Characterizing cross-linguistic regularities beyond semantic similarity

Thomas Brochhagen
Universitat Pompeu Fabra, thomas.brochhagen@upf.edu

Keywords: crosslinguistic regularities, semantic similarity, lexical meaning

A growing body of research suggests that the way meaning is organized in the lexicon across languages can be partially explained through latent semantic knowledge –or proxies thereof– such as meanings' associativity; visual resemblance; taxonomic closeness; affective, or distributional relationships. This kind of approach has shown promise in characterizing both large scale cross-linguistic regularities, such as colexification patterns and semantic shifts (e.g., Xu et al. 2020; Di Natale et al. 2021; Brochhagen & Boleda 2022; Fugikawa et al. 2023), as well as developmental data from children's early linguistic behavior (e.g., Ferreira Pinto & Xu 2021). The aim of the present contribution is, on the one hand, to synthesize these findings, with attention to methodological advances made, as well as shortcomings to improve on. On the other hand, I will present preliminary findings concerning regularities that, due to the aforementioned shortcomings, have fallen outside of the purview of past studies.

Broadly speaking, the aforementioned studies all form part of a recent surge in lexical semantic research that is data-driven, cross-linguistic, and domain-general (i.e., not constrained to, e.g., color or number systems). This surge has been enabled by the creation and consolidation of large-scale databases and resources, such as CLICS^3 (Rzymski et al., 2020) or the Small World of Words project (De Deyne et al., 2018). Methodologically, the main approach taken has been to try to explain cross-linguistic variation through one; multiple; or a blend of proxies of semantic similarity such as the ones mentioned above (e.g., with associativity and taxonomic closeness as predictors in a multi-level regression model). These –quite successful– characterizations of cross-linguistic lexical organization have thereby provided further evidence to the idea that "semantic relatedness", operationalized in different ways, plays a central role in the way meaning is stored and deployed across languages. The main shortcoming of such approaches, by definition, lies in not being able to account for lexical regularities explained through factors that are not well accounted for by (language-specific) psychometrics. For instance, independent research suggests that the environment can come to shape language (e.g., Dediu et al. 2017; Josserand et al. 2021) but this idea has yet to be put to the test at a large cross-linguistic and domain-general scale in lexical semantics. Similarly, multi-lingual resources or pragmatic considerations have played little to no role so far. I here provide first results and discussion on methods to do this.

References

Brochhagen, Thomas & Gemma Boleda. 2022. When do languages use the same word for different meanings? The Goldilocks principle in colexification. *Cognition* doi:10.1016/j.cognition.2022.105179.

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.

Dediu, Dan, Rick Janssen & Scott R Moisik. 2017. Language is not isolated from its wider environment: vocal tract influences on the evolution of speech and language. *Language & Communication* 54. 9–20.

Di Natale, Anna, Max Pellert & David Garcia. 2021. Colexification networks encode affective meaning. *Affective Science* 2(2). 99–111.

Ferreira Pinto, Renato & Yang Xu. 2021. A computational theory of child overextension. *Cognition* 206. doi:10.1016/j.cognition.2020.104472.

Fugikawa, Olivia, Oliver Hayman, Raymond Liu, Lei Yu, Thomas Brochhagen & Yang Xu. 2023. A computational analysis of regularity in semantic change. *Frontiers in Communication* .

Josserand, Mathilde, Emma Meeussen, Asifa Majid & Dan Dediu. 2021. Environment and culture shape both the colour lexicon and the genetics of colour perception. *Scientific Reports* 11(1). 19095.

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* .