Continuous Vector Space Models for Variation and Change in Sparse, Richly Annotated Indo-European Argument Structure Data

Jóhanna Barðdal, Gard B. Jenset, Laura Bruno, Esther Le Mair,
Peter Alexander Kerkhof, Svetlana Kleyner, Leonid Kulikov & Roland Pooth

Quantitative studies of variation and change for historical languages are often hampered by sparsity of attested data but with rich annotation drawing on long traditions of linguistic and philological scholarship (Jenset & McGillivray 2017). Conversely, in natural language processing (NLP) for modern languages, data are plentiful but annotated data are scarce, prompting the use of neural network models that can accurately infer linguistic properties based on distributional information from very large un-annotated corpora (Mikolov et al. 2013a, Mikolov et al. 2013b). These techniques are relevant to historical linguistics because of their ability to handle sparse data and to model highly complex relations. Distributional approaches to variation and change in historical data (Barðdal et al. 2012, Jenset 2013) have previously relied on vector space representations that capture broad patterns but may struggle with highly complex distributional relations with sparse data.

We aim to apply state of the art NLP models to combinations of historical data and database annotations. The data stem from the NonCanCase Database, compiled within the ERC-funded EVALISA project, carried out at Ghent University. The database consists of lexical entries of verbs and compositional predicates that select oblique subjects, mostly accusative and dative, across all the eleven earliest branches of Indo-European, approximately 4.000 types in total. There are great differences in meaning found across these types, including senses expressing the expected experience, perception, cognition, and bodily states, but also involving modality, evidentiality and possession, ranging to different types of happenstance events, expressing success and failure, gain, innate properties, as well as all kinds of hindrances. By combining sparse historical attestations enriched with expert linguistic judgments, and state of the art NLP capable of representing highly complex distributional relations, we aim to demonstrate how studies of variation and change in historical linguistics can benefit from NLP.

References

Barðdal, J., Smitherman, T., Bjarnadóttir, V., Danesi, S., Jenset, G. B., & McGillivray, B. (2012). Reconstructing Constructional Semantics: The Dative Subject Construction in Old Norse-Icelandic, Latin, Ancient Greek, Old Russian and Old Lithuanian. Studies in Language 36(3): 511–547.

Jenset, G. B. (2013). Mapping Meaning with Distributional Methods: A Diachronic Corpus-Based Study of Existential *There*. *Journal of Historical Linguistics* 3(2): 272–306.

Jenset, G. B., & McGillivray, B. (2017). Quantitative Historical Linguistics: A Corpus Framework. Oxford: Oxford University Press.

Mikolov, T., Yih, W. T., & Zweig, G. (2013a). Linguistic Regularities in Continuous Space Word Representations. In hlt-Naacl (Vol. 13, pp. 746-751).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013b). Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781.