Quantifying the dynamics of topical fluctuations in language

Citation for published version:
Karjus, A, Blythe, R, Kirby, S & Smith, K 2020, 'Quantifying the dynamics of topical fluctuations in language', *Language Dynamics and Change*, vol. 10, no. 1, pp. 86-125. https://doi.org/10.1163/22105832-01001200

Digital Object Identifier (DOI):
10.1163/22105832-01001200

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Language Dynamics and Change

BRILL

LANGUAGE DYNAMICS AND CHANGE 10 (2020) 86–125

brill.com/ldc

Quantifying the dynamics of topical fluctuations in language

Andres Karjus
University of Edinburgh, Edingburgh, UK
a.karjus@sms.ed.ac.uk

Richard A. Blythe
University of Edinburgh, Edingburgh, UK
r.a.blythe@ed.ac.uk

Simon Kirby
University of Edinburgh, Edingburgh, UK
simon.kirby@ed.ac.uk

Kenny Smith
University of Edinburgh, Edingburgh, UK
kenny.smith@ed.ac.uk

Abstract

The availability of large diachronic corpora has provided the impetus for a growing body of quantitative research on language evolution and meaning change. The central quantities in this research are token frequencies of linguistic elements in texts, with changes in frequency taken to reflect the popularity or selective fitness of an element. However, corpus frequencies may change for a wide variety of reasons, including purely random sampling effects, or because corpora are composed of contemporary media and fiction texts within which the underlying topics ebb and flow with cultural and socio-political trends. In this work, we introduce a simple model for controlling for topical fluctuations in corpora—the *topical-cultural advection model*—and demonstrate how it provides a robust baseline of variability in word frequency changes over time. We validate the model on a diachronic corpus spanning two centuries, and a carefully-controlled artificial language change scenario, and then use it to correct for topical fluctuations in historical time series. Finally, we use the model to show that the emergence of new words typically corresponds with the rise of a trending topic. This suggests

© ANDRES KARJUS ET AL., 2020 | DOI:10.1163/22105832-01001200

that some lexical innovations occur due to growing communicative need in a subspace of the lexicon, and that the topical-cultural advection model can be used to quantify this.

Keywords

advection – lexical dynamics – language change – language evolution – frequency – topic modeling – corpus-based

1 Introduction[1]

Elements of a language, be they words or syntactic constructions, never exist by themselves, but in some context. Contexts, or topics, tend to change with the times, along with the world that they describe. These changes are expected to be reflected in (representative, balanced) diachronic corpora. If a particular topic—be it computers, cuisine or terrorism—rises or falls in public interest or newsworthiness, it would be reasonable to expect a similar effect in the corpus frequencies of lexical elements relevant to the given topic, particularly content words such as nouns.[2] It follows from this that the changing popularity of some words, apparent from raw corpus frequencies, might well be explained simply by the rise or fall of their most prevalent topics, rather than being a product of other aspects driving language change, such as sociolinguistic prestige or inherent contextual fitness.

This paper seeks to investigate this idea, which we believe is rather intuitive and widely held, yet to our knowledge has not been formalized in a quantitative way. We will argue that by doing so, we arrive at an informative baseline for frequency-based approaches to lexical dynamics and language change in general. In particular, we show its potential for quantifying topic-driven innovations in the lexicon, and its utility in distinguishing selection-driven change from changes stemming from language-external factors, which manifest as topical fluctuations.

[1] A previous, considerably shorter version of this paper outlining the basic model appeared as an extended abstract in the proceedings of the Society for Computation in Linguistics (Karjus et al., 2018b).

[2] We will use the terms 'word', 'lexical item', 'linguistic variant' and 'linguistic element' more or less interchangeably in the following text, depending on the literature or subfield being discussed.

More precisely, we introduce a quantitative measure of topical change that we call *advection*, a term borrowed from physics where it is used to denote the transport of a substance by the bulk motion of a fluid. The analogy is that words are swept along by movements (increases or decreases in frequency) of associated topics. We implement a topical advection measure using a readily interpretable computational technique based on a robust method from distributional semantics. This approach requires very little tuning of global parameters and produces reasonable results given a sufficiently large corpus. As we will show, it is capable of capturing the effect of changing topic frequencies on the frequencies of individual words.

We begin in Section 2 by providing a brief overview of the state of the art of corpus-based evolutionary language dynamics research and identify the difficulties associated with disentangling different contributions to word frequency changes that may be of interest. We introduce the *topical-cultural advection model* in Section 3, and define our measure of advection in terms of the frequency change of words associated with topics. We first show (Section 4.1) that advection is positively correlated with word frequency changes in the Corpus of Historical American English (COHA), indicating that the model successfully captures a component of language change. In Section 4.2 we test the advection model by showing that it correctly associates word frequency changes with a stylistic shift in an artificially-constructed corpus. We then show how it can be used to adjust frequency time series (Section 4.3), and finally (Section 4.4) how it also allows us to quantify the propensity for new words to emerge alongside trending topics.

We conclude that topical advection should be controlled for in any corpus-based research which relies on the (changing) frequencies of lexical items to make claims about patterns or mechanisms of language change. While this paper focuses on language, we believe that the same basic approach could also be utilized in studying the rise and fall of other products of human culture, given appropriate databases or corpora.

2 Background: corpus-based approaches to lexical dynamics and language evolution

A question that often arises in corpus-based evolutionary language dynamics is the causal origin of language change. A key difficulty lies in disentangling the many different possible causes of language change, some of which may be of greater or lesser interest. A number of factors operating on the level of the individual speaker that potentially influence linguistic selection have been proposed and tested, either in experimental settings, simulations, or corpora with speaker metadata—such as the competing pressures of learnability, expressivity, simplicity and efficiency (Kirby et al., 2008; Smith et al., 2013; Carr et al., 2017; Kanwal et al., 2017; Zipf, 1949; Enfield, 2014; Culbertson and Kirby, 2016), egocentricity and content biases (Tamariz et al., 2014), socially conditioned variation (Samara et al., 2017), and various other social effects (Calude et al., 2017; Lev-Ari and Peperkamp, 2014; Labov, 2011). While language change is perpetuated by the utterance selections of individual speakers over time, some factors also influencing selection may be seen as properties of the population, or those of the linguistic system, such as various structural-phonological properties (e.g. Szmrecsanyi, 2016; Ohala, 1983), phonological dispersion and clustering (Dautriche et al., 2016, 2017; Newberry et al., 2017), polysemy (Hamilton et al., 2016a; Calude et al., 2017), social network properties (Baxter et al., 2009; Castelló et al., 2013), top-down language regulation (Daoust, 2017; Ghanbarnejad et al., 2014; Rubin et al., 1977; Amato et al., 2018), community consensus and relative prestige associated with different variants and languages (cf. Pierrehumbert et al., 2014; Abrams and Strogatz, 2003; Hernández-Campoy and Conde-Silvestre, 2012; Labov, 2011). However, some changes may be a result of purely random effects, as individual speakers have access only to a finite sample of utterances (cf. Section 2.2).

In evolutionary terms, this amounts to the problem of teasing apart drift from selection in language change. Even where one can identify a systematic component to a change (selection), factors that might be of interest from a linguistic perspective need to be disentangled from those that are driven by changes in society and culture, or appear due to uneven sampling of genres, registers or topics in a corpus (Szmrecsanyi, 2016; Szmrecsanyi et al., 2014; Hinrichs et al., 2015; Pechenick et al., 2015). Such considerations have come to the fore due to sharp increases in the availability of quantitative data over the last decades. These datasets record how languages are used (corpora), what their distinguishing features are (typological databases) and to what extent languages are used (demographic databases). This development has given rise to the field of *language dynamics*, which has been described as an interdisciplinary approach to language change, evolution, and interlanguage competition, relying on large databases and quantitative modeling, including simulation-based approaches (Wichmann, 2008). Since our contribution applies to corpus research first and foremost, our focus in the following brief review will be on this strand of language dynamics.

2.1 *Previous research*

Large diachronic collections of language use are of greatest utility from the perspective of understanding language change, as from these one can extract trajectories of change and dynamics of competition between communicative variants. One body of research aims to quantify statistical laws of language change over time, those of word growth and decline, and relationships between word frequencies and lexical evolution (Keller and Schultz, 2013, 2014; Feltgen et al., 2017; Pagel et al., 2007; Newberry et al., 2017; Lieberman et al., 2007; Cuskley et al., 2014; Amato et al., 2018). This has also involved claims regarding the effects of real-world events (like wars) on these processes (Wijaya and Yeniterzi, 2011; Petersen et al., 2012; Bochkarev et al., 2014).

There is also an emerging strand of research investigating semantic change and language dynamics from the point of view of meaning, using diachronic corpora and distributional semantics methods. These include the various flavors of Latent Semantic Analysis (Deerwester et al., 1990) and word2vec (Mikolov et al., 2013). This research broadly falls into two categories: methods proposals usually accompanied by exploratory results (Sagi et al., 2011; Gulordava and Baroni, 2011; Wijaya and Yeniterzi, 2011; Jatowt and Duh, 2014; Kulkarni et al., 2015; Hamilton et al., 2016a; Frermann and Lapata, 2016; Schlechtweg et al., 2017; Dubossarsky et al., 2017; Kim et al., 2014; Rosenfeld and Erk, 2018)—and applications of such methods, usually with more specific linguistic questions in mind (Hamilton et al., 2016b; Xu and Kemp, 2015; Perek, 2016; Rodda et al., 2017; Dubossarsky et al., 2016; Dautriche et al., 2016). Notably, all of these approaches are, one way or another, based on (co-occurrence) frequencies of words, and as such are naturally subject to sampling biases potentially introduced by uneven representation of topics and genres in a corpus.

We believe our contribution is also relevant for traditional corpus linguistics, or research more geared towards investigating specific phenomena in some target language(s)—if it involves counting frequencies of words or other elements of speech in diachronic corpora, and using these counts in explanatory models. In all of these cases, it is necessary to deal with factors that serve to confound the explanatory factor of interest, for example, those that are specifically linguistic, such as various language processing and transmission biases. In particular, as noted above, there is a need to separate random and systematic effects, and frequency changes arising from changes in topic and genre across the corpus and over time. We expand on both confounds below.

2.2 *Confound 1: language change involves drift*

It is widely agreed that not all language change is necessarily caused by selection by speakers for certain variants or utterances, but also involves random processes (i.e., drift, or neutral evolution) (Sapir, 1921; Hamilton et al., 2016b; Blythe, 2012; Newberry et al., 2017; Jespersen, 1922; Reali and Griffiths, 2010; Andersen, 1990). Naturally, this should be taken into account in a diachronic study of language. This requires some way of distinguishing changes resulting from drift and those, potentially more interesting ones, resulting from selection.

Our proposal is by no means the first attempt to construct some form of baseline or null model against which potential cases of directed change can be compared. There have been various proposals to carry over the selection and neutral drift paradigm from evolutionary biology, where drift refers to cases for differential replication without selection (cf. Croft 2000). It has been argued that a prerequisite for studying language change through this paradigm would be the construction of well-informed null models (Blythe, 2012). Proposals in this vein tend to rely directly on or draw from Kimura's neutral model of evolution and the Wright-Fisher model (Kimura, 1994; Ewens, 2004). Alleles are equated with linguistic variants and neutral evolution (drift) with (neutral, random) language change (Reali and Griffiths, 2010).

Adopting this framework, Newberry et al. (2017) apply tests developed in genetics for distinguishing drift and selection to frequency time series of competing linguistic variants. In particular, they apply the Frequency Increment Test (Feder et al., 2014), and do so on three test cases of changes in the grammar of the English language. They conclude that this constitutes a systematic approach for distinguishing changes likely resulting from linguistic selection rather than drift (however, cf. Karjus et al., 2018a). With the culturomics proposal (Michel et al., 2011) in mind, Sindi and Dale (2016) propose another model to detect departures from neutral evolution in word frequency variation, based on comparing frequency series with randomly generated baselines.

In a slightly different sense, the notion of '(linguistic) drift' has also been used previously in a computational semantics study (Hamilton et al., 2016b). Drift is defined there as semantic change stemming from (presumably regularly ongoing) change in language—not a reflection of considerable change in the culture that a particular language codifies. The latter is labeled as 'cultural shift', which is claimed to be more common in nouns than verbs. Detecting 'significant' changes in word meaning has also been attempted (Kulkarni et al., 2015), with the two aforementioned approaches using a similar distributional semantics method for determining semantic similarity across time, and the latter employing a similar significance detection method as Feder et al. (2014).

The concept of linguistic drift is also commonly utilized in computational modeling of experimental communication data, where the null model, without communicative biases (such as bias for egocentric coordination or superior expression, cf. Tamariz et al., 2014) would consist of randomized changes, or drift. The question of distinguishing selection from drift has also arisen more widely in cultural evolution, for example, in the contexts of prehistoric pottery (Crema et al., 2016), keywords in academic publishing (Bentley, 2008) and baby names (Hahn and Bentley, 2003).

Another take on neutral evolution was proposed by Stadler et al. (2016), who demonstrated using a simulation model that language change may also self-actuate without selection but via momentum, whereby variants simply become more popular by virtue of having gradually become more popular. This model produces S-shaped frequency change curves, which have been argued to be a characteristic of language change (Blythe and Croft, 2012). Relatedly, a similar S-shaped trajectory was seen in a model where a neutral process of language acquisition interacts with a dynamic social network structure (Kauhanen, 2017)

2.3 *Confound 2: language is not independent of its environment*

No linguistic element exists in isolation: we use language to communicate about salient events in the world, and the language in use in a given time period therefore indirectly reflects the events, concerns and preoccupations of that time. These reflections should be observable in a representative corpus. The potential effect of real-world changes and hot media topics on corpus-based language usage patterns have been noted in multiple recent studies (see below). However, the way this is approached varies between studies with different aims. We observe at least three ways the connection between language use and real-world change has been considered: as a minor by-product of corpora; as an assumption for language-based culture research; and thirdly, as a factor to be necessarily accounted for in linguistic analysis. All of these deserve further discussion.

2.3.1 Topical-cultural impact on corpora as an inconsequentiality

In a study of mathematical approaches to detecting selection (against drift, cf. Section 2.2) Sindi and Dale (2016) observe that words with very similar frequency change patterns also qualitatively belong to similar semantic clusters or topics (e.g., words related to war increasing during periods of war at similar rates). Since their focus is on evolutionary selection dynamics, the topical effect is discussed in passing. Keller and Schultz (2013) look into word formation dynamics and also observe qualitatively that cultural changes seem to be reflected in the dynamics of the larger morpheme families, but do not explore further.

2.3.2 Topical-cultural impact on corpora as an assumption

The field of 'culturomics' is based on the assumption that changes in the sociocultural environment of a language should be reflected in the concurrent usage of its lexical items. Word frequencies in large diachronic collections of texts (such as Google Books) are seen as an interesting way of observing and studying historical real-world changes (Michel et al., 2011; Bentley et al., 2014). It has also been noted that times of change and conflict, such as wars and revolutions, are observable in language dynamics, such as the emergence of new words (Bochkarev et al., 2014, 2015) and word growth rates (Petersen et al., 2012). Petersen et al. (2012) conclude that "[t]opical words in media can display long-term persistence patterns /…/ and can result in a new word having larger fitness than related 'out-of-date' words". Socio-political change can in some cases be observed in the contemporary (distributional) semantics of words, e.g., *Kennedy* being associated with *senator* before and *president* after the year of his election (Wijaya and Yeniterzi, 2011). There have been at least two claims of correlations between changes in language and political processes (Frimer et al. (2015) on the US Congress, Caruana-Galizia (2015) on Nazi Germany), although these have both recently been criticized for methodological errors resulting in spurious correlations (Koplenig, 2017b). The culturomics approach, and research based on the Google Books corpus in particular, has been recently criticized for ignoring important issues such as metadata of the texts underlying the corpus (Koplenig, 2017a) and unbalanced sampling of topics, genres or authors in corpus composition (Pechenick et al., 2015).

2.3.3 Topical-cultural impact on corpora as a problem

While the relationship between topicality and language use allows us to use language as a window into changes in the world, as claimed by practitioners of culturomics, it poses a problem if we want to use fluctuations in those same patterns of language use as a diagnostic for linguistic, rather than sociocultural, change. In recent years a number of authors have drawn attention to the importance of controlling for contextual factors such as genre and topic, with some voicing the concern that studying language change via corpus frequencies of linguistic elements alone could potentially be misleading. We review some of these below.

Lijffijt et al. (2012) are concerned with testing the assumption that a single-genre general purpose corpus should be relatively homogeneous over time.

They find that the period of the English Civil War had an identifiable effect on word frequencies in the Corpus of Early English Correspondence, which they attribute to the over-representation of war-related topics and authors with a military background, violating the assumption of homogeneity. In a corpus study on the English *which-that* alternation, Hinrichs et al. (2015) emphasize the importance of controlling for genre and register, since those alternating variants are associated with different genres. In a study on the evolution of the English genitive markers, Szmrecsanyi (2016)—lamenting the unreliability of corpus frequencies in general—reasons that while a "proper" grammatical change has taken place, "[a] good deal of the diachronic frequency variability in the dataset can be traced back to environmental changes in the textual habitat". They point out that the shifting nature of the topics in the news section of their diachronic English language corpus—in particular, the coverage of non-animate entities such as collective bodies—plays a role in the changing frequencies of *of*-genitives, their object of study.

Topical effects have also been suggested to play a role in word survival dynamics and semantic change. In a synchronic sociolinguistic study of Mãori loanwords in New Zealand English, Calude et al. (2017) point out that simple across-corpus loanword frequencies could be misleading in terms of loanword success, since "certain words and concepts can become more widely used because they might be relevant to certain topics of conversation". Studying the success of loanwords in French news corpora, Chelsey and Baayen (2010) similarly ask if topic matters: is the occurrence of many financial borrowings the result of a high proportion of financial articles in the corpus, or are financial borrowings just more likely to become entrenched? Their conclusion is that, without information on topics, there is simply no way to tell. Investigating the rise and decline of words in online newsgroups, Altmann et al. (2011) find that while diffusion among users (speakers) is the primary determinant of the success of a word, spread across the conversation threads within newsgroups (which could also be seen as "topics") also plays a significant role, with both being better predictors than raw frequency. Using a distributional semantics approach, Rodda et al. (2017) find qualitative support for the idea that the diffusion of Christianity drove semantic change in Ancient Greek, but point to the over-representation of certain genres in their corpus and call for more research on the effects of corpus composition.

Although many corpora do include metadata on genres and registers, fine-grained topics—which may well change rapidly within genres like daily news—are more often than not missing from the picture. Consequentially, there appears to be a widely articulated need across various branches of corpus-based language research for a method to control for topical fluctuations in corpora, as they are recognized to have potentially far-reaching effects on linguistic analyses based on such data, particularly if they make use of frequencies of linguistic elements. The method we introduce below aims to address that issue.

3 The topical-cultural advection model

We begin with the simple intuition that if a topic becomes more prevalent, the words describing it, relating to it and possibly giving rise to it, should become more frequent as well. Similarly, the decline of a topic may drive the decline of words related to it. This effect should be clearer for words specific to certain topics, and less pronounced (or absent altogether) for words with a more general meaning. While we do not claim that our approach offers a remedy to all the concerns reviewed above, we will show that it does provide a simple, easily implemented and intuitive baseline for controlling for topic-related effects arising from sociocultural change or uneven sampling of a corpus. In this section we define the topical-cultural advection model. To aid readability, we defer certain technical details of the implementation to a Technical Appendix.

3.1 *Definition of the model*

In its simplest form, the topic of a target word in the topical-cultural advection model is defined as the set of words that are most strongly associated with the target word in terms of co-occurrence over a particular period of time. The context sets should be re-evaluated for each period subsample in a corpus, to accommodate for natural semantic change of words (which would also entail changes in context).

The advection value of a word in time period t is defined as the weighted mean of the changes in frequencies (compared to the previous period) of those associated words. More precisely, the topical advection value for a word ω at time period t is

$$\text{advection}(\omega; t) := \text{weightedMean}\big(\{\text{logChange}(N_i; t) \mid i = 1, \ldots m\},\ W\big) \quad (1)$$

where N is the set of m words associated with the target at time t and W is the set of weights (to be defined below) corresponding to those words. m is a free parameter (we use the value 75 in the following). The weighted mean is simply

$$\text{weightedMean}(X, W) := \frac{\sum x_i w_i}{\sum w_i} \quad (2)$$

where x_i and w_i are the i^{th} elements of the sequences X and W respectively. The log change for period t for each of the associated words ω' is given by the change in the natural logarithm of its frequencies from the previous to the current period. That is,

$$\text{logChange}(\omega';t) := \ln[f(\omega';t)+s] - \ln[f(\omega';t-1)+s] \tag{3}$$

where $f(\omega';t)$ is the number of occurrences of word ω' in the time period t, and s is a smoothing constant, to avoid $log(0)$ appearing in the expression. The value of s is set to 0 if the relevant frequency $f(\omega') > 0$, or if both $f(\omega';t)$ and $f(\omega';t-1)$ are zero. Otherwise, s is set to the value equivalent of 1 occurrence after frequency normalization. Simply put, we replace zero-frequencies with small values to be able to compute log frequency change from and to 0. Mentions of log frequencies and log change here and below refer to natural logarithms. See the Appendix for details on why log change is favored over percent change.

The crucial ingredient in the model is the set of weights W for the words in N. Here, we adopt the positive pointwise mutual information (PPMI) score (Church and Hanks, 1990). We provide details of how PPMI is calculated in the Technical Appendix. The idea is that PPMI assigns a higher score to words that are strongly associated, based on their co-occurrence with other words. While a very general, high frequency word may occur more often in the vicinity of a target word than some specific, low frequency word, the conceptual association between the target and the general word is likely quite low, as the latter co-occurs with many other words as well—while the topic-specific one likely does not. PPMI captures this notion and downweights co-occurrence counts with such general words. In terms of the advection model, weighting the frequency changes of the context words by their association scores leads to a better model, as context words more strongly associated with the target more likely belong to the same underlying topic.

3.2 *Connections with previous work*

This model builds on the core notions and recent developments in distributional (vector) semantics, where the meanings and topics of words are defined through their vectors of co-occurring words. These vector spaces may be learned directly from data (Mikolov et al., 2013) or be based on term co-occurrence matrices (Deerwester et al., 1990; Pennington et al., 2014). In all of these approaches, two words with similar vectors (across dimension reduced vector spaces, or across the vocabulary of context words) are considered to have similar meaning. A common measure of similarity is the cosine of the angle between the two vectors. Recently, an alternative has been proposed in the form of the APSyn measure (Santus et al., 2016), which involves comparing the rankings of the topmost associated context words instead of the whole vocabulary. The intuition behind APSyn is that only the most associated context words hold relevant information about the target word, while most of the words are likely irrelevant. Santus et al. (2016) demonstrate the capacity of APSyn to perform as well, and in some cases better than the vector cosine. Considering only top ranking contexts is also similar to Hamilton et al. (2016b), who use cosine similarity between word vectors between time periods to measure semantic change, but as a second measure, the extent of the change in a word's similarity to its top nearest neighbors (Hamilton et al., 2016b). We adopt this approach of considering only the top most *m* associated context words here to determine a "topic" for each word, using PPMI as the association score.

It is nevertheless worthwhile to compare our PPMI-weighted approach with a more traditional topic model. To this end, we also implemented the advection measure using Latent Dirichlet Allocation (LDA) (Blei et al., 2003). In this approach, each of its latent k topics (we used $k = 500$) is assigned a frequency change value based on the frequency changes in the vocabulary, weighted by their association with the topic (as a latent topic is essentially a distribution across the vocabulary). The topical advection value of a target word is then the mean of the changes in the topic frequencies, weighted according to the probability a word belongs to each given topic. The details of this calculation are given in the Technical Appendix.

As will be seen below in Section 4.1, the descriptive power of the two models is rather similar. While LDA is widely used, we feel that our simple PPMI-weighted model has certain advantages. In addition to requiring the setting of only a single parameter, it is much less computationally complex (thus faster), and the results are easily interpretable. Specifically, each "topic" of a target is a short list of top context words (meaning the advection value, being the weighted mean of their log frequency change values, is on the same scale as the target word log frequency change values). It is also straightforward to observe the behavior of a target word's topic and calculate its advection value both before and after it has entered the language or gone out of use—by re-using the context word list and the corresponding weights from a period where the target word was already (or still) frequent enough for its topic to be inferred.[3]

[3] Similar extensions for evaluating topics over time exist for the latent topic modeling approach, (cf. Wang and McCallum, 2006; Blei and Lafferty, 2006; Roberts et al., 2013), which we will not be examining in further detail here. Furthermore, Fremann and Lapata (2016) use a Bayesian approach in some aspects similar to classical topic modeling to measure semantic

4 Results of applying the advection model in a number of language change scenarios

We now turn to two large, representative, POS-tagged corpora, in order to get a sense of how well the topical-cultural advection model performs, and proceed to demonstrate a number of useful applications. We preface the results with a few crucial technical details that apply to all the following subsections, and both the PPMI and LDA based models, while leaving a more thorough description of the parameterizations of the models and relevant corpus preprocessing steps to the Technical Appendix.

The word counts for each time period (segment) in a corpus were normalized as frequencies per million words (pmw). Since cultural effects are likely the most pronounced on content words, particularly nouns (see also Hamilton et al., 2016b), we only consider common noun targets in the following analyses. For the context vectors (see Section 3.1), we exclude stop words and use only content words (based on POS tags). We use the top $m = 75$ context words for the PPMI based model. We set a (rather conservative) threshold of a minimum of 100 occurrences per period for words to be included in the model. If a word occurs less than 100 times in a corpus period, it will not be assigned a context vector—thus also no advection value for this period—nor will it be used as a context word. This comes down to a classical statistical sampling problem: if a word only occurs a few times, then its context vector (topic) is more likely to be composed of quite random words, in a random ranking, while if a word is observed numerous times, the ranking of its (recurring) context words becomes more reliable.

This however also means that it is not possible to calculate the advection value for low frequency words like recent innovations and words going out of usage. Since these correspond to periods of particular interest for such words, we experimented with using a 'smoothing' procedure to improve the informativeness of the topics. Specifically, the 'smoothed' data, used for deriving the topics, comprises text from a target period and its preceding period (word counts still correspond to the frequencies in the target period). This procedure increases the chance of inclusion for relevant context words that would otherwise not be present due to being too low frequency in one or both of the periods. Consequently, it also improves the precision of the advection measure for words decreasing in frequency in a given target period.

change in a word as change in its distribution of "contexts" (topics). Their model however appears very demanding in terms of the size of the training corpus.

4.1 *Topical advection and diachronic language change*

We use the Corpus of Historical American English (COHA) (Davies, 2010) as a test set in order to evaluate the extent to which the model is capable of accounting for variance in word frequency changes. The COHA spans two centuries, starting with 1810, is binned into decade-length subcorpora by default, and is meant to be balanced across genres for each period (news, magazines, fiction, non-fiction; but see the Appendix for details).

With 20 decades, there are potentially 19 frequency change points that can be calculated for each target word. There are 7551 unique words in the no-smoothing condition, and 75653 data points. There are 10060 words (107475 data points) in the smoothing condition (concatenated data results in more words being above the minimal threshold to be eligible for the advection calculation).

To test the descriptive power of the two aforementioned implementations of the advection model, PPMI-based and LDA-based, we correlate the log frequency change values of common nouns between successive decades in the COHA corpus to their respective advection values (their log topic frequency change values in the same decades).[4] The results are presented in Fig. 1. The different scales on the axes indicate that words experience more rapid changes in either direction than topics, as one might expect, topic values being averages of context word frequency changes.

We find that, as expected, frequency changes correlate significantly and positively with advection, and that the smoothing operation further improves the correlation. The LDA-based and the PPMI-based models yield similar results. The less complex PPMI-based model (with smoothing) performs even slightly better, describing an average of 30% of variation in noun frequency changes between decades. There is also some variation between decades. The stronger correlations in some decades may be an indication of either a change in discourse in American English, as chronicled in the corpus, or differences in topical sampling between the subcorpora. We find that the strength of this relationship is in turn positively—but only moderately—correlated with observed divergences between distributions of genres in the decade subcorpora (see the Appendix for more details). In short, the advection model tends to describe more variance in word frequency changes between decade pairs which exhibit a larger divergence in their genre distribution (which can be expected to affect the underlying topic distribution).

[4] Importantly, we are not correlating absolute frequencies of words with the absolute frequencies of topics, which could easily lead to spurious correlations (cf. Koplenig and Müller-Spitzer (2016) for recent criticisms).

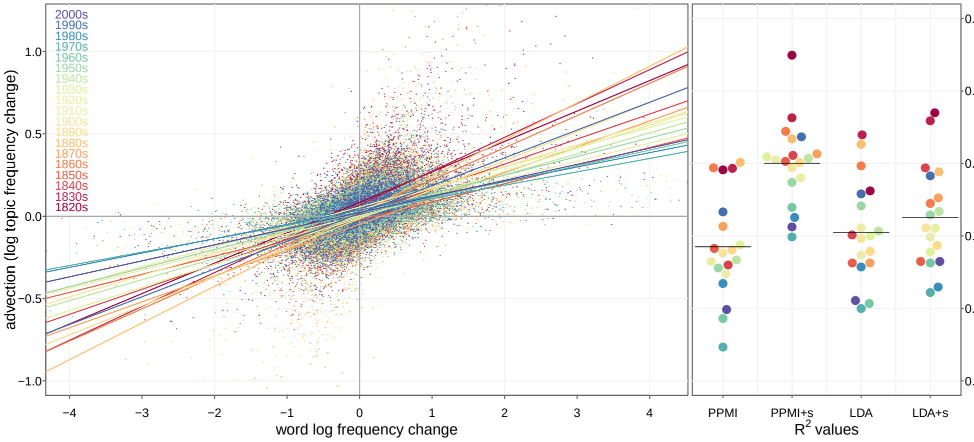

FIGURE 1 Left panel: log frequency changes of nouns and their corresponding topical advection (log topic change) values from two centuries of language change (from the PPMI-based model with topic smoothing). Each of the 107475 dots indicates the frequency change and advection value of one of the 10060 nouns, colored by decade. As such, many words occur multiple times in this figure. Positive values indicate increase, negative ones indicate decrease. Right: R^2 values for correlations for each decade. *+s* indicates models with topical smoothing; the black bars mark the means. The PPMI-based models with smoothing have the highest mean R^2 of 0.25. All $p < 0.001$. This figure illustrates the robust correlation between frequency change and advection. We will be using the same colors to indicate decade subcorpora throughout this paper.

These results clearly show that topical fluctuations can be expected to explain a significant amount of variability in the change in word frequencies, which one might otherwise be tempted to attribute to other processes, such as selection. As such, the topical-cultural advection measure serves as a useful baseline in any quantitative model predicting frequency changes in linguistic elements.

4.2 *Artificially-constructed language change based on genres in a synchronic corpus*

Having established that advection constitutes one (small but significant) contribution to word frequency change in general, we now test whether our model can identify instances where it is the main contribution to change. This is difficult to determine with natural data, as one does not know *a priori* what the drivers of change are (beyond the genre distribution discussed in the previous section). To deal with this problem in a more controlled way, we construct an artificial corpus wherein the main component of change between two subcorpora is a known stylistic shift. We should then find that changes in word frequencies are strongly correlated with topics that are more prevalent in one style than the other.

Specifically, we employ the Corpus of Contemporary American English (COCA) (Davies, 2008), which is the synchronic cousin of COHA. It consists of contemporary American English data from 1990–2012, again labeled by genres. However, in contrast to COHA, COCA is large enough that genre subcorpora from even relatively short time segments contain enough data for training the advection model. This allows us to avoid the potential confound of actual diachronic language change. We used only data from a short time span (2005–2010) in the academic journals and spoken language (TV and radio transcripts) subcorpora to construct an artificial "language change" from academic to spoken style and content, by defining the former subcorpus as one "period" and the latter as the following one.

We then measured the log frequency changes of nouns, as in the previous section, and their respective advection (log topic frequency change) values. Not surprisingly, among the top decreased are words like *subscale*, *coefficient*, *self-efficacy*, *carcinoma*, *pretest*; while words like *tonight's*, *ma'am*, *fiancee*, *everybody*, and *paparazzi* have all increased with the switch in genre. Again, the advection measure correlates positively with frequency change, and describes a notable amount of its variability: in our favored PPMI-based model, we find $R^2 = 0.45$ without smoothing and $R^2 = 0.73$ with smoothing applied.[5] This is to say, the advection model appears to successfully pick up on the genre change, reflected in the high (positive) correlation value—the decrease in academic and increase in spoken style word frequencies corresponding to the fall of the academic and rise of the spoken topics or genres. Importantly from the perspective of validating our model, the R^2 values are higher here than in the analysis of COHA. Presumably there are other forces affecting word frequencies in the COHA besides genre divergences and topic fluctuations; at the same time, the (actual) changes between subsequent decades are likely less stark.

4.3 *Using advection to adjust for topical fluctuations in time series*

Having measured the descriptive power of the advection model and demonstrated how it behaves with re-evaluated topics over time, we now turn to an application of the model to deal with the confounds set out in Section 2.3.3. When it comes to predicting frequency changes of words or any other linguistic elements between periods of time, the advection measure can be included

[5] As there are only two 'periods', smoothing here refers to concatenating the entire spoken and academic subcorpora for the purposes of estimating the topics of each word.

as a control variable in a predictive model (see Section 4.1). In the case of time series analysis (i.e., involving multiple changes over time), it is possible to utilize the advection measure as a form of (in the following example, additive) time series decomposition, by carrying out the following operation. For a given word, for every period data point: subtract the advection value (log topic frequency change) of the target word from the log frequency change value of the target word. This yields a new series of frequency change values where the topical change component has been removed. In this section, we make use of the simple PPMI-based model (with smoothing). The advection values therein are averages over individual word log frequency changes, so the two quantities are on the same natural scale (changes in word frequencies) and can therefore simply be subtracted from each other. See the Appendix for a more technical breakdown of the approach.

The operation described above is similar to seasonal decomposition, a commonly applied approach in (multi-year) time series analysis to control for seasonal ups and downs (e.g., heating costs in cold and warm seasons). In our case, the "seasonality" (topical fluctuations) is not inferred from the time series itself, but calculated independently. Another way of looking at this is as a way of distinguishing the metaphorical "word of the day", one that is selected for, from a word that just comes and goes with the "topic of the day". Adjusting for topics has the potential to be useful in carrying out more objective tests of linguistic selection (cf. Newberry et al., 2017; Sindi and Dale, 2016; Bentley, 2008; Blythe, 2012), by controlling for the topical-cultural element.

Figure 2 illustrates the results of the adjustment operation on the example of a segment of the time series of the word *payment* in COHA. The left side panel depicts the log frequency changes and the subsequent adjustment. The middle panel shows the same data as actual (per-million) word frequencies. Namely, the time series of word frequencies may be subsequently reformed for visualization purposes, after operating on the change points, as the (exponential of the) cumulative sum of the resulting log change values, initialized with the log frequency of the word at the start of the time series. This however requires selecting the arbitrary initialization value for the cumulative sum, which of course shifts the actual frequency values in the reformed series. The same approach can be used to visualize a topic "frequency" time series.

Finally, the right side panel in Fig. 2 illustrates yet another way of looking at word frequency changes through the lens of advection, making use of regression residuals. We ran a linear regression model for each decade (cf. Fig. 1), where frequency change is predicted by advection. Each blue point above and below the zero line marks the residual value of *payment* in each decade. Above zero indicates that the word is doing better than would be expected by its topic

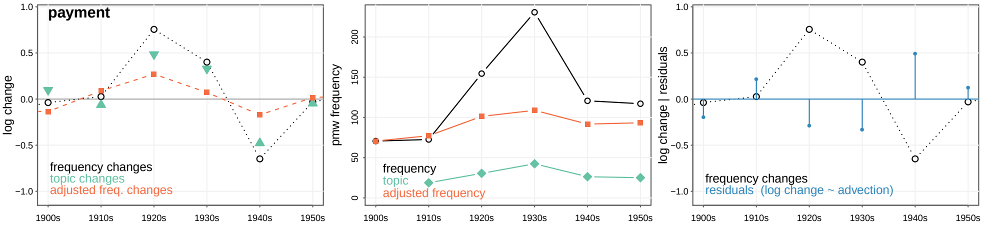

FIGURE 2 Time series of *payment* in the first half of the 20th century. Usage of the word increases considerably in the 1930s, but so does its topic. Black circles: log frequency change values (dotted line), actual frequency (solid line). Green triangles: topic frequency; change values on the left panel, with the triangle pointing up and down corresponding to the adjustment; as relative frequency in the middle panel. Orange squares: frequency changes of the word adjusted by subtracting the log topic frequency changes from the word log frequency changes (left; as a reformed series in the middle panel). Note that the green topic line in the middle panel is plotted for reference and only illustrates topic frequency as a relative measure, being a cumulative sum of the log topic changes, initiated with an arbitrary value. Blue dots below and above zero on the right side panel: residuals of the target word taken from per-decade regression models. The adjustment operation is generally in line with the residuals: frequency gets adjusted upwards when the residual is positive, and downwards when the residual is negative.

(hinting at selection). Conversely, below zero values indicate that the word is used less than would be expected given the prevalence of its topic.

One obvious concern with using the advection measure for a decomposition-like operation—subtracting topic frequency change from word frequency change—is that it might be over-correcting frequency changes and interfere with observing genuine competition in language, whereby one lexical element is replaced with a synonym over time. To investigate this possibility, we constructed a second artificial corpus, based on 11 decades (1900s–2000s) of the preparsed COHA corpus (cf. Section 4.1). The manipulation of the corpus consisted of replacing a set of otherwise stable words with (invented) synonyms in a controlled way. We find that after applying the advection adjustment, the artificially-constructed language change remains untouched, leading us to believe that this adjustment by subtraction does not obscure genuine (although in this case artificial) cases of selection (see the Appendix for a full technical breakdown).

4.4 *Advection predicts lexical innovation*

McMahon (1994) notes that "new words are most likely to survive, and indeed to be created in the first place, if they are felt to be necessary in the society concerned. This is a difficult notion to formalize, but a well-established one". Previous empirical research has linked vocabulary size with communicative need as well. Studying color words in 110 languages across the world, Gibson et al. (2017) argue that the communicative needs rising from the environment where these languages are spoken dictates (to an extent) the color naming systems that emerge. In another cross-linguistic study, Regier et al. (2016) show that the need for efficient communication—which varies across cultures and environments—does seem to drive vocabulary size (in their case, of words for 'ice' and 'snow').

From a historical perspective, this suggests the hypothesis that an increasingly popular topic (i.e. exhibiting positive advection) would be expected to attract new words, providing the detailed vocabulary required—or, conversely, a new word would be expected to exhibit a strong positive advection at its period of first occurrence, compared to the advection values of its topic in previous periods. We are now equipped to test the latter hypothesis.

We identified a test set of 73 "successful" novel common nouns from the COHA that meet the following criteria: our successful novel nouns appear as new words in the 1970s to 2000s, and, importantly, occur with high enough total frequency across (at least some of) these decades for their topics to be reliably modeled (it is in this sense that the nouns are "successful"). Notably, each period of COHA includes a rather large number of new words, but most of them occur at very low frequencies. Figure 3 illustrates the differences in subcorpora sizes across decades in the corpus and the number of new nouns per period.[6]

To remedy the small sample problem particularly relevant to new words (that often start out at low frequencies), we again used the simple "smoothing" technique (see introduction of Section 4), this time concatenating data from all the last four decades for the purposes of constructing the PPMI-based topic vectors. We chose only novel target words from the last few decades of the corpus in order to carry out the following comparison.

[6] Note that these counts correspond to our cleaned version of the corpus (cf. Section 4; this also included the removal of all capitalized words to avoid occurrences of mistagged proper nouns, see the Appendix for details). The numbers of "new" or previously unseen words are likely inflated by the occurrence of spelling mistakes, uncommon words and OCR errors (which commonly end up with the noun tag).

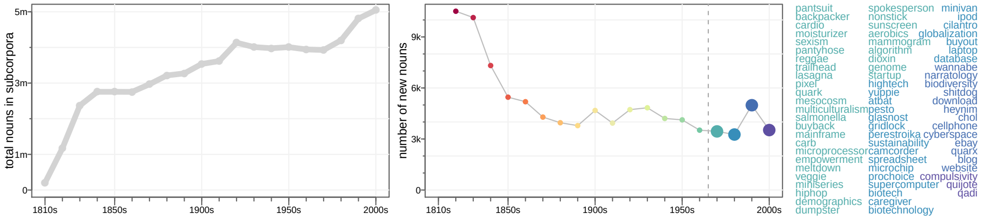

FIGURE 3 Token frequencies of nouns (left) and type frequencies of new nouns (middle panel) in the (preparsed) COHA corpus across period subcorpora. The vertical dashed line on the middle panel indicates the last four decades used to determine the test set of new words in this section; these words are visualized on the right (in corresponding colors).

As each topic consists of a list of words, we computed their advection values (log frequency changes) across ten decades preceding the decade where the target word would first occur in the corpus.[7] In essence, we track how well each topic of each new word is doing throughout a century before the appearance of the innovation. This allows us to measure how many of the (successful) new words belong to topics that exhibit higher advection than before in the period where the new word first appears. For 58 % of novel nouns out of the 73, the advection value of the topic associated with the word was found to be above the upper bound of the 95 % confidence interval of the mean of its advection values over the preceding 10 decades (e.g., *microchip*, cf. Fig. 4). 37 % fell around the means, and only 5 % were below the lower bound of their respective confidence intervals.[8]

We also conducted a t-test in the following manner to test the apparent tendency. We calculated the z-score of the advection value of each of the 73 new words at the decade of first occurrence, using the mean and standard deviation values of the previous decades (separately for each of the new words). A one-sample t-test on this set of z-scores indicated that its mean is significantly ($p < 0.001$) above zero—or in other words, the advection values of new words are on average significantly higher at the time of entry than in preced-

[7] Importantly, the advection calculation only took into account words that actually occur (frequency above 0) in a given decade: 0-to-0 frequency changes are not allowed to bias the earlier advection values to be closer to 0. Although some topic words are also new, most topic words do occur in previous decades.

[8] We also checked if the large number of new words above their mean advection values could possibly be due to some particular semantic cluster of words that might all belong to a similar (trending) topic and thus inflate the results. We computed the APSyn similarity (Santus et al., 2016) on all pairs of the topic vectors of the 73 nouns and found them to be sufficiently dissimilar.

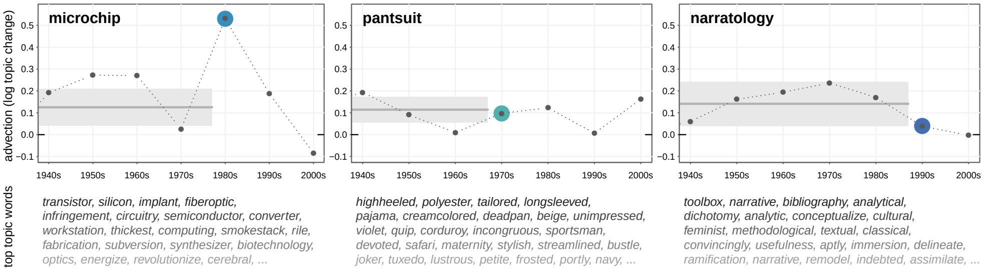

FIGURE 4 Three example novel words. The dashed and dotted dark gray line: the advection (log change) values of the topic of the word; above 0 indicates an increase, below 0 a decrease in the topic (note that this is not the frequency of the word, but the mean log changes in the topic). The brightly colored circle marks the entry decade of the word—this is the advection value that is compared against the mean of the preceding advection values. The mean of preceding decades is indicated with the horizontal solid gray line, with a light gray colored confidence interval. The relevant co-occurring topic words are visualized as clouds below each panel (ordered by their PPMI scores). The word *microchip* is among the 58% of our novel word sample that enter the corpus when its topical advection value is significantly above the mean of the past 10 decades. It is around the mean for *pantsuit*, and below for *narratology*.

ing decades. These findings suggests that the appearance of new words does indeed correspond to the rise of certain topics, or the increasing communicative need for new words. Figure 4 illustrates this effect for three novel words that enter into the corpus at different advection values.

5 Discussion

A language corpus is essentially a sample of aggregated utterance selections by (a sample from) the population of speakers. In principle, factors which have been claimed to drive selection could therefore be tested for in a corpus, as some have been—a diachronic one in case of claims about change dynamics, and synchronic if the claims concern properties of language as such. Models connecting individual-level biases and population-level observations have been recently proposed as well (Kandler et al., 2017; Kandler and Powell, 2018). In the diachronic case, if the analysis was to involve changing frequencies over time, then the topical-cultural advection model would be straightforwardly applicable as a factor of control or baseline change. It could likely also improve tests for selection and drift (cf. Newberry et al., 2017; Sindi and Dale, 2016; Bentley, 2008; Blythe, 2012) by adjusting for the component of fluctuating topics presumably driven by socio-cultural processes or "newsworthiness". While contextual suitability for a topic could be argued to be itself a signal of selection, our model remains applicable, allowing for a quantification of that signal, or to be used as a predictor on its own, as shown in Section 4.4.

In the case of natural language, our technique for measuring topical advection does require a certain amount of data to be reliable (in terms of inference of the topics, cf. Section 4). As such, it is directly applicable to (sufficiently large) corpora, regardless of them consisting of newspapers, books, transcripts, dialogs or interviews. This includes both diachronic corpora (i.e., involving two or more time periods) and synchronic corpora (consisting of distinct subcorpora, cf. Section 4.2). It is less likely to be useful in experimental settings. In principle, the advection model could also be used in other domains of cultural evolution, where there is diachronic data available about the systematic co-occurrence of traits or properties (in lieu of context words) of cultural elements (in lieu of target words, such as nouns in the previous sections).

In a sense, our model also orthogonally complements the momentum model of Stadler et al. (2016). They demonstrate, using a simulation of language evolution, that change can self-perpetuate without selection, when a linguistic variant gains enough momentum in its frequency changes over time. While they model momentum from the frequency change of a variant itself, we model the frequency change of a variant potentially driven by the frequency change in its immediate contextual topic (not itself), or what could be called 'topical momentum'.

6 Conclusions

We presented the topical-cultural advection model, along with two potential implementations, as a straightforward method capable of capturing topical effects in frequency changes of linguistic elements over time. In particular, we demonstrated that the model accounts for a considerable amount of variability in noun frequency changes between decades in a corpus spanning two centuries, retains its capacity when used on an artificially sampled corpus where a change in style and contents has been simulated, and can, to an extent, predict lexical innovation, based on increases in topic frequencies. We also introduced a way of using the advection measure for time series adjustment to distinguish (presumably selection-driven) changes from topical fluctuations (or potentially uneven corpus sampling). We conclude that the topical-cultural advection model adds an important analytical approach to the toolkit for corpus-based lexical dynamics research, or any investigation drawing inference from changing frequencies of linguistic (or other cultural) elements over time.

Acknowledgments

The authors would like to thank Stella Frank for advice with implementing the Latent Dirichlet Allocation version of the model; Jennifer Culbertson for fruitful discussions. The first author of this research was supported by the scholarship program Kristjan Jaak, funded and managed by the Archimedes Foundation in collaboration with the Ministry of Education and Research of Estonia.

References

Abrams, Daniel M. and Steven H. Strogatz. 2003. Modelling the Dynamics of Language Death. *Nature*, 424:900.

Altmann, Eduardo G., Janet B. Pierrehumbert, and Adilson E. Motter. 2011. Niche as a Determinant of Word Fate in Online Groups. *PLOS ONE*, 6(5):1–12.

Amato, Roberta, Lucas Lacasa, Albert Díaz-Guilera, and Andrea Baronchelli. 2018. The Dynamics of Norm Change in the Cultural Evolution of Language. *Proceedings of the National Academy of Sciences*, 115(33):8260–8265.

Andersen, Henning. 1990. The Structure of Drift. In Henning Andersen and Konrad Koerner, editor, *Historical Linguistics 1987. Papers from the 8th International Conference on Historical Linguistics*, pages 1–20. Amsterdam: Benjamins.

Baxter, Gareth J., Richard A. Blythe, William Croft, and Alan J. McKane. 2009. Modeling Language Change: An Evaluation of Trudgill's Theory of the Emergence of New Zealand English. *Language Variation and Change*, 21(02):257–296.

Bentley, R. Alexander. 2008. Random Drift versus Selection in Academic Vocabulary: An Evolutionary Analysis of Published Keywords. *PLOS ONE*, 3(8):1–7.

Bentley, R. Alexander, Alberto Acerbi, Paul Ormerod, and Vasileios Lampos. 2014. Books Average Previous Decade of Economic Misery. *PLoS ONE*, 9(1):e83147.

Blei, David M. and John D. Lafferty. 2006. Dynamic Topic Models. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120, Pittsburgh, Pennsylvania, USA. ACM.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Blythe, Richard A. 2012. Neutral Evolution: A Null Model for Language Dynamics. *Advances in complex systems*, 15(3–4).

Blythe, Richard A. and William Croft. 2012. S-Curves and the Mechanisms of Propagation in Language Change. *Language*, 88(2):269–304.

Bochkarev, V., V. Solovyev, and S. Wichmann. 2014. Universals versus Historical Contingencies in Lexical Evolution. *Journal of The Royal Society Interface*, 11(101).

Bochkarev, V.V., A.V. Shevlyakova, and V.D. Solovyev. 2015. The Average Word Length

Dynamics as an Indicator of Cultural Changes in Society. *Social Evolution and History*, 14(2):153–175.

Calude, Andreea S., Steven D. Miller, and Mark Pagel. 2017. Modelling Loanword Success a Sociolinguistic Quantitative Study of Māori Loanwords in New Zealand English. *Corpus Linguistics and Linguistic Theory*:1–38.

Carr, Jon W., Kenny Smith, Hannah Cornish, and Simon Kirby. 2017. The Cultural Evolution of Structured Languages in an Open-Ended, Continuous World. *Cognitive Science*, 41(4):892–923.

Caruana-Galizia, Paul. 2015. Politics and the German Language: Testing Orwell's Hypothesis Using the Google N-Gram Corpus. *Digital Scholarship in the Humanities*, 31(3):441–456.

Casler, Stephen D. 2015. Why Growth Rates? Which Growth Rate? Specification and Measurement Issues in Estimating Elasticity Values. *The American Economist*, 60(2): 142–161.

Castelló, Xavier, Lucía Loureiro-Porto, and Maxi San Miguel. 2013. Agent-Based Models of Language Competition. *International journal of the sociology of language*, 2013(221):21–51.

Chelsey, Paula and Harald R. Baayen. 2010. Predicting New Words from Newer Words: Lexical Borrowings in French. *Linguistics*, 48(6):1343–1374.

Church, Kenneth Ward and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational linguistics*, 16(1):22–29.

Crema, Enrico R., Anne Kandler, and Stephen Shennan. 2016. Revealing Patterns of Cultural Transmission from Frequency Data: Equilibrium and Non-Equilibrium Assumptions. *Scientific reports*, 6:39122 (2016).

Croft, W. 2000. *Explaining Language Change: An Evolutionary Approach*. Longman, editions.

Culbertson, Jennifer and Simon Kirby. 2016. Simplicity and Specificity in Language: Domain-General Biases Have Domain-Specific Effects. *Frontiers in Psychology*, 6: 1964.

Cuskley, Christine F., Martina Pugliese, Claudio Castellano, Francesca Colaiori, Vittorio Loreto, and Francesca Tria. 2014. Internal and External Dynamics in Language: Evidence from Verb Regularity in a Historical Corpus of English. *PLOS ONE*, 9(8):1–7.

Daoust, Demise. 2017. Language Planning and Language Reform. In *The Handbook of Sociolinguistics*, pages 436–452. Wiley-Blackwell, editions.

Dautriche, Isabelle, Kyle Mahowald, Edward Gibson, Anne Christophe, and Steven T. Piantadosi. 2017. Words Cluster Phonetically beyond Phonotactic Regularities. *Cognition*, 163:128–145.

Dautriche, Isabelle, Kyle Mahowald, Edward Gibson, and Steven T. Piantadosi. 2016. Wordform Similarity Increases With Semantic Similarity: An Analysis of 100 Languages. *Cognitive Science*, 41:2149–2169.

Davies, Mark. 2008. *The Corpus of Contemporary American English: 450 Million Words, 1990–2012*. Available Online at http://corpus.byu.edu/coca. editions.

Davies, Mark. 2010. *The Corpus of Historical American English* (COHA)*: 400 Million Words, 1810–2009*. Available Online at http://corpus.byu.edu/coha. editions.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American society for information science*, 41(6):391.

Dubossarsky, Haim, Yulia Tsvetkov, Chris Dyer, and Eitan Grossman. 2015. A Bottom up Approach to Category Mapping and Meaning Change. *NetWordS 2015 Word Knowledge and Word Usage*:66–70.

Dubossarsky, Haim, Daphna Weinshall, and Eitan Grossman. 2016. Verbs Change More than Nouns: A Bottom-up Computational Approach to Semantic Change. *Lingue e linguaggio*, 15(1):7–28.

Dubossarsky, Haim, Daphna Weinshall, and Eitan Grossman. 2017. Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1147–1156.

Enfield, N.J. 2014. Transmission Biases in the Cultural Evolution of Language: Towards an Explanatory Framework. In Dor, Daniel, Chris Knight, and Jerome Lewis, editors, *The Social Origins of Language*. Oxford University Press, Oxford, editions.

Ewens, W.J. 2004. *Mathematical Population Genetics 1: Theoretical Introduction*. Interdisciplinary Applied Mathematics. Springer New York, editions.

Feder, Alison F., Sergey Kryazhimskiy, and Joshua B. Plotkin. 2014. Identifying Signatures of Selection in Genetic Time Series. *Genetics*, 196(2):509–522.

Feltgen, Q., B. Fagard, and J.-P. Nadal. 2017. Frequency Patterns of Semantic Change: Corpus-Based Evidence of a near-Critical Dynamics in Language Change. *Open Science*, 4(11).

Fremann, Lea and Mirella Lapata. 2016. A Bayesian Model of Diachronic Meaning Change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Frimer, Jeremy A., Karl Aquino, Jochen E. Gebauer, Luke (Lei) Zhu, and Harrison Oakes. 2015. A Decline in Prosocial Language Helps Explain Public Disapproval of the US Congress. *Proceedings of the National Academy of Sciences*, 112(21):6591–6594.

Ghanbarnejad, Fakhteh, Martin Gerlach, José M. Miotto, and Eduardo G. Altmann. 2014. Extracting Information from S-Curves of Language Change. *Journal of The Royal Society Interface*, 11(101).

Gibson, Edward, Richard Futrell, Julian Jara-Ettinger, Kyle Mahowald, Leon Bergen, Sivalogeswaran Ratnasingam, Mitchell Gibson, Steven T. Piantadosi, and Bevil R. Conway. 2017. Color Naming across Languages Reflects Color Use. *Proceedings of the National Academy of Sciences*, 114 (40):10785–10790.

Gulordava, Kristina and Marco Baroni. 2011. A Distributional Similarity Approach to the

Detection of Semantic Change in the Google Books Ngram Corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71. Association for Computational Linguistics.

Hahn, Matthew W and R. Alexander Bentley. 2003. Drift as a Mechanism for Cultural Change: An Example from Baby Names. *Proceedings of the Royal Society of London B: Biological Sciences*, 270(Suppl 1):S120–S123.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016a. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7–12, 2016, Berlin, Germany, Volume 1: Long Papers*, pages 1489–1501.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016b. 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.

Hernández-Campoy, Juan Manuel and Juan Camilo Conde-Silvestre. 2012. *The Handbook of Historical Sociolinguistics*. John Wiley & Sons, editions.

Hinrichs, Lars, Benedikt Szmrecsanyi, and Axel Bohmann. 2015. Which-Hunting and the Standard English Relative Clause. *Language*, 91(4):806–836.

Jatowt, Adam and Kevin Duh. 2014. A Framework for Analyzing Semantic Change of Words across Time. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 229–238. IEEE Press.

Jespersen, Otto. 1922. *Language, Its Nature, Development, and Origin*. H. Holt, editions.

Jurafsky, D. and J.H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall Series in Artificial Intelligence. Pearson Prentice Hall, editions.

Kandler, Anne and Adam Powell. 2018. Generative Inference for Cultural Evolution. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 373(1743).

Kandler, Anne, Bryan Wilder, and Laura Fortunato. 2017. Inferring Individual-Level Processes from Population-Level Patterns in Cultural Evolution. *Royal Society Open Science*, 4(9).

Kanwal, Jasmeen, Kenny Smith, Jennifer Culbertson, and Simon Kirby. 2017. Zipf's Law of Abbreviation and the Principle of Least Effort: Language Users Optimise a Miniature Lexicon for Efficient Communication. *Cognition*, 165:45–52.

Karjus, Andres, Richard A. Blythe, Simon Kirby, and Kenny Smith. 2018a. Challenges in Detecting Evolutionary Forces in Language Change Using Diachronic Corpora. *ArXiv e-prints*, arXiv:1811.01275.

Karjus, Andres, Richard A. Blythe, Simon Kirby, and Kenny Smith. 2018b. Topical Advection as a Baseline Model for Corpus-Based Lexical Dynamics. *Proceedings of the Society for Computation in Linguistics*, 1:186–188.

Kauhanen, Henri. 2017. Neutral Change. *Journal of Linguistics*, 53(2):327–358.

Keller, Daniela Barbara and Jörg Schultz. 2013. Connectivity, Not Frequency, Determines the Fate of a Morpheme. *PLOS ONE*, 8(7):1–8.

Keller, Daniela Barbara and Jörg Schultz. 2014. Word Formation Is Aware of Morpheme Family Size. *PLOS ONE*, 9(4):1–6.

Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal Analysis of Language through Neural Language Models. *ACL 2014*:61.

Kimura, M. 1994. *Population Genetics, Molecular Evolution, and the Neutral Theory: Selected Papers*. Evolutionary Biology. University of Chicago Press, editions.

Kirby, Simon, Hannah Cornish, and Kenny Smith. 2008. Cumulative Cultural Evolution in the Laboratory: An Experimental Approach to the Origins of Structure in Human Language. *Proceedings of the National Academy of Sciences*, 105(31):10681–10686.

Koplenig, Alexander. 2017a. The Impact of Lacking Metadata for the Measurement of Cultural and Linguistic Change Using the Google Ngram Data Sets—Reconstructing the Composition of the German Corpus in Times of WWII. *Digital Scholarship in the Humanities*, 32(1):169–188.

Koplenig, Alexander. 2017b. Why the Quantitative Analysis of Diachronic Corpora That Does Not Consider the Temporal Aspect of Time-Series Can Lead to Wrong Conclusions. *Digital Scholarship in the Humanities*, 32(1):159–168.

Koplenig, Alexander and Carolin Müller-Spitzer. 2016. Population Size Predicts Lexical Diversity, but so Does the Mean Sea Level—Why It Is Important to Correctly Account for the Structure of Temporal Data. *PLoS ONE*, 11(3):e0150771.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically Significant Detection of Linguistic Change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Labov, W. 2011. *Principles of Linguistic Change, Volume 3: Cognitive and Cultural Factors*. Language in Society. Wiley, editions.

Lev-Ari, Shiri and Sharon Peperkamp. 2014. An Experimental Study of the Role of Social Factors in Language Change: The Case of Loanword Adaptations. *Laboratory Phonology*, 5(3):379–401.

Lieberman, Erez, Jean-Baptiste Michel, Joe Jackson, Tina Tang, and Martin A. Nowak. 2007. Quantifying the Evolutionary Dynamics of Language. *Nature*, 449(7163):713–716.

Lijffijt, Jefrey, Tanja Säily, and Terttu Nevalainen. 2012. CEECing the Baseline: Lexical Stability and Significant Change in a Historical Corpus. In Jukka Tyrkkö, Matti Kilpiö, Terttu Nevalainen, Matti Rissanen, editor, *Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources*, Studies in Variation, Contacts and Change in English. Research Unit for Variation, Contacts and Change in English (VARIENG), Helsinki, editions.

McMahon, A.M.S. 1994. *Understanding Language Change*. Cambridge University Press, editions.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014):176–182.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In Burges, C.J.C., L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., editions.

Newberry, Mitchell G., Christopher A. Ahern, Robin Clark, and Joshua B. Plotkin. 2017. Detecting Evolutionary Forces in Language Change. *Nature*, 551(7679):223–226.

Ohala, John J. 1983. The Origin of Sound Patterns in Vocal Tract Constraints. In *The Production of Speech*, pages 189–216. Springer, editions.

Pagel, Mark, Quentin D. Atkinson, and Andrew Meade. 2007. Frequency of Word-Use Predicts Rates of Lexical Evolution throughout Indo-European History. *Nature*, 449(7163):717–720.

Pechenick, Eitan Adam, Christopher M. Danforth, and Peter Sheridan Dodds. 2015. Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution. *PLoS ONE*, 10(10):e0137041.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Perek, Florent. 2016. Using Distributional Semantics to Study Syntactic Productivity in Diachrony: A Case Study. *Linguistics*, 54(1):149–188.

Petersen, Alexander M., Joel Tenenbaum, Shlomo Havlin, and H. Eugene Stanley. 2012. Statistical Laws Governing Fluctuations in Word Use from Word Birth to Word Death. *Scientific Reports*, 2:313 (2012).

Pierrehumbert, Janet B., Forrest Stonedahl, and Robert Daland. 2014. A Model of Grassroots Changes in Linguistic Systems. *ArXiv e-prints*, arXiv:1408.1985.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, editions.

Reali, F. and T.L. Griffiths. 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, Terry, Alexandra Carstensen, and Charles Kemp. 2016. Languages Support Efficient Communication about the Environment: Words for Snow Revisited. *PLOS ONE*, 11(4):1–17.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Edoardo M. Airoldi, and others. 2013. The Structural Topic Model and Applied Social Science. In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*.

Rodda, Martina Astrid, Marco S.G. Senaldi, and Alessandro Lenci. 2017. Panta Rei: Tracking Semantic Change with Distributional Semantics in Ancient Greek. *Italian Journal of Computational Linguistics*, 3:1:11–24.

Rosenfeld, Alex and Katrin Erk. 2018. Deep Neural Models of Semantic Shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 474–484.

Rubin, J., J. DasGupta, B.H. Jernudd, J.A. Fishman, and C.A. Ferguson. 1977. *Language Planning Processes*. Contributions to the Sociology of Language. Mouton, editions.

Sagi, Eyal, Stefan Kaufmann, and Brady Clark. 2011. Tracing Semantic Change with Latent Semantic Analysis. *Current methods in historical semantics*:161–183.

Samara, Anna, Kenny Smith, Helen Brown, and Elizabeth Wonnacott. 2017. Acquiring Variation in an Artificial Language: Children and Adults Are Sensitive to Socially Conditioned Linguistic Variation. *Cognitive Psychology*, 94:85–114.

Santus, Enrico, Emmanuele Chersoni, Alessandro Lenci, Chu-Ren Huang, and Philippe Blache. Testing APSyn against Vector Cosine on Similarity Estimation. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers*, pages 229–238, Seoul, South Korea.

Sapir, E. 1921. *Language. An Introduction to the Study of Speech*. Harcourt, Brace and Company, editions.

Schlechtweg, Dominik, Stefanie Eckmann, Enrico Santus, Sabine Schulte im Walde, and Daniel Hole. 2017. German in Flux: Detecting Metaphoric Change via Word Entropy. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 354–367.

Selivanov, Dmitriy and Qing Wang. 2018. *Text2vec: Modern Text Mining Framework for R*. editions.

Sindi, Suzanne S. and Rick Dale. 2016. Culturomics as a Data Playground for Tests of Selection: Mathematical Approaches to Detecting Selection in Word Use. *Journal of Theoretical Biology*, 405:140–149.

Smith, Kenny, Monica Tamariz, and Simon Kirby. 2013. Linguistic Structure Is an Evolutionary Trade-off between Simplicity and Expressivity. In Markus Knauff, Michael Pauen, Natalie Sebanz and Ipke Wachsmuth, editors, *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, pages 1348–1353. Cognitive Science Society, editions.

Stadler, Kevin, Richard A. Blythe, Kenny Smith, and Simon Kirby. 2016. Momentum in Language Change: A Model of Self-Actuating S-Shaped Curves. *Language Dynamics and Change*, 6(2):171–198.

Szmrecsanyi, Benedikt. 2016. About Text Frequencies in Historical Linguistics: Disentangling Environmental and Grammatical Change. *Corpus Linguistics and Linguistic Theory*, 12(1):153–171.

Szmrecsanyi, Benedikt, Anette Rosenbach, Joan Bresnan, and Christoph Wolk. 2014. Culturally Conditioned Language Change? A Multi-Variate Analysis of Genitive Constructions in ARCHER. In Hundt, M., editor, *Late Modern English Syntax*, Studies in English Language, pages 133–152. Cambridge University Press, editions.

Tamariz, M., T.M. Ellison, D.J. Barr, and N. Fay. 2014. Cultural Selection Drives the Evolution of Human Communication Systems. *Proceedings of the Royal Society B: Biological Sciences*, 281(1788):20140488.

Törnqvist, Leo, Pentti Vartia, and Yrjö O. Vartia. 1985. How Should Relative Changes Be Measured? *The American Statistician*, 39(1):43–46.

Wang, Xuerui and Andrew McCallum. 2006. Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 424–433. ACM.

Wetherell, Charles. 1986. The Log Percent (L%): An Absolute Measure of Relative Change. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 19(1):25–26.

Wichmann, Søren. 2008. The Emerging Field of Language Dynamics. *Language and Linguistics Compass*, 2(3):442–455.

Wijaya, Derry Tanti and Reyyan Yeniterzi. 2011. Understanding Semantic Change of Words over Centuries. In *Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural diversiTy on the Social Web*, pages 35–40. ACM.

Xu, Yang and Charles Kemp. 2015. A Computational Evaluation of Two Laws of Semantic Change. In Noelle, D.C., Dale, R., Warlaumont, A.S., Yoshimi, J., Matlock, T., Jennings, C.D. and Maglio, P.P., editors, *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, pages 2703–2708. Austin, TX: Cognitive Science Society.

Zipf, G.K. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press, editions.

A Technical appendix

A.1 *Notes on preprocessing and parameters*

We take a number of preprocessing steps to ensure a reasonable quality in the inference of the topic vectors that underlie the advection model. Both in the case of COHA and COCA, we exclude stop words (and also a list of known OCR errors) and use only content words (based on corpus POS tags). While COHA and COCA distinguish proper and common nouns in its tagging, we noticed quite a few proper nouns were tagged as common ones, hence we decided to remove all capitalized words (this is particularly relevant in the context of Section 4.4, where we needed to avoid detecting mistagged proper nouns as innovative common nouns). We also reduced variability in spelling by removing hyphens, and replaced all sequences of numbers within content words with a placeholder.

We used a context window of 10 words on both sides of the target word (after the removal of stop words, etc.), linearly weighted by distance, for inferring co-occurrence. The co-occurrence matrices were subsequently weighted, using the positive pointwise mutual information (PPMI) between each target word w and context word c:

$$\mathrm{PPMI}(w, c) := \max \left\{ log_2 \frac{P(w, c)}{P(w)P(c)}, 0 \right\} \tag{4}$$

This is essentially a weighting scheme that gives more weight to co-occurrence values of word pairs that occur together but not so much with other words, and less weight to pairs that co-occur with everything. Since we set a threshold of 100 occurrences per period for a word to be included, we circumvent the known small values bias of PPMI. Since we use positive PMI, all co-occurrence values end up as ≥ 0. See e.g. the textbook by Jurafsky and Martin (2009) for further details and examples.

For the advection model based on vectors drawn from a PPMI-weighted co-occurrence matrix, we use the top $m = 75$ context words as the topic (having observed that very small values lead to less reliable topics, while considerably larger values deteriorate the results in some cases). Importantly, the word counts (that underlie the log change values, which in turn make up the advection values) for each period were normalized to per million frequencies using the total word count in that period (periods corresponding to decades by default in COHA).

A.2 *Algorithmic description of the topical-cultural advection model*

1. Preprocessing steps

   1.1 (optional) Basic text cleaning (using a list of OCR errors, a list of stop and function word tags, words shorter than 3 characters), keep only content words; remove all capitalized words to avoid proper nouns

   1.2 (optional) Affix tags to words in the POS class of interest (e.g., nouns in our case; more tags and more specific tags improve disambiguation, but also increases sparsity)

   1.3 Split texts in the corpus files according to document delimiter tags (e.g., '##' in COHA) to avoid word co-occurrence windows crossing document boundaries

   1.4 Aggregate and store the preprocessed texts according to chosen periods (e.g., decades)

2. Calculate frequency change

   2.1 Count the frequencies of words in each period subcorpus and normalize the counts to obtain comparable (relative) values (subcorpora may be of different size)

   2.2 For each word ω, between each pair of successive time periods t, calculate the log frequency change value: $\text{logChange}(\omega; t) = \ln[f(\omega; t) + s] - \ln[f(\omega; t-1) + s]$ where $f(\omega; t)$ is the number of times word ω appears in the corpus during time period t. Note we use the $+s$ offset to avoid $\ln(0)$, and set the value of s to the equivalent the value corresponding to 1 occurrence after normalizing to per-million counts. s is set to 0 if $f(\omega) > 0$ or if both frequencies are 0.

3. (A) Topics and advection (if using the PPMI vectors based approach)

   3.1 Generate term co-occurrence matrices for each period (e.g., target words as rows and context words as columns), using a context window of some length (we used ±10, and linearly weighted context words by distance within the window)

      3.1.1 (optional) If targeting a specific POS class, filter the matrices by keeping only rows with the previously affixed tag

      3.1.2 (optional) Filter by setting a frequency threshold for a word to be included (we used a threshold of 100 raw occurrences per period or per concatenated dataset, if using smoothing)

   3.2 Apply positive pointwise mutual information (PPMI) weighting to each matrix

   3.3 Retrieve and store relevant context words for each target, in each period (i.e., sort each row of each matrix and store the top m context words, along with their PPMI weights in that row; we used $m = 75$)

3.4 (optional) to apply the “smoothing” operation, concatenate data from pairs of successive periods instead, and apply the previous 3 steps

3.5 For each target word ω, in each period t, calculate its advection value:

3.5.1 The advection values is a weighted mean over the log frequency change values in the set (of length m) of a target's context words N (i.e., the ‘topic’), with their PPMI values as the weights W;
$\text{advection}(\omega; t) := \text{weightedMean}(\{\text{logChange}(N_i; t) \mid i = 1, \ldots, m\},\ W)$, where $\text{weightedMean}(X, W) := \frac{\sum_i x_i w_i}{\sum_i w_i}$

3. (B) Topics and advection (if using the LDA topics based approach)

3.1 Train Latent Dirichlet Allocation (Blei et al., 2003) models for all period subcorpora (we used the following parameters: $\alpha = \beta = 0.1$, $k = 500$, maximum allowed iterations: 5000)

3.2 For each word ω in each period t, calculate its advection value:

3.2.1 Given the k topics, τ, identified by LDA, we determine the number of times $n(\omega, \tau)$ that each word ω appears in each of the topics τ. From this we can define the two conditional distributions $p(\omega|\tau) = n(\omega, \tau) / \sum_{\omega'} n(\omega', \tau)$ and $p(\tau|\omega) = n(\omega, \tau) / \sum_{\tau'} n(\omega, \tau')$. Given a word frequency change $\text{logChange}(\omega; t)$ at time t, its contribution to the change of the topic τ is $\text{logChange}(\omega; t) p(\tau|\omega)$.
To construct the advection of a target word ω, we need to determine the frequency changes of all topics that are coming from words other than ω, i.e., $\text{logTopicChange}(\tau; \omega, \mathrm{t}) = \sum_{\omega' \neq \omega} p(\omega'|\tau) \text{logChange}(\omega'; t) p(\tau|\omega') / [1 - p(\omega|\tau)]$.
Then, $\text{advection}(\omega; t) = \sum_{\tau} \text{logTopicChange}(\tau; \omega, \mathrm{t})\ p(\tau|\omega)$. The last part is thus analogous to point 3.5.1, the change in topic frequency being operationalized as a weighted mean of the changes in word frequencies, with weight from the distribution of words over topics.

4. (optional) Measure the descriptive power of the advection model by correlating the advection value of each word in each period to its respective log frequency change value.

A.3 *Additional remarks on the model and data processing*

A.3.1 For our purposes, logarithmic change is more useful than percentage change

We opt to quantify the changes in word counts between different time period subcorpora, using the measure of logarithmic difference—thus referring to it simply as 'log change' (cf. also Altmann et al., 2011; Petersen et al., 2012). Logarithmic difference between values V_1 and V_2 is defined as $\ln(V_2) - \ln(V_1) = \ln(V_2/V_1)$. This is sometimes also referred to as log percent or L% when the result is multiplied by 100 (Törnqvist et al., 1985; Wetherell, 1986), logarithmic growth rate (Casler, 2015), log points, nepers (centinepers in the case of multiplication with 100), decibels (when using log_{10}), or logarithmic growth rates. Measuring change on a logarithmic scale has three useful related advantages over the often used percentage change, defined as $(V_2 - V_1)/V_1 \cdot 100$. These are symmetry, additivity, and the lack of extreme positive outliers.

The absolute value of log change between two counts is the same regardless of which is used as the reference point. Given a series of log changes, the final (log) frequency is equal to the sum of the initial (log) frequency and the series of log changes. Percentage change is by definition bounded at –100% on the negative end, while increases starting at small values yield very large positive numbers.

Log change has the disadvantage that any 0-counts must be smoothed to avoid negative infinity resulting from $\ln(0)$, while for percent change, smoothing is strictly necessary only for increases from 0 to non-0 (to avoid division by 0), as a decrease from non-0 to 0 is always –100% (regardless of the actual difference between the two values, which in itself may be seen as another disadvantage, depending on the use case). Simple +1 smoothing could be used to avoid this problem by incrementing all frequencies by 1. This leads to some bias when dealing with relatively small values (particularly after normalizing to per million words). We use a slightly more elaborate version where we only change any 0 values involved in frequency change calculations to the value that corresponds to 1 occurrence in the per-million normalized frequency counts, and leave all > 0 values untouched.

Log frequencies are also better suited than raw frequencies (and absolute change) when dealing with word frequencies, smoothing the influence of the small number of extremely frequent words at the top end of the typically Zipfian distribution. We also tested the advection model using absolute frequency changes. Correlating absolute change based advection values with absolute frequency changes yields a practically zero correlation value. When using absolute frequencies for the advection calculation, but correlating these with log frequency changes, the correlations tend to come out as either the same or lower

TABLE 2 Time series decomposition using topical advection on the example of the word *payment*, corresponding to Fig. 2

	1900s	1910s	1920s	1930s	1940s
(a) pmw frequency	69.2	71.2	151.5	226.3	118.3
(b) log freq	4.25	4.28	5.03	5.43	4.78
(c) log change		+0.03	+0.75	+0.4	−0.64
(d) advection		−0.06	+0.45	+0.3	−0.42
(x) adjusted log change		+0.09	+0.3	+0.1	−0.23
(y) reformed series	69.19	75.53	102.08	112.99	90.15

Frequencies (a) are per million words. Log frequency and log change (b, c) refer to natural logarithms. The advection values (d) are based on the PPMI model with corpus topic smoothing. All values are rounded to save space and are therefore not precise. The increases in frequency of *payment* in the 1920s and 1930s, as well as the decrease in the 1940s (cf. row c) coincide with the changes in the averaged frequency of the topic words of *payment*, i.e., topical advection (d). The adjusted log change values (x) reflect the estimated frequency changes of *payment* when topical fluctuations are accounted for.

compared to using log change everywhere (as we do in this paper). In summary, there is little reason to not use log change to measure change. Table 1 illustrates the differences of logarithmic and percentage measures of change in frequencies between two time periods, t_1 and t_2.

A.3.2 Additional remarks on using advection for time series adjustment

Table 2 illustrates the word frequency time series adjustment operation based the topical advection measure, described in Section 4.3. The alphabetic abbreviations in the following equations refer to the rows in Table 2. The decomposition-like adjustment is additive: the adjusted log change values $x = c - d$. The frequency series can be reformed as the exponential of the cumulative sum of the adjusted values, initiated with the log frequency at period 1, a_1:

$$y_i = e^{a_1 + \sum_{j=1}^{j=i} x_j}$$

This could be useful for visualization purposes, as on Fig. 2, but of course the actual values in the reformed series depend on the (arbitrary) initialization value. The values in the resulting reformed (exponentiated) series will never be negative, but may be very small, if topical advection for a given word at a given time point is considerably higher than its frequency change (we observe this to be rarely the case).

TABLE 1 Fictional word counts and the resulting change values using different measures. Note the asymmetry in percentage change values when the counts are flipped. Natural logarithms are rounded to save space.

t_1	1	5	50	1	10	10	10	100	100	100
t_2	10	10	100	100	100	1	5	50	1	10
abs. change	9	5	50	99	90	−9	−5	−50	−99	−90
% change	900%	100%	100%	9900%	900%	−90%	−50%	−50%	−99%	−90%
ln change	2.3	0.69	0.69	4.61	2.3	−2.3	−0.69	−0.69	−4.61	−2.3
$\log_{10}$ change	1	0.3	0.3	2	1	−1	−0.3	−0.3	−2	−1

A.3.3 Time series adjustment does not hide genuine competition

This section further supplement Section 4.3, detailing the artificial corpus construction. The artificial series were inspected to see if the adjustment operation might possibly hinder the detection of actual competition between linguistic elements.

We selected four test nouns of various frequencies that each: occur frequently enough in the corpus during the past century to evaluate their topics; exhibit relative stability across the 11 time periods (1900s–2000s) in terms of their occurrence frequency, as well as meaning (based on the APSyn measure (cf. Section 3.2) on their context word vectors); and have small (absolute) advection values. The words *roof* (frequency at period 1: 163 per million words), *reason* (724), *town* (748), and *face* (1938) satisfied these criteria.

We then generated artificial competing synonyms by replacing a linearly-increasing proportion of the occurrences of each of the four target words with an invented "synonym" (*word′*) in the corpus. We also experimented with an S-shaped increase curve (arguably more characteristic of language change, cf. Blythe and Croft, 2012), which did not change the results. For example, at period 1, the invented synonym *town′* appears nowhere in the manipulated corpus, while in period 2, 10% of the occurrences of *town* are replaced with *town′* in the manipulated corpus, 20% in period 2 and so on up to 100% in period 11. Importantly, the replacement positions in the corpus were sampled at random, in order to simulate a scenario where the two synonyms are used freely (i.e., without regard for any contextual factors like style or genre).

On applying the advection correction to each of the original words and their synonyms, we find their frequency change points are only shifted slightly from their known values. When looking at the advection-adjusted fraction of occurrences of a word or its invented synonym (i.e., relative frequencies), the shifts due to the advection adjustment are barely noticeable. In other words, we find that advection-based adjustment does not seem to obscure genuine (although in this case artificial) cases of selection.

A.4 *Details on correlating advection model power and genre divergence*

As mentioned in Section 4.1, we found that the advection measure correlates positively with divergences between genre distributions in COHA. Data in the decade subcorpora in COHA is subsequently divided into four genres (fiction, magazines, news, non-fiction). We measured the genre distribution of each decade by counting the total number of words in each genre. Genre distributions of successive decades were compared using Kullback-Leibler divergence (to avoid zeros in the calculation, we incremented zero word counts by 1, in the early decades lacking the "news" genre). A value of 0 would indicate an identical distribution. The distribution of the aforementioned genres in the 1950s subcorpus is 50%, 24%, 14% and 12%. The difference to the 1940s is less than 1 percentage point in each genre, yielding a divergence of 0.00002. The largest observed divergence value is 0.13, between 1810s and 1820s, where "magazines" and "non-fiction" both differ by about 16 percentage points.

We find that (the log of) these divergence values correlates positively with the coefficients of determination from the advection model (i.e., the models where advection values are correlated with the word frequency change values). The R^2 values from correlating the divergence values to the R^2 values from the PPMI-based model without and with smoothing, and the LDA-based ones, without and with smoothing, in that order, are: 0.17, 0.41, 0.05, and 0.26. This indicates that the advection model is picking up on the changes between genre sample sizes, but also that discrepancies in genre sampling are likely not the only thing driving the observed changes in COHA over time. Figure 5 visualizes these results.

A.5 *Choice of corpora and methods, and their limitations*

We used fairly large corpora—COHA and COCA—for our analyses, both of which have been described as relatively representative and well balanced in terms of genre. We excluded the first decades of COHA in some cases, due to their smaller size and less balanced nature. Notably, the "news" genre is entirely missing in the first five decades. Mileage of utilizing the advection model with smaller corpora would probably vary, and is of course open for experimentation in terms of the parameters, thresholds and possibly the topical-semantic smoothing as described above. It is not impossible that superior results could be potentially achieved using larger and better balanced corpora and more sophisticated methods of topic modeling with carefully optimized

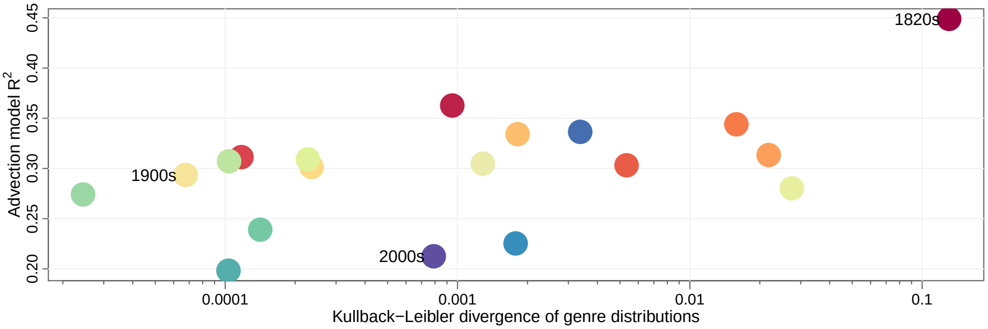

FIGURE 5 Divergence of genre distributions and the descriptive power of the advection measure (in the PPMI-based model, with smoothing). Each dot stands for one decade pair comparison, e.g. the dark purple dot marks the comparison of the 2000s to the preceding 1990s. The colors correspond to the colors in Fig. 1. Note the log scale on the horizontal axis. Decade pairs where the advection model describes more variance in noun frequency changes tend to be the ones with higher divergence in genre distributions.

parameterizations (for example, our exploration of the LDA parameter space was admittedly fairly limited).

A.5.1 Variations in operationalizing the test corpora

The results in Section 4.1 were based on comparing frequency changes between decade-length bins of the COHA. We also experimented with different temporal distances to see if the model behaves considerably differently. We found that with increased distance between the target decade and future decades, the values do improve in the case of some decade subcorpora, but not all, presumably depending on how much the subcorpora differ in terms of their underlying topic distribution. For example, the advection model describes more variance between mid-20th century decades and the 2000s compared to their immediate successors, while the 1810s subcorpus, clearly divergent in its distribution of genres and topics, shows relatively high correlations with all other subcorpora.

We also experimented with applying the advection model to a shuffled corpus to test if there the observed correlation between word frequency changes and topical advection (cf. Section 4.1) could be the result of some overlooked artifact of the model. We used the last decade subcorpus of COHA, but randomized the position of every word in the corpus, and calculated the topical advection value for all the target words, i.e. the weighted mean log context change (PPMI based, without smoothing), but using the randomized contexts. This resulted in $R^2 < 0.001$, $p = 0.4$, indicating that the topical advection measure—if calculated based on natural language use and not on random sequences of words—does yield meaningful information about the frequency change in the topic of a word.

A.5.2 Semantics, semantic change, and corpus smoothing

We re-evaluated the topics of words for every period to accommodate for natural semantic change. In principle this may not be necessary, if the meaning of a word is known to be very stable across time. In this case, the context vector from a single period, or aggregated across periods, could be used. The latter would also remedy the inherent problem of inferring context vectors for low-frequency words.

We note that the advection model should not be affected by the recent critique of distributed semantics by Dubossarsky et al. (2017), who show that semantic change measures based on vector spaces tend to be biased by differences in frequency. In particular, they call into question the entire enterprise of automatically measuring meaning change, attempting to replicate previous studies (Dubossarsky et al., 2015; Hamilton et al., 2016a) and finding that the proposed results either do not hold up or have drastically diminished descriptive power in comparisons against randomized baselines—attributing them to problems in vector space construction methods as well as bias from word frequency.

The same context word vectors we use to determine topic could indeed easily also be used to determine semantic change, by comparing the lists of top context words (cf. Fig. 4) between periods either by directly using the APSyn measure (cf. Section 3.2), or comparing the entire (suitably aligned) PPMI context vectors using vector cosine (in case of the former, care should be taken not to include 0-weight words in the topics, since APSyn only considers the rankings of context words in the vector, not their weights).

However, advection (topic frequency change) is meant to be re-evaluated for each corpus period. As such, semantic change is not directly a concern. We did also demonstrate additional results using what we called “smoothing” (Section 4), or concatenating the data from the target period t and the preceding period $t-1$ for the purpose of inferring topic vectors. In our experiments, this improved the power of advection to predict frequency change. In principle, smoothing could be applied using any number of $t \pm n$ periods; we also experimented with concatenating the entire corpus, and found that the descriptive power of the advection model suffered considerably. We assume semantic change to be the reason, since the context words (using which the advection measure is calculated) relevant to a target in one period may be quite irrelevant from another period, if the use (meaning) of the target differs—leading to uninformative topics.

Notably, the advection model is not expected to work as well with highly polysemous or general words (and homonyms), as it would with words with a more specific meaning (unless the meanings are somehow disambiguated and sense-tagged). The same goes for phrases and multi-word units, which we do not attempt to detect or parse in this contribution. Polysemy and multi-word units, however, are widespread problems across most NLP tasks, not only the one at hand.

A.6 *Notes on implementation, code and data*

The models and calculations presented in this paper were implemented using R 3.5.0 (R Core Team, 2018), and making use of the text2vec package (Selivanov and Wang, 2018). The code and data are available at https://github.com/andreskarjus/topical_cultural_advection_model. The corpora used here can be found at https://corpus.byu.edu.