When do languages use the same word for different meanings? The Goldilocks principle in colexification

Thomas Brochhagen[1] & Gemma Boleda[2]
[1] Universitat Pompeu Fabra, thomas.brochhagen@upf.edu [2] Universitat Pompeu Fabra / ICREA

Keywords: language universals, colexification, cognitive effort, ambiguity, efficient communication

Colexifications are pervasive in language, and often systematic (François, 2008; Jackson et al., 2019; Xu et al., 2020). For instance, the colexification of toe and finger is found in at least 135 languages (Rzymski et al., 2020). We examine the interplay between two competing pressures that affect how languages are shaped, and that may partially explain cross- linguistic patterns in colexification: the cognitive pressure for simplicity, and the communicative pressure for complexity.[1] Indeed, 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 & Chater, 2008; Regier et al., 2015); we provide evidence for such a balance in cross-linguistic colexification patterns. In particular, we propose a Goldilocks principle according to which meanings colexify when they are related enough to foster cognitive economy, and at the same time not too confusable in actual language use. We provide support for this principle using data about over 2200 languages and 1400 meanings harvested from CLICS^3 (Rzymski et al., 2020), the largest cross-linguistic database of colexifications available to date.

We first fit a generalized additive logistic model to the colexification data. The model characterizes how likely a pair of meanings is to colexify in a given language as a function of a data-induced estimate of relatedness: the first principal component (PC1) of two semantic relatedness measures –distributional similarity, using pre-trained embeddings (Grave et al., 2018), and associativity, using association norms (De Deyne et al., 2018). The models are also passed information about how often a pair of meanings colexifies in other languages, to control for the effect of language contact and common linguistic ancestry (Jackson et al., 2019; Xu et al., 2020). Figure 1A shows the marginal effect of semantic relatedness in the model, and Figure 1B provides illustrative sample predictions. As predicted under the Goldilocks principle, the induced pattern is linear at first and changes for high relatedness values; however, the effect may less strong than we anticipated (see shaded area for the credible interval).

We do a second analysis, not reported in full for space reasons, that specifically targets confusability, since our hypothesis is that the observed decrease in colexification likelihood for highly related meanings is due to their confusability. We find that indeed, among related meanings, more confusable meaning pairs like left-right colexify less across languages than other meaning pairs that we argue are not as confusable in actual language use (e.g. toe-foot).

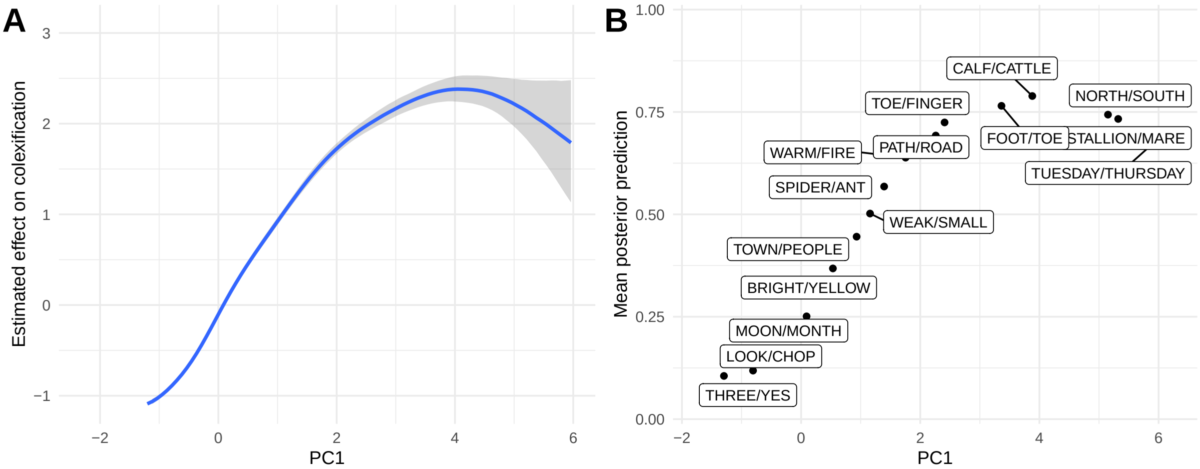

Fig. 1: A: Marginal effect semantic relatedness (PC1, in standardized units). Shading shows 95% credible intervals. A smooth function is inferred from the data and characterizes how the contribution of PC1 to colexification likelihood changes across its values (on the logit scale). B: Example of mean posterior predictions for meaning pairs.

[1] This abstract summarizes the work published in Brochhagen & Boleda (2022).

References

Brochhagen, Thomas & Gemma Boleda. 2022. When do languages use the same word for different meanings? the goldilocks principle in colexification. *Cognition* 226. 105179.

Christiansen, Morten H. & Nick Chater. 2008. Language as shaped by the brain. *BBS* 31(05). doi: 10.1017/s0140525x08004998.

De Deyne, Simon, Danielle J. Navarro, Amy Perfors, Marc Brysbaert & Gert Storms. 2018. The “Small World of Words” English word association norms for over 12,000 cue words. *BRM* 51(3). 987–1006. doi:10.3758/s13428-018-1115-7. `https://doi.org/10.3758%2Fs13428-018-1115-7`.

François, Alexandre. 2008. Semantic maps and the typology of colexification: Intertwining polysemous networks across languages. In *Studies in language companion series*, 163–215. John Benjamins Publishing Company. doi:10.1075/slcs.106.09fra.

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

Jackson, Joshua C., 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* doi:10.1126/science.aaw8160.

Regier, Terry, Charles Kemp & Paul Kay. 2015. *Word meanings across languages support efficient communication* chap. 11, 237–263. John Wiley & Sons, Ltd. doi:https://doi.org/10.1002/9781118346136.ch11. `https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118346136.ch11`.

Rzymski, Christoph, 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.

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