Selection above the baseline:

Quantifying the advection effect in four domains of cumulative culture

ANDRES KARJUS

andres.karjus@hotmail.com

University of Edinburgh, University of Tartu

A commonly used proxy to the selective fitness of elements over time, in language as well as culture, is their frequency, estimated from what are assumed to be representative datasets (such as large diachronic corpora in the case of language). However, naive frequencies may be misleading. In language, elements (e.g., words) may well rise or fall due to their relevant topics being discussed more (or less) than before; in culture, elements may appear or disappear due to larger scale changes in the system they find application in. Conversely, if an element becomes more frequent than would be predicted based on the increase of its related elements (e.g., its topic or genre in the case of language), then that would constitute reasonable grounds to posit selection and identify the need to look for selective biases and causes that may be driving this change.

This notion is straightforwardly quantified in the cultural-topical advection model (Karjus et al. 2018), originally proposed as a control mechanism for topical fluctuations in diachronic language corpora. The term 'advection' here refers to transport by bulk motion - particles or elements being carried along by the flow of other particles.

We apply the advection model to three datasets chronicling changes in three different domains of culture: cuisine, cinema, and board games, as well as the original domain of language (using the following datasets, respectively: Feeding America, IMDb, BoardGameGeek, and the Corpus of Historical American English). We find that the advection effect accounts for 30-80% of variation in frequency changes over time, depending on the domain and the time periods under observation. We further demonstrate how departures from the prediction (regression residuals) can be used to highlight cases of selection that occur above the cultural-topical baseline.

References

BoardGameGeek. Dataset available via https://www.boardgamegeek.com/xmlapi

Davies, Mark (2010). The Corpus of Historical American English (COHA): 400 million words, 1810-2009. Available online at http://corpus.byu.edu/coha/

Feeding America: The Historic American Cookbook Dataset. East Lansing: Michigan State University Libraries Special Collections. https://www.lib.msu.edu/feedingamericadata/

IMDb datasets. http://www.imdb.com/interfaces/

Karjus, Andres, Richard A. Blythe, Simon Kirby, and Kenny Smith (2018). "Topical advection as a baseline model for corpus-based lexical dynamics". In: Proceedings of the Society for Computation in Linguistics, 1, pp. 186–188.