The interaction between cognitive ease and informativeness shapes the lexicons of natural languages

Thomas Brochhagen* **Gemma Boleda*†**

*Universitat Pompeu Fabra
†ICREA
{firstname.surname}@upf.edu

Introduction

Across languages, it is common for words to be associated with multiple meanings. Moreover, certain meanings are expressed by the same form more often than others (Jackson et al., 2019; Xu et al., 2020). For instance, the colexification –i.e., the conventional association of multiple meanings with the same form– of TOE and FINGER is found in at least 135 languages (Rzymski et al., 2020). These languages are spoken throughout the world and span multiple unrelated language families.

Recent research suggests that semantic relatedness increases colexification likelihood (Xu et al., 2020). Semantic memory may favor colexifying meanings that are easy to relate to one another. This, in turn, may aid vocabulary acquisition, lexical retrieval and interpretation. Building on these findings, we investigate the interplay between this and another major force: pressure for the lexicon to be informative, in the sense of supporting accurate information transfer (e.g., Regier et al., 2015). We hypothesize that languages strike a balance between these two forces. In particular, we expect colexification likelihood to increase with semantic relatedness, until a point is reached at which meanings are too related; for these highly related meanings, we expect pressure for informativeness to counteract the increasing trend, because these meanings would not be easy to disambiguate even in context. We find support for this hypothesis in two large scale analyses.[1]

Analysis 1

To study the relationship between semantic relatedness and colexification, we fit three generalized additive logistic models to colexification data spanning over 1200 languages and more than 1400 meanings, totaling 203056 data points. This data comes from CLICS^3 (Rzymski et al., 2020), the largest cross-linguistic database of colexifications available to date. The models characterize how likely a pair of meanings is to colexify in a given language as a function of one of three data-induced estimates of relatedness: distributional similarity, using pre-trained embeddings (Grave et al., 2018); associativity data (De Deyne et al., 2018); and the first principal component of these two measures (PC1). Both distributional and associative information are based on Dutch and English glosses of the meanings found in CLICS^3; that is, Dutch and English words are used as surrogates for meanings to estimate the latter's relatedness. Since language contact and common linguistic ancestry influence colexification (Jackson et al., 2019; Xu et al., 2020), the models are also passed information about how often a pair of meanings colexifies in other languages. This information is weighted by the phylogenetic/geographic distance to the response language. An indicator codifies whether a relatedness estimate stems from Dutch or English data.

Model comparison using approximate leave-one-out cross-validation suggests that PC1 is the best predictor of colexification, with a difference of -715 in expected log pointwise predictive density to the second highest ranked model. Figure 1 shows its estimated marginal effects. These results largely support to our hypothesis: colexification increases with relatedness until meanings are "too related", which makes their colexification decrease. Note, however, that the data are also consistent with a plateau rather than a decrease for highly related meanings (see shaded area in the figure). This is still consistent with the main hypothesis – informativeness counteracting simplicity for highly related meanings–, with a smaller effect of informativeness than we had expected.

[1] The manuscript that this abstract is based on is found at https://psyarxiv.com/efs4p

Proceedings of the Society for Computation in Linguistics (SCiL) 2022, pages 217-219.
Held on-line February 7-9, 2022

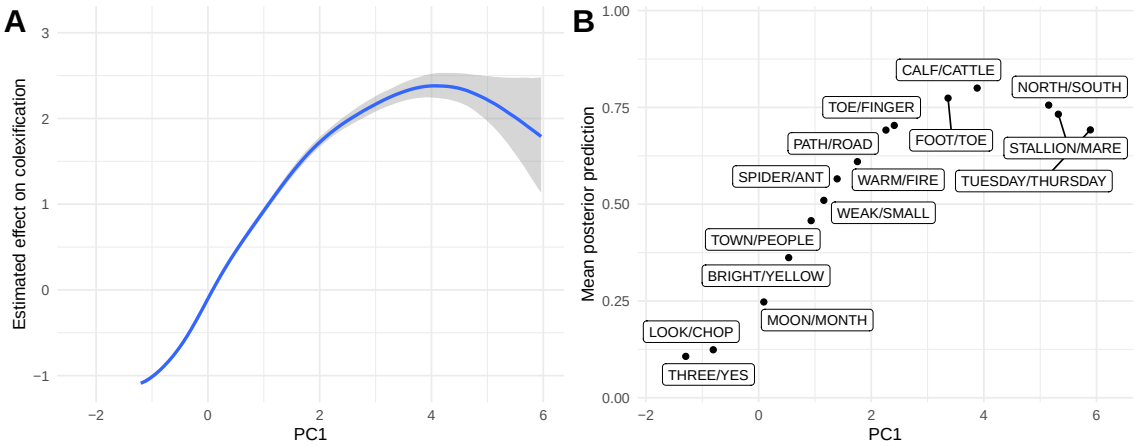

Figure 1: A: Marginal effects of standardized PC1. Shading shows 95% credible intervals. The smooth function $s(\cdot)$ characterizes how PC1's contribution to colexification likelihood changes across values. B: Mean posterior predictions for exemplary meaning pairs across PC1 values.

Analysis 2

Our hypothesis specifically predicts that the decrease in colexification likelihood for highly related meanings is due to their confusability. We next probe confusability more directly, focusing on the kind of relationship meanings stand in.

Pressure for informativeness should make colexifying opposites (e.g., LEFT and RIGHT) less likely than colexifying meanings in other kinds of relationships. Opposite meanings express contrasts, being maximally similar in every respect but one (e.g., Kliegr and Zamazal, 2018). Therefore, losing the distinction they encode can be expected to be particularly harmful in communicative terms. We compare opposites to meaning pairs standing in two semantic relations that do not necessarily lead to high confusability: part-whole (e.g., TOE-FOOT) and subsumption (e.g., CALF-CATTLE).

Colexification rates were estimated from 1416 meanings and 2279 languages from CLICS[3]. Semantic relations are from WordNet (Fellbaum, 2015), using English words as proxies for meanings. Pairs in none of the three relations were classified as 'none/other'. As expected, this group has the lowest mean percentage of colexification (0.06, with a 95% CI of $[0.06, 0.06]$), followed by opposites (1.4 $[1.3, 1.5]$), then by subsumption (3.1 $[3.0, 3.3]$) and part-whole pairs (3.7 $[3.5, 3.8]$). These results suggest, first, that standing in one of the three relations increases the odds for meanings to colexify compared to 'none/other'; and second, that not all relations are equally conducive to colexification, with opposites being less likely to colexify.

We thus again find that relatedness makes colexification more likely, but that the need to distinguish confusable meanings can counteract this trend. Under our interpretation, simplicity makes colexification likelihood for opposites increase, whereas informativeness makes them decrease, resulting in their position in the middle compared to the other relations.

Conclusions

A growing body of research supports the idea that languages are efficient in the sense that they strike a good balance between informativeness and simplicity (e.g., Christiansen and Chater, 2008; Regier et al., 2015). Our large scale analyses suggest such a balance in the lexicon. We find that colexification likelihood increases with semantic relatedness, until an inflection point is reached, after which it decreases or flattens out (Analysis 1). This shift may be a consequence of a need for meanings to be distinguishable in context (Analysis 2).

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 715154). This paper reflects the authors' view only, and the EU is not responsible for any use that may be made of the information it contains.

References

Morten H. Christiansen and Nick Chater. 2008. Language as shaped by the brain. *BBS*, 31(05).

Simon De Deyne, Danielle J. Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2018. The "Small World of Words" English word association norms for over 12,000 cue words. *BRM*, 51(3):987–1006.

Christiane Fellbaum. 2015. *WordNet*. OUP.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proc. LREC*.

Joshua C. Jackson, Joseph Watts, Teague R. Henry, Johann-Mattis List, Robert Forkel, Peter J. Mucha, Simon J. Greenhill, Russell D. Gray, and Kristen A. Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. *Science*.

Tomáš Kliegr and Ondřej Zamazal. 2018. Antonyms are similar: Towards paradigmatic association approach to rating similarity in SimLex-999 and WordSim-353. *DKE*, 115:174–193.

Terry Regier, Charles Kemp, and Paul Kay. 2015. *Word Meanings across Languages Support Efficient Communication*, chapter 11. John Wiley & Sons, Ltd.

Christoph Rzymski, Tiago Tresoldi, Simon J Greenhill, Mei-Shin Wu, et al. 2020. The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Sci. Data*, 7(1):1–12.

Yang Xu, Khang Duong, Barbara C Malt, Serena Jiang, and Mahesh Srinivasan. 2020. Conceptual relations predict colexification across languages. *Cognition*.