Semantic Networks

Johann-Mattis List (University of Passau)

1 Semantics and Semantic Change

It is well known and not surprising for practitioners of historical linguistics that semantics and semantic change are topics that are very difficult to handle systematically. The reason for this lies in what Sperber (1923: 1) calls the *psychological factors* of meaning, which are much more difficult to grasp and describe than it is to give logical definitions of certain concepts.

Apart from the general question where to allocate semantic change (in the domain of the lexicon or the domain of pragmatics, or as a transition between the two, see (Traugott 2012)), the reason for the problems one faces when dealing with semantic change can be found in the structural differences between sign form and sign meaning and the resulting processes by which both entities change. While the formal part of the linguistic sign is characterized by its sequential structure and sound change is characterized by the *alternation* of segments, the meaning part is better described as some kind of *conceptual network*, and semantic change is not based on an alternation but on the *accumulation* and *reduction* of potential referents,[1] for example by a reorganization of the sign's *reference potential* (List 2014: 36). Although change in meaning is traditionally considered to be notoriously irregular and unpredictable, with scholars emphasizing that "there is [...] little in semantic change which bears any relationship to regularity in phonological change" (Fox 1995: 111), it is also obvious that a large number of observed pathways of semantic change can be observed to occur independently in many different language families of the world. In some sense, we face the same problems we also found for the handling of regular sound change patterns. If we want to study pathways of semantic change cross-linguistically, we will need to find a way to make our data comparable. That this can be cumbersome and difficult could be observed for the Catalogue of Semantic Shifts (Zalizniak 2018, Zalizniak et al. 2012), which originally presented a larger collection of observed semantic change processes, but ultimately has problems to provide a rigorous specification of the different meanings that were tracked.[2]

How can we imagine this process of accumulation and reduction to take place, and what is meant by "reference potential"?

2 Multilingual Approaches to Semantic Change

We have repeatedly seen and discussed how notoriously difficult it is to study semantic change systematically, given that, once it comes to "meaning, one has as a guide only a certain probability based on common sense, on the personal evaluation of the linguist, and on the parallels that he can cite" (Wilkins 1996: 264). Interestingly, however, the often-invoked differences between semantic change and sound change become much less striking when we stop to think about sound change as something ultimately *regular*. In the last session, we have discussed the regularity of sound change a lot, and one of the important aspects was that the apparent regularity is nothing else than a change on a higher level, not at the level of the word alone, a change of the phoneme system, as emphasized

[1] This can already be found in the work of Herman Paul (1846–1921), who emphasizes that there is always an "extension or restriction of the extent of the meaning" and that "only the succession of extension and restriction allows the emergence of a new, from the original one completely different meaning" (Paul 1880 [1886]: 66, my translation).

[2] To my knowledge, the authors are currently working on a new version that will hopefully cope with the problems of the older version and also provide an increase in data (see `http://datsemshift.ru`).

early by Bloomfield (1933 [1973]: 351). If we look at the *substance* of sound change, at concrete patterns, and the incredible number of different sound segments which scholars propose to have found in certain languages (Anderson et al. 2018), however, sound change does not seem much more chaotic then semantic change. On the contrary: if it is possible to establish a first reference catalogue of phonetic transcriptions, and if we trust that the initial work done in the Concepticon project has been done thoroughly enough, and if we further keep in mind that diachronic patterns often can also be observed synchronically, we may be able to work on feasible solutions to at least approximately reconstruct basic semantic structure from cross-linguistic data.

How does semantic change surface in synchronic linguistic data?

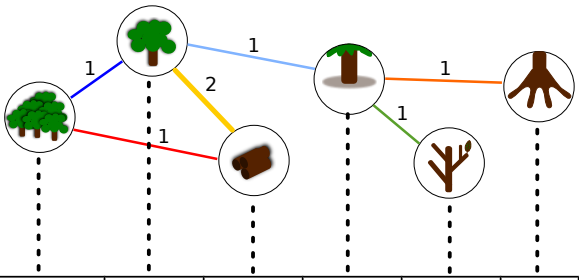
Polysemy, Homophony, and Colexification

Polysemy and homophony are two seemingly contrary concepts in linguistics. However, in the end they describe both the same phenomenon, namely that a word form in a given language can have multiple meanings. François (2008) therefore suggests to replace the two interpretative terms by the descriptive term colexification. Colexification in this context only means that an individual language "is said to colexify two functionally distinct senses if, and only if, it can associate them with the same lexical form" (ibid.: 171).

How can the distinction between interpretative and descriptive terminology be understood?

Colexification Networks

If one has enough data, it is considerably easy to construct *concept networks* from cross-linguistic colexifications (Cysouw 2010). The starting point are semantically aligned word lists for a large amount of different languages from different language families. By counting, in how many languages, or in how many language families a certain colexification recurs, we can further *weight* the edges of the network, as shown in Figure 1.



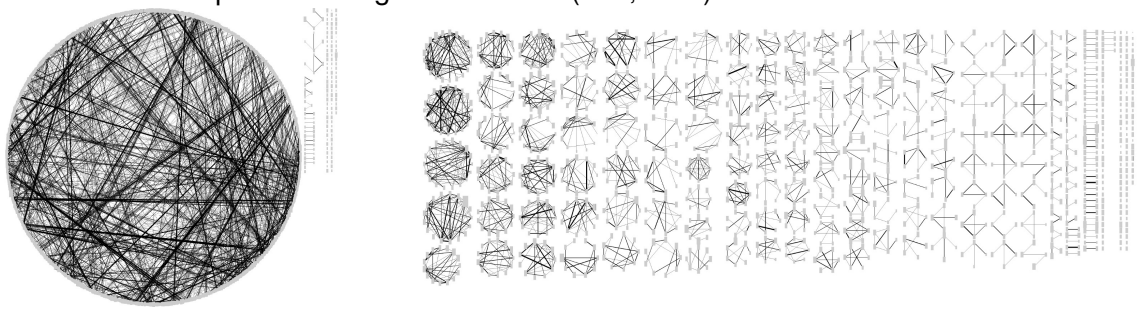
	forest	*tree*	*wood*	*stem*	*branch*	*root*
French	fɔʀɛ, bwɑ	aʀbʀə	bwɑ	tʀɔ	bʀɑʃ	ʀasin
Russian	lʲes	dʲerɪva	dʲerɪva	stvɔl	vʲetvʲ	kɔrɪnʲ
Croatian	ʃuma	staːblɔ	drɔv	staːblɔ	graːna	kɔrɪɛn
Yukaghir	aːnmonilʲe	saːl	saːl	tʃilge	tʃilge	warulu:
Yaqui	dʒuja	dʒuja	kuta	naːwa	budʒa	naːwa

Figure 1: Reconstructing colexification networks from multi-lingual wordlists.

Is there any straightforward way to derive directed graphs from weighted, undirected colexification networks?

Analyzing colexification networks

Taking a colexification network alone does not necessarily help us in answering questions regarding semantic change or human cognition. This is due to the increasing complexity of colexification networks, the more concepts and languages we add. The graphic below, for example, shows a network which has been constructed from an analysis of 195 languages covering 44 language families (List et al. 2013). What we need is a network analysis which uses specific algorithms to analyse the structure of the network more properly. In concrete, analyses for *community detection* can help us to partition the networks into groups which correspond to important *semantic fields*. The term *community* was first coined in social network analysis, where it was used to identify communities of people in social networks. In a broader sense, a community refers to "groups of vertices within which the connectionso are dense but between which they are sparser" (Newman 2004: 4). In List et al. (2013), we used the algorithm by Girvan and Newman (2002) to analyse the network on the left. The result is given in the graphic on the right, where the originally almost completely connected network has been partitioned into 337 communities, with 104 being relatively big (5 and more nodes, covering a rather large parts of the 1289 concepts in our original database (879, 68%).



(a) complete networks **(b) analysed network**

Figure 3: Comparing clustered and unclustered colexification networks.

Below a community from the network is shown, in which meanings which center around "tree" and "wood" have been grouped together. What can we learn from the network? What can't we learn?

Database of Cross-Linguistic Colexifications

CLICS^3 (`https://clics.clld.org`, Rzymski et al. 2020) is an online database of colexifications in about 2000 language varieties of the world. CLICS^3 is the third installment of the Database of Cross-Linguistic Classifications, following the second version published two years before (List et al. 2018), and an even earlier version from 2014 (List et al. 2014), which introduced the interactive representation of cross-linguistic colexification patterns (Mayer et al. 2014) which is still one of the major reasons why CLICS is so popular. While the original CLICS database was low in terms of cross-linguistic coverage and difficult to maintain, the strict adherence to the format specifications based on the CLDF initiative made it possible to grow the data drastically, from originally 221 language varieties in the original version up to 1220 varieties in second version (List et al. 2018), up to more than 2000 varieties in the third installment (Rzymski et al. 2020).[3]

2.1 Data Curation and Aggregation in CLICS^3

The major advancement of CLICS^3 was a new framework for data curation and aggregation, entirely built on the CLDF strategies. Essentially, this workflow consists of four major stages, which can be

[3] We have a new update for CLICS^4 in preparation, which will, however, no longer grow the number of languages covered, but rather concentrate on the quality of the data.

carried out independently from each other. These stages include the *mapping of concepts* to Concepticon (List et al. 2022b), the *referencing of sources* in the original data, the *linking of languages* to Glottolog (Hammarström et al. 2021), and the *cleaning of lexical entries* using a dedicated suite of Python scripts (later published as part of the Lexibank workflow List et al. 2022a). Once data are prepared in this form and rendered in PDF, aggregating data from different sources into a larger database is extremely straightforward. Since the investigation of colexification patterns furthermore does not require to compare word forms *across* languages, but only *inside*, no further normalization (e.g., of the transcriptions) is needed.[4]

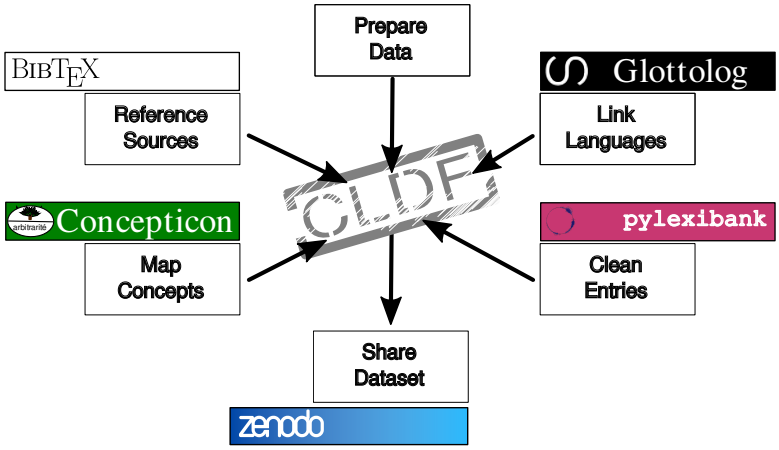


Figure 4: Workflow for data aggregation and curation in CLICS³.

What pitfalls should one avoid when trying to clean lexical entries?

2.2 Examples

The visualization framework used in CLICS is based on an interactive, force-directed, graph layout, written in JavaScript. The basic idea behind this visualization is to allow users to inspect both all the data underlying a given colexification (ideally up to allowing to trace the original datasets, the word forms, and the original elicitation glosses), while at the same time offering a bird's eye view on the global distribution of a given colexification pattern. This is illustrate in the screenshot in Figure 2, where the cluster around words for "tree" and "wood" is shown.

[4] The upcoming fourth installment of the CLICS database, however, will have fully transcribed word forms for a then slightly smaller amount of language varieties, since we decided that transcribed, unified transcriptions offer for more possibilities to analyze the data consistently.

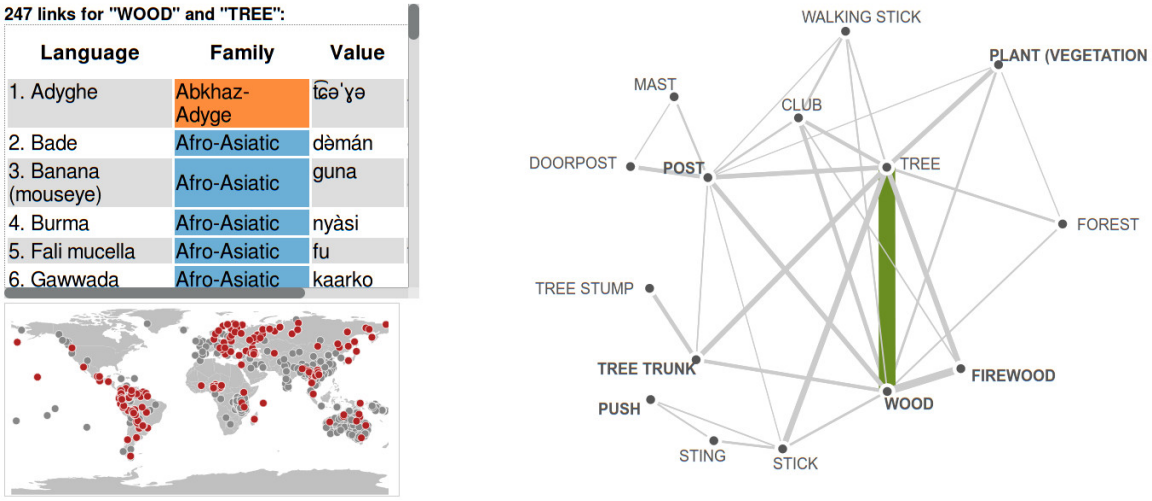


Figure 2: Screenshot from the CLICS² database (see infomap_2_WOOD).

What exactly does this visualization tell us?

3 Beyond Colexification Networks

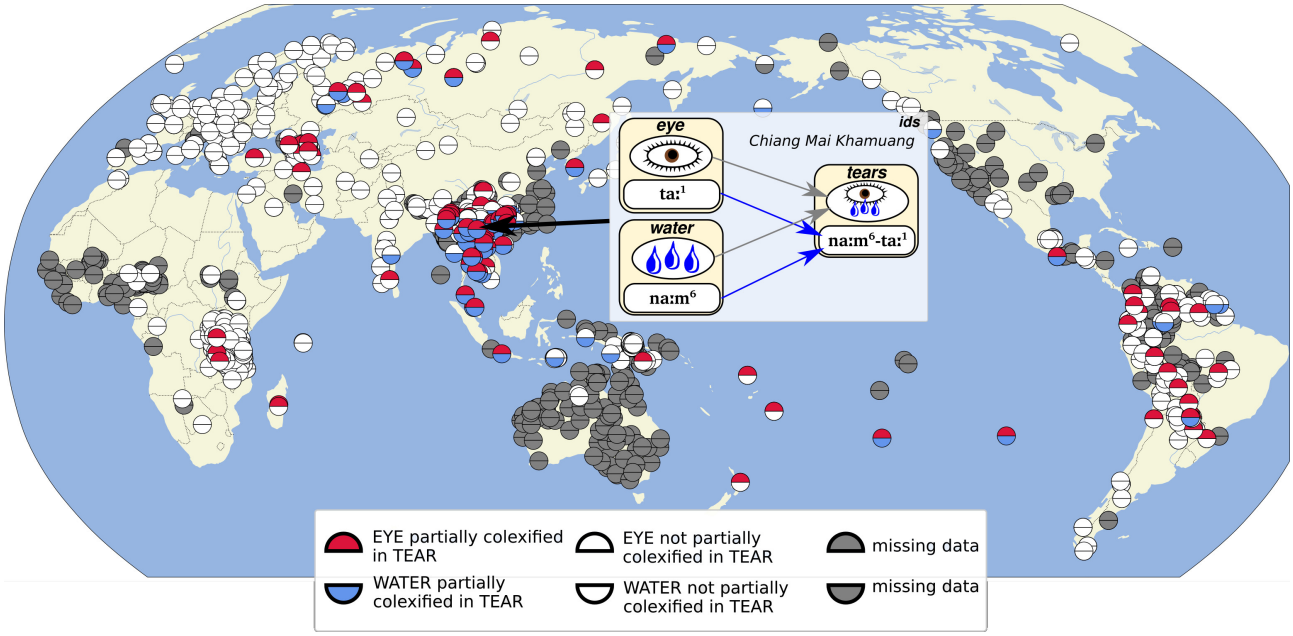
In contrast to the problem of sound change, the identification, the inference of cross-linguistically recurring polysemies can be rather straightforwardly done, by avoiding any distinction between polysemy and homophony in a first place, and then searching for those patterns which recur often enough in big colexification networks. Colexification networks as proposed in the CLICS³ database, however, do not solve all problems. First of all, they are a convenient way to present the data to linguists who are interested in the investigation of polysemy patterns due to their individual research. The colexification data as it was assembled with help of our improved CLDF data curation workflows, however, offer much more potential for future investigations. This is shown, for example, by Gast and Koptjevskaja-Tamm (2018) who study areal aspects of polysemy patterns, as well as by (Georgakopoulos and Polis 2018), who present new ideas to add a diachronic dimension. Additionally, there is a lot of potential for studies that *use* the colexification data in order to check linguistic, cognitive, and psychological theories and hypotheses.

What theories could, for example, be tested, with the help of polysemy patterns?

Lexibank and CL Toolkit

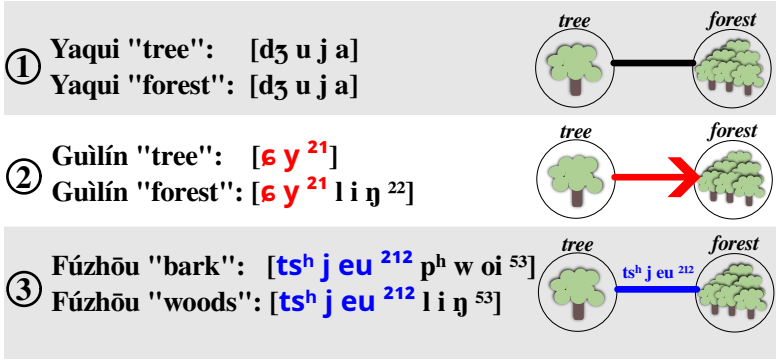
With the publication of the Lexibank database, we have shown how both phonological and lexical features can be automatically extracted from large aggregated collections of CLDF word lists (List et al. 2022a). For 30 exemplary lexical features, we also illustrate how they can be computed with the help of CL Toolkit (List and Forkel 2021), a package that facilitates the representation of features in code. All 30 lexical features defined in this form are based on *colexifications*, but not all features are based on *full colexifications*, but we also look for two types of partial colexifications, one based on the identification of common *substrings*, and one based on the identification of *part-of* relations (called *affix colexifications* in our study). This technique allows us to define individual colexification patterns and then search for them directly in the data in order to see how many languages show these patterns, and how many languages do not show them.

The figure shows the affix colexification for words for EYE being in an affix relation with words for TEAR and words for WATER being in an affix relation with words for TEAR in the Lexibank sample of languages. What do we find regarding the distribution of languages showing both patterns?



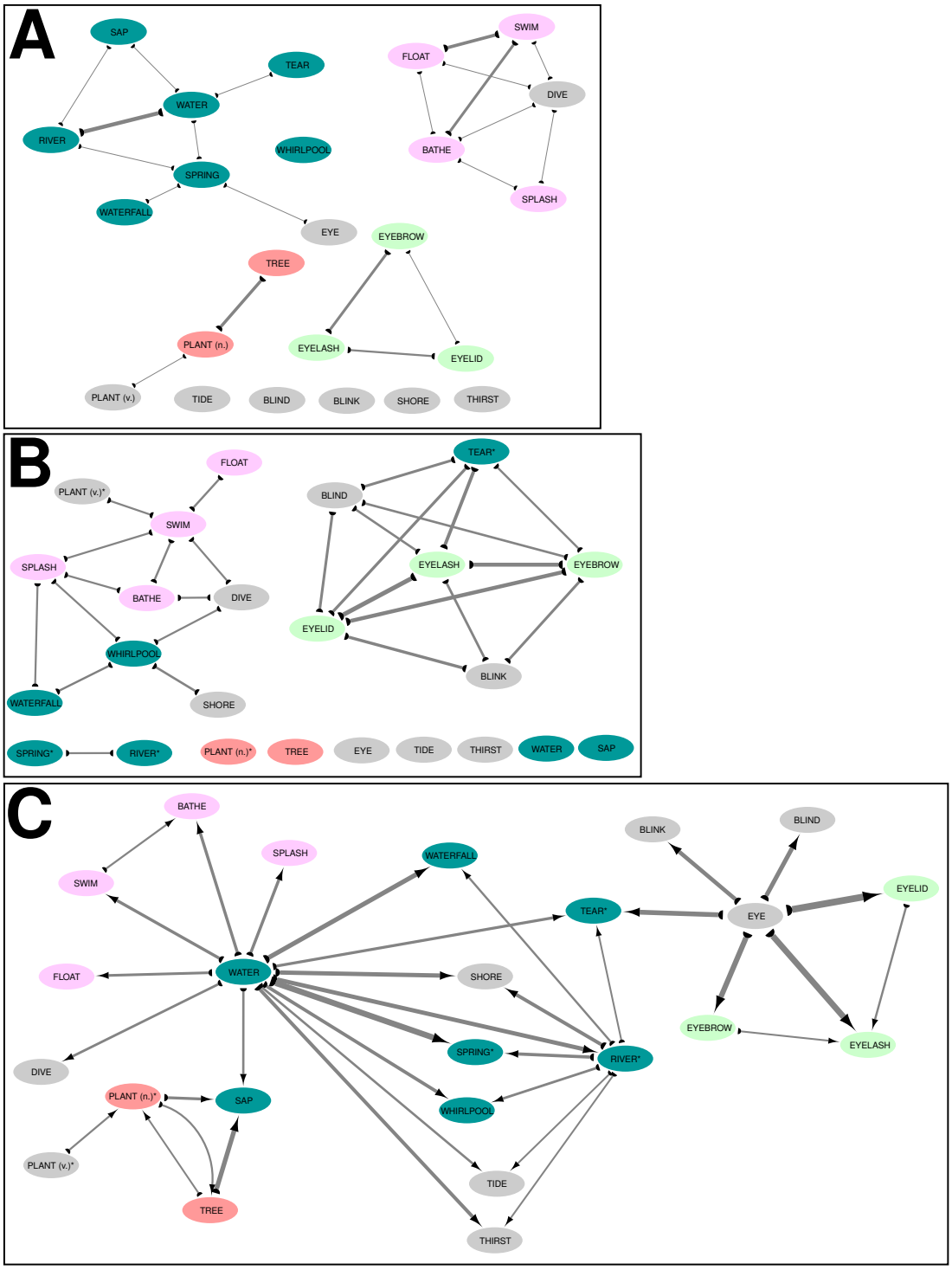
From CLICS to CLIPS

In a study under review (List 2023), we go one step further in trying to derive three kinds of colexification networks, including full, overlap, and affix colexifications from CLDF wordlists. While traditional colexification networks have been defined and used for a long time now (specifically as part of CLICS[1], CLICS[2] and CLICS[3]), the new pilot study defines two kinds of partial colexifications, following the earlier relations proposed in List et al. (2022a), by defining two specific kinds of partial colexifications, namely part-of relations and substring relations. While part-of relations should be modeled in directed networks, with the direction indicating what word is part of the other word, substring relations should be modeled in undirected networks, analogously to "full" colexification networks. To keep computation time at a reasonable level, the study introduces specific subtypes of part-of and substring relations: the affix relation (one word must be either a prefix or a suffix of the other word) and the overlap relation (two words can share a substring, but the substring must be either a prefix or a suffix in both strings).



The results indicate that all three types of colexification networks are fundamentally different, while they are still semantically meaningful. Moreover, when modeling affix colexifications, we find that the

weighted in-degree of these colexification networks correlates moderately (0.42, $r < 0.0001$) with the weighted degree of overlap colexification networks, while the weighted out-degree correlates moderately (0.50, $r < 0.0001$) with the weighted degree of full colexification networks (using Spearman rank correlations, Spearman 1904). These findings can be interpreted in such a way that they point to the tendency that concepts which are generally colexified very often are also frequently re-used as compounds or affixes in complex words. This shows that one could take the out-degree of affix colexifications as evidence for the phenomenon of *lexical root productivity* (the term is inspired by a discussion with Alexandre François, see List 2019a and List 2019b).



The correlation between the in-degree in affix colexification networks, that means, the tendency of words to re-appear in compounds, and the degree of overlap colexification networks, that means, the tendency of concepts to be expressed by a compound word, is not very surprising. It shows, however, that overlap colexifications can be used to compute the *compoundhood* of concepts, a property, that has only rarely been investigated for a larger number of languages.

Can we "see" the differences with respect to the in-degree and the out-degree of affix colexification networks in the figure (C) above when comparing them with full colexifications (A) and overlap colexifications (B)?

References

Anderson, C., T. Tresoldi, T. C. Chacon, A.-M. Fehn, M. Walworth, R. Forkel, and J.-M. List (2018). "A Cross-Linguistic Database of Phonetic Transcription Systems." *Yearbook of the Poznań Linguistic Meeting* 4.1, 21–53.

Bloomfield, L. (1933 [1973]). *Language*. London: Allen & Unwin.

Cysouw, M. (2010). "Semantic maps as metrics on meaning." *Linguistic Discovery* 8.1, 70–95.

Fox, A. (1995). *Linguistic reconstruction. An introduction to theory and method*. Oxford: Oxford University Press.

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

Gast, V. and M. Koptjevskaja-Tamm (2018). "The areal factor in lexical typology. Some evidence from lexical databases." In: *Aspects of linguistic variation*. Ed. by D. Olmen, T. Mortelmans, and F. Brisard. Berlin and New York: de Gruyter, 43–81.

Georgakopoulos, T. and S. Polis (2018). "The semantic map model: State of the art and future avenues for linguistic research." *Language and Linguistics Compass* 12.2. e12270 LNCO-0727.R1, e12270–n/a.

Girvan, M. and M. E. Newman (2002). "Community structure in social and biological networks." *Proceedings of the National Academy of Sciences of the United States of America* 99.12, 7821–7826.

Hammarström, H., M. Haspelmath, R. Forkel, and S. Bank (2021). *Glottolog. Version 4.4*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: `https://glottolog.org`.

List, J.-M. (2014). *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.

— (2019a). "Open problems in computational diversity linguistics: Conclusion and Outlook." *The Genealogical World of Phylogenetic Networks* 6.12.

— (2019b). "Typology of semantic promiscuity (Open problems in computational diversity linguistics 10)." *The Genealogical World of Phylogenetic Networks* 6.11.

— (2023). *Inference of Partial Colexifications from Multilingual Wordlists*.

List, J.-M. and R. Forkel (2021). *CL Toolkit. A Python Library for the Processing of Cross-Linguistic Data [Software Library, Version 0.1.1]*. Geneva: Zenodo.

List, J.-M., R. Forkel, S. J. Greenhill, C. Rzymski, J. Englisch, and R. D. Gray (2022a). "Lexibank, A public repository of standardized wordlists with computed phonological and lexical features." *Scientific Data* 9.316, 1–31.

List, J.-M., S. J. Greenhill, C. Anderson, T. Mayer, T. Tresoldi, and R. Forkel (2018). "CLICS^2. An improved database of cross-linguistic colexifications assembling lexical data with help of cross-linguistic data formats." *Linguistic Typology* 22.2, 277–306.

List, J.-M., T. Mayer, A. Terhalle, and M. Urban (2014). *CLICS: Database of Cross-Linguistic Colexifications. Version 1.0*. Marburg: Forschungszentrum Deutscher Sprachatlas. URL: `https://lingpy.org/clics/`.

List, J.-M., A. Terhalle, and M. 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*. "IWCS 2013" (Potsdam, 03/19–03/22/2013). Association for Computational Linguistics. Stroudsburg, 347–353.

List, J.-M., A. Tjuka, C. Rzymski, S. J. Greenhill, and R. Forkel (2022b). *CLLD Concepticon [Dataset, Version 3.0.0]*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: `https://concepticon.clld.org/`.

Mayer, T., J.-M. List, A. Terhalle, and M. 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.

Newman, M. E. J. (2004). "Analysis of weighted networks." *Physical Review E* 70.5, 056131.

Paul, H. (1880 [1886]). *Principien der Sprachgeschichte*. 2nd ed. Halle: Max Niemeyer. prinzipiendersp01paulgoog: `ia`.

Rzymski, C. et al. (2020). "The Database of Cross-Linguistic Colexifications, reproducible analysis of cross- linguistic polysemies." *Scientific Data* 7.13, 1–12. URL: `https://clics.clld.org`.

Spearman, C. (1904). "The Proof and Measurement of Association between Two Things." *The American Journal of Psychology* 15.1, 72–101.

Sperber, H. (1923). *Einführung in die Bedeutungslehre*. Bonn and Leipzig: Kurt Schroeder.

Traugott, E. C. (2012). "Pragmatics and language change." In: 249–565.

Wilkins, D. P. (1996). "Natural tendencies of semantic change and the search for cognates." In: *The comparative method reviewed. Regularity and irregularity in language change. The comparative method reviewed. Regularity and irregularity in language change*. Ed. by M. Durie. With an intro. by M. D. Ross and M. Durie. New York: Oxford University Press, 264–304.

Zalizniak, A. A. (2018). "The Catalogue of Semantic Shifts: 20 years later." *Russian Journal of Linguistics* 22.4, 770–787.

Zalizniak, A. A., M. Bulakh, D. Ganenkov, I. Gruntov, T. Maisak, and M. Russo (2012). "The catalogue of semantic shifts as a database for lexical semantic typology." *Linguistics* 50.3, 633–669.