Ponsonnet 2025); and so forth.

Every one of these semantic fields has its own properties and dynamics, and it would go beyond the limits of this chapter to summarise all these publications. At least, this preliminary inventory gives an idea of the abundance of research being carried out on a broad array of semantic domains. The list remains open, and new domains will be explored in years and decades to come.

## 3.2 Experimental and psycholinguistic approaches

Some publications adopt a general angle as they choose to observe the very phenomenon of colexification, whether from a cognitive or experimental perspective, across multiple semantic domains.

Xu et al. (2020) compare 1,310 meanings across 246 languages, in order to identify potential correlations between colexification and other parameters. They find that the best predictor for colexification frequency is "conceptual associativity" – a metric obtained through psycholinguistic tests of word association tasks (*Human Brain Cloud* data). For example, such tests show that speakers spontaneously associate the concepts TEACH and EXPLAIN; this strong associativity correlates with high levels of cross-linguistic colexification for that pair of concepts. Their conclusion supports a principle of cognitive economy: languages favour colexifications that minimize cognitive effort in production and comprehension.

Karjus et al. (2021), using artificial language games, demonstrate that lexification patterns are determined, not just by conceptual associativity, but also by the need for communicative efficiency. While speakers naturally colexify similar concepts (e.g. ABDOMEN–BELLY), this tendency is overridden when communicative need demands distinction. In a social or pragmatic context that would tend to distinguish *abdomen* from *belly*, the likelihood of colexification diminishes.

Lexification patterns thus reflect a balance between conceptual similarity and communicative pressure. A similar conclusion is reached by Brochhagen & Boleda (2022).

### 3.3 The areal distribution of lexification patterns

Several studies have used (co)lexification patterns as a tool to describe the semantic profile of certain families or linguistic areas.

Schapper et al. (2016) explore a large area known as "Sahul" – encompassing Australia, New Guinea, and nearby territories. They demonstrate that the three concepts TREE, FIREWOOD and FIRE show a strong tendency to colexify together in this area, more than in other regions of the world. Figure 2 shows the high prevalence of the {FIREWOOD = FIRE} colexification (white circles), or even the full triplet {TREE = FIREWOOD = FIRE} (black triangles). This areal feature might be the trace of a shared linguistic history in a remote past.

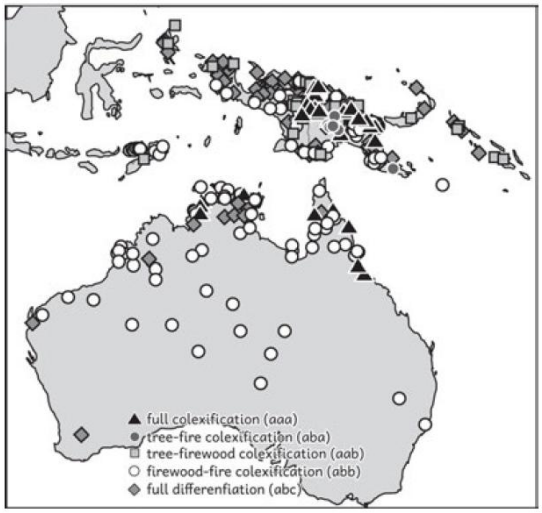


Figure 1 – Patterns of colexification involving the three concepts TREE–FIREWOOD–FIRE among Papuan languages.

Similarly, Schapper (2019) observes the colexification {SMELL = KISS} across Southeast Asia. The high prevalence of this semantic link here transcends family boundaries, as it can be observed in Austroasiatic, Austronesian, Tai-Kadai and Tibeto-Burman languages. This most likely results from a history of substratal or horizontal diffusion, tied to a specific cultural practice: the habit of "smell-kissing", i.e. expressing one's affection – especially towards infants – by touching their cheek with one's nose. This is an instance of colexification anchored in a shared cultural habit, widespread in this particular region.[8]

In a similar vein, Ding & Dong (2024) show how an ancient cultural connection between DRAGON and THUNDER has resulted in their colexification in many Sino-Tibetan (ST) languages. They propose that the pattern originated from North China, and might reconstruct all the way to Proto-ST – before it spread to non-ST languages (Mongolic, Turkic) through contact.

In the Amazon region, Aikhenvald & Dixon (2026) demonstrate that one dialect of Tariana, a North Arawak language, has come to colexify the two celestial bodies {SUN = MOON}, due to contact with East Tukanoan languages that do the same. Another dialect of Tariana lacks that colexification, this time under the influence of the closely related language Baniwa of Içana.

In sum, numerous publications show how lexification patterns – whether colex or dislex – tend to spread from neighbour to neighbour through language contact, to such an extent that they can help define entire Sprachbund areas. They can be interpreted as diagnostic of shared history, recent or ancient, among populations in social contact.

Beyond the case studies listed above, certain works demonstrate this important point at a global scale. One reference in this respect is Gast & Koptjevskaja-Tamm (2018), "The areal factor in lexical typology". Building on the data in CLICS and ASJP, the authors lead a large-scale quantitative study crossing lexification data with spatial information; this allows them to identify statistically significant areal clusters. They detect several recurrent patterns of lexification, such as {EAR = LEAF} in Eastern Africa, or {STONE = MOUNTAIN} in parts of Africa and Australia.

The geographic distribution of lexification patterns continues to inspire many studies. It has become a promising angle for detecting areal signals, and formulating hypotheses of shared history, in a way that complements phonological or grammatical criteria. We will come back to such issues in §5.4.

### 3.4 On the design of semantic maps

In certain publications, areal signals are taken into account, yet do not constitute the main focus of research. Thus Georgakopoulos et al. (2022a) investigate universal and macro-areal patterns in the lexicon, in the domain of perception and cognition (e.g. SEE, HEAR, KNOW, UNDERSTAND). Combining datasets from CLICS (cf. Figure 1 above) and WordNet, the authors detect certain areal signals – e.g. the colexification {HEAR = KNOW} in South America – but also bring certain universal patterns to light. For example, they notice that the perception modalities SEE and HEAR hardly ever colexify together: rather, they connect separately to cognitive meanings, such as KNOW or UNDERSTAND. Also, the latter meanings always link to verbs of non-controlled perception (SEE, HEAR) rather than their controlled counterparts (LOOK, LISTEN). Both of these universal patterns are visible in Figure 3.

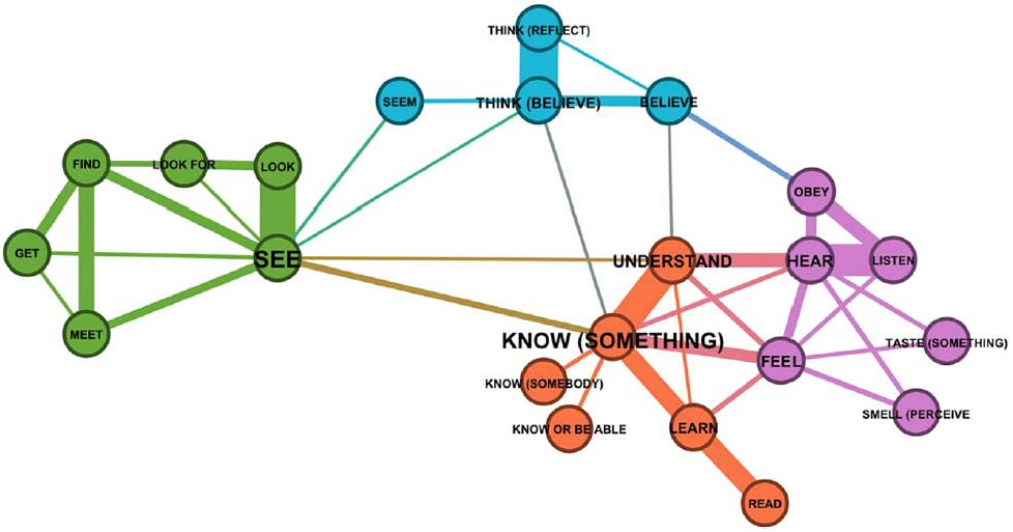


Figure 3 – Semantic map (based on CLICS2) showing universal colexification patterns in the domains of perception and cognition (Georgakopoulos et al. 2022a:458). Direct links connect cognition verbs (KNOW, UNDERSTAND, THINK) to non-controlled perception (SEE, HEAR).

Over the years, Thanasis Georgakopoulos and his colleagues have investigated numerous aspects of lexification patterns. In terms of semantic domains, they have explored not only cognition and perception as we just saw, but also emotions (Georgakopoulos & Polis 2022b), time concepts (Georgakopoulos & Polis 2021 – see Fig. 5 below), space and motion events (Georgakopoulos et al. 2020; Georgakopoulos & Nikiforidou 2026). Beyond this diversity of semantic fields, Georgakopoulos' contributions are characterised by a constant attention paid to issues of theory and methodology in lexical typology, as they explore different mathematical models or frameworks to analyse the data and display results of lexification studies. Thus Georgakopoulos et al. (2022a), mentioned above, argue in favour of multiple visualisation techniques, including semantic maps (Figure 3), colexification networks, conceptual lattices, correlation plots; they also show the value of using dimensionality reduction (UMAP) when comparing macro-areas. Other publications discuss methodological questions of graph structure and topology, clustering, modularity, centrality, and dimensionality reduction.

This body of research has made Georgakopoulos – with his coauthors – an expert in the general theory of semantic maps (Georgakopoulos 2019; Georgakopoulos & Polis 2018, 2022b). He argues that semantic maps should not be drawn manually, but always derived from the data through explicit algorithms and reproducible procedures. Using robust methods in graph construction also means that maps can be used as constrained graphs capable of predicting new patterns, e.g. suggesting implicational universals of the type "*if a language colexifies X and Y, it will likely also colexify X and Z*"; this possibility was imagined by François (2008:175), but operationalised through Georgakopoulos' research.

# 4 Lexification patterns, from synchrony to diachrony

This overview of research in lexical typology has, so far, mostly adopted a synchronic angle. Indeed, the distribution of meaning across lexical units can be conceived primarily as a structural notion, characteristic of a lexical system at a given point in time. When François (2008) defined colexification, the initial focus was on *strict colexification* – namely, the coexpression of two meanings by the same word in synchrony. However, that paper also mentioned – under the label *loose colexification* – various other types of links that can exist between two meanings:

> "These may include the linking of two senses by a single lexeme across different periods of its semantic history […], their association in the form of doublets […] or other etymologically related forms […], the impact of lexical derivation […] or composition […], and so on." (François 2008:171)

While these other relations differ from colexification strictly speaking, they do keep the trace of the special links that can bring senses together.

The next pages will present some lines of research that tackle lexification patterns, yet do not restrict themselves to strict colexification in synchrony. I will introduce the DatSemShift project, and its concept of *semantic shift*, which can subsume a whole array of formal links between meanings. We will then focus on semantic change and the light it can shed on lexification patterns; I will mention recent work around diachronic semantic maps, including the idea of *lexical tectonics*. Finally, Section 5 will define a new tool for observing the dynamics of lexification patterns: *dialexification* – itself the source of a new database called $\mathcal{E}\mathcal{v}\mathcal{o}$Sem.

## 4.1 *DatSemShift*, the database of semantic shifts

Years before the scholarly developments on colexification, one of the pioneer projects in lexical typology was the *Catalogue of Semantic Shifts* (CSSh), led by Moscow-based linguist Anna Zalizniak [§2.1]. Initiated in 1998, the project later took the form of an online database: DatSemShift, the *Database of semantic shifts in languages of the world* (Zalizniak 2008, 2018; Zalizniak et al. 2012).

The research undertaken by Zalizniak and her colleagues revolves around the notion of *semantic shift* (семантический переход). This refers to a special connection between two meanings A and B, that can be empirically observed through its formal manifestations. The ways in which a semantic shift manifests itself, called "realisations", include synchronic polysemy (colexification {A=B}), diachronic change {A→B}, cognates, or morphological derivation.[9] In fact, the same semantic shift can show up through different realisations (Zalizniak 2008:228). For example, the map in Figure 4 shows how the link {TO GRASP, SEIZE – TO UNDERSTAND} can manifest itself in various ways across languages:[10]

- through polysemy, e.g. in English (*to grasp*, *to get* – both of which can also mean 'to understand'); in Russian (схватывать *sxvatyvat'* 'to grasp; to understand')
- through semantic evolution,
  e.g. Lat. *capio* 'to catch' → It. *capire* 'to understand';
  Old East Slavic поняти *pojati* 'to take, to get' → Rus. понять *ponjat'* 'to understand'
- through cognate links,
  e.g. Rus. хапать *hapat'* 'to grasp, seize' : Czech *chápat* 'to understand'
- through morphological derivation,
  e.g. Grm *greifen* 'to grasp' → *begreifen* 'to understand'.

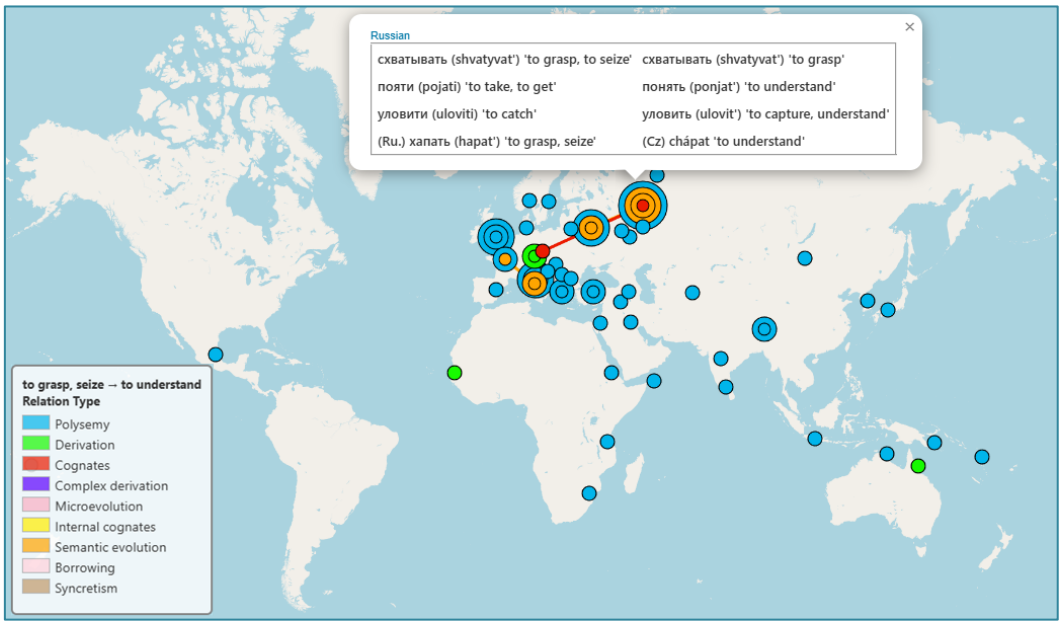


Figure 4 – Map from DatSemShift, illustrating the semantic shift {TO GRASP, SEIZE – TO UNDERSTAND} and its geographical distribution. The colors indicate the different "realisations", i.e. the form taken by this shift (polysemy, derivation, etc.).

DatSemShift currently contains about 5,000 unique meanings and 10,000 semantic shifts, covering 50,000 realizations from 2,400 languages worldwide.[11] While many examples were entered and verified manually over the years, an automated module was recently introduced, using NLP technologies, to continue enriching the database (Gruntov & Rykov 2023).

## 4.2 Approaching semantic change

Let us now focus more specifically on semantic change in the lexicon, and see how it can shed light on the proximity between meanings. One common assumption is that semantic change involves the very same kind of links as the ones uncovered by synchronic observations, of the sort we have discussed so far.

Several scholars (e.g. Wilkins 1996:267; Geeraerts 1997; Gfeller et al. 2026) have thus argued that synchronic polysemies can be a valuable source of information when studying semantic change. The idea is that a historical shift from sense A to sense B must have involved, if only temporarily, a stage of synchronic polysemy {A = B} (Sweetser 1990; cf. Evans & Wilkins 2000); therefore, typological trends in synchronic polysemy could provide a suitable approximation of potential paths of lexical change. In the same spirit, Dellert (2016) proposes to leverage on colexification data to determine "directional tendencies in semantic evolution":

> "Synchronic polysemies provide us with a snapshot of semantic evolution in action, which we can exploit as observational data in order to retrieve hints about tendencies in their historical development." (Dellert 2016)

In this perspective, the synchronic graphs of colexification we saw earlier would be interpreted as a likely representation of the semantic space where lexical change takes place. Thus Figure 3 above, while based on synchronic data, could be used to predict that a word meaning KNOW is more likely to have its origin in a verb SEE or HEAR than, say, come from LOOK or LISTEN. This general hypothesis – namely, that semantic networks based on synchronic data can help predict facts of diachronic change – would deserve to be confirmed based on attested historical data.

Precisely, generations of philologists and historical linguists have also collected abundant data about diachronic change. Some languages are attested in writing over several millennia, allowing us to observe actual cases of change in the lexicon. While we saw that Georgakopoulos and his colleagues have published on synchronic patterns of lexification, part of their research has also consisted in tackling the historical dimension. Georgakopoulos & Polis (2021), studying data from Ancient Greek and Ancient Egyptian, thus propose to develop "Lexical Diachronic Semantic Maps". Their idea is to exploit the power of graph representations to encode not just synchronic polysemy, but also the dynamics of lexification patterns. Looking into the domain of time words (DAY, MONTH, HOUR, YEAR, SEASON…), Figure 5 shows how semantic "multigraphs" (i.e. "graphs with multiple edges between nodes, that can be either directed or undirected") can represent potential paths of semantic change.

Figure 5 – A semantic "multigraph" (from Georgakopoulos & Polis 2021:43) representing potential paths of semantic extension in the lexical domain of time, based on a diachronic corpus of texts in Ancient Egyptian and Ancient Greek.

Such graphs make it possible to capture the directionality of certain changes. For example, their corpus analysis highlights directed paths such as {SPRINGTIME → SUMMER → HARVEST → AUTUMN}; whereas the link between TIME and HOUR appears to go in both directions. Such promising conclusions, here based on the internal history of two languages, would deserve to be tested over a larger empirical sample, to see how common they are across language families.

## 4.3 Splits and mergers: Lexical tectonics

The same authors edited a collective volume around semantic maps and their potential for representing diachrony (Georgakopoulos & Polis 2022). In that volume, François (2022a) proposes that semantic evolution can be analysed in terms of lexification patterns:

(5) *Semantic change* takes place when patterns of lexification (the distribution between senses and forms) undergo a change in time.

A polysemous word, which once lexified senses {$s_1$=$s_2$=$s_3$}, may evolve in such a way that it later encodes {$s_2$=$s_3$=$s_4$}. Some of the meanings that used to be colexified are now dislexified, and vice versa. Leaning on empirical data from Romance and from Oceanic, François outlines a general model of historical lexicology. He identifies five types of structural innovations in the lexicon, among which the main two are the split and the merger:

- *lexical split*: a former colexification {$s_1$=$s_2$} gives way to dislexification {$s_1$≠$s_2$}
- *lexical merger*: a former dislexification {$s_3$≠$s_4$} gives way to colexification {$s_3$=$s_4$}

A historical linguist reconstructing semantic change in the history of a language family will identify cases of splits and mergers. For example, Classical Latin used to contrast {MOTHER'S BROTHER ≠ FATHER'S BROTHER}, respectively as *avunculus* vs. *patruus*; yet over time, the two kin categories were merged in all Romance languages (Coseriu 1964:159). The innovative colexification {MOTHER'S BROTHER = FATHER'S BROTHER} is now encoded through reflexes of *avunculus* – e.g. Romanian *unchi*, Old Fr. *oncle* (the source of Eng. *uncle*) – but also through forms inherited from Late Latin *thius*, like Sp. *tío* or It. *zio*.

Splits and mergers can be displayed using a set of semantic maps, ordered chronologically.[12] Such maps are useful not only when representing change that is attested historically, but also when reconstructing the evolution of a word based on the principles of the comparative method. Figure 6 thus reconstructs the path of a lexical innovation that took place in Dorig, an Oceanic language spoken in northern Vanuatu.

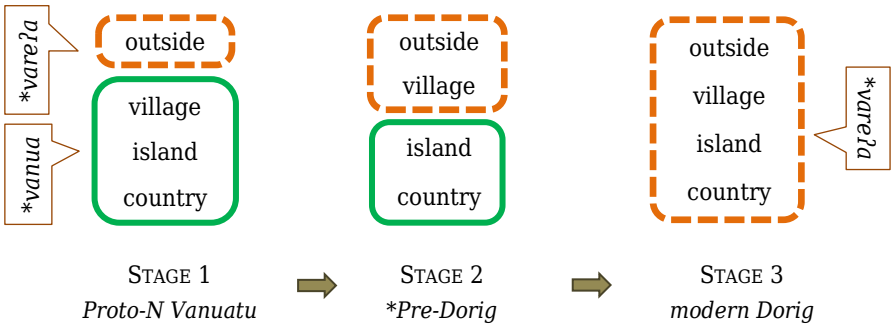


Figure 6 – Language-internal relexification via semantic realignment. In Dorig, the root *vareʔa entered the territory of *vanua via the sense 'village', but eventually replaced that root in its entire polysemy 'village, island, country' (after François 2022a:114).

Proto-Northern Vanuatu can be reconstructed with a root **vanua* that colexified {VILLAGE = ISLAND = COUNTRY};[13] but that root was lost in Dorig, which now uses a form /vrɛ/. The latter word goes back to an etymon **vareʔa*, reconstructable as meaning 'outside, public space' (Clark 2009: 218); this initial configuration is represented as Stage 1 in Figure 6. It is likely that **vareʔa* began encroaching on the meaning 'village' (the 'outside' with respect to one's house), where it replaced **vanua*; Stage 2 reconstructs the innovative lexical split {VILLAGE ≠ ISLAND} that must have resulted from this semantic replacement. Yet that new distribution of meanings vs. forms was not stable. Under the continued influence of the neighbouring languages that had kept the structures of Stage 1, Dorig eventually realigned its lexification structures with them, in the form of a lexical merger {VILLAGE = ISLAND = COUNTRY} – but this time it did so with the new etymon **vareʔa* (Stage 3). This is how a series of lexical splits and mergers can result in wholesale "language-internal relexification" (François 2022a:114).

As lexical structures change over time, words migrate, so to speak, from neighbour sense to neighbour sense, before they settle into new territories. François (2022a) likens this slow

process to geological movements, and labels it *lexical tectonics*. In principle, given enough fine-grained data, many cases of historical change in the lexicon could lend themselves to a similar graphic representation, in which semantic gains and losses could be represented as a step-by-step progression across semantic space.

# 5 Dialexification: Retrieving semantic links through etymology

Our final section will focus on an ongoing project of mine, that aims at synthesising our knowledge of lexical change and etymology, and at using it as leverage to determine the likelihood of change between meanings. This project builds on the new concept of *dialexification*, and is embodied in a new online resource: $\mathcal{E}\upsilon\sigma$Sem, a database of etymological links.

The "lexical tectonics" I outlined in §4.3 describes a general process whereby lexical units evolve over time, changing their semantic profile, gaining new meanings and losing old ones. Through different processes of semantic extension (metaphor, metonymy...) and restriction, words progressively change their lexification patterns as they migrate from neighbour to neighbour in semantic space. But while that general diachronic process is universal, what do we know about individual changes? Do they exhibit random variation across languages, or can we identify typological tendencies? If we take a given concept as a point of departure, can we guess what possible meanings it is most likely to take? In order to address this question, we need empirical methods that can help us identify a meaning's closest "neighbours" in semantic space. Such an endeavour would provide a cartography of potential extensions – a universal background against which local instances of semantic change could be assessed.

The $\mathcal{E}\upsilon\sigma$Sem project proposes to create an online database of known etymological links. This scientific programme is reminiscent of the lexification studies we've discussed in Sections 2–3, except it is designed to integrate diachronic and synchronic data together at once. But before I introduce the resource, let me present its main tenets.

## 5.1 Capturing etymological links: The case for dialexification

A word from an ancient language can be the source of descendant words, or "reflexes", in later languages. This may happen through direct inheritance within a given family, or through borrowing. Given a source word or etymon, the set of all its reflexes forms a "cognate set". For example, the Latin verb *clāmāre* 'exclaim, cry out' is the source of a cognate set that includes Italian *chiamare*, Spanish *llamar*, Portuguese *chamar* – which all mean 'call, name' – but also French *clamer* 'proclaim', and the English verb *claim* borrowed from it.

If the source word is attested in texts, we can know its meaning(s) with reasonable certainty, which allows us to establish etymological links in diachrony. For example, we know that the Latin noun *spathula* – originally the diminutive of *spatha* 'straight sword' – was the name given to a variety of flat tools, such as a putty knife or a spatule. Later, by analogy of shape, it would refer to the anatomical 'scapula' or shoulder blade (Apicius, 5$^{th}$ c. AD). In Romance languages, reflexes of *spathula* then shifted, by metonymy, to refer to the 'shoulder' (Catalan *espatlla*, French *épaule*), while some languages pushed further to refer to the 'back' (Spanish *espalda*, Italian *spalle*). These empirical facts allow us to establish the following path of semantic change:

(SMALL SWORD →) SPATULA → SCAPULA → SHOULDER → BACK

Here, our access to the semantics of the source word (Lat. *spathula*) is a precious ally in identifying the sequence of semantic shifts.

That said, the *Evo*Sem project plans to compare etymological links across the globe, in as many language families as possible. For many of them, the common ancestor is not a language that is attested in writing, but a protolanguage, reconstructed by inference from modern data. The comparative method (see Rankin 2003; Weiss 2014) does provide sharp tools to propose reconstructions – words conventionally marked by an asterisk to signal their hypothetical status – yet both protoforms and their protoglosses are necessarily speculative. In order to build exclusively from observable data, we need to design an approach that can capture etymological links, even when the meaning of their initial source is not known. The concept of dialexification, which I will introduce now, precisely aims at capturing such links even in the absence of temporal directionality.

Let us take an example from the Austronesian phylum, a family of languages with virtually no ancient attestation. Language comparison tells us that Hawaiian *honua* 'land, earth' is cognate with Tongan *fonua* 'country', and also with Malay *benua* 'empire', Hiw [Van.][14] *vönyö* 'island', Selaru [Indon.] *hnu* 'village', Marovo [Sol.] *vanua* 'house', Longgu [Sol.] *vanua* 'people' – not to mention Kapampangan [Phil.] *banwá* 'sky; year' or Pangasinan [Phil.] *banwá* 'sun'.[15] The common source behind this cognate set, at the level of Proto-Malayo-Polynesian (PMP), can have its phonological shape reconstructed with little difficulty, as an etymon **banua* (Blust 1987); but what about its proto-meaning? Common – though not necessarily good – practice among historical linguists is to provide protoforms with a minimalist protogloss, often as abstract and monosemous as possible, rather than reconstruct a polysemy. Thus Blust et al. (2025) gloss **banua* 'inhabited land, territory supporting the life of a community'. Yet in reality, it is likely that **banua* must have already shown a complex polysemy from the very start (Pawley 2005:218).[16] The latter must have included many of the senses attested in descendant languages – though probably not all of them: knowing which ones exactly would be a challenge.

The main point here is that a reconstructed gloss, as sound as it may seem, does not have the same epistemic status as philological data. The practice of semantic reconstruction (cf. Blust 1987; Pawley 1985, 2005; Urban 2015) is an important endeavour, but one that requires a solid empirical basis before reliable hypotheses can be formulated. This should lead us to concentrate on the factual evidence we have access to – the forms and meanings of actually attested words – and make them the centre of our attention. Whereas the diachronic path {SPATULA → SCAPULA → SHOULDER} can be safely established based on empirical data, it would be imprudent to posit an evolution {*INHABITED LAND → ISLAND}, for we cannot be sure of the etymon's original meaning. And yet, we would not want to forgo the valuable semantic insights that the rich **banua* cognate set can bring to lexical typology.

What we want is to capture the special link that Austronesian languages draw between all these meanings {LAND, COUNTRY, EMPIRE, ISLAND, VILLAGE, HOUSE, PEOPLE, SUN, SKY, YEAR}. Yet we need to do so without relying on a speculative protogloss, and on imaginary semantic paths. In other terms, we want to study etymological links, while allowing ourselves to be agnostic about the direction of change.

The special relation that exists between these senses doesn't really have an established name, other than "meanings attached to historically related forms"; in other contexts, these would be described as "false friends". The term *colexification* cannot apply here, because we are not observing a single word in one language, but a set of related words in different languages. In order to capture that relationship, I propose the term *dialexification*:[17]

(6) Dialexification is the pattern whereby different meanings are attached to the same cognate set.

(7) Two meanings are dialexified if they are attached to cognate words: that is, words – typically from different languages – that share the same etymology.

Dialexification can be described as a form of "semantic cognacy". It is analogous to colexification, but this time defined at the level of a whole cognate set, rather than an individual word in a given language. We can thus say that the two meanings ISLAND and PEOPLE are dialexified in the Austronesian family, (in the cognate set) under the PMP etymon **banua*.

The original motivation for the prefix *dia-* (from Greek διά 'across') was the idea that the semantic link in question is not confined to a single language, but – so to speak – can span "across" different languages. Another motive for the prefix is that the semantic paths have necessarily arisen through historical processes of change – even though we don't always know their direction. In other words, *dialexification* can be understood, in a way, as a shortcut for "DIAchronic (co)LEXIFICATION" – or more exactly, "lexification pattern arisen through diachronic change". That said, a key property of dialexification is that it is not necessarily a vertical relation from an etymon to its descendants: it is conceived as a possibly undirected, ahistorical relationship that links senses with each other, across a cognate set.

## 5.2 Methodological choices around dialexification

In sum, when the shared ancestor of a cognate set is reconstructed, the decision is to ignore its protogloss altogether, and focus on the sibling-like relation that exists among descendant meanings. As a corollary, dialexification is, in principle, agnostic of the directionality of change: this dimension is worthy of attention whenever it can be known, but is deliberately neutralised as we first wish to establish the network of cognate senses.

I propose to use braces and a colon {A : B} as a conventional way to indicate that meanings A and B are dialexified – contrasting with strict colexification {A = B} [§2.2]. To come back to our last example, we can write that the cognate set under PMP **banua* dialexifies the senses {VILLAGE : HOUSE : PEOPLE : YEAR : …}. By convention, we can shorten this formulation by saying that those senses are dialexified "by" PMP **banua*. This shortcut is acceptable, as long as we keep in mind that this is not a statement about the meaning of the etymon itself: it is only an empirical observation about the set of words that descend from it. Likewise, we can state that (the cognate set under) Latin *spathula* dialexifies the meanings { BACK : SCAPULA : SHOULDER : SPATULA } – here listed alphabetically, temporarily disregarding their ordering in time.

Importantly, colexification can be treated as a special case of dialexification, and is entirely contained within it. Indeed, all the meanings of a given word logically belong to the broader cognate set that includes that word. Thus we saw that Kapampangan *banwá* colexifies {SKY = YEAR}: this colex pair will therefore be contained in the broader set of senses dialexified under **banua*, namely {LAND : (…) SUN : SKY : YEAR}. All in all, dialexification is capable of capturing much broader semantic sets than what colexification can on its own. For the **banua* cognate set, we already cited ten concepts – more than what any of its reflexes may colexify alone.[18]

One important caveat concerns the scope of what we should define as a cognate set. The notion that two words are "etymologically related" is vague, and is often used to just mean "sharing the same root". For example, the Latin root **plic-* 'to fold' is at the source of many English words, including *display*, *deploy*, *employ*, *imply*, *apply*, *reply*, *complex*, *simple*, *multiple*… It is not wrong to say that the English verb *display* (< O.Fr. *desploiier* < Lat. *dis-plicare*) is etymologically related to Latin *plicare* 'to fold' – through a meaning 'unfold, unroll'. This semantic connection is obviously of great interest, but it would only be a case of "dialexification" if we took that term in quite a loose sense, that would ignore derivation or compounding. If we were

to conclude with the formula {FOLD → DISPLAY}, we would be overlooking the significant semantic contribution brought by the added morphology – in this case, the reversive Latin prefix *dis-*, capable of turning 'fold' into 'unfold'. In sum, just like we distinguish *loose* vs. *strict* colexification [§4], it is suitable to define *strict dialexification* as a relation among forms that reflect exactly the same etymon:

(8) *Strict dialexification* is the pattern whereby different meanings are attached to the same cognate set, defined as the set of forms that reflect the exact same etymon, with no added or removed derivational morphology.

As a corollary, an important distinction must be drawn between the root and the etymon. The ROOT is a minimal lexical constituent that forms the common denominator of many words, if we ignore the rest of their morphological make-up – e.g. the Latin root **plic-* 'to fold' (< PIE **pleḱ-* 'fold, plait'). By contrast, the ETYMON is a morphologically well-formed word, endowed with all its derivational (and inflectional) affixes. Every morphologically specific etymon is the source of its own cognate set, distinct from other etyma even if they share the same root. Strict dialexification does not operate at the level of the root, but of each etymon taken separately – see Table 1.

| Root | etymon | reflexes (sample) | dialexified senses (sample) |
|---|---|---|---|
| Latin ***plic-** 'fold' | plicare | It. *piegare*; Fr. *plier*, *ployer*; Rom. *pleca*… | FOLD : BEND : COVER : DEPART : FINISH |
| | implicare | It. *impiegare*; Fr. *employer*; Eng. *employ, imply*… | ENTANGLE : INVOLVE : USE : EMPLOY : HIRE |
| | displicare | Fr. *déplier*, *déployer*; Eng. *deploy*, *display* … | UNFOLD : DISPLAY : EXHIBIT : DEMONSTRATE |
| | replicare | It. *ripiegare*; Fr. *répliquer*; Eng. *reply*, *replicate*… | FOLD UP : REPLICATE : REPEAT : REPLY : ANSWER |

*Table 1* – Strict dialexification is observed for each etymon, i.e. a word endowed with its derivational morphology. Examples with four Latin etyma sharing the root **plic-*.

Finally, dialexification includes not just meanings obtained by strict inheritance, but also borrowings. Thus, Latin *implicare* is reflected by regularly inherited words like It. *impiegare* or Fr. *employer*; but there is no reason to exclude from that same cognate set the learned borrowings made by the same languages, resulting in etymological doublets (It. *implicare*, Fr. *impliquer*…); or those made by non-Romance languages, such as English *employ* and *imply*.

In sum, dialexification captures all the meanings that belong to the same etymological chain – whether these meanings are still colexified today, or have ended up separated by history and geography. A single cognate set may include a large number of reflexes, and far-reaching semantic extensions, accumulated over centuries of linguistic diversification from a common ancestor. Dialexification is a powerful tool to retrieve large networks of etymological links, based exclusively on empirical observations.

### 5.3 *Ετυ*Sem, a catalogue of etymological links

In order to conclude this overview of lexification studies, I will present *Ετυ*Sem, an online catalogue of etymological links.[19] Since I launched the idea in 2022, this project has allowed me to collaborate with Siva Kalyan, David Kletz, Martial Pastor, Mathieu Dehouck and Konstantin

Henke. The database was first released to the public in 2023, and was the object of a first publication (Dehouck et al. 2023); this section will summarise its essential features.

The *Evo*Sem database brings in one place tens of thousands of polysemous cognate sets from different families around the world, and displays their internal polysemy. For every etymon, the cognate set is displayed in the form of a semantic diagram labelled *etymograph* – a force-directed graph similar to the ones used by CLICS for colexification [§2.3]. Figure 7 shows the etymograph page of the PIE etymon **wṓkʷs* ('voice'+), the source of Eng. *voice*. Besides the graph itself, the page also includes a number of figures and tables, namely: a geographical map locating all the languages of the cognate set; one table (right) listing all reflexes of the etymon with their glosses, with links to our sources; and another table (bottom-right) recapitulating all the meanings dialexified in that set, ranking them by frequency of attestation within the set.[20]

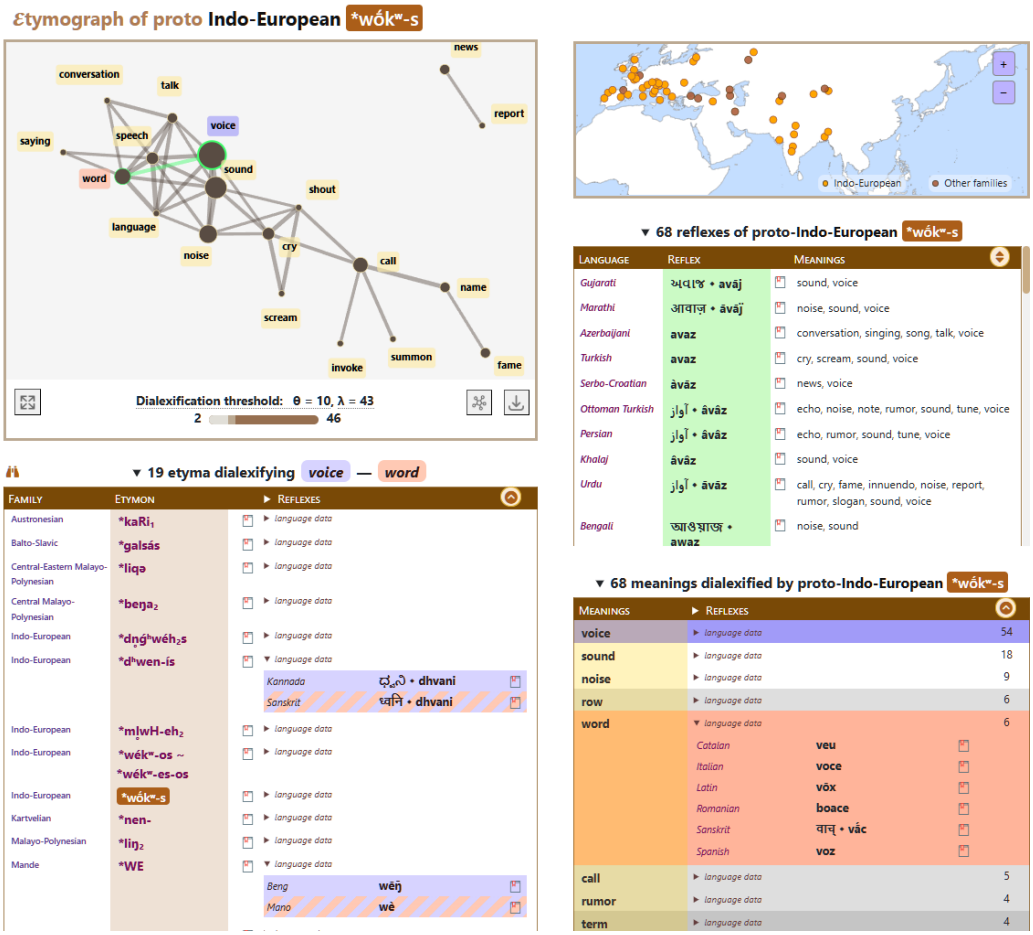


Figure 7 – Screenshot from the *Evo*Sem online database: Etymograph page of the PIE etymon **wṓkʷs* 'voice+', showing the 68 meanings dialexified by (i.e. attached to) its cognate set. Selecting a pair of senses, here {VOICE : WORD}, brings up a dialex table (bottom-left) listing all cognate sets with the same dialex pattern.

While the detailed display of every cognate set is valuable per se, the main power of *Evo*Sem lies in its comparative dimension. For any pair of meanings (or "concepts"), the database counts the number of cognate sets, across the entire database, that dialexify them together. This metric provides the "dialex score" $\delta$ (delta) for that pair. Thus in Figure 7, the pair {VOICE : WORD} selected on the graph yields a score $\delta = 19$: this means that these two concepts are dialexified by 19 different etyma across the world. These are listed in the dialex table

(bottom-left of Figure 7), where each cognate set can be opened and viewed separately. The blue colour refers to sense$_1$ (here VOICE), the pink one to sense$_2$ (WORD); the striped pattern means that Sanskrit *dhvani* and Mano *wè* colexify the two concepts.

The dialex score δ, calculated for every pair of concepts, is an important metric provided by $\mathcal{Evo}$Sem. On etymographs, every edge (link between two nodes) reflects the number δ through its thickness: the thicker an edge, the more cross-linguistically widespread the dialex pattern.[21] Under the etymon **wṓkʷs*, we find pairs of senses that score high across the whole database, reflecting a high degree of semantic closeness between the two concepts – e.g. $\delta_{\{VOICE:WORD\}} = 19$, $\delta_{\{VOICE:LANGUAGE\}} = 22$, $\delta_{\{VOICE:SOUND\}} = 31$, $\delta_{\{LANGUAGE:SPEECH\}} = 46$. Other links are weaker, because their dialexification is cross-linguistically less common: e.g. $\delta_{\{SOUND:CALL\}} = 8$, $\delta_{\{FAME:NEWS\}} = 4$, $\delta_{\{FAME:SOUND\}} = 3$. The latter links are not visible on Figure 7, because its "dialexification threshold" (below the graph) was set at $\theta = 10$: that is, in order to enhance the graph's legibility, the user has here chosen to display only links with $\delta \geq 10$. (The default threshold is $\theta = 2$.) The δ score underlies the self-organisation of the graph itself, as it determines whether two nodes will attract each other or drift apart. As the user chooses a higher threshold, fewer links (λ) materialise in the figure, allowing entangled graphs to open up.

The diameter of nodes in the graph is tied to a particular etymograph, as it shows the frequency of a gloss in this cognate set: thus Figure 7 shows that 54 reflexes of PIE **wōkʷs* mean 'voice', while only 6 mean 'word'. By contrast, the dialex score δ is independent of the etymon chosen: for any cognate set that dialexifies together *sense*$_1$ and *sense*$_2$, their link will be displayed with the same thickness on all graphs, reflecting the same cross-linguistic δ metric. This is an important point, because it means that δ has universal relevance. The $\mathcal{Evo}$Sem catalogue, taken at a given time of its development, is thus able to provide valuable insights into the semantic proximity between any two concepts – or more exactly, what may be called their "etymological proximity".

The universal metric δ is a component in the construction of our etymographs, for every cognate set; but it is also the key to a general semantic search engine, called $\mathcal{Evo}$Concepts. For any meaning (captured by an English gloss), the engine provides a list of its semantic "neighbours", i.e. the other meanings that are dialexified with it at least twice in our database. For instance, Figure 8 shows the results of a search on VOICE.[22] The engine finds that the concept VOICE has 129 semantic neighbours on $\mathcal{Evo}$Sem (left column), and lists them ranked by cross-linguistic frequency (δ score): e.g. the dialexification {VOICE : SOUND} is attested in δ=31 cognate sets. Concepts further down the list are less often associated with VOICE – but still instructive, precisely due to their relative rarity: e.g. SONG, NECK, GOSSIP, MOUTH… The user can colour-code certain meanings, and bring up all the relevant etyma on the table's right side. For example, Figure 8 shows that {VOICE : NECK} is a semantic link attested in five cognate sets anchored in the Pacific (Austronesian, Oceanic) and in Africa (Khoe-Kwadi, Mande); it is not attested in Europe, continental Asia or the Americas. Every cognate set can be opened within the table, or deployed in its own etymograph page.

Although $\mathcal{Evo}$Concepts uses English glosses (words or phrases), this does not mean that the language data were obtained through elicitation based on English concepts. Our sources are comparative dictionaries whose structure is always semasiological: they start from vernacular words, and like most dictionaries, describe their emic meanings using definitions and glosses that just happen to be in English. This approach makes sure that the semantic connections that are being described are not those of the English language, but of the vernacular lexicons. Once the dataset has been built, the $\mathcal{Evo}$Concepts search engine adopts the onomasiological angle,

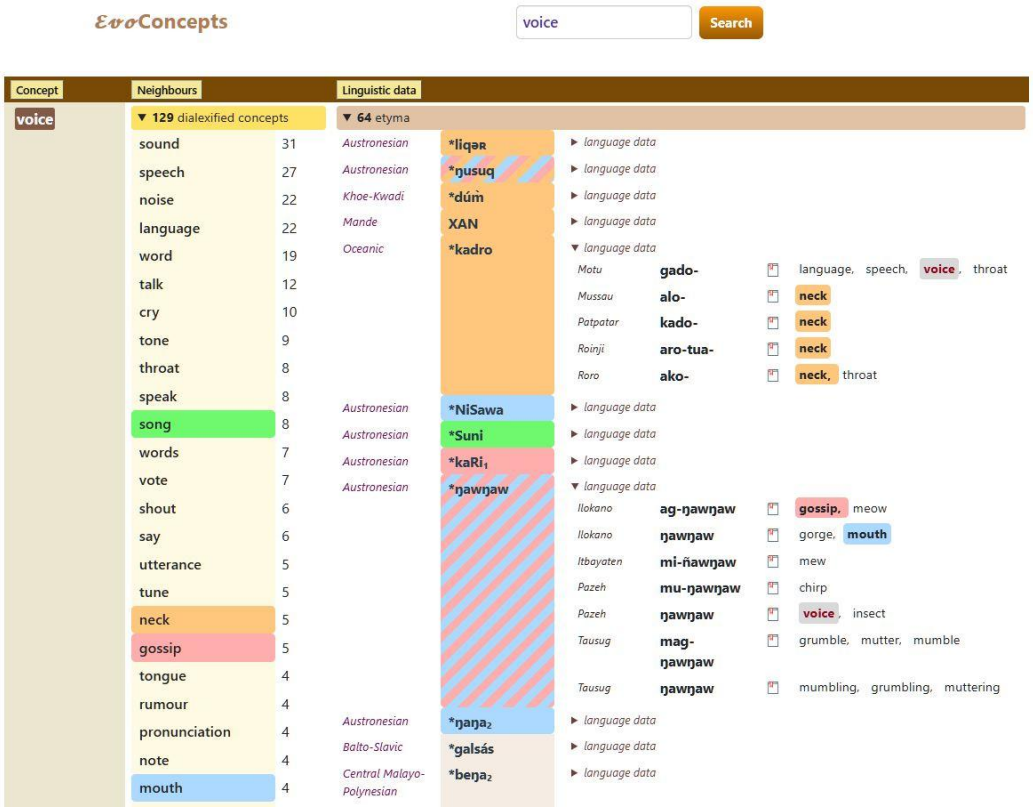


Figure 8 – Screenshot from the *Evo*Concepts search engine. Starting from a given concept (e.g. VOICE), the table lists its closest neighbours (left side). On the right side, it also gives access to all the cognate sets that include the meanings which the user has selected and colour-coded. A clickable icon gives access to our sources.

which we saw is essential to lexical typology [§2.2]. The English glosses, obtained by running definitions through a parsing script (Dehouck et al. 2023: 70–71), function as comparative concepts (Haspelmath 2010), which makes them a robust tool for cross-linguistic comparison.

Every element of information on *Evo*Sem (etymon, reflexes, senses…) can be traced back to our sources. These are several online databases that provide reliable information on etymological links around the world. One such source is the ACD, or *Austronesian Comparative Dictionary* (Blust et al. 2025); another is STEDT, the *Sino-Tibetan Etymological Dictionary and Thesaurus* (Matisoff 2016). A large portion of our dataset is provided by the etymological sections of Wiktionary, an open-access, online collaborative dictionary particularly well documented for Indo-European, Semitic, Turkic, Uralic languages; we incorporate the data through Wiktextract (Ylonen 2022). In order to enrich our family coverage, in 2025 we created a sister project to *Evo*Sem, named *Evo*Lex, an online series of comparative dictionaries for less documented language families of the world, in collaboration with experts of individual families. We have already published dictionaries of Khoe-Kwadi, a family from Austral Africa (Fehn et al. 2025), and of Mande, from Western Africa (Vydrin 2026); other dictionaries are currently in preparation. The *Evo*Lex project enhances the phylogenetic diversity of our *Evo*Sem database.

As of mid 2026, *Evo*Sem includes a total of 260,022 lexical forms from 3,345 languages – ancient and modern. These are distributed across 29,664 cognate sets, whose etyma belong to 170 protolanguages. Altogether, the number of meanings covered by *Evo*Concepts is 30,359. EvoSem 1.0, the version currently online, focused exclusively on reconstructed etyma – like PMP **banua* [§5.2] or PIE **wōkʷs*. A new version is in preparation at the time of writing, scheduled

to be released in late 2026: in addition to those starred reconstructions, EvoSem 2.0 will also include cognate sets stemming from attested words, in a broad array of source languages – including Ancient Greek, Latin, Arabic, Persian, Sanskrit or Chinese, among many. As a corollary, the 2.0 release will be much larger than the current version.

## 5.4 A tool of future discoveries

Finally, I will briefly mention a few ways in which 𝓔𝓿𝓸Sem can be useful to scholarship on lexical typology.

One first use of 𝓔𝓿𝓸Sem is to provide a new empirical method to assess semantic similarity among concepts. It joins other databases such as DatSemShift or CLICS as a useful resource to all linguists – whether cognitive or historical linguists, semanticists, typologists – who are interested in measuring semantic distances, and in mapping the conceptual space as it is shaped by human languages. It will then be a matter of more investigation to compare whether the sort of "etymological proximity" as measured by 𝓔𝓿𝓸Sem is congruent with the observations based on synchronic data. It is quite possible that the mechanics of etymology and semantic change defines a new kind of network, and a specific organisation of our semantic space. Because 𝓔𝓿𝓸Sem itself includes abundant data on every language in synchrony, it constitutes a vantage point from where to automatically compare lexical links internal to each language (colexification), vs. those that are only observable through cognacy ('pure' dialexification).

One use of 𝓔𝓿𝓸Sem concerns more specifically historical linguists who carry out lexical reconstruction in particular families. Should they hypothesise a potential semantic connection based on similarity of forms, yet find it dubious, they will find here a place where they can check its plausibility. Such connections as {VILLAGE : SKY}, {WATER : YEAR}, {FLAME : SPEECH}, {COUNTRY : WIFE} would seem iffy at first sight – but the 𝓔𝓿𝓸Sem catalogue establishes that they are actually dialexified at least twice in the world. Symmetrically, the tool provides sufficient empirical evidence to detect when a semantic hypothesis is weak and unlikely. In addition, it should be possible, for any two concepts, to calculate the minimal number of steps that separate them: e.g. it takes two steps to link SEA to BRIDGE (SEA – PATH – BRIDGE), or KING to SERVANT (KING – HUSBAND – SERVANT).

In line with ongoing research on colexification [§3.3], 𝓔𝓿𝓸Sem – specifically 𝓔𝓿𝓸Concepts – can also prove helpful in detecting the areal distribution of certain patterns. Thus, an exploratory study (Dehouck et al. 2025) showed that many dialex patterns – e.g. {WATER : LAKE} – were quite evenly distributed across the world's families, while others show an areal or phylogenetic bias. Using a probabilistic methodology based on binomial distribution, we compared the distribution of dialex patterns across different families, looking for probabilistic outliers. For example, the Indo-European (IE) and Pama-Nyungan (PN, Australia) families each have 7 cognate sets that include the meaning TONGUE; among these, the sense FLAME is also found in 1 etymon for IE, but 6 for PN. In other words, the dialex pattern {TONGUE : FLAME} "overdialexifies" in PN (6/7 score) compared to other families: this semantic link is thus an areal feature of Australia. (The dialexification is here based on a metaphor of flames as 'tongues of fire'; the further connection of TONGUE with SPEECH explains the existence of a rare dialex pattern {FLAME : SPEECH}.) The same Pama-Nyungan family also overdialexifies {FIRE : FIREWOOD} (with a 6/6 score), thereby corroborating the initial observations of Schapper et al. (2016) [see Fig. 2 in §3.3]. The areal signal can also be negative. For example, the well-known {TONGUE : LANGUAGE} connection is widespread across Eurasia (AGr. γλῶσσα, Lat. *lingua*, Kurdish *ziman*…) and Africa

(Hausa *harshe*, Ewe *aɖe*, Zulu *ulimi*…), but it is virtually unknown in the Americas and the Pacific. Likewise, we saw how Figure 8 points to a skewed areal distribution for {NECK : VOICE}.

Finally, given that 𝓔𝓿𝓸Sem 2.0 is bound to include data from attested source languages – and not just reconstructions – it will provide valuable empirical information on attested cases of semantic change: from Ancient to Modern Greek, from Latin to Romance languages and beyond, from Arabic or Persian to the many languages that borrowed from them… Due to the focus of 𝓔𝓿𝓸Sem 1.0 on reconstructed etyma, we had decided to circumvent protoglosses due to their speculative status [§5.1]; this methodological precaution was even the trigger for defining the very notion of dialexification – an etymological link that can be agnostic of the direction of change. Now, the wealth of historical data retrievable from 𝓔𝓿𝓸Sem 2.0 will make it methodologically safe to observe actual paths of change. For a large number of dialex patterns, it will be possible to observe their historical ordering, and observe whether all paths agree in their directionality [see §4.2]. Just like some cases of sound change can be highly directional (e.g. *s → h, but ??*h → s) while others are more ambiguous (e.g. *r ↔ *l), likewise an empirical question would consist in determining, in lexical change, which paths are more likely to follow a directional slope, and which ones are more ambivalent. When a cognate set dialexifies {SEE : KNOW}, which meaning is more likely to be the source: SEE or KNOW? Such questions are central to historical reconstruction, and need to be addressed through a testable, non-speculative methodology. The large dataset of 𝓔𝓿𝓸Sem 2.0 should make it easier to study this question.

These are just a few examples of different ways the database could be exploited. 𝓔𝓿𝓸Sem has the potential to inspire many new studies of lexical typology in years to come.

## 6 Conclusion

The development of lexical typology in the last two decades has shown how much can be gained by shifting our attention from language-specific lexical categories to the relations that hold between meanings and forms across languages. The onomasiological turn taken by scholars in the late 2000s proposed to anchor cross-linguistic observation on individual meanings, taken as comparative concepts; this released the possibility of carrying out typological comparison on a large scale. Colexification has stood the test of time, in providing powerful ways to capture a number of cross-linguistic patterns. Yet its focus on synchronic lexical systems leaves open a further question: What can we learn from the historical relations among words that have sprung from a common source, yet no longer belong to the same language?

A possible answer may come from the notion of *dialexification* proposed here, and the new attention paid to *cognate sets* rather than individual words. While they are traditionally built to serve as evidence for studying sound change or for reconstructing etyma, cognate sets really have a much deeper significance: they preserve the memory of far-reaching networks of form–meaning associations that can span over continents and millennia. The semantic study of cognate sets, here labelled dialexification, has the potential to bring etymological information into the typological study of lexical semantics. Unlike the notion of semantic change, dialexification is deliberately designed to be agnostic about the direction of semantic paths, whenever this is needed to preserve its empirical tenets. The aim of dialexification is to detect semantic links, ancient or modern, through the traces they have left in today's languages. The creation of 𝓔𝓿𝓸Sem, a large database organised around this principle, opens a new methodological space in which synchronic colexification and diachronic dialexification can be studied together, compared across language families and areas, and used to investigate how associations between meanings emerge, persist, split, merge, and disappear.

Finally, patterns of lexification are arguably one instance of a more general question: namely, how form–meaning relations are packaged in languages. Lexical typology has mostly been taking individual words or lexicalised phrases as its unit of observation: this is indeed the angle that best lends itself to the efficient mining of comparative dictionaries and other lexical datasets. However, polysemies and semantic extensions are also found in other corners of language systems: in culturally-entrenched conceptual metaphors, idioms and phraseology (Schapper 2021:343), and of course in grammar and grammaticalisation (e.g. Traugott & Heine 1991; Bisang & Malchukov 2017). Advocates of Construction Grammar (Fillmore et al. 1988, Goldberg 1995) have argued that meaning really pertains to a broader continuum that encompasses grammar and lexicon: it is embodied in conventionalised form–meaning pairings known as *constructions*. In this perspective, current methods in lexical typology could be expanded and adapted to investigate the typology of constructional patterns. Given two semantic functions, lexical or grammatical, we could systematically test whether they are formally coexpressed or not in different languages – that is, whether they are *co-constructed* or *dis-constructed*.[23]

Beyond lexical typology, such a proposal could open up inspiring new directions for semantic typology more generally. Yet if we wish to define an object broader than the lexicon, we will also need to develop methodological tools and resources capable of tackling cross-linguistic comparison on a large scale, in a way comparable to what has already been achieved in the domain of lexical typology.

## Abbreviations

*Languages*

AGr: Ancient Greek – Eng: English – Fr: French – Grm: German – It: Italian – Lat: Latin – O.Fr: Old French – PIE: Proto-Indo-European – Ptg: Portuguese – Rom: Romanian – Rus: Russian – Sp: Spanish – ST: Sino-Tibetan.

*Countries*

Indon: Indonesia – Sol.: Solomon Islands – Phil: Philippines – Van: Vanuatu

## References

Aikhenvald, Alexandra Y. & R. M. W. Dixon. 2026. The mighty moon: colexification and change in northwest Amazonia. *Journal of Linguistics* 1–28.

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, et al. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1). 1–18.

Berlin, Brent & Paul Kay. 1969. *Basic color terms: their universality and evolution*. Berkeley, Calif.: Univ. of California Press.

Bisang, Walter & Andrej L. Malchukov (eds.). 2017. *Unity and diversity in grammaticalization scenarios* (Studies in Diversity Linguistics). Berlin: Language Science Press.

Blank, Andreas. 1999. Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change. In Andreas Blank & Peter Koch (eds.), *Historical Semantics and Cognition* (Cognitive Linguistics Research), 61–90. Berlin, New York: Mouton de Gruyter.

Blust, Robert A. 1987. Lexical reconstruction and semantic reconstruction: The case of Austronesian "house" words. *Diachronica* 4(1–2). 79–106.

Blust, Robert, Stephen Trussel, Robert Forkel & Alexander D. Smith. 2025. Austronesian Comparative Dictionary. Zenodo.

Bréal, Michel. 1897. *Essai de sémantique (Science des significations)*. 1st edn. Paris: Hachette.

Clark, Ross. 2009. **Leo Tuai: A comparative lexical study of North and Central Vanuatu languages.* (Pacific Linguistics, 603.) Canberra: Australian National University.

Coseriu, Eugenio. 1964. Pour une sémantique diachronique structurale. In *Travaux de linguistique et de littérature*, vol. II, 1, 139–186. Strasbourg: Université de Strasbourg.

Cruse, David A. 2001. *Lexical semantics* (Cambridge Textbooks in Linguistics). Reprint. Cambridge: Cambridge Univ. Press.

Dehouck, Mathieu, Alexandre François, Siva Kalyan, Martial Pastor & David Kletz. 2023. EvoSem: A database of polysemous cognate sets. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, 66–75. Singapore: Association for Computational Linguistics.

Dehouck, Mathieu, Alexandre François & Tom Poulton. 2025. Disentangling areal and hereditary signals in colexification and dialexification patterns. Workshop *Disentangling contact and inheritance in lexical semantics* (Societas Linguistica Europaea). Université Bordeaux Montaigne.

Dellert, Johannes. 2016. Using causal inference to detect directional tendencies in semantic evolution. In Seán Roberts, Christine Cuskley, Luke McCrohon, Lluís Barceló-Coblijn, Olga Feher & Tessa Verhoef (eds.), *The Evolution of Language: Proceedings of the 11th International Conference*, 9. New Orleans.

Di Natale, Anna, Max Pellert & David Garcia. 2021. Colexification networks encode affective meaning. *Affective Science* 2.2: 99-111.

Ding, Hongdi & Sicong Dong. 2024. Colexification of “thunder” and “dragon” in Sino-Tibetan languages. *Asia Pacific Translation and Intercultural Studies* 1–23.

Du Marsais, César Chesneau. 1730. *Des tropes, ou Des diferens sens dans lesquels on peut prendre un mème mot dans une mème langue. Ouvrage utile pour l’intelligence des auteurs, & qui peut servir d’introduction à la rhétorique & à la logique*. Paris: chez la Veuve de Jean-Batiste Brocas.

Evans, Nicholas. 2010. Semantic typology. In Jae Jung Song (ed.), *The Oxford Handbook of Linguistic Typology* (Oxford Handbooks in Linguistics), 504–533. Oxford: Oxford University Press.

Evans, Nicholas & David Wilkins. 2000. In the mind’s ear: The semantic extensions of perception verbs in Australian languages. *Language* 76(3). 546–592.

Fehn, Anne-Maria, Anna Smirnitskaya, Tshisimogo Leepang, Hirosi Nakagawa, Admire Phiri & Jorge Rocha. 2025. *A comparative dictionary of Khoe-Kwadi languages*. $\mathcal{E}\mathcal{v}\mathcal{o}\mathbb{L}\mathbb{e}\mathbb{x}$ Comparative Dictionaries. [online: https://evolex.huma-num.fr/]

Fillmore, Charles, Paul Kay & Mary Catherine O’Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of ‘let alone.’ *Language* 64(3). 501–538.

Fontanier, Pierre. 1825. *Manuel classique pour l’étude des tropes, ou Élémens de la science du sens des mots*. 3rd edn. Paris: Maire-Nyon.

François, Alexandre. 2008. Semantic maps and the typology of colexification: Intertwining polysemous networks across languages. In Martine Vanhove (ed.), 163–215.

—— 2013. Shadows of bygone lives: The histories of spiritual words in northern Vanuatu. In Robert Mailhammer (ed.), *Lexical and Structural Etymology* (Studies in Language Change), 185–244. Berlin, Boston: DeGruyter Mouton.

—— 2021. Olfactory words in northern Vanuatu: *Langue* vs. *parole*. In Łukasz Jędrzejowski & Przemysław Staniewski (eds.), *The Linguistics of Olfaction: Typological and diachronic approaches to synchronic diversity* (Typological Studies in Language), vol. 131, 277–304. Amsterdam: Benjamins.

—— 2022a. Awesome forces and warning signs: Charting the semantic history of *tabu words in Vanuatu. *Oceanic Linguistics* 61(1). 212–255.

—— 2022b. Lexical tectonics: Mapping structural change in patterns of lexification. *Zeitschrift für Sprachwissenschaft* 41(1). 89–123.

—— 2023. About dialexification. Paper presented at the *Colloquium on language typology, on the occasion of Martin Haspelmath’s 60th birthday*, Universität Potsdam. March 2023.

François, Alexandre & Siva Kalyan. 2023. Dialexification: A tool for studying cross-linguistic patterns of semantic change. Paper presented at the 16th International Cognitive Linguistics Conference (ICLC16), Heinrich-Heine-Universität, Düsseldorf. Aug 2023.

François, Jacques. 2013. Semantische Karten für die vergleichende lexikologische Analyse europäischer Sprachen. *Journal for EurolinguistiX* 10. 15–28.

Gast, Volker & Maria Koptjevskaja-Tamm. 2018. The areal factor in lexical typology. In Daniël Olmen, Tanja Mortelmans & Frank Brisard (eds.), *Aspects of Linguistic Variation*, 43–82. Berlin, Boston: De Gruyter.

Geeraerts, Dirk. 1997. *Diachronic Prototype Semantics: A contribution to historical lexicology*. Oxford: Oxford University Press.

Georgakopoulos, Thanasis. 2019. Semantic Maps. In *Linguistics*. Oxford University Press.

Georgakopoulos, Thanasis, Eitan Grossman, Dmitry Nikolaev & 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.

Georgakopoulos, Thanasis, Eliese-Sophia Lincke, Kiki Nikiforidou & Anna Piata. 2020. On the polysemy of motion verbs in Ancient Greek and Coptic: Why lexical constructions are important. *Studies in Language* 44(1). 27–69.

Georgakopoulos, Thanasis & Kiki Nikiforidou. 2026. Constructions in semantic maps: Evidence from the polysemy of motion verbs. *Journal of Greek Linguistics* 26(1). 99–129.

Georgakopoulos, Thanasis & Stéphane Polis. 2018. The semantic map model: State of the art and future avenues for linguistic research. *Language and Linguistics Compass* 12(2). e12270.

—— 2021. Lexical Diachronic Semantic Maps. Mapping the evolution of time-related lexemes. *Journal of Historical Linguistics*.

—— 2022. New avenues and challenges in semantic map research (with a case study in the semantic field of emotions). *Zeitschrift für Sprachwissenschaft* 41(1). 1–30.

Gfeller, Kim, Sabine Stoll, Chundra Cathcart & Paul Widmer. 2026. Investigating lexical change through crosslinguistic colexification patterns. *Language Dynamics and Change* 16(1). 1–30.

Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago: The Univ. of Chicago Press.

Gruntov, Ilya & Elisei Rykov. 2023. Computer-assisted detection of typologically relevant semantic shifts in world languages. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2023."* RSUH.

Haspelmath, Martin. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In Michael Tomasello (ed.), *The new psychology of language*, vol. 2, 211–243. New York: Erlbaum.

—— 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86(3). 663–687.

—— 2023. Coexpression and synexpression patterns across languages: Comparative concepts and possible explanations. *Frontiers in Psychology* 14. 1236853.

Jackson, Joshua Conrad, Joseph Watts, Teague R. Henry, Johann-Mattis List, Robert Forkel, Peter J. Mucha, Simon J. Greenhill, Russell D. Gray & Kristen A. Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. *Science* 366(6472). 1517–1522.

Jedrzejowski, Lukasz & Przemyslaw Staniewski (eds.). 2021. *The linguistics of olfaction: Typological and diachronic approaches to synchronic diversity* (Typological Studies in Language 131). Amsterdam: John Benjamins.

Kopecka, Anetta & Bhuvana Narasimhan (eds.). 2012. *Events of putting and taking: a crosslinguistic perspective* (Typological Studies in Language 100). Amsterdam: John Benjamins.

Koptjevskaja-Tamm, Maria (ed.). 2015. *The linguistics of temperature* (Typological Studies in Language 107). Amsterdam: John Benjamins.

Koptjevskaja-Tamm, Maria. 2022. Semantic maps and temperature: Capturing the lexicon-grammar interface across languages. *Zeitschrift für Sprachwissenschaft* 41(1). 125–177.

Koptjevskaja-Tamm, Maria, Martine Vanhove & Peter Koch. 2007. Typological approaches to lexical semantics. *Linguistic Typology* 11(1). 159–186.

Lakoff, George & Mark Johnson. 1980. Conceptual metaphor in everyday language. *The Journal of Philosophy* 77(8). 453–486.

Langacker, Ronald. 1987. *Foundations of Cognitive Grammar*. Vol. 1 : "Theoretical prerequisites". Stanford: Stanford University Press.

List, Johann-Mattis, Michael Cysouw & Robert Forkel. 2016. Concepticon: A resource for the linking of concept lists. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation, May 23-28, 2016, Portorož, Slovenia*, 8.

List, Johann-Mattis, Katja Bocklage, & Thanasis Georgakopoulos. 2026. The ambiguity of the colexification term. In *Computer-Assisted Language Comparison in Practice*, 9.2: 85-92 [first published on 19/08/2026], URL: https://calc.hypotheses.org/9547.

Majid, Asifa. 2012. Current Emotion Research in the Language Sciences. *Emotion Review* 4(4). 432–443.

Majid, Asifa, James S. Boster & Melissa Bowerman. 2008. The cross-linguistic categorization of everyday events: A study of cutting and breaking. *Cognition* 109(2). 235–250.

Matisoff, James A. 2016. STEDT: The Sino-Tibetan Etymological Dictionary and Thesaurus. University of California, Berkeley. https://stedt.berkeley.edu/

Newman, John (ed.). 2009. *The linguistics of eating and drinking* (Typological Studies in Language 84). Amsterdam Philadelphia: John Benjamins Publishing Company.

Norcliffe, Elisabeth & Asifa Majid. 2024. Verbs of perception: A quantitative typological study. *Language* 100(1). 81–123.

Östling, Robert. 2016. Studying colexification through massively parallel corpora. In Päivi Juvonen & Maria Koptjevskaja-Tamm (eds.), *The Lexical Typology of Semantic Shifts* (Cognitive Linguistics Research 58), 157–176. Berlin: Walter de Gruyter.

Paul, Hermann. 1880 [2010]. *Prinzipien der Sprachgeschichte* (Konzepte der Sprach- und Literaturwissenschaft 6). 10th edn. Berlin: De Gruyter.

Pawley, Andrew K. 1985. Proto-Oceanic terms for 'person': A problem in semantic reconstruction. *Oceanic Linguistics Special Publications* 20. 92–104.

—— 2005. The meaning(s) of Proto Oceanic *panua. In Claudia Gross, Harriet D. Lyons & Dorothy A. Counts (eds.), *A polymath anthropologist: Essays In honour of Ann Chowning*, 211–223. Auckland: University of Auckland.

Pericliev, Vladimir. 2015. On Colexification among Basic Vocabulary. *Journal of Universal Language* 16(2). 63–93.

Ponsonnet, Maïa. 2025. The semantic typology of expressive interjections: Colexifications in pain, disgust and joy interjections across languages. *Lingua* 324. 103979.

Rakhilina, Ekaterina & Tatiana Reznikova. 2016. A Frame-based methodology for lexical typology. In Päivi Juvonen & Maria Koptjevskaja-Tamm (eds.), *The Lexical Typology of Semantic Shifts*, 95–129. Berlin, Boston: De Gruyter.

Rakhilina, Ekaterina V., Tatiana I. Reznikova & Daria Ryzhova (eds.). 2022. *The typology of physical qualities* (Typological Studies in Language 133). Amsterdam: John Benjamins.

Rankin, Robert. 2003. The Comparative Method. In Brian D. Joseph & Richard D. Janda (eds.), *The Handbook of Historical Linguistics*, 183–212. Oxford: Blackwell.

Ryzhova, Daria, Ekaterina Rakhilina, Tatiana Reznikova & Yulia Badryzlova. 2024. Lexical systems with systematic gaps: verbs of falling. *Folia Linguistica* 58(1). 191–226.

Rzymski, Christoph, Tiago Tresoldi, Simon J. Greenhill, Mei-Shin Wu, Nathanael E. Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, et al. 2020. The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data* 7(1). 13

San Roque, Lila, Kobin H. Kendrick, Elisabeth Norcliffe & Asifa Majid. 2018. Universal meaning extensions of perception verbs are grounded in interaction. *Cognitive Linguistics* 29(3). 371–406.

Saussure, Ferdinand de. 1916 [1972]. *Cours de linguistique générale*. Paris: Payot.

Schapper, Antoinette. 2019. The ethno-linguistic relationship between smelling and kissing: A Southeast Asian case study. *Oceanic Linguistics* 58(1). 92–109.

—— 2021. Baring the bones: the lexico-semantic association of bone with strength in Melanesia and the study of colexification. *Linguistic Typology* 26(2). 313–347.

Schapper, Antoinette, Lila San Roque & Rachel Hendery. 2016. Tree, firewood and fire in the languages of Sahul. In Päivi Juvonen & Maria Koptjevskaja-Tamm (eds.), *The Lexical Typology of Semantic Shifts*, 355–422. Berlin, Boston: De Gruyter.

Stathi, Katerina. 2023. *Granularity in the verbalization of events and objects: a cross-linguistic study* (Studies in Language Companion Series volume 233). Amsterdam: Benjamins.

Sweetser, Eve E. 1990. *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure* (Cambridge Studies in Linguistics 54). Cambridge: Cambridge University Press.

Talmy, Leonard. 1983. How language structures space. In Herbert Pick & Linda Acredolo (eds.), *Spatial orientation: theory, research and application*. New York: Plenum Press.

—— 1985. Lexicalization patterns: Semantic structure in lexical forms. In Timothy Shopen (ed.), *Language Typology and Syntactic Description*, vol. 3, 36–149. Cambridge: CUP.

Tjuka, Annika, Robert Forkel, Christoph Rzymski & Johann-Mattis List. 2025. Advancing the Database of Cross-Linguistic Colexifications with new workflows and data. In *Proceedings of the 16th International Conference on Computational Semantics*, 1–15. Düsseldorf: ACL.

Tjuka, Annika, Robert Forkel, Christoph Rzymski & Johann-Mattis List. 2026. CLICS$^4$: An improved database of cross-linguistic colexifications. Zenodo. DOI 10.5281/zenodo.19608447.

Traugott, Elizabeth C. & Bernd Heine (eds.). 1991. *Approaches to grammaticalization* (Typological Studies in Language 19). 2 vols. Amsterdam: Benjamins.

Traugott, Elizabeth Closs & Richard B. Dasher. 2001. *Regularity in Semantic Change* (Cambridge Studies in Linguistics 97). Cambridge: Cambridge University Press.

Urban, Matthias. 2015. Lexical semantic change and semantic reconstruction. In Claire Bowern & Bethwyn Evans (eds.), *The Routledge handbook of historical linguistics* (Routledge Handbooks in Linguistics), 374–392. New York: Routledge.

Vanhove, Martine (ed.). 2008. *From Polysemy to Semantic Change: Towards a Typology of Lexical Semantic Associations* (Studies in Language Companion Series 106). Amsterdam: John Benjamins.

Vydrin, Valentin. 2026. *A comparative dictionary of Mande languages*. $\mathcal{E}\mathcal{v}\mathcal{o}\mathbb{L}\mathbb{e}\mathbb{x}$ Comparative Dictionaries. [online: https://evolex.huma-num.fr/]

Wälchli, Bernhard & Michael Cysouw. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics* 50(3).

Weiss, Michael. 2014. The Comparative Method. In Claire Bowern & Bethwyn Evans (eds.), *The Routledge Handbook of Historical Linguistics*, 127–145. New York: Routledge.

Wilkins, David. 1996. Natural tendencies of semantic change and the search for cognates. In Mark Durie & Malcolm Ross (eds.), *The comparative method reviewed*, 264–304. Oxford: OUP.

Ylonen, Tatu. 2022. Wiktextract: Wiktionary as machine-readable structured data. In *Proceedings of the 13th Language Resources and Evaluation Conference*, 1317–1325. Marseille: European Language Resources Association.

Zalizniak, Anna A. 2008. A catalogue of semantic shifts: Towards a typology of semantic derivation. In Vanhove (ed.), 217–232.

—— 2018. The Catalogue of Semantic Shifts: 20 Years Later. *Russian Journal of Linguistics* 22(4). 770–787.

Zalizniak, Anna A., Maria Bulakh, Dmitrij Ganenkov, Ilya Gruntov, Timur Maisak & Maxim Russo. 2012. The catalogue of semantic shifts as a database for lexical semantic typology. *Linguistics* 50(3).

## Notes

[1] I am grateful to the editors of this volume for their invitation, and to Konstantin Henke for his useful comments on a previous version of this chapter. This work pertains to the scientific project "Historical and comparative linguistics" of the French program HéLiCéO, funded by CNRS, ENS and ANR (ANR-24-RRII-0001). The present study is dedicated to the memory of Andrew Pawley (1941–2026), with whom I had the privilege to hold regular conversations about etymologies across the world's families. These inspiring moments no doubt played an important role in my desire to further explore this domain over the years.

[2] Even though I am the author of the 2008 article on colexification, I will refer to it in the third person, like with other references I cite: this is because I view it as one reference among many that have contributed, over the years, to the domain of lexical typology. I will only resort to first person in the final sections of this chapter (Section 0), as I present my current ideas and projects.

[3] Throughout this study, I will use "senses" and "meanings" interchangeably.

[4] "Les concepts (...) sont purement différentiels, définis non pas positivement par leur contenu, mais négativement par leurs rapports avec les autres termes du système".

[5] Thus Tjuka et al. (2026:1) define colexification as "a cover term that unifies the notions of polysemy, homophony and underspecification, referring to cases where a single word form in a given language expresses multiple senses (François, 2008)".

[6] While the prefix *co(n)-* expresses similarity or union, the prefix *dis-* encodes the idea of difference or separation: compare *converge* vs. *diverge*, *concord* vs. *discord*, *consent* vs. *dissent*. Latin *dis-* must not be confused with the prefix *dys-* of Greek origin, encoding difficulty, as in *dyspnea* or *dyslexia*.

[7] Link: https://clics.clld.org (accessed 8 May 2026).

[8] The lexical connection SMELL–KISS is also sporadically attested in Melanesia, along with the cultural habit it reflects (François 2021:279).

[9] Among these four types of relation, the first three (polysemy, diachronic change, cognate links) will be subsumed under the notion of *dialexification*, to be defined in Section 0.

[10] Link to the map: https://datsemshift.ru/map242 (accessed 27 July 2026).

[11] See https://datsemshift.ru/stat (accessed 27 July 2026).

[12] Jacques François (2013) proposed similar maps for some polysemies of French and German.

[13] This form **vanua* reflects an etymon **banua* of Proto-Malayo-Polynesian, to be discussed in §0.

[14] For abbreviations of country names, see the Appendix.

[15] See the entry **banua* https://acd.clld.org/cognatesets/25129 in the *Austronesian comparative dictionary* (Blust et al. 2025). The same root was mentioned (under the form **vanua*) in Figure 6 above.

[16] The method presented by Pawley (2005) – partly based on semantic maps – to reconstruct a complex network of senses for **banua* later inspired me to reconstruct, in a similar way, the polysemy of other culturally significant Proto-Oceanic etyma – such as **qata-mate* 'ghost, spirit, ogre, shaman+' (François 2013:214) or **tabu* 'taboo, forbidden, sacred, numinous+' (François 2022b).

[17] While I have used the term on previous occasions, both spoken (François 2023, François & Kalyan 2023) and written (Dehouck et al. 2023), the present chapter is the first where I define and discuss the notion at length.

[18] The entry **banua* https://tiny.cc/EvoSem_PMP_banua in the $\mathcal{Evo}$Sem database lists many more meanings dialexified under that etymon.

[19] It is available at https://evosem.huma-num.fr.

[20] See https://tiny.cc/EvoSem_PIE_wokws (accessed 20 July 2026).

[21] This is reminiscent of how colexification frequency is represented on CLICS [see §0]. A difference is that, if the same colexification is shared by several words that are cognate, it will be counted only once in $\mathcal{Evo}$Sem.

[22] See https://tiny.cc/EvoConcepts_voice.

[23] These terms were suggested to me in conversations with Konstantin Henke, co-designer of $\mathcal{Evo}$Sem 2.0.