Computer-Assisted Language Comparison in Practice

Tutorials on Computational Approaches to the History and Diversity of Languages

# The Origins of Cross-Linguistic Colexifications

Annika Tjuka
Department of Linguistic and Cultural Evolution
Max Planck Institute for Evolutionary Anthropology

In recent years, studies exploring the phenomenon of colexification across languages have steadily increased in number. Colexification occurs if a word has multiple meanings, regardless of whether the meanings are related (*dish* 'plate; meal') or unrelated (*bank* 'financial institution; part of a river'). The investigation of cross-linguistic colexifications yields many interesting findings that are important for different research fields. Psychologists and cognitive scientists are interested in the overarching principles that establish a connection between meanings and how speakers categorize the environment around them. Historical linguists are concerned with diachronic processes that lead to semantic shifts and what these can tell us about language evolution. Typologists engage in the study of language contact scenarios and how linguistic areas are formed. All these processes are entwined with one another and disentangling them is a challenge. This blog post is the first step into a deeper exploration of the origins of cross-linguistic colexifications and discusses the four processes underlying this phenomenon.

## Introduction

The term "colexification" was first introduced by François (2008) and is a cover term for polysemy, vagueness, and homophony. These phenomena have been established in linguistics and discussed since the early beginnings of lexical typology (Ullmann 1953). The distinction between polysemy, vagueness, and homophony, especially in a semantic analysis of diverse languages, is not always clear-cut so working with a broader term has advantages. It is therefore useful to speak of "cross-linguistic colexifications" to indicate that the underlying data establishing a colexification come from different languages. The assumption is that if two unrelated languages use the same lexical item for two different concepts, the concepts can be said to colexify. For example, Wolof and Indonesian use one word for the concepts FOOT and LEG indicating that FOOT and LEG are cross-linguistically colexified.

*Computer-Assisted Language Comparison in Practice*
Volume 6, Number 1
https://calc.hypotheses.org/5001/

Published on 16/01/2023

Cross-linguistic colexifications were first adopted in typology for the creation of semantic maps (François 2008). Semantic maps are a graphical representation to illustrate the relation between recurring meaning expressions in a language (Haspelmath 2003). They can be applied in typological research of different semantic domains such as emotions (e.g., Georgakopoulos & Polis 2022). One of the main challenges of this methodological approach is the premise that the meanings of words are comparable across languages. This is of course a general issue for lexical typology which often uses an etic approach, i.e., research based on standardized measures (Evans 2010). In analogy to Haspelmath's (2010) grammatical categories, which introduced the notion of comparative concepts, collections of standardized concepts like the Concepticon (List et al. 2016) attempt to mitigate these challenges through a bottom-up approach and assessment by multiple language experts. The creation of the Concepticon has also led to an adjunct database in which cross-linguistic colexifications are collected: CLICS[3] (Rzymski et al. 2020). This database has been successfully applied in a number of studies in typology (e.g., Gast & Koptjevskaja-Tamm 2019, 2022), computational linguistics (Bao et al. 2021; Di Natale et al. 2021), and cognitive science (Jackson et al. 2019; Brochhagen & Boleda 2022).

Understanding the diverse semantic patterns found in the languages of the world is the main aim of these cross-cultural investigations and they add to an otherwise Western-centric scientific landscape (Blasi et al. 2022, Henrich et al. 2010). The study of cross-linguistic colexifications is of interest to a diverse set of researchers because their origins are related to three fields: cognition, history, and language contact. Untangling the different processes is not an easy task, so the aim of the following sections is to give an overview and point to some relevant literature. Note that this blog post is not aimed to give a full account of the entire literature, but should be seen as a starting point for further explorations of the topic.

## Process 1: Recognizing Cognitive Principles

When first encountered, nothing seems odd about a language that uses the same word for the concepts SKIN (of a human) and BARK (of a tree). The two concepts are similar in their function and the meanings are thus related. In the second, third, or even hundredths languages in which the same connection between the two concepts occurs, the cross-linguistic colexification becomes interesting and one wonders if a universal process is at play. In linguistics, numerous of these recurring cross-linguistic colexifications have been explored in detailed analyses of different semantic domains. Especially in earlier lexical typological studies, the aim was to identify universals in the linguistic systems of color (Berlin & Kay 1969) or body parts (Brown 1976, Andersen 1978). The underlying assumption of these studies is that speakers recognize a shared similarity between two

concepts and thus, use the same word to express them. Examples of these supposedly widespread cross-linguistic colexifications, for example, FIRE-FIREWOOD, FOOT-LEG, ARM-HAND, EYE-FACE, SKIN-BARK, EGG-TESTICLES feature also in more recent investigations (e.g., Urban 2011, Schapper et al. 2016). While the role of cognition is clearly acknowledged for the establishment of these widespread cross-linguistic colexifications, based on linguistic evidence, one can only speculate about the exact cognitive mechanisms. A program that is informed by linguistics but uses the tools of cognitive science (similar to the program established by Stephen C. Levinson in the domain of space) would be able to work out those cognitive mechanisms.

Another line of research in which cross-linguistic colexifications have received a lot of attention lately is related to the notion of communicative efficiency (Gibson et al. 2019). Contrary to the assumption that ambiguity is detrimental in a language system (Chomsky 2002), Piantadosi et al. (2012) proposed that ambiguity is a necessary property of language and improves efficient communication. Building on this information-theoretic approach, a study of cross-linguistic colexifications in 250 languages showed that related meanings are more frequently colexified (Xu et al. 2020). The findings indicate that the association strength between concepts leads to the cross-linguistic colexification patterns that we see in the languages of the world. However, there is a trade-off between simplicity and informativeness that languages need to balance in order to make communication efficient. Brochhagen and Boleda (2022) demonstrated that languages rarely colexify meanings that are highly related. This is supported by the experimental findings of Karjus et al. (2021) who showed that domains less relevant to a speech community are simpler showing that communicative needs are influenced by a given cultural importance of a particular domain.

Apart from these seemingly universal cognitive principles, the patterns of cross-linguistic colexifications are also influenced by common ancestry and language contact. This was shown by Jackson et al. (2019) who investigated cross-linguistic colexifications in the domain of emotions and found that languages have different associations between emotional concepts. Persian speakers connect the concepts of GRIEF with REGRET whereas Dargwa speakers associate it with ANXIETY. While there was an underlying universal core reflected in the importance of the categories of valence and activation, the subtle differences in how emotions are conceptualized in diverse languages offer important insights into the origins of cross-linguistic colexifications.

## Process 2: Changing Meanings over Time

Coming back to the example of the cross-linguistic colexification between SKIN and BARK. This colexification occurs in 213 language varieties (see https://clics.clld.org/edges/763-1204) and provides clues about a historical pattern.

Polysemy has been established to be an integral part of semantic change (e.g., Brown & Witkowski 1983; Wilkins 1996; Koch 2016). The idea is that meanings undergo a process of change that involves an intermediate stage of polysemy. So the original word for SKIN is extended based on a semantic similarity to BARK leading to the colexification of SKIN-BARK. In the final stage, the language is said to develop two individual words for SKIN and BARK. Urban (2011) sketches a scenario with an additional intermediate stage of overt marking which includes a complex form 'tree-skin' for BARK that then leads to the colexification of SKIN-BARK. Thus, "[s]ynchronic polysemy becomes crucial in the investigation of semantic changes because it acts as a proof of the plausibility that two meanings are semantically related and that one meaning could give rise to the other" (Wilkins 1996).

The first aim of the investigation of semantic change is to identify commonalities across languages and the second is to work out the directionality of the semantic change. Thus, cross-linguistic colexification patterns can be seen as a first step in the analysis of semantic change since the recurring colexification of two concepts indicates that they are semantically related. By applying diachronic semantic maps, François (2021) analyzed which forms are associated with which meaning at a given time and showed that lexical competition initiates semantic change. The second step, defining the directionality, is more challenging. Witkowski and Brown (1985) proposed that the direction of a semantic change is based on the salience of the referent so that the word for HAND extends to ARM since the hand is a more salient referent. This assumption is supported by the fact that the word for HAND more commonly occurs as an unmarked form indicating that the process goes from unmarked to marked (Witkowski & Brown 1985). In establishing the unidirectional shifts in the domain of body parts, Wilkins (1996) argued for a natural tendency of semantic change from visible part to visible whole and a non-arbitrariness of the process. He then applied the analysis to identify cognates (i.e., words in different languages that derive from the same ancestor language) in Australian languages based on etymological dictionaries. In contrast, Urban (2011) presents an alternative approach based on synchronic data that uses preferences for word-formation devices to work out the direction of a semantic change.

Identifying the pathways of a semantic change is closely connected to the cognitive processes discussed above. While historical linguists focus on cross-linguistic similarities, they are able to recognize recurring connections and hypothesize about the nonlinguistic cognitive mechanism underlying these patterns. However, there is no exchange of key findings between historical linguistics and cognitive science although it would enable researchers to explain the existence of historical developments due to particular cognitive mechanisms. This shortcoming is recognized by historical linguists (e.g., Wilkins 1996) and proposals for connecting knowledge about semantic change with speaker judgments are put forward (e.g., Koch & Marzo 2007). However, there is

yet no research program that undertakes this mission. One of the most promising approaches, which unfortunately came to a sudden end, was that of Peter Koch, who used diachronic cognitive onomasiology, which combines onomasiology and cognitive linguistics in the light of a diachronic perspective.

## Process 3: Speaking across Linguistic Borders

Sorting out the origins of cross-linguistic colexifications becomes even more complicated if we consider the process of language contact. Here, it is important to distinguish between processes in which words are borrowed or loaned from another language compared to general semantic properties that are shared across languages in a particular geographic area (for details on areal lexico-semantics, see Koptjevskaja-Tamm & Liljegren 2017; Schapper & Koptjevskaja-Tamm 2022). The latter will be the focus of the following paragraphs. One of the subtypes of lexico-semantic parallels is "polysemy copying" (Heine & Kuteva 2003), also called "polysemy calquing" (Koptjevskaja-Tamm & Liljegren 2017, see also Urban 2012). This phenomenon occurs if a language copies a colexification of another language into its own lexicon. Thus, identifying areal patterns of particular cross-linguistic colexifications can help in establishing a linguistic area. Here, the distinction between loose and strict colexifications becomes important (François 2008).

Studies investigating loose colexifications are based on dense samples and make use of a fine-grained semantic analysis (e.g., Urban 2010; Schapper et al. 2016; Schapper 2022; Urban 2022). In a cross-linguistic analysis of words for the concept SUN which is expressed by complex terms roughly translated as 'eye of the day', Urban (2010) showed that this lexico-semantic pattern occurs predominantly in languages of Southeast Asia and Oceania, especially in the language families Austroasiatic, Tai-Kadai, and Austronesian. Contrary to Urban's analysis, Blust (2011) showed that the pattern is more widespread, occurring in many languages outside of the proposed area, and argued for treating the linguistic expression eye of the day for SUN as a linguistic universal. Another areal semantic pattern discussed in detail is the colexification between the concepts TREE, FIREWOOD, and FIRE in Australian and Papuan languages. Schapper et al. (2016) demonstrated that the colexification FIREWOOD-FIRE is common in the languages of Sahul but not the connection with TREE. By looking at individual cross-linguistic colexifications these studies offer important insights into how contact can shape the lexicon of languages in the same geographic region.

Another approach is to use strict colexifications and work with networks of pairwise colexifications to identify areal patterns based on quantitative methods (Gast & Koptjevskaja-Tamm 2019, 2022; Georgakopoulos et al. 2022). The first study utilizing

this approach was by Gast and Koptjevskaja-Tamm (2019) who compared patterns of cross-linguistic colexifications in two databases: CLICS$^2$ (List et al. 2018) and ASJP (version 17, Wichmann et al. 2016). They identified various areal patterns and showed that the databases diverged in their results which is in part due to intrinsic biases in the underlying data. The study can be seen as a point of departure for further detailed analysis of the patterns in the proposed areas. This was done in a prequel study by Gast and Koptjevskaja-Tamm (2022) revealing the extent of variation in the persistence and diffusibility of colexification patterns in European languages. While zooming in on the domains of perception and cognition, Georgakopoulos et al. (2022) presented analyses of different statistical measures and applied them to three datasets. This bottom-up approach yielded interesting findings, for example, that knowledge is more frequently associated with vision and understanding with hearing in the CLICS$^2$ data (Georgakopoulos et al. 2022). By supplementing the strict colexifications found within language islands in a particular area with short phrases to elicit the context in which they occur, Souag (2022) presented a detailed contact scenario of African languages based on quantitative evidence. This study explicitly illustrated the processes that lead languages to form a linguistic area and the usefulness of cross-linguistic colexifications to establish such a scenario.

Regardless of the methodological approach, the study of cross-linguistic colexifications from an areal perspective is important for disentangling the origins of colexifications since it allows one to determine colexifications that are shared due to a cultural similarity rather than a genealogical relatedness or a universal cognitive principle.

## What about Coincidence?

One process that has been neglected from the discussion so far (in this blog post and in many studies that use the term colexification) is *coincidence*. A hypothetical colexification between the concepts BELLY and PEN found in one language is likely to be a chance colexification rather than representing a universal, historical, or areal process. There are multiple examples of unrelated meanings being expressed by the same word in languages. These instances of homophony could blur the picture of true semantic similarity. Thus, studies on cross-linguistic colexifications commonly focus on patterns that appear in languages belonging to different language families to ensure that the probability of finding the origin of a chance colexification is limited. However, this procedure raises the question of why scholars are adopting the term colexification rather than polysemy.

The term colexification was explicitly proposed for the analysis of lexical semantic patterns across diverse languages (François 2008). From my point of view, there are two

motivations that cause colexification to be the popular choice across research fields. On a theoretical level, the proposal by François (2008) unites the phenomena of polysemy, homophony, and vagueness which even in an analysis of an individual language can be difficult to differentiate. On a methodological level, François (2008) introduced empirical cross-linguistic observations as the basis for distinguishing meanings. This approach comes with the methodological advantage of identifying lexical patterns based on cross-linguistic data rather than picking out distinctive semantic patterns based on a researcher's observation. By implementing restrictions on the networks of a given dataset, true polysemy patterns can be established and represented with the caveat of ignoring some less frequent colexifications (List et al. 2013). These cross-linguistic colexification networks can be used to explore community structures and analyze general processes across diverse languages.

## Conclusion

The study of cross-linguistic colexifications holds a wide range of theoretical and methodological advances in the future that will expand our understanding of word meanings. Especially exciting is the integration of a theoretical construct from linguistics in adjacent research fields such as cognitive science and computational science. The more we explore the different origins of colexifications, the better we will connect the knowledge about the three processes and arrive at a holistic view of linguistic diversity. For this, we need an interdisciplinary approach and I look forward to working on this endeavor in the future.

## References

Andersen, Elaine S. 1978. Lexical universals of body-part terminology. In Joseph H. Greenberg (ed.), *Universals of Human Language: Word Structure, vol. 3*, 333–368. Stanford, California: Stanford University Press.

Bao, Hongchang, Bradley Hauer & Grzegorz Kondrak. 2021. On universal colexifications. In Piek Vossen & Christiane Fellbaum (eds.), *Proceedings of the 11th Global WordNet Conference*, 1–7. University of South Africa: Global WordNet Association. https://www.aclweb.org/anthology/2021.gwc-1.1.

Berlin, Brent & Paul Kay. 1969. *Basic color terms: Their universality and evolution*. Berkeley, USA: University of California Press.

Brochhagen, Thomas & Gemma Boleda. 2022. When do languages use the same word for different meanings? The Goldilocks principle in colexification. *Cognition* 226. 1–8. https://doi.org/10.1016/j.cognition.2022.105179.

Brown, Cecil H. & Stanley R. Witkowski. 1983. Polysemy, lexical change and cultural importance. *Man* 18(1). 72–89. https://doi.org/10.2307/2801765.

Brown, Cecil H. 1976. General principles of human anatomical partonomy and speculations on the growth of partonomic nomenclature. *American Ethnologist* 3(3). 400–424. https://doi.org/10.1525/ae.1976.3.3.02a00020.

Chomsky, Noam. 2002. An interview on minimalism. In Adriana Belletti & Luigi Rizzi (eds.), *On Nature and Language*, 92–161. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511613876.005.

Di Natale, Anna, Max Pellert & David Garcia. 2021. Colexification networks encode affective meaning. *Affective Science* 2. 99–111. https://doi.org/10.1007/s42761-021-00033-1.

François, Alexandre. 2008. Semantic maps and the typology of colexification: Intertwining polysemous networks across languages. In Martine Vanhove (ed.), *From polysemy to semantic change: Towards a typology of lexical semantic associations*, 163–215. Amsterdam/Philadelphia: John Benjamins Publishing. https://doi.org/10.1075/slcs.106.09fra.

François, Alexandre. 2021. Lexical tectonics: Mapping structural change in patterns of lexification. *Zeitschrift fur Sprachwissenschaft* 41(1). 89–123. https://doi.org/10.1515/zfs-2021-2041.

Gast, Volker & Maria Koptjevskaja-Tamm. 2019. The areal factor in lexical typology. In Daniel Van Olmen, Tanja Mortelmans & Frank Brisard (eds.), *Aspects of linguistic variation*, 43–82. Berlin/New York: Walter de Gruyter. https://doi.org/10.1515/9783110607963-003.

Gast, Volker & Maria Koptjevskaja-Tamm. 2022. Patterns of persistence and diffusibility in the European lexicon. Linguistic Typology, 26(2), 403-438. https://doi.org/10.1515/lingty-2021-2086.

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. https://doi.org/10.1515/lingty-2021-2088.

Georgakopoulos, Thanasis & Stéphane Polis. 2022. New avenues and challenges in semantic map research (with a case study in the semantic field of emotions). *Zeitschrift fur Sprachwissenschaft (Semantic Maps (Special Issue))* 41(1). 1–30. https://doi.org/10.1515/zfs-2021-2039.

Haspelmath, Martin. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In Michael Tomasello (ed.), *The New Psychology of Language: Cognitive and Functional Approaches To Language Structure, Volume II*, 217–248. New York, NY, US: Psychology Press. https://doi.org/10.4324/9781410606921-11.

Heine, Bernd & Tania Kuteva. 2003. On contact-induced grammaticalization. *Studies in Language* 27(3). 529–572. https://doi.org/10.1075/sl.27.3.04hei.

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. 1517–1522. https://doi.org/10.1126/science.aaw8160.

Karjus, Andres, Richard A. Blythe, Simon Kirby, Tianyu Wang & Kenny Smith. 2021. Conceptual similarity and communicative need shape colexification: An experimental study. *Cognitive Science* 45(9). 1–30. https://doi.org/10.1111/cogs.13035.

Koch, Peter & Daniela Marzo. 2007. A two-dimensional approach to the study of motivation in lexical typology and its first application to French high-frequency vocabulary. *Studies in Language* 31(2). 259–291. https://doi.org/10.1075/sl.31.2.02koc.

Koch, Peter. 2008. Cognitive onomasiology and lexical change: Around the eye. In Martine Vanhove (ed.), *From polysemy to semantic change: Towards a typology of lexical semantic associations*, 107–137. Amsterdam/Philadelphia: John Benjamins Publishing. https://doi.org/10.1075/slcs.106.07koc.

Koch, Peter. 2016. Meaning change and semantic shifts. In Paivi Juvonen & Maria Koptjevskaja-Tamm (eds.), *The lexical typology of semantic shifts*, 21–66. Berlin, Germany: De Gruyter Mouton. https://doi.org/10.1515/9783110377675-002.

List, Johann-Mattis, Simon J. Greenhill, Cormac Anderson, Thomas Mayer, Tiago Tresoldi & Robert Forkel. 2018. CLICS$^2$: An improved database of cross-linguistic colexifications assembling lexical data with the help of cross-linguistic data formats. *Linguistic Typology* 22(2). 277–306. https://doi.org/10.1515/lingty-2018-0010.

Piantadosi, Steven T., Harry Tily & Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition* 122(3). 280–291. https://doi.org/10.1016/j.cognition.2011.10.004.

Schapper, Antoinette, Lila San Roque & Rachel Hendery. 2016. Tree, firewood and fire in the languages of Sahul. In Paivi Juvonen & Maria Koptjevskaja-Tamm (eds.), *The lexical typology of semantic shift*s, 355–422. Berlin/New York: Walter de Gruyter.

Schapper, Antoinette. 2022. Baring the bones: The lexico-semantic association of bone with strength in Melanesia and the study of colexification. *Linguistic Typology* 26(2). 313–347. https://doi.org/10.1515/lingty-2021-2082.

Schapper, Antoinette & Maria Koptjevskaja-Tamm. 2022. Introduction to special issue on areal typology of lexico-semantics. *Linguistic Typology* 26(2). 199–209. https://doi.org/10.1515/lingty-2021-2087.

Souag, Lameen. 2022. How a West African language becomes North African, and vice versa. *Linguistic Typology* 26(2). 283–312. https://doi.org/10.1515/lingty-2021-2083.

Xu, Yang, Khang Duong, Barbara C. Malt, Serena Jiang & Mahesh Srinivasan. 2020. Conceptual relations predict colexification across languages. *Cognition* 201. 1–9. https://doi.org/10.1016/j.cognition.2020.104280.

Ullmann, Stephen. 1953. Descriptive semantics and linguistic typology. *WORD* 9(3). 225–240. https://doi.org/10.1080/00437956.1953.11659471.

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

Urban, Matthias. 2011. Asymmetries in overt marking and directionality in semantic change. *Journal of Historical Linguistics* 1(1). 3–47. https://doi.org/10.1075/jhl.1.1.02urb.

Urban, Matthias. 2012. *Analyzability and semantic associations in referring expressions: A study in comparative lexicology*. Leiden, Netherlands: Leiden University. PhD thesis. https://scholarlypublications.universiteitleiden.nl/handle/1887/19940.

Urban, Matthias. 2022. Red, black, and white hearts: 'heart', 'liver', and 'lungs' in typological and areal perspective. *Linguistic Typology* 26(2). 349–374. https://doi.org/10.1515/lingty-2021-2081.

Wichmann, Søren, Cecil H. Brown & Eric W. Holman (eds.). 2016. *The ASJP Database (Version 17)*. Jena, Germany: Max Planck Institute for the Science of Human History. https://asjp.clld.org/.

Wilkins, David P. 1996. Natural tendencies of semantic change and the search for cognates. In Mark Durie & Malcolm Ross (eds.), *The comparative method reviewed: Regularity and irregularity in language change*. Oxford/New York: Oxford University Press.

Witkowski, Stanley R. & Cecil H. Brown. 1985. Climate, clothing, and body-part nomenclature. *Ethnology* 24(3). 197–214. https://doi.org/10.2307/3773610.