Topical advection as a baseline model for corpus-based lexical dynamics

Andres Karjus  
*Centre for Language Evolution, School of Philosophy, Psychology and Language Sciences, University of Edinburgh,*  
a.karjus@sms.ed.ac.uk

Richard A. Blythe  
*Centre for Language Evolution, School of Philosophy, Psychology and Language Sciences, University of Edinburgh; School of Physics and Astronomy, University of Edinburgh,*  
r.a.blythe@ed.ac.uk

Simon Kirby  
*Centre for Language Evolution, School of Philosophy, Psychology and Language Sciences, University of Edinburgh,*  
simon.kirby@ed.ac.uk

Kenny Smith  
*Centre for Language Evolution, School of Philosophy, Psychology and Language Sciences, University of Edinburgh,*  
kenny.smith@ed.ac.uk

Follow this and additional works at: [https://scholarworks.umass.edu/scil](https://scholarworks.umass.edu/scil)  
Part of the Computational Linguistics Commons, and the Other Linguistics Commons

**Recommended Citation**  
DOI: [https://doi.org/10.7275/R5RR1WFX](https://doi.org/10.7275/R5RR1WFX)  
Available at: [https://scholarworks.umass.edu/scil/vol1/iss1/22](https://scholarworks.umass.edu/scil/vol1/iss1/22)
1 Introduction

An important question in the field of corpus-based evolutionary language dynamics research is concerned with distinguishing selection-driven linguistic change from neutral evolution, and from changes stemming from language-external factors (cultural drift). A commonly used proxy for the popularity or selective fitness of an element is its corpus frequency. However, a number of recent works have pointed out that raw frequencies can often be misleading. We propose a model for controlling for drift in contextual topics in corpora - the topical-cultural advection model - and demonstrate that this simple measure is capable of accounting for a considerable amount of variability in word frequency changes in a corpus spanning two centuries of language use.

2 Background and motivation

There have been various proposals to carry over the selection and neutral drift paradigm from evolutionary biology, (where drift stands for differential replication without selection, cf. Croft (2000)), and apply similar tests to language data (Reali and Griffiths, 2010; Blythe, 2012; Ahern et al., 2016; Sindi and Dale, 2016). While previous research has been mostly concerned with distinguishing selection from drift in terms of frequencies, ours is a model for controlling for topical drift (somewhat similarly to Hamilton et al. (2016b), who contrast cultural and linguistic change). Clearly no linguistic element exists in isolation, without context. In order to objectively model the success or decline of an element, its context (or topic) should be taken into account. The potential effect of cultural processes and hot media topics on language usage patterns, as attested in corpora, have been often noted in recent studies. However, the way such phenomena are viewed varies: while ‘culturomics’ and related approaches treat word frequency changes as a way to study historical real-world changes (Michel et al., 2011), both on their own and as effects on language dynamics (Bochkarev et al., 2014; Petersen et al., 2012), a number of linguists have voiced concerns about relying on frequencies for linguistic inference without controlling for corpus composition in terms of register, genre and topic (Chelsey and Baayen, 2010; Lijffijt et al., 2012; Hinrichs et al., 2015; Szmrecsanyi, 2016; Calude et al., 2017).

3 The cultural-topical advection model

The cultural-topical advection model formalizes the following intuition: if a topic becomes more prevalent, then the words describing it, relating to it and possibly giving rise to it, should become more frequent as well, and vice versa with decline (with a clearer effect on topic-specific words). The term advection is borrowed from physics, denoting transport of particles by bulk motion or flow.

We define the ‘topic’ of a word as the set of words that are most strongly associated with the target word in a given period. This is inspired by the recent proposal of the APSym distributional semantics similarity metric, which is based on the intersection of the most strongly associated (mutual information weighted) co-occurring context words (Santus et al., 2016). The advection value of a word in a given period \( w_t \) is defined as the weighted mean of the (smoothed) log changes in frequencies \( \log(w_{freq_i} + 1) - \log(w_{freq_{i-1}} + 1) \) of the ordered set of associated words \( N_i \), weighted by their association score (i.e.,

\[
\text{wMean}(\{\log \text{Change}(N_i) \mid i = 1, \ldots m\}, W_{1:m}),
\]
where set of weights \( W \) corresponds to the PPMI association scores of the words in the set \( N \) and \( m \) is the number of context words to use.

We also implemented the advection measure using Latent Dirichlet Allocation (Blei et al., 2003), a more traditional topic model. The results in terms of the descriptive power of the model were rather similar. In an LDA-driven advection model, each topic is assigned a frequency change value, based on the (weighted) frequency changes of the words in the topics; the topical advection value of a target word is the (topic-word association weighted) mean of the change values of its topics. In contrast with LDA, our PPMI-weighted top-relevant-context-words based model requires almost no optimization of parameters (only choosing the \( m \)), is considerably simpler (and thus faster), and the results are easily traceable and interpretable, as each “topic” of a target is just a short list of top context words.

4 Results

We used the Corpus of Historical American English (COHA) (Davies, 2010) in order to get a sense of how well the topical-cultural advection model performs. Since cultural effects are likely the most pronounced on nouns (cf. also Hamilton et al. (2016a)), we only model the advection of common nouns; we use only content words from the co-occurrence vectors (of window size 5; \( m = 75 \)), and set a (rather conservative) threshold of 100 occurrences per period for words to be included in the advection model, to maintain reliable semantics. We also experimented with adding “smoothing” to the input data to the topic models, in the sense of concatenating text from a target period and its preceding period, in order to better capture diminishing topics and words.

To test the descriptive power of the advection model, we correlate the log frequency changes of nouns to their respective advection (topic log change) values. For the first test, we include data points on frequency change across 19 decades (1820-2000) of all nouns that occur above the threshold at least once; frequency change data points from 19 decades (more data points in the smoothed versions: concatenated data results in more words being above the minimal threshold). Bottom half, separated by double line, marked with ‘: models using the persistent subset.

<table>
<thead>
<tr>
<th></th>
<th>no smooth</th>
<th>smooth</th>
</tr>
</thead>
<tbody>
<tr>
<td>n unique words</td>
<td>7539</td>
<td>10076</td>
</tr>
<tr>
<td>n data points</td>
<td>75494</td>
<td>107096</td>
</tr>
<tr>
<td>PPMI vectors</td>
<td>0.2</td>
<td>0.31</td>
</tr>
<tr>
<td>LDA</td>
<td>0.17</td>
<td>0.25</td>
</tr>
<tr>
<td>n unique words</td>
<td>2004</td>
<td>2004</td>
</tr>
<tr>
<td>n data points</td>
<td>38076</td>
<td>38076</td>
</tr>
<tr>
<td>PPMI vectors′</td>
<td>0.26</td>
<td>0.38</td>
</tr>
<tr>
<td>LDA′</td>
<td>0.17</td>
<td>0.34</td>
</tr>
</tbody>
</table>

Table 1: The \( R^2 \) values of the two methods with and without smoothing. Top: models using all words that occur above the threshold at least once; frequency change data points from 19 decades (more data points in the smoothed versions: concatenated data results in more words being above the minimal threshold). Bottom half, separated by double line, marked with ‘: models using the persistent subset.

5 Conclusions

We conclude that advection can be considered a reasonably strong baseline for describing changes in word frequency. Obviously, the implementation is open to improvements and experimentation with the parameters. It would be fairly straightforward to use this approach as time series decomposition, by subtracting the advection value from the frequency change value, and reforming the frequency time series as a cumulative sum of the resulting values. This has potential to be useful in carrying out more objective tests of linguistic selection akin to Ahern et al. (2016), by removing or controlling for the topical-cultural element. As a baseline, it could be useful to models incorporating further effects of language change, such as structural-phonological properties (Szmrecsanyi, 2016) and content biases (Tamariz et al., 2014), polysemy (Hamilton et al., 2016b), socially conditioned variation (Samara et al., 2017),
network properties (Pierrehumbert et al., 2014) and other sociolinguistic effects (Calude et al., 2017).

In principle, the advection approach also could be used in other domains of cultural evolution, where there is diachronic data available about the co-occurrence of traits or properties (in lieu of context words) of cultural elements (in lieu of words).

References


