Finding Case Constructions: Topological Data Analysis of Very Large Corpora

Steven J. Clancy, Sara Kališnik Verovšek, Quang Nhat Le, Joseph Borkowski, Nicholas Tomlin
Harvard University, Brown University, Brown University, Harvard University, Brown University
sclancy@fas.harvard.edu, sara_kalisnik_verovsek@brown.edu, quang_nhat_le@brown.edu,
borkowski@fas.harvard.edu, nicholas_tomlin@brown.edu

Keywords: constructions, Slavic languages, quantitative methods, corpus linguistics, case systems, deep learning

The largely atheoretical approach developed within natural language processing (NLP) in the past few decades would seem to bear out Goldberg's (2006) claim that linguistic structure is "constructions all the way down". These approaches have allowed the empirical data in increasingly large corpora to speak for itself, resulting in highly successful unsupervised learning of orthographic systems, phonemes, morphemes, and words from raw language corpora. If construction grammar frameworks are correct, then understanding even complex syntactic and semantic features, such as the case system of a language, should be a matter of building on constructions identified at other levels of language. Earlier computational approaches to syntax and semantics defaulted to a formalist framework for lack of a better model, a decision that may be responsible for subsequent delays in progress, but more recent research in NLP has succeeded exclusively by looking at words and their neighbors within very large unlabeled corpora. Recent work using word2vec (Mikolov et al 2013) and GloVe (Pennington et al 2014) have yielded impressive results through the representation of words as relatively low-dimensional vectors that successfully capture a sense of semantic similarity as well as specific relationships such as male-female counterpart nouns, singular-plural, positive-comparative-superlative degrees, and word paradigms. Our study leverages a variety of tools to crack open the meaning in the Slavic languages, where a high level of morphological complexity complicates the work of previous NLP studies that have largely been conducted on English.

In the realm of data visualization, early semantic maps such as those for indefinite constructions (Haspelmath 1997) provided for contiguous structured semantic relationships without definite structure or measurable semantic distance. The use of multidimensional scaling in Croft and Poole (2008) replicated Haspelmath's semantic map and provided measurable semantic distance with more definite structure and introduced a means of addressing large-scale problems that would have been humanly intractable using qualitative methods. However, these techniques have often yielded largely unreadable two-dimensional graphs that have over-simplified the high-dimensional data. The present study aims to address these problems and further advance the field in extracting semantic and syntactic meaning from a very large corpus of Russian. We combine the use of Goldsmith's *Linguistica* (2001) and other methods to break the corpus into stems and inflectional morphology. We then vectorize the words in the corpus for topological data analysis (Carlsson 2009) using a tool called *Mapper*, which uses a collection of vectors X and a reference function f: $X \rightarrow R$ (the set of real numbers) to produce a graph. The ability to tailor the map f to the specific data at hand allows *Mapper* to be used as a method for exploratory data analysis. *Mapper* has already had marked success in identifying novel subtypes of breast cancer (Nicolau et al 2011) and in defining the characteristics of NBA basketball players via performance statistics (Lum et al 2013). This study marks the first time *Mapper* has been used to study linguistic data.

References

Carlsson, Gunnar. 2009. Topology and data. Bulletin of the American Mathematical Society, Vol. 46, No. 2, pp. 255-308.

Croft, W., and K.T. Poole. 2008. Inferring universals from grammatical variation: multidimensional scaling for typological analysis. Theoretical Linguistics 34.1-37.

Goldberg, Adele. 2006. Constructions at Work: The Nature of Generalization in Language. Oxford University Press, Oxford, UK.

Goldsmith, John. 2001. Unsupervised learning of the morphology of a natural language. Computational Linguistics 27:153-198. [The Linguistica Project — http://linguistica.uchicago.edu]

Haspelmath, Martin. 1997. Indefinite pronouns. Oxford: Oxford University Press.

Lum, P.Y., G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson & G. Carlsson. 2013. Extracting insights from the shape of complex data using topology. Scientific Reports 3, Article number: 1236.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.

Nicolau, Monica, Arnold J. Levine, and Gunnar Carlsson. 2011. Topology Based Data Analysis Identifies a Subgroup of Breast Cancers with a Unique Mutational Profile and Excellent Survival. PNAS 108.17: 7265–7270.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. <http://nlp.stanford.edu/projects/glove/>