Letting meaning surface: a corpus-based study of Estonian perception verbs

Mariann Proos
University of Tartu
mariann.proos@ut.ee

Keywords: corpus linguistics, cognitive semantics, polysemy, multiple correspondence analysis, Estonian

Corpus-based methods of language research enable us to make use of large sets of language data we have at our command. However, studying meaning from a corpus can be an elusive task. One method that has been successfully used for both polysemy as well as synonymy research is behavioural profile analysis (e.g. Proos, 2019, Divjak & Fieller, 2014; Divjak & Gries, 2006). This method is based on the idea that meaning can be derived from and accessed through information about co-occurrences with other language items. In a corpus sample, each sentence is annotated for a number of both morpho-syntactic as well as semantic information. For polysemy, this annotation also includes, for every sentence, designating a specific meaning to the polysemous item that is under analysis. However, annotating meaning is questionable for quite a few reasons. As the sense of the language element is not something visible, it is left to the researcher to "assign" senses to all of the tokens. This can only be done with taking into account the context of the specific token. Thus, this process can become circular – the dependent variable is influenced by the independent variables from the start, creating a bias. Moreover, in this way, senses are treated as discrete language units, but following the theory of polysemy in cognitive and usage-based semantics, this is not the case (Geeraerts, 2010).

To escape this fallacy, Glynn (2016) has proposed to study meaning bottom-up: letting the meanings surface from the data. One way to do this is using multiple correspondence analysis (MCA). Glynn (2016) has demonstrated an analysis of *to annoy* with MCA as a proof-of-concept. In this paper, the method is extended to 14 Estonian perception verbs on large-scale (14000 sentences) bottom-up meaning study. The aim is twofold – showcasing multiple correspondence analysis as a valid method for polysemy research, and making conclusions about the polysemy of Estonian perception verbs. Preliminary results with the perception verb *nägema* 'to see' show that multiple correspondence analysis allows for a meaningful representation of the polysemy. Using MCA, it was possible to show a potential structure of polysemy of *nägema* 'to see', as well as determine which factors/variables account for the meaning variation. Quantitative meaning research is an important direction in current linguistics. The kind of bottom-up research fleshed out above has not yet reached the same popularity as behavioural profile analysis. Although strides are being made towards using newer methods, like machine learning technologies (e.g. Beekhuizen et al., 2018) for studying meaning, this line of analysis, while promising, is still in its infancy (Desagulier, 2018). Bottom-up corpus research is a source of valuable information about the possibilities of quantitative semantics, but at the moment there is a lack of in-depth research. This paper starts to fill that gap by offering results from a large-scale meaning study of a specific selection of verbs – the main perception verbs of Estonian.

References

Beekhuizen, B., Milić, S., Armstrong, B. C., & Stevenson, S. (2018). What Company Do Semantically Ambiguous Words Keep? Insights from Distributional Word Vectors. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Desagulier, G. (2018). Can word vectors help corpus linguists? Retrieved from https://halshs.archives-ouvertes.fr/halshs-01657591v2

Divjak, D., & Fieller, N. (2014). Cluster analysis. Finding structure in linguistic data. In D. Glynn & J. A. Robinson (Eds.), *Corpus Methods for Semantics. Quantitative studies in polysemy and synonymy* (pp. 405–441). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Divjak, D., & Gries, S. T. (2006). Ways of trying in Russian: clustering behavioural profiles. *Corpus Linguistics and Linguistic Theory*, *2–1*, 23–60.

Geeraerts, D. (2010). *Theories of Lexical Semantics*. Oxford: Oxford University Press.

Glynn, D. (2016). Quantifying polysemy: Corpus methodology for prototype theory. *Folia Linguistica*, *50*(2), 413–447.

Proos, M. (2019). Polysemy of the Estonian perception verb "nägema" 'to see.' In L. J. Speed, C. O'Meara, L. San Roque, & A. Majid (Eds.), *Perception Metaphors* (pp. 231–252). John Benjamins Publishing Company.