Linguistic Typology 2018; 22(2): 277–306

Johann-Mattis List*, Simon J. Greenhill, Cormac Anderson,
Thomas Mayer, Tiago Tresoldi, Robert Forkel

CLICS^2: An improved database of cross-linguistic colexifications assembling lexical data with the help of cross-linguistic data formats

https://doi.org/10.1515/lingty-2018-0010
Received March 07, 2018; revised May 17, 2018

Abstract: The Database of Cross-Linguistic Colexifications (CLICS), has established a computer-assisted framework for the interactive representation of cross-linguistic colexification patterns. In its current form, it has proven to be a useful tool for various kinds of investigation into cross-linguistic semantic associations, ranging from studies on semantic change, patterns of conceptualization, and linguistic paleontology. But CLICS has also been criticized for obvious shortcomings, ranging from the underlying dataset, which still contains many errors, up to the limits of cross-linguistic colexification studies in general. Building on recent standardization efforts reflected in the *Cross-Linguistic Data Formats* initiative (CLDF) and novel approaches for fast, efficient, and reliable data aggregation, we have created a new database for cross-linguistic colexifications, which not only supersedes the original CLICS database in terms of coverage but also offers a much more principled procedure for the creation, curation and aggregation of datasets. The paper presents the new database and discusses its major features.

***Corresponding author: Johann-Mattis List [ˈjoːhan ˈmatʰɪs lɪstʰ],** Department of Linguistic and Cultural Evolution, MPI-SHH, Jena, Germany, E-mail: mattis.list@shh.mpg.de
Simon J. Greenhill [ˈsaimən ˈgriːnhɪl], Department of Linguistic and Cultural Evolution, MPI-SHH, Jena, Germany; ARC Centre of Excellence for the Dynamics of Language, Australian National University, Canberra, Australia, E-mail: greenhill@shh.mpg.de
Cormac Anderson [ˈkʰʊrməkʰ ˈændərsən], Department of Linguistic and Cultural Evolution, MPI-SHH, Jena, Germany, E-mail: anderson@shh.mpg.de
Thomas Mayer [ˈtʰoːmas ˈmajɐ], Independent Researcher, Munich, Germany, E-mail: thommy.mayer@gmail.com
Tiago Tresoldi [tiˈago treˈsɔldi], Department of Linguistic and Cultural Evolution, MPI-SHH, Jena, Germany, E-mail: tresoldi@shh.mpg.de
Robert Forkel [ˈʁoːbɐtʰ ˈfɔɐkʰl̩], Department of Linguistic and Cultural Evolution, MPI-SHH, Jena, Germany, E-mail: forkel@shh.mpg.de

1 Introduction

When linguists succeed in identifying similar patterns across various languages, a number of different types of explanations for these similarities are possible. Common patterns might result from *coincidence*; from *natural reasons* with a basis in human cognition, psychology or physiology, or indeed the nature of our environment; or they might derive rather from historical processes, among which it is customary to differentiate *inheritance* and *contact*. Traditionally, coincidence is not considered linguistically interesting, while natural reasons are the primary focus of work in *linguistic typology*, as well as providing the theoretical grounding for various universal frameworks of grammatical architecture, while inheritance and contact fall into the purview of *historical linguistics*. However, to draw a clear dividing line between these two subfields is not necessarily useful or even tenable, as the research results of each feed back into the other, and unpicking which of the various explanations for a given phenomenon is most convincing is not always an easy task. Nowhere is this as clear as in the domain of *lexical typology*.

Languages differ in how they label the universe and sometimes these labels clash in interesting and informative ways, such that one word form has multiple meanings. This may result from coincidence, termed *homophony*, whereby multiple meanings for one word form arise accidentally as two word forms come to sound alike, as in French *paix* ‘peace’ vs. *pet* ‘fart’, which are both pronounced as [pɛ]. In contrast to this are cases of *polysemy*, in which one word form comes to have multiple related senses, as in Russian дерево *dérevo*, which can denote both ‘tree’ and ‘wood’.

Cases of polysemy may be cross-linguistically frequent, in which case an explanation can likely be found in natural factors, be they linked to some aspect of human psychology or cognition, or the inherent structure of the natural environment (e.g. ‘rain’ and ‘water’, the above example of ‘tree’ and ‘wood’, or the common colexification of ‘moon’ and ‘month’). On the other hand, where a polysemic pattern is relatively rare cross-linguistically, this is likely to point to a historical explanation in common inheritance or contact. For example, many Austronesian and Papuan languages in eastern New Guinea and northern Australia use the same term for both ‘fire’, ‘firewood’, and ‘tree’. As this pattern is rare world-wide, this hints that there might be some deep connection between these groups across the Torres Strait (Schapper et al. 2016). Another case is given by Urban (2010), who notes that the word for ‘sun’ can typically be translated as ‘eye of the day’ in many Austroasiatic, Tai-Kadai, and Austronesian languages. In spite of the fact that a diachronic development based on a similar equation is attested in Indo-European (e.g. Old Irish *súil* ‘eye’, from the PIE root *seh$_2$l-, thus cognate with Latin *sōl* ‘sun’, see Vaan 2008: 570; Classical Armenian արեգակն *aregakn*, a compound of *arew* ‘sun’ and *akn* ‘eye’, see Olsen 2002), the relative cross-linguistic rarity of this pattern and its prevalence in Southeast Asia suggests an explanation in terms of historical factors.

Deciding on a natural or historical explanation (i.e. distinguishing between homophony and polysemy) may be relatively straightforward for small groupings of languages for which detailed etymological and historical knowledge is available, but it becomes increasingly difficult on a larger scale, and impossible where detailed historical information is unknown. To circumvent this problem, scholars have increasingly begun to use the agnostic cover term *colexification*, where two senses in a given language *colexify* if the language uses the same lexical form for both (François 2008). Taking a colexification approach enables scholars to approach questions of lexical semantics from the perspective of the data: if a pattern of colexification of certain meanings in one language is replicated across different languages or linguistic areas, that is indicative (if not diagnostic) of polysemy, rather than homophony (List et al. 2013). However, if frequency is to be used in this way as a proxy to infer polysemy, reliable large-scale cross-linguistic colexification resources are required. The revised CLICS database outlined in this paper is one such resource.[1]

A key underpinning of all colexification studies, whether explicitly or implicitly, are *networks*, which play a crucial role in the investigation of cross-linguistic colexification patterns. First, they offer a convenient way to visualize the complexity of recurring semantic associations along with a number of high-quality tools for the interactive exploration of network data (Smoot et al. 2011; Bastian et al. 2009). Second, thanks to recent advances in the empirical study of networks (Newman 2010), many aspects of network structures are well understood, and a multitude of methods and statistics are available (Csárdi & Nepusz 2006; Hagberg 2009), making it easy for scholars to apply them in their research.

The application of colexifications in the form of a network is straightforward. Following Cysouw (2010), lexical *comparative concepts* (Haspelmath 2010) are represented as nodes in a network, while edges connect colexified concepts. The

[1] Historically, the idea of colexifications goes back to the concept of *semantic maps* (Haspelmath 2003), which was most prominently used by typologists to study grammaticalization patterns (van der Auwera & Malchukov 2005; Cysouw 2007; Forker 2015), before it inspired scholars to study phenomena of lexical typology in a similar manner (see the very detailed overview in Georgakopoulos & Polis 2018). Since semantic maps, however, imply rather specific techniques for analysis, which are not necessarily directly required when studying phenomena of lexical typology (as, for example, reflected in the articles introduced by Koptjevskaja-Tamm 2012), we prefer to look at colexifications derived from cross-linguistic data in form of a network.

problem with this approach is that it does not allow one to represent whether a given word form colexifies more than two concepts. In order to represent this kind of information, more complicated network structures are needed, like *hypergraphs*, in which one edge can connect more than two nodes,[2] or *bipartite networks* (Newman 2010: 122), in which nodes are divided into two types and edges can only be drawn between different types (see Hill & List 2017 for an example of bipartite colexification networks). A further enhancement comes from using different *edge weights* where the weights reflect the frequency of a given colexification in a given dataset, while *node weights* represent the overall occurrence of a given concept (List et al. 2013). Figure 1 gives examples for different possibilities for representing cross-linguistic colexification data in the form of networks.

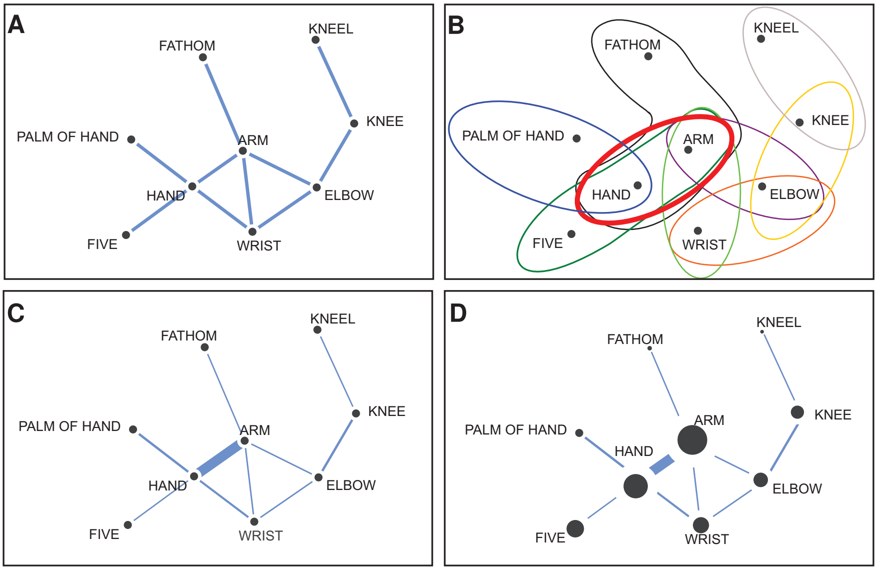


Figure 1: Examples for hypothetical cross-linguistic polysemy networks. A shows an unweighted graph. B shows a hypergraph, C represents a network with weighted edges (edge-width representing relative weight), and D shows a network with weighted edges and weighted nodes (node size representing relative node weight).

[2] This network representation is most frequently used for the drawing of isogloss maps in areal linguistics and dialectology. See Georgakopoulos et al. (2016) for examples involving colexifications. Unfortunately, hypergraphs are very difficult to visualize. Since all nodes in the graph sharing the same feature need to be shown as collected by drawing a boundary around them (excluding all nodes not sharing that very feature), a specific re-arranging of nodes and boundaries is required, which is a challenge both for computational and manual approaches to visualization.

The promise of a network-based approach to colexification (see List et al. 2013; Mayer et al. 2014 for colexification analyses and Rosvall & Bergstrom 2008 for general purpose studies) led to the publication of the *Database of Cross-Linguistic Colexifications* (CLICS, List et al. 2014), which provided cross-linguistic colexification patterns for 1280 concepts across 220 language varieties. While this version of the CLICS database was a valuable resource it also had a number of serious shortcomings. In particular, it had little data, including only 220 languages spoken primarily in South America and Eurasia, and what data were available were hard to check, curate and extend.

In this paper we describe an updated release of CLICS (henceforth called CLICS^2) based around a new framework that attempts to solve these problems, while at the same time scaling up the available data, thus facilitating future research into colexifications. The most important points of improvements we see are:

(A) separating data from display,

(B) making exhaustive and principled use of existing *reference catalogs* like Concepticon (List et al. 2016, for concepts) and Glottolog (Hammarström et al. 2017, for languages) along with recently proposed standardization efforts for cross-linguistic data (Forkel et al. 2017),

(C) curating data and code with help of a transparent Application Programming Interface (API), and

(D) regularly releasing data in release cycles of at least one per year (Haspelmath & Forkel 2015).

In following these design guidelines, we have developed a new database of cross-linguistic colexifications which supersedes the old CLICS database not only in size, both in terms of the number of language varieties and the number of comparative concepts represented, but also with respect to the ease of data curation and the flexibility of the API.

2 How to compare semantics across languages?

Semantic comparison across languages is notoriously difficult. A naive approach to identifying colexification across languages would be to simply map identical translation glosses in wordlists and dictionaries to each other. However, this can easily lead to errors. For example, Chén (1996) originally intended to directly translate Swadesh's (1952) list of 200 items for Chinese dialects. However, the item *dull (knife)* was mistranslated as *dāi, bèn* 呆, 笨 'dull, stupid' in the Chinese questionnaire. Chén's version of the Swadesh 200 item list became quite influential in China and was re-used in a number of studies (Ben Hamed & Wang 2006; Wang & Wang 2004).

Another approach would be to harvest data for clearly attested polysemies across the languages of the world. Scholars could collect instances of colexifications which they deem interesting, and by using careful hand-collected data, these scholars could control for word meaning and homophony from the start. Such an approach would have the advantage of being extremely flexible in terms of the concepts and the languages investigated. However, the amount of work required for a project of such a nature makes it unfeasible to assemble and curate a global database of polysemies. One project that has attempted this is *DatSemShifts* (Bulakh et al. 2013), which attempts to provide an exhaustive resource on attested instances of semantic shifts across the languages of the world. In its 2015 form,[3] the *DatSemShifts* database listed as many as 2424 distinct glosses for comparative concepts. The glosses were, however, only minimally specified, which makes it difficult for users to both find a certain concept and to understand what concept they are dealing with, *cross-linguistically*.

2.1 Harvesting cross-linguistic data with Concepticon

In order to make semantic association patterns comparable across the world's languages, it is clear that we *must* base our analysis on a rigorous collection of comparative concepts. The *Concepticon* reference catalog project (List et al. 2016, 2018) is an attempt to provide consistent links across the multitude of lexical questionnaires that linguists have used to elicit words. Concepticon works by defining specific concept sets based on published datasets (whether these are from field work, or from historical or typological studies), and then linking the *labels* used by researchers to these defined conceptsets. For example, the "dull, blunt" vs. "dull, stupid" error is solved by linking the "dull" label in the list of Chén (1996) to the concept set STUPID^{1518} while linking that of Swadesh (1952) to the concept set BLUNT^{379}.[4] If possible, these links are further checked against the original data, that is, the words in the target languages that were elicited in the end, in order to make sure that what is glossed as *blunt* is indeed reflecting the comparative concept *stupid*.

[3] The database was originally freely accessible at http://datsemshifts.ru but is currently under construction. The version we refer to is the one we accessed on December 29, 2015.

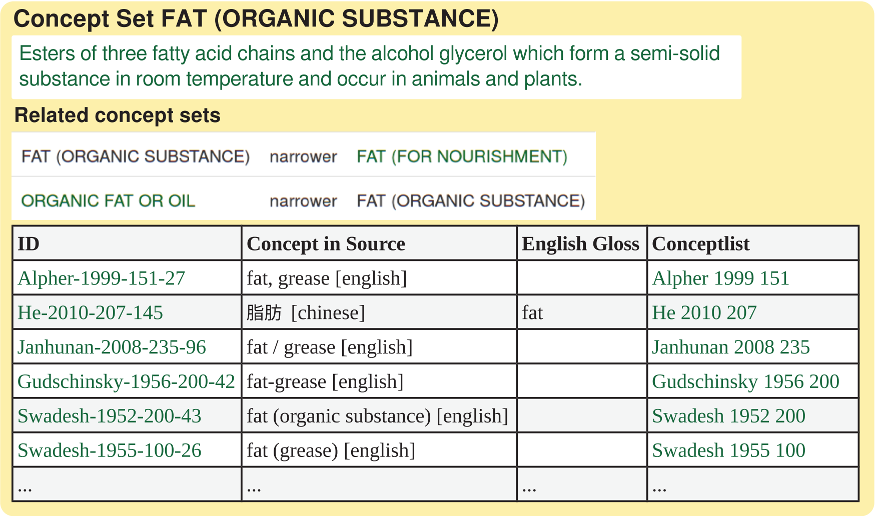
[4] Superscript numbers indicate the identifier used by Concepticon.

The immediate advantage of linking data to the Concepticon is that it enables us to merge data from different sources quickly and safely. We now know which word lists contain lexemes for STUPID^{1518} and which contain lexemes for BLUNT^{379}. And we can now directly ask which languages contain lexemes that colexify STUPID^{1518} and BLUNT^{379}, and colexify these lexemes with any of the other 3144 concept sets defined in Concepticon. In order to avoid errors, we have striven for rigor and strictness when linking concepts to Concepticon. Concepticon does not tolerate "fuzzy" matchings and deliberately avoids linking one elicitation gloss in a single dataset to more than one concept set in the Concepticon resource. If no ideal concept set could be found to link a given elicitation gloss in a questionnaire, it was left unlinked rather than linking to a semantically "close" concept.

Concepticon concept sets can further be linked among each other with help of a simplifying ontology that identifies concept sets which are *broader* or *narrower* with respect to their denotation range. The concept set ARM OR HAND^{2121}, for example, is useful for languages such as Russian or Irish, where the words рука *rukà* and *làmh* respectively refer not only to the part of the arm which other languages denote as *hand*, but also to the entire upper limb. The concept set ARM OR HAND^{2121} is thus considered broader than either HAND^{1277} or ARM^{1637}. While scholars might object to this procedure, preferring to represent a comparative concept reflecting the semantics of Russian рука *rukà* or Irish *làmh* by linking them to both HAND and ARM, it is important to emphasize that this practice, which may seem counterintuitive from the perspective of a given language, is critical if one wishes to guarantee a rigorous mapping of word elicitation glosses in questionnaires to lexical comparative concepts. If a given questionnaire contains the gloss *arm/hand* (as we can find across many questionnaires which have been used to assemble a large number of data points) and we linked it to both ARM^{1277} and HAND^{1637}, we would lose the essential information that the original questionnaire was asking for the word expressing the concept that covers both concept sets in a single term. Since the ontology allows us to derive the information that ARM OR HAND is semantically broader than ARM and HAND, we can choose, over the course of our analysis, to link the elicitation gloss *arm/hand* to both narrower concept sets.

Each Concepticon gloss is also linked to additional metadata e.g. a semantic field, ontological categories (reflecting the more language-specific notion of part of speech), as well as additional metadata derived from norm datasets in psycholinguistics and natural language processing, including age-of-acquisition information for individual languages (Kuperman et al. 2012), ontologies like WordNet (Princeton University 2010), or word frequency counts, again for individual languages (Brysbaert & New 2009).

To illustrate how lexical comparative concepts are organized in the Concepticon, Figure 2 provides a small excerpt of the data which is linked to the concept set FAT (ORGANIC SUBSTANCE)^{323}. As we can see from the figure, this concept set itself is narrower than the concept set ORGANIC FAT OR OIL^{2551}, which shows that many languages do not explicitly distinguish oil from fat. On the other hand, the concept set is broader than FAT (FOR NOURISHMENT)^{2095}. The definition at the top of the figure indicates that the comparative concept should only be linked to those elicitation glosses which target the organic substance as opposed to potential non-organic variants. The five exemplary elicitation glosses in the table at the bottom of the figure are only a small excerpt of what can be found in the whole data linked by the Concepticon. According to the current version (Concepticon-1.1.0, List et al. 2018), the concept set FAT (ORGANIC SUBSTANCE) recurs in 107 different questionnaires and surfaces in the form of 34 distinct elicitation glosses. As we can see from the five examples in the figure, elicitation glosses can vary drastically, not only because they may be given in different languages, but also because authors do not always pay much attention to consistency. Thus, Swadesh used two different glosses, *fat (organic substance)* in his list of 1952, and *fat (grease)* in his later list from 1955, but when inspecting other articles written by Swadesh, we can see that these were absolutely not the only two variants he used, and we find *fat* in Swadesh (1950) and *grease* in Swadesh (1971).



Concept Set FAT (ORGANIC SUBSTANCE)

Esters of three fatty acid chains and the alcohol glycerol which form a semi-solid substance in room temperature and occur in animals and plants.

Related concept sets

FAT (ORGANIC SUBSTANCE)	narrower	FAT (FOR NOURISHMENT)
ORGANIC FAT OR OIL	narrower	FAT (ORGANIC SUBSTANCE)

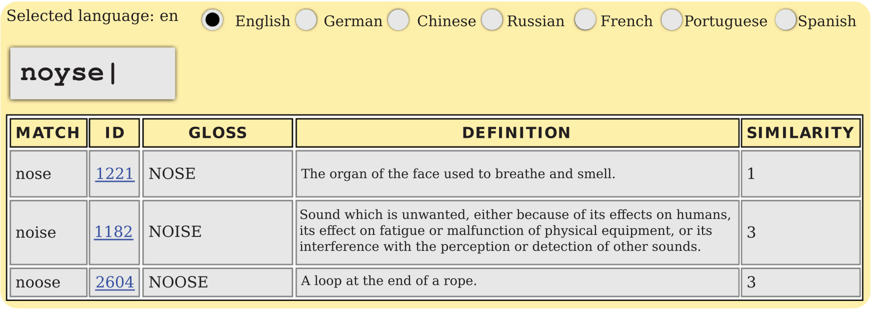
ID	Concept in Source	English Gloss	Conceptlist
Alpher-1999-151-27	fat, grease [english]		Alpher 1999 151
He-2010-207-145	脂肪 [chinese]	fat	He 2010 207
Janhunan-2008-235-96	fat / grease [english]		Janhunan 2008 235
Gudschinsky-1956-200-42	fat-grease [english]		Gudschinsky 1956 200
Swadesh-1952-200-43	fat (organic substance) [english]		Swadesh 1952 200
Swadesh-1955-100-26	fat (grease) [english]		Swadesh 1955 100
...	...	...	...

Figure 2: Example for the representation of data in the Concepticon project. Note that the three different separators in the elicitation glosses for *fat/grease* are given as such in the data, thus reflecting the high degree of inconsistency we find in linguistic practice.

2.2 Using, expanding, and improving Concepticon

We have created a simple automated mapping algorithm to quickly link new wordlists into Concepticon. After applying this algorithm, all links are manually checked to avoid embarrassing errors. In order to further facilitate the task of concept mapping, we wrote a small web-based standalone application which serves as a straightforward lookup tool, including a fuzzy search, across all 3042 concept sets which are currently defined and which can currently be used in seven languages (English, German, Chinese, French, Spanish, Russian, and Portuguese). This web-application, which can be used offline from common web browsers, is provided along with the supplementary material (SI:A) accompanying this paper (see below for further information on the supplementary material). It can also be accessed at http://calc.digling.org/concepticon, where the most recent version is listed. Figure 3 gives a brief example illustrating how it can be used.

Given the complexity of the lexical semantics of natural languages, it is obvious that resources like Concepticon or datasets that link to it will contain uncertainties, less-than-perfect links, and even straightforward errors.[5] Concepticon is designed to be easily correctable and welcomes contributions and additions submitted in form of GitHub issues[6] or email inquiries.[7] Reference catalogs such as Concepticon are community efforts that can only be enhanced if the



Selected language: en ● English ○ German ○ Chinese ○ Russian ○ French ○ Portuguese ○ Spanish

noyse|

MATCH	ID	GLOSS	DEFINITION	SIMILARITY
nose	1221	NOSE	The organ of the face used to breathe and smell.	1
noise	1182	NOISE	Sound which is unwanted, either because of its effects on humans, its effect on fatigue or malfunction of physical equipment, or its interference with the perception or detection of other sounds.	3
noose	2604	NOOSE	A loop at the end of a rope.	3

Figure 3: Example for the web-based lookup tool for Concepticon mapping.

[5] We recently found, for example, a link of German *schaukeln* ‘rock (somebody or something)’ to STONE OR ROCK^{2125}, reflecting a typical case of sloppy automatic linking that can usually only be avoided by manual refinement.

[6] See https://github.com/clld/concepticon-data/issues for details.

[7] By sending an email to concepticon@shh.mpg.de.

research community actively takes part in improving them and it is advantageous for our field if scholars help correct existing problems.

3 A new database of cross-linguistic colexifications

The first version of the CLICS database was based on four different sources which were publicly available at the time of publication and offered sufficient coverage in terms of comparative concepts. With more and more large questionnaires being linked to the Concepticon resource, it has become possible to easily harvest further data and add it to create an improved colexification dataset. With CLDF as a basic representation format that can be easily manipulated with help of the CLDF API written in the Python language (see Section 4.3), all that needs to be done is to assemble different datasets, convert them to CLDF by linking language varieties to Glottolog and elicitation glosses to Concepticon, and analyze them with the standard algorithms which were already present at the release of the CLICS database.

Table 1 lists all datasets that were selected for the first version of our improved colexification database (http://clics.clld.org). All datasets are *multilingual wordlists* in the sense of List (2014, 23f). They are based on a collection of (elicitation) *glosses* that are translated into different language *varieties*. In our CLDF representation of the data, we have linked the elicitation glosses to Concepticon Concept sets, and the language varieties to Glottocodes. Since not all elicitation glosses could be successfully linked to Concepticon, the number of links to Concepticon and the number of original elicitation glosses in the respective datasets shown in the table often vary, showing fewer Concepticon links than elicitation glosses in the original data. The number of varieties and Glottolog entries also varies, but for different reasons, since it can happen that two or more varieties are linked to the same Glottocode, either because the varieties can be seen as identical but stemming from different datasets, or because the Glottocodes for the subvarieties of a given language or dialect are not yet available in the Glottolog project.

As can easily be seen, the data crucially improves upon the old database in terms of languages. The overlap in terms of concepts, however, is less promising, since many of the lists we assembled are in the range of 300 to 500 concepts. Since these sum up to 2487 different concept sets in total, while none of the original lists provides that many concepts, it is also clear that our new sample is considerably skewed, with only a few concepts recurring across all datasets. However, the framework we have implemented will enable us to rapidly increase the size of this database and improve the coverage via a series of “rolling releases”.

Table 1: Overview of datasets converted to CLDF for the new database of cross-linguistic colexifications. Note that the fact that we list 498 distinct glosses for the Bai dataset but 499 Concepticon concept sets, is due to an ambiguity in the English gloss “old”, which occurs two times in this dataset, one time referring to OLD (AGED)^{2122}, and another time to OLD (USED)^{2113}. This information is available in the Chinese glosses, but not in the English ones.

#	Dataset	Source	Range	Glosses	Concepticon	Varieties	Glottocodes	Families
1	allenbai	Allen (2007)	Bai (ST)	498	499	9	3	1
2	bantubvd	Greenhill and Gray (2015)	Bantu	430	415	10	10	1
3	beidasinitic	Běijīng Dàxué (1964)	Sinitic (ST)	905	700	18	18	1
4	bowernpny	Bowern and Atkinson (2011)	Pama-Nyungan	348	338	170	168	1
5	hubercolumbian	Huber and Reed (1992)	Colombian	374	343	69	65	16
6	ids	Key and Comrie (2016)	World-wide	1310	1305	321	276	60
7	kraftchadic	Kraft (1981)	Chadic	434	428	67	60	3
8	northeuralex	Dellert and Jäger (2017)	North-Eurasian	1016	940	107	107	21
9	robinsonap	Robinson and Holton (2012)	Alor-Pantar	398	393	13	13	1
10	satterthwaitetb	Satterthwaite-Phillips (2011)	Sino-Tibetan	423	418	18	18	1
11	suntb	Sūn (1991)	Sino-Tibetan	1004	905	48	48	1
12	tls	Nurse and Phillipson (1975)	Tanzanian	1589	808	120	97	1
13	tryonsolomon	Tryon and Hackmann (1983)	Solomon Islands	324	311	111	96	5
14	wold	Haspelmath and Tadmor (2009)	World-wide	1460	1457	41	41	24
15	zgraggenmadang	Z’graggen (1980abcd)	Madang	380	306	98	98	1
	TOTAL / Overlap				2487	1220	1028	90

Furthermore, interested users can easily use our framework to analyze more balanced sub-samples of the data (for technical details, see Section 4.5) or even feed their own data into our framework.

3.1 General statistics

Our new database automatically derives colexifications for 1220 language varieties (of which 1029 are distinguished by Glottolog) and a total of 2487 distinct Concepticon concept sets. Based on a strict threshold that only accepts a given colexification if it occurs in at least three different language families (as defined by Glottolog),[8] this results in a total of 2638 different colexification patterns, corresponding to 66140 individual instances of colexifications in the languages in our sample. Partitioning the colexification network with the help of the Infomap algorithm (Rosvall & Bergstrom 2008) resulted in 248 different *communities* consisting of more than one concept.[9] Given that some of the concepts in our data are never colexified (probably due to the sparseness of the data) our colexification network consists of 1534 different concepts.

In Table 2, we list the most frequently recurring colexifications in our database, sorted by the attestation per language family along with detailed counts on distinct languages and distinct words that attest the colexification. As can be seen from the table, the results are not particularly surprising. It is well-known that many languages use the same word for MOON and MONTH or for WOOD and TREE. That kinship terms are particularly heavily colexified clearly results from underspecification and does not reflect any potential instances of semantic shift. After manually inspecting the data to look for potential errors, we are quite confident that the unbalanced distribution of our data did not lead to any major errors, providing the same look and feel as the original CLICS database, while offering a drastically increased quantity of data.

[8] The threshold does not have any deeper scientific value. We selected it, since it resulted in communities of a reasonable size that guarantee a smooth look and feel when inspecting the network visualizations, not because we believe in any magic number that would provide us with true polysemies. Since we provide the software to create the networks and clusters along with the data, interested users can test different thresholds and parameter settings for the same data.

[9] In contrast to the earlier version of CLICS, we no longer follow Dellert (2014) in normalizing the edge weights with respect to the frequency by which concepts are reflected across all languages, since we found that the granularity of the communities produced when taking the number of language families as weight is sufficient for our purposes of displaying the data in interesting chunks.

Table 2: The ten most frequently recurring colexifications encountered in our database.

ID A	Concept A	ID B	Concept B	Families	Languages	Words
1370	MONTH	1313	MOON	56	289	294
1803	WOOD	906	TREE	55	211	310
1258	FINGERNAIL	72	CLAW	50	209	216
2267	SON-IN-LAW (OF MAN)	2266	SON-IN-LAW (OF WOMAN)	49	262	285
2265	DAUGHTER-IN-LAW (OF MAN)	2264	DAUGHTER-IN-LAW (OF WOMAN)	47	235	262
1608	LISTEN	1408	HEAR	47	102	105
763	SKIN	629	LEATHER	46	233	255
2259	FLESH	634	MEAT	46	222	232
1599	WORD	1307	LANGUAGE	45	94	98
1228	EARTH (SOIL)	626	LAND	43	158	181

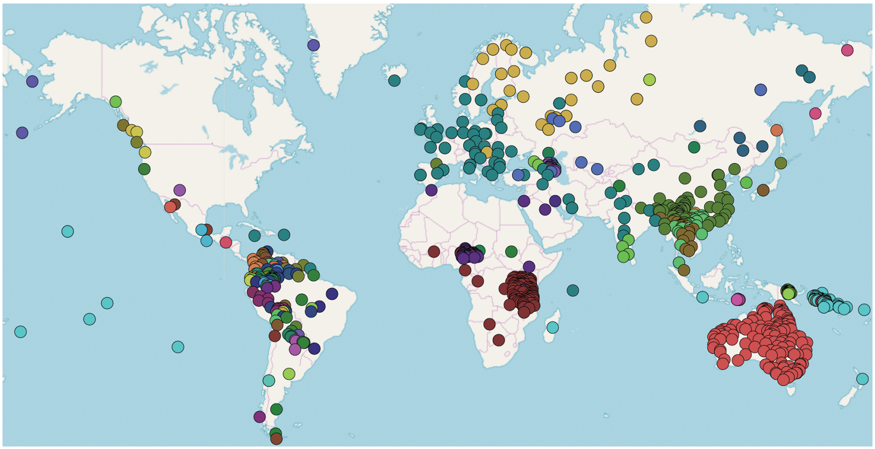


Figure 4: The current coverage of our colexification database in terms of language varieties. Colors indicate the major genetic subgroupings of the languages, following the Glottolog classification.

The increase in data is also reflected when one inspects the geographic distribution of languages covered in our colexification database. As can be seen from the map in Figure 4, our colexification data has drastically reduced the number of empty areas, which were so characteristic of the original CLICS database. However, this does not mean that the data could not be further improved. We can still find many areas on the map, especially in Africa and North America, but also in the Pacific and South Asia, in which coverage is poor to non-existent. We hope to improve the geographical coverage in further releases.

3.2 Exploring colexifications through web-based applications

Our application (hosted at http://clics.clld.org) allows users to explore the data from various additional perspectives, including geographic maps, the inspection of the data of individual languages, or the distribution of concepts for which we find translations in our data. In addition, each data point can be traced back to its original source, allowing the users to rigorously check whether the automatic findings we present can be confirmed through qualitative research. In order to illustrate the new application, a number of examples follow.

3.2.1 Bird's eye view of the colexification data

Before going into details, a bird's eye view of the data is given in Figure 5. While we can see that most of the nodes in our graph form a large connected component, we can also see that the community detection algorithm singles out communities, adding structure which would otherwise be difficult to spot.

3.2.2 Colexifications of SAY, WORD, and LANGUAGE

Of the 248 communities containing at least two concepts in the data, the largest community consists of more than 20 nodes, centering around the concept SPEAK^{1623}, which has the largest number of links in this subgraph (see Figure 5 regarding the position of the network in our big graph). The detailed network is shown in Figure 6. This subgraph also contains the link between LANGUAGE^{1307} and WORD^{1599}, ranking at position nine in the collection of most frequently recurring links shown in Table 2.[10] Those concepts in the figure which are shown in bold font show external links recurring in at least three different language families, clearly suggesting an explanation in terms of natural factors. The concept LANGUAGE, for example, further links to the concept MOUTH^{674} which is placed in a community with BEAK^{73} as the central concept. The concept set SAY^{1458} further

[10] Note that due to ongoing work on the database, this figure may change with future versions of the application, although we are confident that the major trends are unlikely to change any time soon.

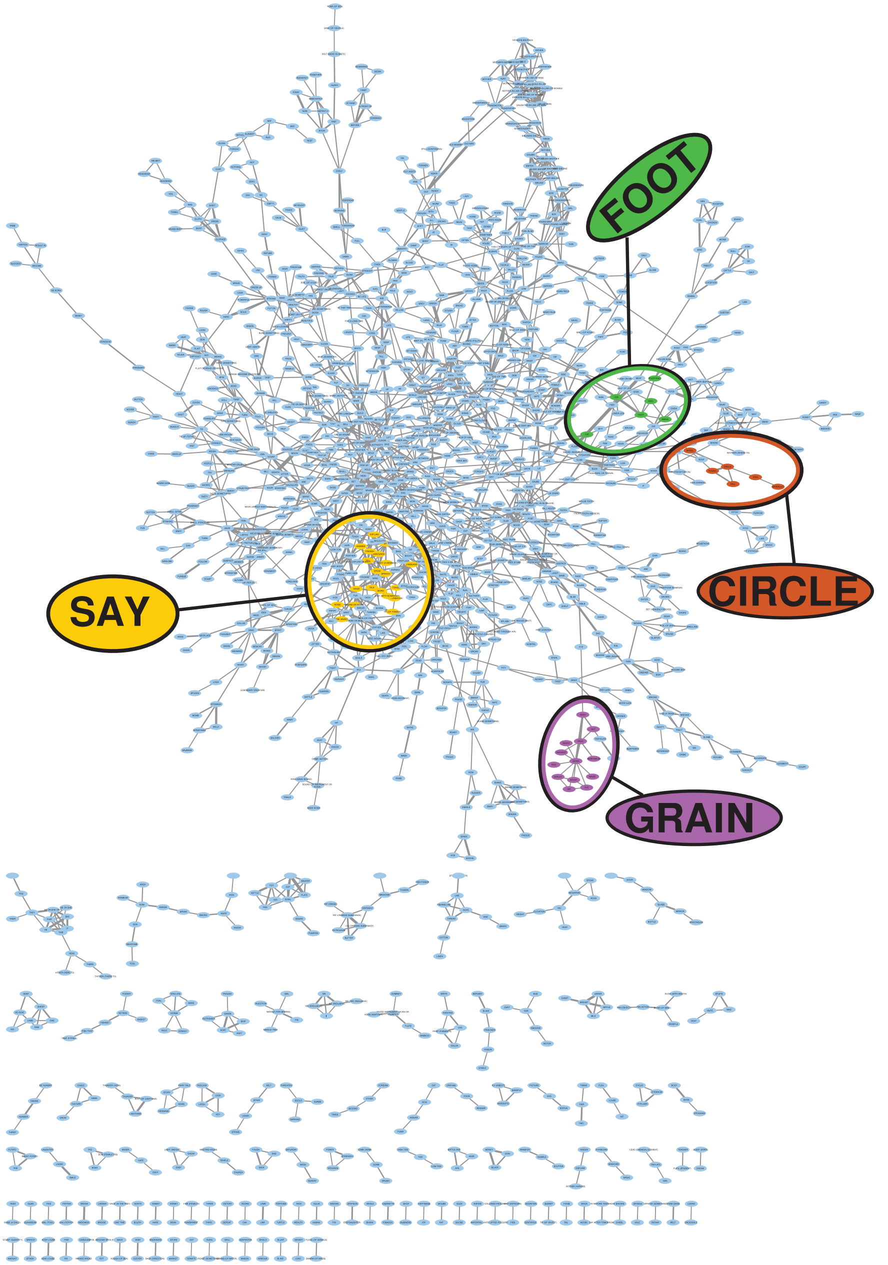


Figure 5: Bird's eye view of our new colexification data in CLICS^2. The graph shows all connected components (113) with some of the communities highlighted.

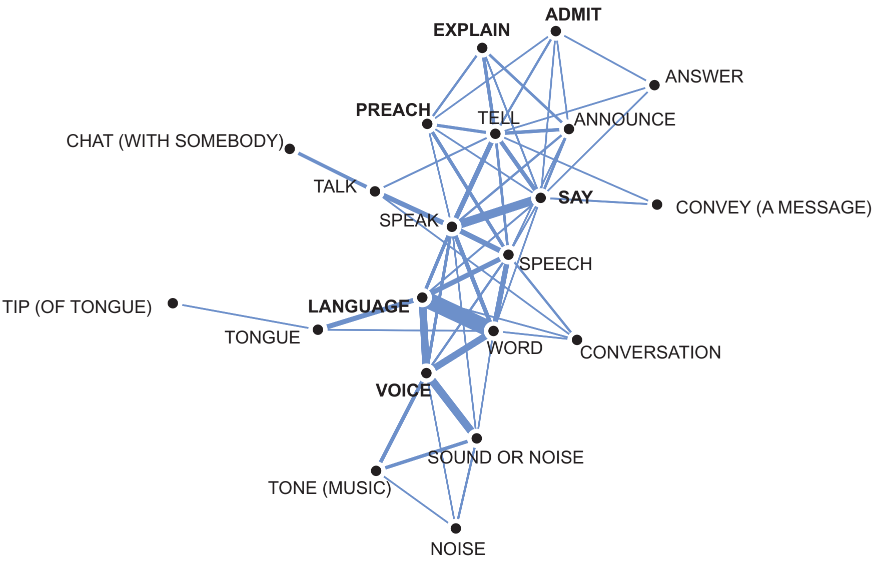


Figure 6: The largest community in our sample, with the concept SAY[1458] showing most connections to the other concepts.

links to 11 different concepts from other communities, notably PROMISE[1675] (central concept OATH[1712]), CALL BY NAME[180] (central concept SHOUT[175]), and DO OR MAKE[2575] (central concept BUILD[1840]). When hovering over the concept in the application, a pop-up provides this information, and users can directly open the respective community to which a given concept with external edges links.

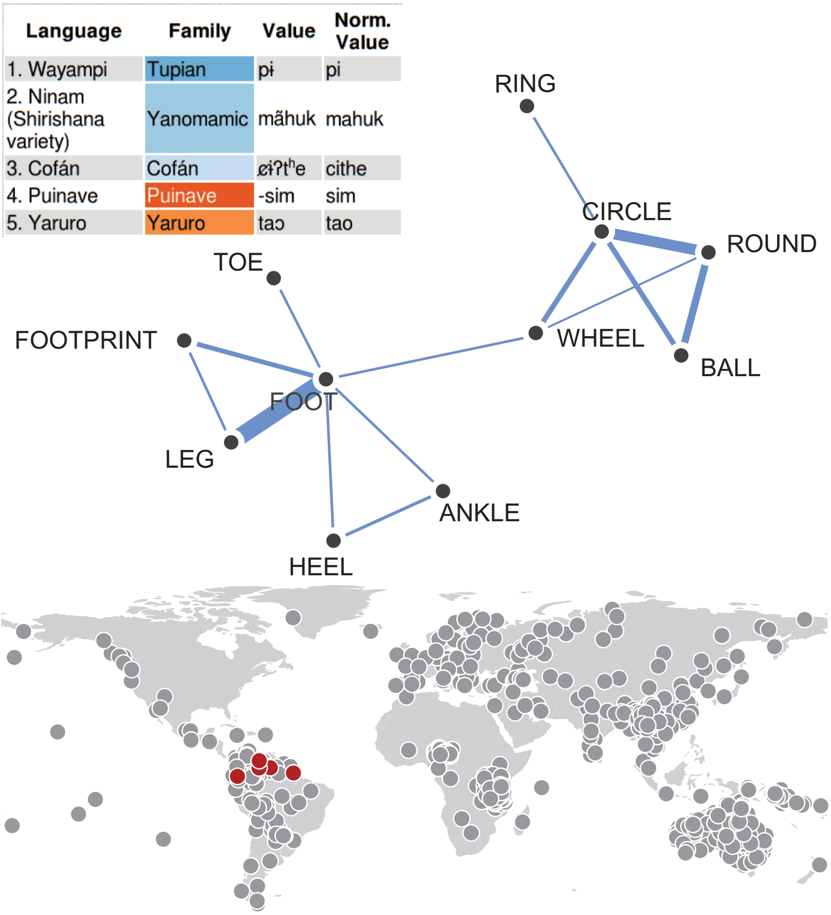
Concepts which could equally well be assigned to different communities are quite common in cross-linguistic colexification data. While this may result from using an inappropriate algorithm for community detection that partitions the network into too small sets of nodes, it also reflects the general indeterminacy of concepts which can often be assigned to different domains. According to our automatic analysis, for example, SAY[1458] plays a role in four different semantic domains, which could be labeled as *neutral speech* (the community shown in Figure 6), *concrete action* (community around DO OR MAKE), PROMISE (community around OATH), and *articulated speech* (community around CALL BY NAME). Unfortunately, our data is not tagged for semantic fields or semantic domains. If it were, we could automatically derive those concepts which are in transitional areas and not easy to assign to only one domain. Much more work will have to be done in the future, both on existing resources such as the Concepticon, and on datasets in CLDF format, as well as our colexification database, in order to exhaust its full potential.

3.2.3 Colexifications of WHEEL and FOOT

As a further example, let us consider a case of regional colexification that was already mentioned by Mayer et al. (2014) and can also be found in our new colexication database: the colexification of FOOT^{1301} and WHEEL^{710} in some South-American languages. In contrast to the example of SAY, WORD, and LANGUAGE in the preceding section, this colexification does not reflect a global pattern which could be identified when looking into the partitions based on the Infomap community detection analysis, which places WHEEL and FOOT into distinct communities. An additional view of the colexification data, introduced by Mayer et al. (2014), however, allows one to find areal patterns, provided they are frequent enough and recur across different language families. This view (called *subgraph* by Mayer et al. 2014) presents the subgraph derived from the closest neighbors of a given query concept. Neighbors of the starting concept are identified by setting a frequency threshold. In consecutive steps, more nodes (the neighbors of the neighbors) can be added to the subgraph, depending on the size of the network, which should not exceed a certain number of nodes to allow for convenient inspection.

Thus, while the colexification between WHEEL and FOOT does not show up in our community analysis, we find it in the subgraph view, as shown in Figure 7. As we can see from the different concrete word forms reflecting the colexification, we are not dealing with a direct borrowing that spread among the languages. Instead, the colexification either reflects an instance of *loan transfer* (in the terminology of Weinreich 1974) or an indirect metaphorical extension. What may substantiate the latter hypothesis is the fact that the WHEEL-FOOT colexification is not restricted to Southern America, but seems to be also reflected in some African languages located on the Eastern coast of Africa (Gilman 1986; Heine & Fehn 2017), but our current version of CLICS^2 does not contain data on these particular languages. The explanation for this particular colexification can thus be sought in historical factors, as a metaphorical extension linked to the introduction of the wheel as a widespread technology in a colonial context.

This again has immediate implications for ongoing debates on *linguistic paleography*. First, the WHEEL-FOOT metaphor shows that concrete historical events may be reflected in languages. Second, however, it also shows that we need to be very careful when evaluating this evidence. As we can see from the subgraph in Figure 7, there are plenty of colexifications for CIRCLE^{1467} and WHEEL in our data as well (our data counts 26 concrete colexifications across 11 different language families). Assuming that societies usually have a way to express the concept ‘circle’, while ‘wheel’ may be missing, our data suggests that the most



Language	Family	Value	Norm. Value
1. Wayampi	Tupian	pɨ	pi
2. Ninam (Shirishana variety)	Yanomamic	mãhuk	mahuk
3. Cofán	Cofán	ȼɨʔt^{h}e	cithe
4. Puinave	Puinave	-sim	sim
5. Yaruro	Yaruro	taɔ	tao

Figure 7: Detecting regional colexification patterns with help of the subgraph explorer. The colors for the languages in the top-left table indicate their genetic relationship. As can be seen, all languages belong to different language families (with three of them being isolates), although they are geographically close.

straightforward strategy to express a new concept ‘wheel’ starts from the word for ‘circle’. Since this can easily happen independently, as we can again see from our data, these findings might be of importance for on-going debates on the origin of terms for ‘wheel’, especially in Indo-European (Hock 2017; Anthony & Ringe 2015). Further studies on lexical typology, including studies on independently recurring patterns of semantic shift as well as the frequency of loan transfer, are required before this linguistic data can be reliably used to reconstruct ancestral cultures. Our extended colexification data may serve as a starting point for these investigations.

4 Technical background

This section provides interested readers with technical details regarding our improved database of cross-linguistic colexifications. More information can also be found in the supplementary material submitted with this study.

4.1 Shortcomings of the previous version of CLICS

The original CLICS database by List et al. (2014) was compiled in a mostly automatic manner. The data were assembled from four different sources (Key & Comrie 2007; Group 2008; Haspelmath & Tadmor 2009; Borin et al. 2013) and were mostly already linked to the same set of comparative concepts, originally based on Buck (1949). The lexical entries were automatically cleaned, using regular expressions and similar standard techniques for text manipulation, and then compared for colexifications. The resulting network was analyzed with help of the Infomap algorithm for community detection (Rosvall & Bergstrom 2008) in order to decrease the complexity, single out colexifications that point to instances of homophony, and split the semantic network into a meaningful set of subgraphs, representing areas of high semantic affinity, close to the notion of *semantic fields* (Anttila 1990).

Since then CLICS has enjoyed considerable popularity among scholars working in the field of lexical typology. On the one hand, this is reflected in studies that mention the database in a favorable manner (Östling 2016; Georgakopoulos & Polis 2018; Šipka 2015), as an inspiration source for similar or enhanced analyses (Söderqvist 2017; Brochhagen 2015; Dellert 2016; Pericliev 2015; Gast & Koptjevskaja-Tamm 2018), or as a potential dataset for additional studies (Youn et al. 2016). On the other hand, it is reflected in a couple of studies that make direct use of the data provided in CLICS (Schapper et al. 2016; Koptjevskaja-Tamm & Liljegren 2017; Staffanson 2017; Regier et al. 2016). It seems that the general strategy of using an interactive interface that allows scholars to explore the actual data, offering both a bird's eye view on colexification patterns while keeping in touch directly with the original data fulfils a certain need for studies on lexical typology, serving as an example for *computer-assisted* as opposed to purely *computer-based* frameworks (List 2016).

As mentioned, the earlier CLICS database suffered from a number of shortcomings. Not only was the *coverage* in terms of languages rather small, with a sample of only 220 languages heavily biased towards Southern America and North Eurasia (Östling 2016), but the data was also not easy to *expand*, as new word lists would have demanded a considerable amount of overlap with the 1280 comparative concepts in CLICS. A factor further complicating the expansion of the existing CLICS database was the difficulty in its *curation*. Given that the data came from independent sources, and that the small number of developers did not have the linguistic expertise to check all wordlists systematically, it was impossible to correct errors in the data itself. Although such curation would have also gone against the original policy of the database, insofar as it was originally built on the idea of providing a different view of already curated datasets, it constituted a serious problem for further development. An additional problem was the *transparency* of the algorithms underlying CLICS: while the source code was online and freely available, it was difficult to use in its previous state, as it was largely undocumented and provided an unfortunate mix of code written for data deployment and code written for data analysis.

4.2 Software implementation

While the original CLICS framework was deployed via PHP, we have integrated the original code for data visualization (Mayer et al. 2014) into a Python-based CLLD application (Forkel & Bank 2018). CLLD offers not only more granular access to the data, but also provides the look-and-feel of well-known typological databases like the *Atlas of Pidgin and Creole Language Structures* (Michaelis et al. 2013) or the *World Atlas of Language Structures* (Dryer & Haspelmath 2013). The new framework allows users to inspect the inferred colexifications online via the CLLD framework, and also allows them to curate their own colexification datasets, to analyze, inspect, and even deploy them, thanks to a standalone application which they can use to convert their own colexification data to an interactive visualization very similar to the look and feel of the original CLICS database.

The new database of cross-linguistic colexifications comes along with a simple interface to compute the colexifications and network statistics. To identify and plot colexifications, we follow the strategy employed by Mayer et al. (2014), but have significantly refactored the code, leading to a drastic increase of speed when searching for colexifications. In contrast to the regular expressions by which the data was automatically cleaned in the original CLICS framework, we have decided to use an even more rigorous approach by stripping off all metalinguistic information that can often be found in linguistic datasets (morpheme boundary markers, brackets for scholars' comments or to indicate pronunciation alternatives) and representing all lexical entries internally with help of ASCII letters. While this carries the danger of introducing errors, our tests indicate that most of these problems can be singled out by only considering colexifications with a frequency above a certain average. Furthermore, since we show the original values as they appear in the data to the users, scholars wishing to work with the data in concrete form can easily check with the original entry or even go back to the original sources of each colexification that we identify with our automated procedure.

We provide platform-independent versions of the Python code which can be used on a Mac, Windows, or Linux computers running either Python 2 or 3. The package provides useful command line tools which we describe in detail at the projects GitHub page at https://github.com/clics/clics2. The software package providing the colexification API is further hosted with Zenodo (see https://doi.org/10.5281/zenodo.1299093 for the most recent version), as well as the 15 datasets (see https://zenodo.org/communities/clics/).

In addition to the CLLD application that provides the data online, we also provide code that exports colexification data to a standalone application, purely based on JavaScript, which can be used locally (and offline) in a web browser, or shared online using a static web server.[11] We assume that this service may turn out to be useful especially for users who want to run their own datasets through our framework, but don't have the technical means or expertise to set up a complex CLLD application. The new CLICS^2 software package also offers information how the standalone application can be computed.

4.3 Cross-linguistic data formats

In order to increase the comparability of cross-linguistic data and to ease the curation and reuse of existing datasets, we follow the standards and recommendations of the Cross Linguistic Data Framework (CLDF). CLDF is a standard to capture different data types often encountered in cross-linguistic research, such as *wordlists*, *typological features*, *parallel texts*, and *dictionaries* (Forkel et al. 2017). The major features of the CLDF specification are: (A) a simple text format for data-storage, based on CSV (comma-separated values), extended by recent recommendations by the World Wide Web Consortium, allowing metadata to be incorporated and data to be linked across multiple CSV files (Pollock et al. 2015; Tennison et al. 2015), (B) a flexible software API which allows validation of whether a given dataset conforms to the specifications, (C) an ontology which

[11] In offline form, only Mozilla Firefox is supported at the moment, but when the data is put on a web server, it can be used from any browser.

allows frequently recurring objects and properties in comparative linguistics to be recognized, and (D) the rigorous integration of reference catalogs in order to increase the comparability of data across datasets.

4.4 Cross-linguistic reference catalogs

CLDF strongly encourages the usage of *reference catalogs*, such as Glottolog or Concepticon, when preparing linguistic data. The advantage of organizing language varieties not only by their common name, but also by adding the identifiers offered by Glottolog are obvious. Since Glottolog harvests various types of information regarding language varieties all over the world, ranging from geographical coordinates via references in the literature up to genealogical classifications, scholars linking the languages in their data to Glottolog identifiers can automatically dispose of this information when carrying out additional studies. Disadvantages may result from incorrect links to Glottolog or from problems that experts may encounter when checking the information provided by Glottolog. If, for example, scholars do not agree with the genealogical classification provided by Glottolog, they may prefer to add their own classification to their dataset. However, even if specific information turns out to be erroneous or not satisfying enough for scholars to help in their application, it is still useful to try to provide a link to Glottolog, as it will make it much easier for other scholars to find their data. Furthermore, Glottolog is curated in public and changes can be proposed and made in a transparent manner in the form of GitHub issues[12] or by contacting the editors via email.

As outlined in Section 2, these advantages also hold for the usage of Concepticon as a reference catalog for comparative concepts. While scholars may still use and embrace their individual questionnaires with their preferred elicitation glosses and concept definitions, linking them to Concepticon guarantees that their data is cross-linguistically comparable and easily accessible to other researchers as well.

4.5 Choosing a representative sample using Average Mutual Coverage

A crucial issue when assembling different datasets in the way this is done in our updated version of the CLICS database is whether the overlap in terms of comparative concepts across datasets is sufficient and representative enough.

[12] See https://github.com/clld/glottolog/ for details.

While the data aggregated from the 15 different datasets should be interesting enough for manual inspection and analysis, some analysis strategies (for example those based on hypothesis testing, as mentioned by Roberts 2018) will require balanced and representative samples.

A seemingly straightforward way to identify a representative sample would be to determine the subset that provides the optimal overlap in terms of languages and concepts. In order to get a better idea of how skewed the data in our updated version of CLICS actually is, we carried out a detailed investigation of the different subsets of the data, employing the concept of *average mutual coverage* (AMC) in multilingual wordlists, as provided in the most recent version of the LingPy software package (List et al. 2017: Version 2.6).

Here, the AMC of a given wordlist is defined as the average of the number of concepts shared between all pairs of languages in a given wordlist divided by the number of concepts in total. Assuming we have a concept list of 100 concepts and three different languages A, B, and C, which have translations for 90, 70, and 60 concepts each, we can determine the average mutual coverage by first checking the individual overlap among the languages (which is not necessarily equal to the number of the concepts translated in the "smaller" of two varieties), and then divide these numbers by the total amount of concepts. If we assume a mutual coverage of 65 between A and B, of 55 between A and C, and 45 between B and C, we can sum up and average the mutual coverage between all pairs. In this example, this would result in an AMC score of 0.55 ($\frac{0.65+0.55+0.45}{3}$).[13]

Therefore, if users want a representative sample, they can make use of our AMC statistics, which are provided along with the source code to compute them in the supplementary material, to extract their sample of choice of the datasets we provide in full (SI:B).

We can use this metric to evaluate how well-balanced a given selection of languages and concepts is. This can be done by dividing the data into different subsets in which concepts and languages are consecutively deleted from the data, using their average size (number of concepts, or number of languages which provide a translation for a concept) as a criterion. We can then compute the AMC for each of these subsets and plot the data in order to see how the AMC scores change when reducing the number of languages and concepts in the data.

We carried out this analysis both for our new collection of datasets in CLDF format and the data underlying the original CLICS database. The results can be seen in Figure 8. As can be seen from these plots, our new data collection is heavily unbalanced, with extremely low mutual coverage scores for samples of a

[13] Since we use 100 concepts as an example, the mutual coverage for each language pair is simply divided by 100.

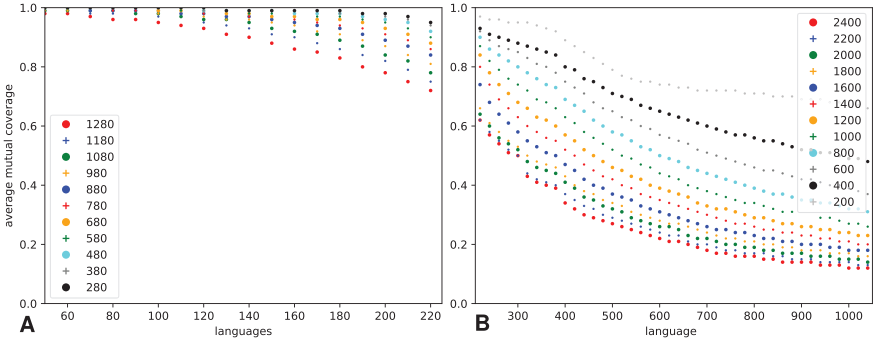


Figure 8: Average mutual coverage in A CLICS, and B our new colexification database. The legend shows the number of concepts for which the AMC has been calculated.

large number of concepts and languages. Comparing our data with the AMC statistics for different subsets of the original CLICS database, however, we can also see that our data supersedes the original CLICS coverage for a subset of about 1200 concepts and about 300 languages.

5 Concluding remarks

Our new colexification database provides a powerful new tool for investigating colexification patterns on both global and regional scales. Thanks to a substantial increase in the data used to identify the patterns, the inference of colexifications is far more robust than before. The new framework based on the CLDF specification and intensive use of reference catalogs has dramatically increased the transparency and replicability of analyses – and our stated policy towards open data and a regular floating release scheme will extend and grow the database further in the future. Our framework can easily be extended following our collaboration guidelines, or co-opted for analysis of alternative datasets as necessary. We see our framework as a central tool for future work in lexical typology.

Author Contributions: JML wrote the first draft. SJG, CA, and TT revised the first draft. JML and RF wrote the Python code for the application. RF wrote the code for the CLLD application. TM wrote the visualization routines in the standalone and the CLLD application. TT, SJG, JML, and RF provided datasets in CLDF format. All authors revised the last draft and agree with its final version.

Funding: This research would not have been possible without the generous support by many institutes and funding agencies. As part of the CLLD project (http://clld.org) and the GlottoBank project (http://glottobank.org), the work was supported by the Max Planck Society, the Max Planck Institute for the Science of Human History, and the Royal Society of New Zealand (Marsden Fund grant 13-UOA-121, RF, JML, and SJG). JML was funded by the DFG research fellowship grant 261553824 Vertical and lateral aspects of Chinese dialect history (2015-2016). JML and TT were funded by the the ERC Starting Grant 715618 Computer-Assisted Language Comparison (http://calc.digling.org). SJG was supported by the Australian Research Council's Discovery Projects funding scheme (project number DE 120101954) and the ARC Center of Excellence for the Dynamics of Language grant (CE140100041).

Acknowledgment: We thank three anonymous reviewers for challenging and constructive advice. We are grateful to Martin Haspelmath, who brought up the idea to submit this paper to Linguistic Typology and always supported our data-driven approach to lexical typology. We thank Russell Gray for encouraging us to pursue our data-driven perspective on historical linguistics and language typology. We also thank Stéphane Polis, Thanasis Georgakopoulos, and Anne-Maria Fehn for helpful comments on earlier versions of this paper. Last but not least, we would like to express our gratitude to all scholars who made their data publicly available, and specifically to Peter Bouda, Claire Bowern, Johannes Dellert, Michael Cysouw, Gerhard Jäger, Harald Hammarström, Jelena Prokić, Damian Satterthwaite-Phillips, and Malcom Ross for their help in rendering their data in our formats.

References

Allen, Brian. 2007. *Bai Dialect Survey*. Dallas: SIL International.

Anthony, David W. & Don Ringe. 2015. The Indo-European homeland from linguistic and Archaeological perspectives. *Annual Review of Linguistics* 1. 199–219.

Anttila, Raimo. 1990. Field theory of meaning and semantic change. In Günter Kellerman & Michael D. Morrissey (eds.), *Diachrony within synchrony: Language history and cognition*, 23–83. Bern: Peter Lang.

van der Auwera, Johan & Andrej Malchukov. 2005. A semantic map for depictive adjectivals. In N. P. Himmelmann & E. Schultze-Berndt (eds.), *Secondary predication and adverbial modification. The typology of depictive constructions*, 393–423. Oxford: Oxford University Press.

Bastian, Mathieu, Sebastien Heymann & Mathieu Jacomy. 2009. Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the third international AAAI conference on weblogs and social media*, Association for the Advancement of Artificial Intelligence. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

Ben Hamed, Mahe & Feng Wang. 2006. Stuck in the forest: trees, networks and Chinese dialects. *Diachronica* 23. 29–60.

Borin, Lars, Bernard Comrie & Anju Saxena. 2013. The intercontinental dictionary series – a rich and principled database for language comparison. In Lars Borin & Anju Saxena (eds.), *Approaches to measuring linguistic differences*, 285–302. Berlin and Boston: De Gruyter Mouton.

Bowern, Claire, Patience Epps, Russell Gray, Jane Hill, Keith Hunley, Patrick McConvell & Jason Zentz. 2011. Does Lateral Transmission Obscure Inheritance in Hunter-Gatherer Languages? *PLoS ONE* 6(9). e25195. doi:10.1371/journal.pone.0025195. http://dx.doi.org/10.1371%2Fjournal.pone.0025195.

Brochhagen, Thomas. 2015. Improving coordination on novel meaning through context and semantic structure. In *Proceedings of the sixth workshop on cognitive aspects of computational language learning*, 74–82.

Brysbaert, Marc & Boris New. 2009. Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* 41. 977–990.

Buck, Carl Darling. 1949. *A dictionary of selected synonyms in the principal Indo-European languages. A contribution to the history of ideas*. Chicago and Illinois: University of Chicago Press.

Bulakh, M., Dimitrij Ganenkov, Ilya Gruntov, T. Maisak, Maxim Rousseau & A. Zalizniak (eds.). 2013. *Database of semantic shifts in the languages of the world*. Moscow: RGGU. http://semshifts.iling-ran.ru/.

Běijīng Dàxué, 北京大学(ed.). 1964. Hànyǔ fāngyàn cìhuì 汉语方言词汇[Chinese dialect vocabularies]. Běijīng 北京: Wénzì Gǎigé 文字改革.

Chén, Bǎoyà 陈保亚. 1996. *Lùn yǔyán jiēchù yǔ yǔyán liánméng* 论语言接触与语言联盟 *[Language contact and language unions]*. Běijīng 北京: Yǔwén 语文.

Csárdi, Gábor & Tamás Nepusz. 2006. The igraph software package for complex network research. *InterJournal. Complex Systems* 1695. http://igraph.org.

Cysouw, Michael. 2007. Building semantic maps: The case of person marking. In Bernhard Wälchli & M. Miestamo (eds.), *New challenges in typology*, 225–248. Berlin and New York: Mouton de Gruyter.

Cysouw, Michael. 2010. Semantic maps as metrics on meaning. *Linguistic Discovery* 8(1). 70–95.

Dellert, Johannes. 2014. Lifting a large multilingual dictionary to the level of concepts. Talk held at the Workshop on historical and empirical evolutionary Linguistics 15–16 February 2014. Eberhard-Karls-Universität Tübingen.

Dellert, Johannes. 2016. Using causal inference to detect directional tendencies in semantic evolution. In Seán G. Roberts, Christine Cuskley, Luke McCrohon, Lluís Barceló-Coblijn, Olga Fehér and Tessa Verhoef (eds.), The Evolution of Language: Proceedings of the 11th International Conference (EVOLANGX11), Online at http://evolang.org/neworleans/papers/139. html.

Dellert, Johannes & Gerhard Jäger. 2017. *Northeuralex (version 0.9)*. Tübingen: Eberhard-Karls University Tübingen.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/.

Forkel, Robert & Sebastian Bank. 2018. CLLD: A toolkit for cross-linguistic databases. doi:10.5281/zenodo.1186271. https://doi.org/10.5281/zenodo.1186271.

Forkel, Robert, Simon J. Greenhill & Johann-Mattis List. 2017. *Cross-Linguistic Data Formats (CLDF)*. Jena: Max Planck Institute for the Science of Human History. http://cldf.clld.org.

Forker, Daniela. 2015. Towards a semantic map for intensifying particles: Evidence from Avar. *STUF – Language Typology and Universals* 68(4). 485–513.

François, Alexandre. 2008. Semantic maps and the typology of colexification: intertwining polysemous networks across languages. In Martine Vanhove (ed.), *From polysemy to semantic change*, 163–215. Amsterdam: Benjamins.

Gast, Volker & Maria Koptjevskaja-Tamm. 2018. The areal factor in lexical typology: Some evidence from lexical databases. In Daniel Van Olmen, Tanja Mortelmans and Frank Brisard (eds.), *Aspects of linguistic variation*, 43-82. Berlin: de Gruyter Mouton.

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–n/a. doi:10.1111/lnc3.12270. http://dx.doi.org/10.1111/lnc3.12270. E12270 LNCO-0727.R1.

Georgakopoulos, Thanasis, Daniel A. Werning, Jörg Hartlieb, Tomoki Kitazumi, Lidewij E. van de Peut, Annette Sundermayer & Gaëlle Chantrain. 2016. The meaning of ancient words for 'earth': An exercise in visualizing colexification on a semantic map. *eTopoi. Journal for Ancient Studies* 6. 1–36.

Gilman, Charles. 1986. African areal characteristics: Sprachbund, not substrate? *Journal of Pidgin and Creole Languages* 1(1). 33–50.

Greenhill, Simon J & Russell D Gray. 2015. Bantu basic vocabulary database.

Group, Logos (ed.). 2008. *Logos Dictionary*. Modena: Logos Group. http://www.logosdictionary.org/index.php.

Hagberg, Aric. 2009. NetworkX. High productivity software for complex networks. Distributed by the author. http://networkx.lanl.gov/index.html.

Hammarström, Harald, Robert Forkel & Martin Haspelmath. 2017. *Glottolog*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://glottolog.org.

Haspelmath, Martin. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In Michael Tomasello (ed.), *The new psychology of language*, 211–242. Mahwah, NJ: Lawrence Erlbaum.

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories. *Language* 86(3). 663–687.

Haspelmath, Martin & Robert Forkel. 2015. *CLLD – Cross-Linguistic Linked Data*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://clld.org.

Haspelmath, Martin & Uri Tadmor (eds.). 2009. *World Loanword Database*. Munich: Max Planck Digital Library.

Heine, Bernd & Anne-Maria Fehn. 2017. An areal view of Africa. In Raymond Hickey (ed.), *The Cambridge handbook of areal linguistics* Cambridge handbooks in language and linguistics, 424–445. Cambridge University Press. doi:10.1017/9781107279872.016.

Hill, Nathan W. & Johann-Mattis List. 2017. Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages. *Yearbook of the Poznań Linguistic Meeting* 3(1). 47–76.

Hock, Hans Henrich. 2017. Indo-European linguistics meets Micronesian and Sunda-Sulawesi. *Wellington Working Papers in Linguistics* 23. 63–67.

Huber, R. Q. & R. B. Reed. 1992. *Vocabulario comparativo: palabras selectas de lenguas indígenas de Colombia [Comparative vocabulary. Selected words from the indigeneous* *languages of Columbia]*. Santafé de Bogota: Asociatión Instituto Lingüístico de Verano.

Key, Mary Ritchie & Bernard Comrie (eds.). 2007. *IDS – The intercontinental dictionary series*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Key, Mary Ritchie & Bernard Comrie. 2016. *The intercontinental dictionary series*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Koptjevskaja-Tamm, Maria. 2012. New directions in lexical typology. *Linguistics* 50(3). 373–394.

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

Kraft, Charles H. (ed.). 1981. *Chadic wordlists*. Berlin: Dietrich Reimer.

Kuperman, Victor, Hans Stadthagen-Gonzalez & Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods* 44(4). 978–990. http://dx.doi.org/10.3758/s13428-012-0210-4.

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

List, Johann-Mattis. 2016. Computer-Assisted Language Comparison: Reconciling computational and classical approaches in historical linguistics. Tech. rep. Max Planck Institute for the Science of Human History Jena.

List, Johann-Mattis, Michael Cysouw & Robert Forkel. 2016. Concepticon. A resource for the linking of concept lists. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the tenth international conference on language resources and evaluation*, 2393–2400. European Language Resources Association (ELRA).

List, Johann-Mattis, Michael Cysouw, Simon Greenhill & Robert Forkel. 2018. *Concepticon. A resource for the linking of concept list*. Jena: Max Planck Institute for the Science of Human History. http://concepticon.clld.org/.

List, Johann-Mattis, Simon Greenhill & Robert Forkel. 2017. *LingPy. A Python library for quantitative tasks in historical linguistics*. Jena: Max Planck Institute for the Science of Human History. doi:https://doi.org/10.5281/zenodo.1065403. http://lingpy.org.

List, Johann-Mattis, Thomas Mayer, Anselm Terhalle & Matthias Urban (eds.). 2014. *CLICS: Database of cross-linguistic colexifications*. Marburg: Forschungszentrum Deutscher Sprachatlas. http://clics.lingpy.org.

List, Johann-Mattis, Anselm Terhalle & Matthias Urban. 2013. Using network approaches to enhance the analysis of cross-linguistic polysemies. In *Proceedings of the 10th International Conference on Computational Semantics – Short Papers*, 347–353. Stroudsburg: Association for Computational Linguistics.

Mayer, Thomas, Johann-Mattis List, Anselm Terhalle & Matthias Urban. 2014. An interactive visualization of cross-linguistic colexification patterns. In *Visualization as added value in the development, use and evaluation of linguistic resources. Workshop organized as part of the International Conference on Language Resources and Evaluation*, 1–8. LREC.

Michaelis, Susanne Maria, Philippe Maurer, Martin Haspelmath & Magnus Huber. 2013. *The Atlas of Pidgin and Creole language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Newman, Mark. E. J. 2010. *Networks. An Introduction*. Oxford: Oxford University Press.

Nurse, Derek & Gérard Phillipson. 1975. *Tanzania Language Survey*. Dar es Salaam: Department of Foreign Languages and Linguistics, University of Dar es Salaam.

Olsen, Brigitte. A. 2002. Thoughts on Indo-European compounds - inspired by a look at Armenian. *Transactions of the Philological Society* 100(1). 233–257.

Östling, Robert. 2016. Studying colexification through massively parallel corpora. In Päivi Juvonen & Maria Koptjevskaja-Tamm (eds.), *The lexical tpypology of semantic shifts*, 157–176. Berlin and Boston: De Gruyter Mouton.

Pericliev, Vladimir. 2015. On colexification among basic vocabulary. *Journal of Universal Language* 16(2). 63–93.

Pollock, Rufus, Jeni Tennison, Gregg Kellogg & Ivan Herman. 2015. Metadata vocabulary for tabular data. Techreport World Wide Web Consortium (W3C). https://www.w3.org/TR/tabular-metadata/.

Princeton University. 2010. WordNet. A lexical database for English. Online Resource. https://wordnet.princeton.edu/.

Regier, Terry, Alexandra Carstensen & Charles Kemp. 2016. Languages support efficient communication about the environment: Words for snow revisited. *PLOS ONE* 11(4). 1–17. doi:10.1371/journal.pone.0151138. https://doi.org/10.1371/journal.pone.0151138.

Roberts, Seán G. 2018. Robust, causal, and incremental approaches to investigating linguistic adaptation. *Frontiers in Psychology* 9. 166. doi:10.3389/fpsyg.2018.00166. https://www.frontiersin.org/article/10.3389/fpsyg.2018.00166.

Robinson, Laura C & Gary Holton. 2012. Internal classification of the Alor-Pantar language family using computational methods applied to the lexicon. *Language Dynamics and Change* 2(2). 123–149.

Rosvall, Martin & Carl T. Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Science. U.S.A.* 105(4). 1118–1123.

Satterthwaite-Phillips, D. 2011. *Phylogenetic inference of the Tibeto-Burman languages or on the usefuseful of lexicostatistics (and "megalo"-comparison) for the subgrouping of Tibeto-Burman*. Stanford: Stanford University Phd thesis.

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 and Boston: De Gruyter Mouton.

Söderqvist, Kajsa. 2017. *Colexification and semantic change in colour terms in Sino-Tibetan and Indo-European languages*. Lund: University of Lund Bachelor's thesis.

Smoot, Michael E., Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang & Trey Ideker. 2011. Cytoscape 2.8. *Bioinformatics* 27(3). 431–432.

Sūn, Hóngkāi 孙宏开 (ed.). 1991. *Zàngmiǎnyǔ yǔyīn hé cíhuì* 藏缅语语音和词汇 *[Tibeto-Burman phonology and lexicon]*. Běijīng: Zhōngguó Shèhuì Kēxué 中国社会科学 [Chinese Social Sciences Press].

Staffanson, Martina. 2017. *Mitt hjärta är bittert. En lexikal typologisk studie om smaktermer*. Stockholm: Institut för Lingvistik Term paper.

Swadesh, Morris. 1950. Salish internal relationships. *International Journal of American Linguistics* 16(4). 157–167.

Swadesh, Morris. 1952. Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* 96(4). 452–463.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21(2). 121–137.

Swadesh, Morris. 1971. *The origin and diversification of language: Edited post mortem by Joel Sherzer*. Chicago: Aldine.

Tennison, Jeni, Gregg Kellogg & Ivan Herman. 2015. Model for tabular data and metadata on the web. Techreport World Wide Web Consortium (W3C). https://www.w3.org/TR/tabular-data-model/.

Tryon, Darrel T. & Brian D. Hackman. 1983. *Solomon Islands languages. An internal classification* (C 72). Canberra: Pacific Linguistics.

Urban, Matthias. 2010. 'Sun'='Eye of the Day': A linguistic pattern of Southeast Asia and Oceania. *Oceanic Linguistics* 49(2). 568–579.

Vaan, Michiel. 2008. Indo-European linguistics. *LINGUA* 118(8). 1228–1232.

Šipka, Danko. 2015. *Lexical conflict. Theory and practice*. Cambridge: Cambridge University Press.

Wang, Feng & William S.-Y. Wang. 2004. Basic words and language evolution. *Language and Linguistics* 5(3). 643–662.

Weinreich, Uriel. 1974. *Languages in contact. With a preface by André Martinet*. The Hague and Paris: Mouton 8th edn.

Youn, Hyejin, Logan Sutton, Eric Smith, Cristopher Moore, Jon F. Wilkins, Ian Maddieson, William Croft & Tanmoy Bhattacharya. 2016. On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences of the United States of America* doi:10.1073/pnas.1520752113.

Z'graggen, John A. 1980a. *A comparative word list of the Mabuso languages, Madang Province, Papua New Guinea*. Canberra, Australia: Pacific Linguistics.

Z'graggen, John A. 1980b. *A comparative word list of the Northern Adelbert Range languages, Madang Province, Papua New Guinea*. Canberra, Australia: Pacific Linguistics.

Z'graggen, John A. 1980c. *A comparative word list of the Rai Coast languages, Madang province, Papua New Guinea*. Canberra, Australia: Pacific Linguistics.

Z'graggen, JA. 1980d. *A comparative word list of the Southern Adelbert Range languages, Papua New Guinea*. Canberra, Australia: Pacific Linguistics.

Supplemental Material: The online version of this article offers supplementary material (https://doi.org/10.1515/lingty-2018-0010).