HOW SENSES ARE GROUPED WITHIN WORD FORMS OVER TIME: AN ANALYSIS OF DIACHRONIC COLEXIFICATION PATTERNS

Lucía Pitarch*[1] and Thomas Brochhagen[2]

*lpitarch@unizar.es
[1]Universidad de Zaragoza
[2]Universitat Pompeu Fabra

Languages commonly express multiple meanings with single forms (Piantadosi, Tily, & Gibson, 2012). However, while ambiguity is pervasive, the reasons why some meanings are more likely to be conflated than others –i.e., why they *colexify* more often (François, 2008)– are still under investigation (Xu, Duong, Malt, Jiang, & Srinivasan, 2020; Brochhagen & Boleda, 2022).

Originally the term *colexification* referred to several senses sharing the same word form (e.g. the senses of *right* as 'direction' and 'being correct' colexify). Later studies broadened this term to apply to concepts instead of senses (Xu et al., 2020, e.g.), relying on CLICS3 (Rzymski et al., 2020), the largest database on cross-linguistic colexifications available to date. CLICS3 provides English glosses for meanings lexicalized in source languages (e.g., *sentir* in Catalan expresses both FEEL and LISTEN). When different glosses (e.g., LISTEN and FEEL) share the same word form in the original language, as *sentir* in Catalan, those concepts colexify. This approach enabled many quantitative studies by being more easily and automatically computable. However, relying on English as meta-language leads to a decrease in the granularity of the original term (François, 2008, cf.). For the present study, we linked the data from CLICS3 for romance languages to BabelNet (Navigli & Ponzetto, 2012), one of the largest multilingual sources for structured semantic data. In BabelNet words are split into senses which have a universal identifier shared by all languages. We searched pairs that colexified in CLICS3 in a certain language in BabelNet, and retrieved the information related to the BabelNet senses shared by both colexifying wordforms in CLICS3. By doing so we describe colexifications not only at the concept level, nor through English as a mediator, but rather at the sense level with universal sense abstractions. Beyond enriching the available data, this study offers a novel, quantitative and diachronic, perspective on colexification. It asks what form- and concept-based information is predictive of the maintenance or loss of colexifications over time. For instance, while the concepts FEEL and LISTEN did not colexify in Latin (with the different word forms *sentire* and *audire* respectively), in Catalan they do, and the same

word form *sentir* is used to refer to both concepts. Yet, in other languages such as Spanish, Latin's semantic organization is maintained.

We focus on the following intralinguistic features: word length, form confusability,[1] part of speech, number of lexifications per sense (in how many other words of a language it appears) and semantic relations (e.g., part/whole and subsumption). The data was processed by grouping senses belonging to the same word forms into pairs, and then fit using logistic regression models with the above features as predictors and, as response, whether a given colexification from Latin was maintained diachronically in at least one of its daughters. That is, whether sense pairs expressed in Latin with the same word are also still colexified in other romance languages. We fit two different models. In the smallest, more imbalanced, model, encompassing all 782 colexifications annotated with semantic relations, only semantic relation information and form confusability were estimated to impact prediction. In fact, the information that two senses are semantically related (no matter the type of relation) was so informative that all colexifications were predicted as maintained, incurring a 10% error rate. The larger model covers all 406777 pairs, but does not include relational information. The most informative feature for diachronic colexification maintenance in this model was form confusability, followed by the number of lexifications of the synset. The rarer a synset's lexification is, the more colexifications it appears in will be maintained. In both models the more different a word is, on average, from other words in its lexicon –i.e., the less mistakable it is– the more the colexifications it hosts are maintained. Both part of speech and word length were the least informative predictors for diachronic colexification maintanance. Word length's lesser impact compared to other features may be due to its overlap with form confusabillity. As for part of speech, Latin's structure may play a role, as its syntactic functions are mainly expressed by case rather than by functional words (e.g., prepositions). Further study of other language families with different peculiarities should shed further light on these matters. The accuracy obtained with both cross-validated models lies between 80% and 90%.

The enrichment of BabelNet data is ongoing (e.g., (Declerck & Bajčetić, 2021) for phonology or (Nayak, Majumder, Goyal, & Poria, 2021) for relation extraction). In the future we aim to improve our diachronic research with this new information to gain a deeper insight on linguistic change.

Acknowledgements

This research has been partially supported by DGA/FEDER.

[1] Mean Levenshtein distance between a form and the lexicon it draws from. The latter was estimated by scraping unique words from 1250 random wikipedia pages for each language. Initially we focus on orthography but we are currently working on enriching the data with phonological resources.

References

Brochhagen, T., & Boleda, G. (2022). *When do languages use the same word for different meanings? The Goldilocks principle in colexification.* Cognition.

Declerck, T., & Bajčetić, L. (2021). Towards the addition of pronunciation information to lexical semantic resources. In *Proceedings of the 11th global wordnet conference* (pp. 284–291). University of South Africa (UNISA): Global Wordnet Association.

François, A. (2008). Semantic maps and the typology of colexification: Intertwining polysemous networks across languages. In *Studies in language companion series* (pp. 163–215). John Benjamins Publishing Company.

Navigli, & Ponzetto. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, *193*, 217–250.

Nayak, T., Majumder, N., Goyal, P., & Poria, S. (2021). Deep neural approaches to relation triplets extraction: A comprehensive survey. *Cognitive Computation*, *13*(5), 1215–1232.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291.

Rzymski, C., Tresoldi, T., Greenhill, S. J., Wu, M.-S., Schweikhard, N. E., Koptjevskaja-Tamm, M., Gast, V., Bodt, T. A., Hantgan, A., Kaiping, G. A., et al.. (2020). The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data*, *7*(1), 1–12.

Xu, Y., Duong, K., Malt, B. C., Jiang, S., & Srinivasan, M. (2020). Conceptual relations predict colexification across languages. *Cognition*, *201*.