

The Ambiguity of the Colexification Term

Johann-Mattis List / Katja Bocklage / Thanasis Georgakopoulos

MCL Chair / MCL Chair / School of English

University of Passau / University of Passau / Aristotle University of Thessaloniki

The term *colexification* has become an important concept in lexical typology in particular and linguistic typology in general. Surprisingly, it is not consistently used by all scholars in the field. Instead, two main interpretations have emerged, a broad version that uses *colexification* as a cover term for polysemy, vagueness, and homophony, and a narrow version that restricts the term to polysemy and vagueness. We discuss the ambiguity and reflect about the usefulness of introducing alternative terms instead.

1 Introduction

The introduction of the term *colexification* to the field of lexical typology by François (2008), can be seen as marking a turning point in the field of lexical typology. Before, lexical typology was a niche discipline pursued by a small number of scholars, concentrating on very specific domains of the lexicon of human languages, such as color, kinship, and body. After the introduction of the term in a widely recognized volume on lexical typology by Vanhove (2008), devoted to the role that polysemy plays in investigating semantic shift, both the available data and the methodology by which the data could be analyzed increased drastically. In 2010, Cysouw (2010) showed, how recurrent polysemies can be visualized and analyzed as a network. In 2011, Urban (2011) presented the idea that certain directional tendencies of semantic change can be inferred from the existence of examples in which semantic similarities are overtly marked in individual language varieties. In 2014, the *Database of Cross-Linguistic Colexifications* (CLICS, <https://clics.clld.org>) was published (List et al., 2014), building on Cysouws' idea of polysemy networks which were derived from a collection of 300 comparative wordlists that were automatically searched for *colexifications*, that is, cases where different concepts in the same language were expressed by identical word forms.

With these new resources and ideas to visualize and explore semantic relations in large-scale accounts of cross-linguistic data, lexical typology grew into a popular subfield of linguistic typology (Koptjevskaja-Tamm et al. 2007, Koptjevskaja-Tamm et al. 2015, Sjöberg et al. forthcoming). While previous studies had compared how the lexicons of human languages organize meaning in different ways, had relied on case studies and small-scale examples involving hand-selected languages, it was now possible to study meaning organization through the lens of a large part of the world's languages with the help of technically and theoretically considerably simple means.

2 From Polysemy to Colexification

François originally defined *colexification* as a relation that holds for the senses that are expressed by identical word forms in one and the same language.

A given language is said to colexify two functionally distinct senses if, and only if, it can associate them with the same lexical form. (François, 2008, p. 170)

From this definition, it is clear that colexification is intended as a *cover term* for the previously established terminology around *polysemy*, *vagueness*, and *homophony*. In linguistic terminology, there are several terms that can be used to *specific* cases where two senses are associated with the same lexical form, such as *polysemy* and *vagueness*, referring to senses associated with the *same* form, and *homophony*, referring to senses associated with different forms that came to sound alike at some point in the history of a language.

From the perspective of linguistic terminology, the term was useful in two regards. Since, on the one hand, it replaces largely *explanative terminology* with *descriptive terminology* (see Jacques & List, 2019, p. 141 on this distinction), it facilitates, on the other hand, the investigation of lexical semantics across languages, since it shifts the burden of proof from the stage of data collection to the stage of data analysis. Explanative terminology (one could also say “diagnostic terminology”) does not only describe a phenomenon, but also tries to explain it at the same time. Polysemy is usually *explained* as the result of semantic change when individual forms acquire more senses, while homophony is *explained* as the result of sound change leading to phonological merger in individual forms of a language. Determining that two senses are colexified in a given language does not explain *why* this is the case, but merely notes this as a fact inferred from empirical observations. This has drastic consequences for the investigation of lexical semantics across languages. While any cross-linguistic study on polysemy, vagueness, or homophony, would have had to distinguish these phenomena

beforehand, reducing the study to the overall phenomenon of distinct senses being expressed by the same form enabled large-scale investigations based on standardized comparative wordlist collections.

In such a facilitated workflow, one would no longer have to search individual languages for instances of polysemy in a fixed set of meanings, but could rather search for *colexifications*, that is, cases where two or more concepts in a comparative wordlist are expressed by identical translational equivalents. How these cases could be *classified* would be a second step, where early research immediately showed that cases of polysemy and vagueness tend to recur across different language families, while cases of homophony tend to be restricted to individual languages and individual language families (List et al., 2013). As a result, the distinction between polysemy and vagueness on the one hand and homophony on the other hand, can be readily established *after* colexification patterns have been assembled for a sufficiently large number of languages and language families.

This procedure was the basic idea behind the CLICS database, which provides an automated procedure to harvest colexification patterns from cross-linguistic wordlist collections that was modified and improved across the database's four different major versions (List et al., 2014, 2018; Rzymiski et al., 2020; Tjuka et al., 2025). When using colexification as a cover term for polysemy, vagueness, and homophony, CLICS assembles individual examples from different languages of all these phenomena and later allows to single out cases of homophony by focusing on the most frequently observed *patterns*.

3 Intended Definition of Colexification

Recent discussions at the workshop on *Semantic Shifts and the Dynamics of Lexification Patterns* organized by Alexandre François, Anna Zalizniak, and Anna Smirnitskaya as part of the *16th International Conference of the Association for Linguistic Typology in Lyon* (July 1-3, 2026) made it clear that the term *colexification* is used in two different ways by linguists from different backgrounds (or “schools”). On the one hand, there are linguists who work either in the tradition of François’ original work or in the context of the *Database of Semantic Shifts* (<https://datsemshift.ru>, Zalizniak et al., 2023), a large-scale catalogue of various attested kinds of semantic shifts that have been collected for more than 20 years now directly from the linguistic literature (Zalizniak, 2018). On the other hand, there are linguists working in the context of the CLICS database. The former tradition restricts the term *colexification* to cases of polysemy and vagueness, excluding cases of homophony from the definition. In the context of the work of the latter tradition,

colexification was always a cover term for *all* cases where one form has multiple senses in a given language.

Given that the tradition restricting colexification to polysemy and vagueness is represented by the scholar who introduced the term colexification, there are good reasons to follow the narrower definition of colexification, even if the original text introducing the term does not explicitly exclude homophony as one of the aspects covered by the term. Even in the case of the CLICS database, which was profiting so much from the broader version of colexification in order to justify its workflow, the narrow version of colexification might be useful and even help to clarify past studies. From these studies on emotion concepts (Jackson et al., 2019), body part terms (Tjuka, 2024; Tjuka et al., 2024), perception and cognition (Georgakopoulos et al. 2022), and lexical creativity (Brochhagen et al., 2023), it is quite clear that the main intention of the database is to *infer* colexifications in the narrow sense, filtering out cases of homophony by rather simple quantitative means.

However, treating colexification as a mere cover term for vagueness and polysemy means at the same time, that the two major advantages of the term, the shift from explanative to descriptive terminology and the facilitation that this shift provides for quantitative approaches, would vanish, at least in parts. It would also mean that terminological frameworks that attempt to view colexification in the larger context of *coexpression* would have to be revised, especially since cases of *cogrammification* as defined by Haspelmath (2023) also do not ask *how* a pattern emerged, but rather conclude that a pattern can be *determined* from the data. Additionally, it may be questionable whether it is actually possible to strictly distinguish between homophony on the one hand and polysemy and vagueness on the other hand in all cases (consider the ambiguity of English *ear*, as referring to a body part or the part of a plant, which is often seen as some kind of polysemy, although it is the result of phonological merger). Thus, despite François' original intention, one may argue that the broad version of colexification is actually in line with the established practice in linguistic typology.

As often, when it comes to terminology, it seems that the field will have to live with some compromise solutions that emerge from the research practice. Abandoning the broad definition of colexification would require us to employ a new term, which may be seen as a minor problem, but given the conflict with the practice in general discussions on coexpression and semantic maps, we would be forced to revise the terminology in this area more generally and broadly. Retaining the broad definition for colexification as the norm has the disadvantage that in most scholars' *practice*, homophony is essentially ignored, no matter whether they see colexification in its broad or narrow version. It

seems for the time being, that it is best to leave everything as is, while being prepared to mention the ongoing conflict with the terminology explicitly, where needed.

4 Beyond Colexification

While François agreed in the discussion about the term *colexification* at the workshop at the ALT conference, that the original paper (François, 2008) does *not* explicitly exclude homophony from the definition, reading the original paper with the revised narrow definition in mind clarifies those parts that go beyond *strict colexification* proper.

In particular, “strict colexification” (same lexeme in synchrony) should be carefully distinguished from “loose colexification” (covering all other cases mentioned here). (François, 2008, p. 171)

Strict colexification is contrasted with *loose colexification*. Under the latter term François (2008) subsumes a variety of other, more or less similar, relations between concepts mediated through their word forms. Among these are *diachronic semantic change*, in which a word form *w* expresses a concept *A* at a point in time *t1* but refers to concept *B* at a later point in time *t2* (at which stage concept *A* may either still be attested or have been lost). Cases involving diachronic semantic change are comparable to strict colexification in that the concepts relate through an identical word form; unlike strict colexification, however, these word forms belong to different historical stages. Owing to this similarity Georgakopoulos & Polis (2021) do not treat the two cases as entirely distinct, but instead introduced the term *diachronic strict colexification*, which captures both the diachronic dimension of the phenomenon and the fact that the same form is involved.

While the broad definition of colexification may be seen as problematic here, since cases in which word forms overlap in parts, are the norm rather than the exceptions in natural languages, even if the word forms involved do not display any etymological connections, it is clear from the examples given by François that colexification is used in the *narrow* sense. This means that François does not consider cases where two word forms show some coincidental overlap in some parts, but rather presupposes that loose colexifications can only be determined if one can prove the historical unity of the shared material in question.

While it seems basically convincing to restrict colexification along these lines, recent research devoted to so-called *partial colexifications*, that is, colexification patterns involving *parts* of the linguistic form (List et al., 2022), seem to show again, that the notion of colexification as a descriptive term has concrete advantages for computational

studies. Thus, List (2023) showed how the idea of *synchronic asymmetries* in *overt marking* proposed by Urban (2011) can be operationalized in a quantitative framework and applied to larger collections of comparative wordlists by searching automatically for those cases in which one word occurs at the beginning or the end of another word in a given language. Since these substring relations are called *affixes* in computer science, List proposes to term these relations *affix colexifications*, as a specific case of *partial colexification*, emphasizing that the inference of these relations across different forms in a given language does not necessarily reflect valid semantic relations, but showing that potentially valid forms can again be inferred by filtering *patterns* across multiple languages and language families. Partial colexification can be seen as more restrictive than loose colexification, given that the latter includes etymological relations, while the former explicitly points to the synchronic identity of *parts* of the linguistic form in a given language.

That cross-linguistically recurring affix colexifications can give hints to valid and potentially interesting semantic relations is further illustrated by a study of Bocklage et al. (2025). In an empirical evaluation, the authors show that filtering affix colexifications by frequency of occurrence across language families can indeed help minimizing the number of false positives.

Abandoning the term *partial colexification* due to its conflict in coverage with the term *loose colexification* may again seem to be the easiest way to deal with the terminological problems here. But again, we must observe that first, the definition of loose colexifications does not explicitly exclude cases of coincidence in form overlap. Second, abandoning the broad colexification definition would yield conflicts with the general coexpression terminology in linguistic typology, given that coexpression cannot rule out coincidence as easily as lexical typology can rule out homophony.

As a result, it may be useful to emphasize that the terms *loose colexification* and *partial colexification* are not compatible, given that they reflect different scholarly traditions, the former being at least in part explanative (definitely diagnostic), and the latter being deliberately descriptive.

5 Conclusion

From what has been discussed so far, it will probably be clear that the terminology with respect to and around the term *colexification* has made some infelicitous turns so far. As a result, we are – yet another time – faced with terminological traditions in our field that display certain conflicts among scholars, differences in opinions, and also sheer misunderstandings. On the one hand, the young history that the term *colexification* experienced so far, can be seen as representative for typical terminology struggles in

science. On the other hand, one should not ignore the momentum of serendipity reflected in the fate of the term *colexification*, given that the reinterpretation of the partially explanative term as a purely descriptive term has given scholars the chance to develop simple but very effective quantitative approaches to lexical typology. Future research will show if additional solutions can be found. For now, all we can recommend is that scholars who work with colexifications make clear in which tradition they allocate their terminology.

References

- Bocklage, K., Georgakopoulos, T., Dam, K. P. van, Ciucci, L., Blum, F., Kučerová, A., Rubehn, A., Stephen, A., Snee, D., & List, J.-M. (2025). Testing the Potential of Automatically Inferred Affix Colexifications for Linguistic Typology. in *Humanities Commons* (pp. 1–35). <https://doi.org/10.17613/a06m1-c9939>
- Brochhagen, T., Boleda, G., Gualdoni, E., & Xu, Y. (2023). From language development to language evolution: A unified view of human lexical creativity. *Science*, 381(6656), 431–436. <https://doi.org/10.1126/science.ade7981>
- Cysouw, M. (2010). Drawing Networks from Recurrent Polysemies. Comment on “polysemous qualities and universal networks” by loïc-michel perrin (2010). *Linguistic Discovery*, 8(1), 281–285.
- François, A. (2008). Semantic maps and the typology of colexification: intertwining polysemous networks across languages. in M. Vanhove (Ed.), *From polysemy to semantic change* (pp. 163–215). Benjamins.
- Georgakopoulos, T. and Polis, S. (2021): Lexical diachronic semantic maps. Mapping the evolution of time-related lexemes. *Journal of Historical Linguistics* 11(3). 367-420. <https://doi.org/10.1075/jhl.19018.geo>
- Georgakopoulos, T., Grossman, E. , Nikolaev, D., and S. 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>
- Haspelmath, M. (2023). Coexpression and synexpression patterns across languages: comparative concepts and possible explanations. *Frontiers in Psychology*, 14, 1–12. <https://doi.org/10.3389/fpsyg.2023.1236853>
- Jackson, J. C., Watts, J., Henry, T. R., List, J.-M., Mucha, P. J., Forkel, R., Greenhill, S. J., Gray, R. D., & Lindquist, K. (2019). Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472), 1517–1522. <https://doi.org/10.1126/science.aaw8160>
- Jacques, G., & List, J.-M. (2019). Save the trees: Why we need tree models in linguistic reconstruction (and when we should apply them). *Journal of Historical Linguistics*, 9(1), 128–166. <https://doi.org/10.1075/jhl.17008.mat>
- Koptjevskaja-Tamm, M., Vanhove, M., and Koch, P. (2007). Typological approaches to lexical semantics. *Linguistic Typology*, 11(1), 159–185. <https://doi.org/10.1515/LINGTY.2007.013>
- Koptjevskaja-Tamm, M., Rakhilina, E., and M. Vanhove (2015): The Semantics of Lexical Typology. In Nick Riemer (ed.), *The Routledge Handbook of Semantics* (pp. 434–454). London: Routledge.

- List, J.-M. (2023). Inference of partial colexifications from multilingual wordlists. *Frontiers in Psychology*, 14(1156540), 1–10. <https://doi.org/10.3389/fpsyg.2023.1156540>
- List, J.-M., Forkel, R., Greenhill, S. J., Rzymiski, C., Englisch, J., & Gray, R. D. (2022). Lexibank, A public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, 9(316), 1–31. <https://doi.org/10.1038/s41597-022-01432-0>
- List, J.-M., Greenhill, S. J., Anderson, C., Mayer, T., Tresoldi, T., & Forkel, R. (2018). CLICS². An improved database of cross-linguistic colexifications assembling lexical data with help of cross-linguistic data formats. *Linguistic Typology*, 22(2), 277–306. <https://doi.org/10.1515/lingty-2018-0010>
- List, J.-M., Mayer, T., Terhalle, A., & Urban, M. (2014). *CLICS: Database of Cross-Linguistic Colexifications. Version 1.0* (version 1.0.0). Forschungszentrum Deutscher Sprachatlas. <http://clics.lingpy.org>
- List, J.-M., Terhalle, A., & Urban, M. (2013). Using network approaches to enhance the analysis of cross-linguistic polysemies. *Proceedings of the 10th International Conference on Computational Semantics – Short Papers*, 347–353.
- Rzymiski, C., Tresoldi, T., Greenhill, S., Wu, M.-S., Schweikhard, N. E., Koptjevskaja-Tamm, M., Gast, V., Bodt, T. A., Hantgan, A., Kaiping, G. A., Chang, S., Lai, Y., Morozova, N., Arjava, H., Hübner, N., Koile, E., Pepper, S., Proos, M., Epps, B. V., ... List, J.-M. (2020). The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data*, 7(13), 1–12. <https://doi.org/10.1038/s41597-019-0341-x>
- Sjöberg, A., Georgakopoulos, T., and Koptjevskaja-Tamm, M. (forthcoming): Typology and universals of word meaning. In D. Geeraerts & D. Glynn (Eds.), *The Cambridge handbook of lexical semantics*. Cambridge University Press.
- Tjuka, A. (2024). Objects as human bodies: cross-linguistic colexifications between words for body parts and objects. *Linguistic Typology*, 0(0). <https://doi.org/10.1515/lingty-2023-0032>
- Tjuka, A., Forkel, R., & List, J.-M. (2024). Universal and cultural factors shape body part vocabularies. *Scientific Reports*, 14(10486), 1–12. <https://doi.org/10.1038/s41598-024-61140-0>
- Tjuka, A., Forkel, R., Rzymiski, C., & List, J.-M. (2025). Advancing the Database of Cross-Linguistic Colexifications with New Workflows and Data. *Proceedings of the 16th International Conference on Computational Semantics*, 1–15. <https://aclanthology.org/2025.iwcs-main.1>
- Urban, M. (2011). Asymmetries in overt marking and directionality in semantic change. *Journal of Historical Linguistics*, 1(1), 3–47.
- Vanhove, M. (Ed.). (2008). *From polysemy to semantic change*. Benjamins.
- Zalizniak, A. A. (2018). The Catalogue of Semantic Shifts: 20 years later. *Russian Journal of Linguistics*, 22(4), 770–787. <https://doi.org/10.22363/2312-9182-2018-22-4-770-787>
- Zalizniak, A. A., Smirnitskaya, A., Russo, M., Mikhailova, T., Bobrik, M., Gruntov, I., Orlova, M., Bibaeva, M., & Voronov, M. (2023). *Database of Semantic Shifts [Version from 20/12/2023]*. Institute of Linguistics at the Russian Academy of Sciences. <https://datsemshift.ru/>

Funding Information
This project was supported by the ERC Consolidator Grant ProduSemy (PI Johann-Mattis List, Grant No. 101044282, see https://doi.org/10.3030/101044282). Views and opinions expressed are however those of the author(s) only and do not

necessarily reflect those of the European Union or the European Research Council Executive Agency (nor any other funding agencies involved). Neither the European Union nor the granting authority can be held responsible for them.