IDENTIFYING LINGUISTIC SELECTION AND INNOVATION WHILE CONTROLLING FOR CULTURAL DRIFT

Andres Karjus*[1], Richard A. Blythe[1,2], Simon Kirby[1], and Kenny Smith[1]

*Corresponding Author: akarjus@exseed.ed.ac.uk
[1]Centre for Language Evolution, University of Edinburgh, UK
[2]School of Physics and Astronomy, University of Edinburgh, UK

1. Introduction

Human languages evolve on a cultural timescale: new linguistic variants are continually being innovated, some variants are selected for, and some are selected against. These processes of selection constitute the engine of language change and, on longer timescales, language evolution. Previous research on lexical competition and selection has generally been based on either limited subsets of natural data (Verkerk, Calude, & Pagel, 2014; Ahern, Newberry, Clark, & Plotkin, 2016; Calude, Miller, & Pagel, 2017) or simulations (Reali & Griffiths, 2010; Blythe & Croft, 2012; Stadler, Blythe, Smith, & Kirby, 2016).

Identifying genuine instances of innovation and selection from natural language corpora on a broad scale is challenging. Language does not exist in a vacuum: discourse topics tend to reflect contemporary social, cultural and political issues. The relative frequency of a word at a given time period is therefore not necessarily an objective measure of its selective fitness at the time, but potentially the result of its related topic(s) being currently more discussed or reported on (cf. also Chelsey & Baayen, 2010; Lijffijt, Sily, & Nevalainen, 2012; Szmrecsanyi, 2016). Here we present the *topical-cultural advection model*, a method which allows us to measure and control for the rise and fall of discourse topics on the frequencies of individual elements (e.g., words).

2. The topical-cultural advection model

The term *advection* is borrowed from physics, denoting the transport of substance (particles) by bulk motion (of fluids), i.e., being carried along by something else (in the linguistic case, words are carried along by topics). We quantify topical fluctuations by measuring change in words associated with the target word by co-occurrence. The advection value of a word in a time period corresponds to the weighted mean of the log frequency changes of its top associated words (its "topic"), with the association scores as weights (building on Santus, Chersoni,

DOI:10.12775/3991-1.046

Lenci, Huang, & Blache, 2016; Hamilton, Leskovec, & Jurafsky, 2016). A positive advection value for a given target word therefore indicates its topic is increasing in popularity, and vice versa.

3. Results

We tested the descriptive power of the model in two corpora. In the Corpus of Historical American English (COHA), spanning 20 decades, the (linear) effect of topic changes describes up to 31% of the variance in noun frequency changes.

We also created an artificial corpus in order to test the ability of the topical advection model to identify and correct for false positives when searching for cases of selection. To do this, we synthesized a 26-period corpus, based on two genres, 'spoken' and 'academic', from the Corpus of Contemporary American English: we sampled from the two genres, incrementally increasing one and decreasing the other, simulating a change from academic to spoken style and content. Academic topics are presumably different from spoken topics, but both genres use contemporary English; as such, there should be relatively little evidence of selection in this artificial corpus after controlling for topics. We used a test akin to the Fitness Increment Test (Feder, Kryazhimskiy, & Plotkin, 2014) to identify words undergoing apparent change of frequency, or selection, in this corpus, running the test once without and once with the control for topical advection. In the naive model which did not control for topical fluctuations, 46% of the 5762 persistent nouns in the corpus were selected for (or against), a high rate of false positives. Among the 2624 nouns undergoing putative selection, adjustment for topical fluctuations caused 32% of them to lose the trend, while causing 2% of the nouns to gain a trend (new false positives). In other words, the adjustment using the advection model was capable of eliminating about a third of the false positive cases of selection.

Finally, we tested the model's ability to predict innovation in language. It has been suggested that communicative need (in part) drives vocabulary size (Regier, Carstensen, & Kemp, 2016; Gibson et al., 2017); if so, an increasingly popular topic (i.e. exhibiting positive advection) might attract new words, providing the detailed vocabulary required. We identified 133 successful new nouns entering COHA at the latter half of the 20th century, and found that the advection values of the topics of 55% of the new words were significantly higher at the time they were introduced compared to the previous 10 decades; 38% were around the mean, and 7% below the mean of the advection values of the previous decades. This suggests that the majority of these new words were indeed introduced to talk about increasingly popular topics.

4. Conclusions

We propose the topical-cultural advection model as a reasonable baseline in modelling language change and evolution in general, as a method for removing the topical component in the changing frequencies of linguistic elements in order to better assess their selective fitness, as well as a baseline for considering the fitness of topics in terms of their conductivity to innovations.

References

Ahern, C. A., Newberry, M. G., Clark, R., & Plotkin, J. B. (2016). Evolutionary forces in language change. *arXiv preprint arXiv:1608.00938*.

Blythe, R. A., & Croft, W. (2012). S-curves and the mechanisms of propagation in language change. *Language*, *88*(2), 269–304.

Calude, A. S., Miller, S., & Pagel, M. (2017). Modelling loanword success a sociolinguistic quantitative study of Mori loanwords in New Zealand English. *Corpus Linguistics and Linguistic Theory*, *0*(0).

Chelsey, P., & Baayen, H. R. (2010). Predicting new words from newer words: Lexical borrowings in French. *Linguistics*, *48*(6), 1343–1374.

Feder, A. F., Kryazhimskiy, S., & Plotkin, J. B. (2014). Identifying signatures of selection in genetic time series. *Genetics*, *196*(2), 509–522.

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., Gibson, M., Piantadosi, S. T., & Conway, B. R. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, *2016*, 2116–2121.

Lijffijt, J., Sily, T., & Nevalainen, T. (2012). CEECing the baseline: Lexical stability and significant change in a historical corpus. In *Studies in Variation, Contacts and Change in English* (Vol. 10). Research Unit for Variation, Contacts and Change in English (VARIENG).

Reali, F., & Griffiths, T. L. (2010). Words as alleles: connecting language evolution with Bayesian learners to models of genetic drift. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1680), 429–436.

Regier, T., Carstensen, A., & Kemp, C. (2016). Languages Support Efficient Communication about the Environment: Words for Snow Revisited. *PLOS ONE*, *11*(4), 1–17.

Santus, E., Chersoni, E., Lenci, A., Huang, C.-R., & Blache, P. (2016). Testing APSyn against Vector Cosine on Similarity Estimation. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation, PACLIC 30, Seoul, Korea, October 28 - October 30, 2016.*

Stadler, K., Blythe, R. A., Smith, K., & Kirby, S. (2016). Momentum in Language Change: A Model of Self-Actuating S-shaped Curves. *Language Dynamics and Change*, *6*(2), 171–198.

Szmrecsanyi, B. (2016). About text frequencies in historical linguistics: disentangling environmental and grammatical change. *Corpus Linguistics and Linguistic Theory*, *12*(1), 153–171.

Verkerk, A., Calude, A. S., & Pagel, M. (2014). Selection in the lexicon. In *Evolution of Language: Proceedings of the 10th International Conference (EVOLANG10)* (pp. 547–548). World Scientific.