Discursive variation explains colexification: A lexicon-wide case study on the DoReCo corpus

Barend Beekhuizen
University of Toronto; barend.beekhuizen@utoronto.ca

Keywords: lexical semantic typology; colexification; computational methodologies; discursive motivation; corpus-based studies

Crosslinguistic variation in patterns of how words group together (or: colexify, cf. François 2008) meanings that are expressed with distinct words in other languages provides us with insight in the functional pressures that shape the structure of the lexicon. One factor that has been explore is the so-called need probability of a concept, or the likelihood that a speaker of a specific language expresses a concept in language use. Kemp et al. (2018) report that a greater need probability in a language goes hand in hand with a lower rate of colexification, the motivation being that having a greater use for a concept warrants that concept having its own label. In this talk, we build on these inquiries into the discursive motivation of colexification, innovating in three substantial ways. First, while previous work considers individual case studies, we present a lexicon-wide study. Second, our study is corpus based, rather than based on secondary, non-discursive data (such as dictionaries and word list). Finally, we consider aspects of discursive organization beyond usage probability.

In particular, we look at contextual diversity, predicting that the more diverse the corpus contexts are in which a pair of concepts occurs, the less likely that pair is colexified: the lexical item would be 'spread too thin' across the contexts of use of the two concepts. This factor is motivated by the observation that knowing a concept does not entail knowing how to apply it in language use, and as such, crosslinguistic variation in the rules of application are expected (Goodwin 1994, Enfield 2014). We furthermore consider, per doculect, how distinct the linguistic contexts of pairs of concepts are from each other. This informs us, similarly, about the need to keep the concepts apart: the more similar the usage contexts of two concepts are to each other, the more likely it is that the term expressing one concept can express the other without much confusion. We predict that greater separability of the contexts coincides with lower rates of colexification.

We use the DoReCo corpus (Seifart et al. 2022), a typologically diverse sample of spoken language from 51 doculects. We operationalize 'concepts' as English lemmas in the free translation (acknowledging the problems with this approach). For every sufficiently frequent concept, we retrieve likely translation-equivalent tokens using Wälchli's (2014) L-Algorithm. Next, we determine whether a doculect colexifies pairs of concepts, by measuring the similarity of the translation-equivalent word tokens of the concepts. This yields 79 concept pairs that are colexified in at least one doculect. See (1) for examples (with proportions of languages colexifying them):

(1) *wife-woman* (.23)
speak-talk (.21)
hear-listen (.14)
river-water (.13)
stick-tree (.10)
bring-carry (.05)

Next, we inquire if our discursive factors (need probability, contextual diversity, separability) allow us to predict whether a doculect colexifies a particular pair of concepts. By themselves, all factors are found to be predictive. In a multiple logistic regression, however, the effect of need probability is mostly obscured by the other two contextual factors, showing that rather than the mere discursive frequency of a concept, it is the ways in which concepts are deployed in discourse that predicts whether a language colexifies them.

References

Enfield, Nick .J. 2014. *The utility of meaning: What words mean and why*. OUP Oxford.

François, Alexander. 2008. Semantic maps and the typology of colexification. In Martine Vanhove (ed.) *From polysemy to semantic change: Towards a typology of lexical semantic associations*, Amsterdam, Benjamins. pp.163-216.

Goodwin, Charles. 1994. Professional vision. *American Anthropologist*, 96(3), pp.606-633.

Kemp, Charles, Xu, Yang & Regier, Terry. 2018. Semantic typology and efficient communication. *Annual Review of Linguistics*, 4(1), pp.109-128.

Seifart, Frank, Ludger Paschen & Matthew Stave (eds.). 2022. *Language Documentation Reference Corpus* (DoReCo) 1.2. Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). DOI:10.34847/nkl.7cbfq779

Wälchli, Bernhard. 2014. Algorithmic typology and going from known to similar unknown categories within and across languages. In: Szmrecsanyi, Benedikt and Wälchli, Bernhard (eds.) *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech.* De Gruyter, pp. 355-393.