Competition, selection and communicative need in language change:

an investigation using corpora, computational modelling and experimentation

Andres Karjus

Doctor of Philosophy
School of Philosophy, Psychology & Language Sciences
University of Edinburgh
2020

Declaration

I declare that this thesis was composed by myself and that the work presented here is my own, except where explicitly stated otherwise. This work has not been submitted for any other degree or professional qualification. My and my co-authors contributions to the co-authored papers are explicitly stated in each Chapter which includes a paper. The reproduction of the papers herein is compatible with the open-access publication policies of the respective journals.

Paper 1. Andres Karjus, Richard A. Blythe, Simon Kirby, and Kenny Smith (2020a). "Challenges in Detecting Evolutionary Forces in Language Change Using Diachronic Corpora". *Glossa: a journal of general linguistics* 5.1, p. 45. DOI: 10.5334/gjgl.909

Paper 2. Andres Karjus, Richard A. Blythe, Simon Kirby, and Kenny Smith (2020c). "Quantifying the Dynamics of Topical Fluctuations in Language". *Language Dynamics and Change* 10.1, pp. 86–125. DOI: 10.1163/22105832-01001200

Paper 3. Andres Karjus, Richard A. Blythe, Simon Kirby, and Kenny Smith (2020b). "Communicative Need Modulates Competition in Language Change". *arXiv preprint* arXiv:2006.09277

Andres Karjus

11th of August 2020

Abstract

Constant change is one of the few truly universal cross-linguistic properties of living languages. In this thesis I focus on lexical change, and ask why the introduction and spread of some words leads to competition and eventual extinction of words with similar functions, while in other cases semantically similar words are able to companionably co-exist for decades.

I start out by using extensive computational simulations to evaluate a recently published method for differentiating selection and drift in language change. While I conclude this particular method still requires improvement to be reliably applicable to historical corpus data, my findings suggest that the approach in general, when properly evaluated, could have considerable future potential for better understanding the interplay of drift, selection and therefore competition in language change.

In a series of corpus studies, I argue that the communicative needs of speakers play a significant role in how languages change, as they continue to be moulded to meet the needs of linguistic communities. I developed and evaluated computational methods for inferring a number of linguistic processes — changes in communicative need, competition between lexical items, and changes in colexification — directly from diachronic corpus data. Applying these new methods to massive historical corpora of multiple languages spanning several centuries, I show that communicative need modulates the outcome of competition between lexical items, and the colexification of concepts in semantic subspaces.

I also conducted an experiment in the form of a dyadic artificial language communication game, the results of which demonstrate how speakers adapt their lexicons to the communicative needs of the situation. This combination of methods allows me to link actions of individual speakers at short timescales to population-level findings in large corpora at historical timescales, in order to show that language change is driven by communicative need.

Non-technical summary

All living languages change over time. I am interested in how people use words, and how their usage evolves. Sometimes when speakers adopt a new word or phrase — a borrowing like *faux pas* or a coinage such as *megxit* — it may end up replacing some older expression, while other times the new element in the lexicon does not cause another one to disappear. I want to figure out if that is something random or something predictable. Most of this thesis focuses on testing one particular hypothesis, that it is the collective communicative needs of the speakers of a language that determine the outcomes of such “competition" between words, and the shape of a language in general. To put it another way, I suspect languages keep changing, because the world keeps changing, and each generation adapts their language(s) to describe that world around them in the most effective and informative way.

I use a combination of methods to probe this idea. One of them involves very large collections of texts (also known as “corpora") from multiple languages, spanning tens or even hundreds of years, or in the case of my Twitter corpus, only a year but covering millions of tweets. Since this is too much data to work through by hand, I make use of artificial intelligence that learns the meanings of words from context and digs up examples of competition. I also use computational simulations, by shuffling real corpora in a controlled manner, and also by creating miniature populations of words that then compete against one another. Finally, I invited some people to a (virtual) lab and asked them to learn small artificial languages and use them communicate with each other. The corpus-based approaches give an idea of what happens in populations of speakers over time; experiments allow me to probe how individual speakers react and adapt when placed in different communicative situations.

This combined approach yields converging evidence supporting the idea that adaptive communicative needs of speakers play an important role in language evolution. Living languages are perpetually shaped and moulded by their speakers in a way that makes sure they remain relevant and effective tools of communication in an ever-changing world. In other words, languages change, because it is useful for them to change.

Acknowledgements

You need what I like to call "some evolutionary bollocks" at the start, and foreground your main contribution.
— Anonymous Elephant, 12.09.2017

As I took a seat, I noticed a big plastic bag of leafy salad, sitting in an upside down cycling helmet on the table. A bike smelling of earth and oil in the corner, an artist's depiction of the tree of language families hanging on the wall. Through the half-open window, the grey Scottish autumn sky, and the faint chatter of excited freshers crowding the street below. It had been a few weeks since I had arrived in Edinburgh. Now, with the reality of doing a PhD in linguistics having sank in, I found myself face to face with a bit of an existential crisis probably not that uncommon in the humanities: why are we even doing what we are doing? How do you motivate yourself to do what you do? And I don't actually remember exactly how Kenny answered this question, but it looks like it worked (it was something along the lines of coming across an unresolved problem or question, and just having a go at solving it, which is pretty much what I've been doing since).

This thesis would not exist without Kenny Smith, Richard Blythe, and Simon Kirby. I've been very lucky to have had the chance to work with these brilliant people, but perhaps equally importantly, to have supervisors who are, well, just normal sane people, who have interests and lives also outside of academia, and sometimes just say,

It turns out tea and cake with the wife trumps your PhD. Sorry! I will be in next Tuesday so we can catch up then.

Kenny is someone incredibly sharp when it comes to figuring stuff out; and the number of little edits and comments he has left on my drafts is probably in the hundreds if not thousands. I am grateful for all of them, even the ones telling me that I, in fact, cannot have every single plot looking like a solar system. Richard has an outsider's perspective on linguistics, and also knows immeasurably more math than I ever will. Both of these things, along with his unlimited optimism, have been very helpful in tackling the topics in this thesis. Also, getting to say "linguistics, but one of my supervisors is a physicist" is handy for livening up the otherwise formulaic so-what-do-you-study conversations when meeting new people.

This is the "supervisorial exclusive we": we totally have to try doing this… and by "we", we mean "you".

Simon has the uncanny ability to keep generating great ideas seemingly out of thin air, be it in art or science. The trick, as advised by former PhD student of the CLE, is to keep writing them all down, but to make sure one does not commit to implementing all of them at once, if one is aiming to finish a PhD any time soon. Finally, special thanks go to Jenny Culbertson, who is not my supervisor, but has nevertheless been very helpful by asking me important questions like "Whattt?!" and "Why??!" at (almost every) group supervision meeting.

There are a number of discussion which have taken place either right here in Edinburgh, at conferences, or over email, which have ended up improving some aspect of this thesis. I'd like to thank the following people for that: Mitchell Newberry, Yang Xu, Barbara C. Malt, Mahesh Srinivasan, Steven Piantadosi, Gary Lupyan, Haim Dubossarsky, Dominik Schlechtweg, William Hamilton, Enrico Santus, Jack Grieve, Anne Kandler, Alison Feder, Natalia Levshina, Janet Pierrehumbert, Christine Cuskley, Isabelle Dautriche, Jasmeen Kanwal, Matt Spike, Stella Frank, Monica Tamariz, Rob Truswell, Jonas Nölle, Fausto Carcassi, Peeter Tinits, Anna Jon-And, Robert Hawkins, Seán Roberts, Mark Atkinson, Thomas Müller, James Winters, Bodo Winter, and Bill Thompson. I also had the opportunity to co-supervise two MSc students, Charlotte Brimer and Tianyu Wang, and discussions with them on their topics led to a couple of crucial improvements in my own thesis. Since this is the post-viva copy, I would also like to take the opportunity to thank my examiners, Yang Xu and Graeme Trousdale, for the interesting and stimulating discussion that was the thesis defense.

Undoubtedly, these four years would have been very different without all the garbage fire hotdogs, mixed veg at the Kitchen, Dagda Fridays, the Firbush and CLE PhD retreats, and just random chats and coffee breaks in Teviot or the DSB 7th floor common room over leftover catering sandwiches. In other words, I feel incredibly lucky to have landed in a research group, a PhD cohort, as well as a postgrad office all filled with nice people. So hey, thanks for having been there Jonas, Fausto, Tom, José, Svenja, Tamar, Fiona, Ash, Cathleen, Marieke, Carmen, Eva-Maria, Jade, Candice, Wenjia, Lisa, Alex, Louis, Jon, John, Vini, Marc, Fang, Annie, Henry, and everybody else who lived in DSB1.15 at some point in the past four years.

My time as a PhD student at Edinburgh Uni would also have been very different if it wasn't for the highland trips with the EUHWC, dance classes and Fiestas at the EUSS, the salsa nights at Boteco, Monday swing at the EUSDS, or Edinbop's book shop lindy hop events. Special thanks go to the wonderful dance teachers Jamie and Katie, Orlando and Julia, and the tireless committees of these societies who continue making great things happen.

During my PhD, I've also been organizing occasional R workshops on data visualization at various venues, from stand-alone events to conference sideshows and summer schools. Thanks to everybody who invited me, including PPLS Writing Centre coordinator James Donaldson for getting me started by offering me to do an hour-long R taster for the Centre in 2017. Also, speaking of R, a considerable portion of the code that powers this research has been built on Stack Overflow Q&As; so thank you, anonymous mad internet people.

It must be said that I would not even have been able to come to Edinburgh, if it wasn't for the Archimedes Foundation of Estonia, which granted me the generous postgraduate scholarship to do so, and the faculty members at the University of Tartu who supported my application at the time: Martin Ehala, Ann Veismann, Renate Pajusalu, and Jaak Vilo. Cheers! Also, I'm grateful to both the School of Philosophy, Psychology & Language Sciences, and the Go Abroad Fund of the University of Edinburgh for providing me with conference travel support during my PhD.

Finally, I remain thankful to my now ex-partner Liisi for having been there for me throughout most of my PhD, and a friend after acquiring the prefix. And I'm thankful to my parents and brothers for always having been supportive of my choices, in particular this weird academic journey of mine, by now spanning a decade and a half and nearly a dozen countries.

Anonymous Elephant: *I was on the train - going to have dinner and put*
kids to bed then I will come back to this!
PS this is kenny, not an elephant

Contents

Chapter 1

Introduction

All living languages change over time. This is one of the few true cross-linguistic universals. Most change starts out small: some speakers switch to using an alternative lexeme or morpheme, start pronouncing them differently, or saying them in a different order. When more and more such elements get changed in a given variety of a language, it becomes less and less intelligible to its related varieties, and at some point will be regarded as separate language (a status change proverbially hastened by the procurement of an army and a navy), while texts written in the old variety become more and more alien to new readers as time passes and changes compound.

In some sense this is very strange: it would be easier to communicate with older generations, both directly with the alive ones, and indirectly with the earlier ones through written texts, if language just stayed the same. It would also be much easier to communicate, trade and coexist with our neighbours if only our dialects had not diverged a few centuries earlier and become unintelligible to each other. Yet this is not how human language works.

Although some change is likely incidental (known as linguistic drift or neutral change; see Chapter 2), and some may be a result of top-down language planning in some communities, this thesis is driven by the idea that most change is, in one way or another, the result of adaption in the evolutionary sense (cf. Croft 2000; Winters et al. 2015). I argue that non-neutral change is brought on by generations of speakers who are essentially performing maintenance on their languages, making sure the languages continue to serve their communities as optimal tools of communication. I will go on to test hypotheses driven by this idea using both historical corpus data as well as experimentation using artificial languages.

One way to view language is as the outcome of a constant struggle between pressures that can be broadly divided in two. On the one hand, speakers need their communication system to be simple, easily learnable and in general efficient in terms of how much effort it takes to remember, use and understand (cf. Zipf 1949; Christiansen and Chater 2008; Kanwal et al. 2017; Gibson et al. 2019). On the other hand, they also need it to be complex, varied, expressive and informative enough so that it can serve the purposes it exists for (Labov 1982; Kirby et al. 2015; Kemp et al. 2018; Carr et al. 2020). These are all more general, higher-order global needs required to be satisfied for successful communication, for any language to serve its purpose and survive. In this thesis however, I will be focusing on communicative needs in the more specific, situational, 'lower-order' sense. These relate to more local, specific elements like subsystems of grammar or syntax, or in the case of most of my research, semantic subfields or topics. These of course all differ between cultures and languages, as what people need to express in their daily discourse varies in time and space.

The reasons are many for changes in communicative need in this sense. A new entity may appear in the environment of the language community and require referring to — so a word is coined, derived or borrowed — or lost, if the entity is no longer of relevance, like tools superseded by improved technology. As the socio-cultural landscape of a community changes, some concepts may need distinguishing in finer detail and require more words to do so (see Chapter 4), or some constructions may need streamlining if used more frequently than before. In short, languages are continually being adapted to the social, political, cultural and natural environments of their communities in order to stay relevant and not be replaced by another one (cf. Boas 1911; Sapir 1921; Martinet 1952; Coulmas 1989; Lupyan and Dale 2010; Christensen et al. 2016). Language also adapts to itself, the complex system that it is (Beckner et al. 2009) — when one component changes, others may need adjustment for the entire system to remain optimal for communication as a whole.

Not all communicative needs are strictly about the content of an utterance: a language is also a conveyor of social and cultural identity (see Joseph 2004), and speakers need their utterances not only to inform but also to have the intended illocutionary force (see Searle 1969). Such metalinguistic needs can be expected to play a role in the selection for elements borrowed from more prestigious varieties, the discarding of those associated with undesirable ones, or speakers changing their pronunciation to sound more like (or unlike) somebody else. Therefore I view all such cases too as adaptions to the various communicative needs of the time and place where a language is being used.

Communicative need in the local, lower-order sense can act as catalyst for the higher-order pressures. For example, if some pair of similar concepts requires distinguishing from one another (e.g. for cultural reasons), like two shades of colour or kinship relations, then the pressure for simplicity can be relaxed in favour of expressivity, lexifying those with individual words. If it is not particularly relevant for the language community to regularly distinguish between blue and a slightly lighter shade of blue, or the maternal and paternal grandfather, then expressivity gives way to simplicity and efficiency, and these examples will be just termed *blue* and *grandfather* (cf. Kemp et al. 2018). As socio-cultural motivations change, language eventually follows.

While many pathways of change in language are to some extent regular and predictable (cf. Traugott and Dasher 2001; Bybee 2002; van Gelderen 2009), many take time on the order of centuries (cf. McMahon 1994; Strang 2015). Some words, usually the most frequent ones, can also remain stable in meaning and form for centuries (if not millennia; Pagel et al. 2013). Yet new words are being continuously coined and borrowed — and though some may be short-lived, others can compete with and replace old variants within the span of decades (see Chapter 4). This faster turnover rate makes words easier to study using diachronic corpora than for example syntactic or morphological phenomena. Additionally, the meanings or functions of words are by far the easiest to model using the automated distributional approaches I employ here to study vast amounts of data that would be inconceivable to handle by hand.

This is largely the approach taken in this thesis: I study changes in word frequencies in historical text corpora (see Chapter 3), inferring their changing meanings and contexts directly from textual data using language-agnostic machine learning tools (Chapters 3, 4). This means all the methods developed and described in this thesis are readily applicable to languages other than those few considered here, and with some adjustment to domains of language beyond that of the lexicon. I also employ computational simulations to test and validate a number of methodologies where needed (notably Chapter 2), and a communication experiment (Chapter 5) to investigate the effects of communicative need on the individual level (as opposed to the population-level aggregates that are corpora).[1]

1.1 Situating the research

In terms of scientific fields or disciplines, this thesis could be viewed as being situated on the intersection of a few. The primary ones would be flavours of linguistics, with various prefixes. Computational, in the sense that most of this research relies heavily on big corpus data, statistics and machine learning tools. Historical linguistics, in the sense that I am looking at diachrony, change over time; and evolutionary, in the sense that I focus on the dynamics and cognitive pressures driving the change. From another point of view, this thesis deals primarily with words and changes in their meaning and usage, and is therefore part of lexicology and semantics. But then Chapter 2 interfaces directly with population genetics, Chapter 3 builds on a metaphor based on fluid dynamics, and Chapter 5 makes use of an experimental paradigm common in experimental psychology and psycholinguistics, but also in experimental semiotics.

I hope that the methods, or methods based on the ones developed here, will also be eventually applied, and the findings discussed, in a number of fields. The advection and competition models (Chapters 3 and 4), but also the discussion on time series binning (Chapter 2), have potential to be useful in (computationally-oriented) historical and evolutionary linguistics, and as discussed above, potentially also in domains of language other than lexicology, and perhaps also beyond the study of language.

1.2 Thesis roadmap

This thesis is structured as follows. In Chapter 2, I evaluate a recently proposed statistical method for differentiating linguistic selection and drift (Newberry et al. 2017). This is an important component of understanding language change: if a change can reliably be attributed to drift, then it would be superfluous to look for further causes like competition between ele-

[1] All the analyses and models are implemented in R (various versions; R Core Team 2016-2020), making heavy use of the packages text2vec (Selivanov and Wang 2018), Matrix (Bates and Maechler 2018), stringdist (van der Loo 2014) for NLP tasks; lme4 (Bates et al. 2015), mgcv (Wood 2011) for statistics, and the tidyverse (Wickham et al. 2019) for data manipulation and visualisation. Almost all the code to run and replicate the analyses has been made publicly available (links in the relevant Chapters).

ments, cognitive pressures or sociolinguistic reasons. I employ extensive computational simulations to mimic language change scenarios to test the applicability of the method, but conclude that it still requires improvement to be reliably applicable to historical corpus data, or at least the kinds of questions I aim to investigate in this thesis. However, my findings suggest that the approach in general, when properly evaluated, could have considerable future potential for better understanding the interplay of drift, selection and therefore competition in language change. The simulations and probing of corpus data in this study also informs many of the technical decisions later in this thesis.

Chapter 2, as well as Chapters 3 and 4 consist of papers in various stages of the publishing process. I have retained their original layouts, meaning they will display their own page numbers (in the headers), while the page numbering of the thesis continues consistently in the footer.

In Chapter 3, I propose a new method for quantifying topical fluctuations in diachronic corpora, called the topical advection model. I show that it can be used as a baseline predictor for frequency changes of lexical items over time, which roughly follow the prevalence of the topics of conversation they are used in. I argue that the same approach can be used as a proxy to changes in communicative needs in a population over time, and show that positive topical fluctuations correlate with the introduction of new words to the lexicon. This methodology forms the basis on which the next two corpus studies build on.

Chapter 4 focuses on lexical competition, with the following premise. When a variant is selected for by it speakers, it can either end up replacing its alternative(s), or enrich its local semantic space without causing others to be discarded. I propose that communicative need may explain some of this variability. To test this, I develop a general method for quantifying competition between linguistic elements in diachronic corpora, and use the advection model from Chapter 4 to operationalize communicative need. I demonstrate using a variety of corpora from different languages that near-synonymous words are more likely to directly compete if they belong to a topic of conversation whose importance to language users is constant over time, possibly leading to the extinction of one of the competing words. By contrast, in topics which are increasing in importance, near-synonymous words can coexist without competing. I therefore argue that in addition to direct competition between words, lexical change can be driven by competition between topics or semantic subspaces, something which should be further studied in future research.

Chapter 5 introduces an artificial language experimental paradigm to test the effect of communicative need on individual lexification choices in discourse, building on and also replicating the predictions of a recently published typological study demonstrating a tendency for languages to colexify similar concepts (Xu et al. 2020). Most diachronic corpora are essentially (relatively small and often edited) population-level aggregate samples of utterances from various speakers and sources. It is not trivial to infer individual-level decision processes from such data (see Chapter 2). Experiments provide a way to probe biases are pressures in speaker behaviour in a controlled manner, but linking these to the larger changes observable over longer time scales is not trivial either. I attempt to do that by conducting a historical corpus study — using a technical setup similar to that of Chapter 4, supported by a corpus-based measure of colexification that I developed — and show that diachronic changes in colexification correlate with changes in communicative need, echoing the results of the experimental study.

The final chapter summarises the thesis and lays out pathways for future research on these topics. In short, in this thesis I aim to make three contributions to language sciences: I introduce a number of novel computational methods to study language change in corpora and discuss how to evaluate such methods; show that communicative need plays a significant role in language change over time; and demonstrate experimentally how variable communicative needs lead to different speaker lexification behaviours.

Chapter 2

Challenges in detecting evolutionary forces in language change using diachronic corpora

As described in Section 1.2 above, this Chapter is concerned with the evaluation of a recently proposed method, the Fitness Increment Test (FIT), for testing time series derived from corpora, with the goal of determining whether a change in frequencies is likely a result of linguistic drift or selection for a variant by speakers. This is an important question: one aim of this thesis is to develop a model of linguistic competition (see Chapter 4), yet competition only makes sense in the presence of selection. If a variant becomes prevalent and another disappears simply due to random drift, there is no sense in trying to model the given case as competition. In the bigger picture, this is also touches wider questions in historical and evolutionary linguistics, such as, how much of change is actually caused by directed selection, or do some languages or domains within languages change more due to drift or selection.

However, after carrying out simulations with artificial language change scenarios based on the Wright-Fisher model, I find that the FIT, while a promising approach, is not yet quite robust enough that it could be readily adopted in the explanatory models I aim to construct in this thesis. Still, carrying out this research proved useful in terms of thinking about the technical choices inevitable involved in corpus-based research, as discussed below.

2.1 Author contributions

The following paper has been published in *Glossa: a journal of general linguistics*. The reproduction of this online version, over the subsequent pages, is in accordance with the publication licence. I carried out the analysis, wrote the paper, created the figures, and handled the submission process. Kenny Smith, Richard A. Blythe and Simon Kirby provided advice on the design of the study and the analysis, as well as edits and comments on the paper.

The paper, as reproduced here, comes with its own header, but thesis page numbers continue throughout, in the footer. Note that the reference "Karjus et al. 2020" refers to the paper that forms Chapter 3 in this thesis, which was in the online-only Advance Article stage at *Language Dynamics and Change* at the time this paper was published.

2.2 Karjus et al. (2020): Challenges in detecting evolutionary forces in language change using diachronic corpora

Glossa a journal of general linguistics

Karjus, Andres, et al. 2020. Challenges in detecting evolutionary forces in language change using diachronic corpora. *Glossa: a journal of general linguistics* 5(1): 45. 1–25. DOI: https://doi.org/10.5334/gjgl.909

RESEARCH

Challenges in detecting evolutionary forces in language change using diachronic corpora

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

[1] Centre for Language Evolution, University of Edinburgh, UK

[2] School of Physics and Astronomy, University of Edinburgh, UK

Corresponding author: Andres Karjus (a.karjus@sms.ed.ac.uk)

Newberry et al. (Detecting evolutionary forces in language change, *Nature* 551, 2017) tackle an important but difficult problem in linguistics, the testing of selective theories of language change against a null model of drift. Having applied a test from population genetics (the Frequency Increment Test) to a number of relevant examples, they suggest stochasticity has a previously under-appreciated role in language evolution. We replicate their results and find that while the overall observation holds, results produced by this approach on individual time series can be sensitive to how the corpus is organized into temporal segments (binning). Furthermore, we use a large set of simulations in conjunction with binning to systematically explore the range of applicability of the Frequency Increment Test. We conclude that care should be exercised with interpreting results of tests like the Frequency Increment Test on individual series, given the researcher degrees of freedom available when applying the test to corpus data, and fundamental differences between genetic and linguistic data. Our findings have implications for selection testing and temporal binning in general, as well as demonstrating the usefulness of simulations for evaluating methods newly introduced to the field.

Keywords: language evolution; language change; selection; drift; corpus-based; temporal binning

1 Introduction

All natural languages change over time. The way each new generation of speakers pronounces their words is subtly different from their parents, new words replace old ones, marginal grammatical paradigms become the norm, and norms dissolve. Many authors have suggested that language change, like other evolutionary processes, involves both directed selection as well as stochastic drift (Sapir 1921; Jespersen 1922; Andersen 1990; McMahon 1994; Croft 2000; Baxter et al. 2006; Van de Velde 2014; Steels & Szathmáry 2018). Systematically quantifying the relative contribution of these two processes — particularly with reference to individual time series — is an open problem.

There are a number of ways in which selective biases may influence language change. For example various cognitive biases have been postulated as important in the evolution of language (Haspelmath 1999; Croft 2000; Kirby, Cornish & Smith 2008; Fay et al. 2010; Smith, Tamariz & Kirby 2013; Enfield 2014; Tamariz et al. 2014) and one might therefore expect to see manifestations of these in instances of language change. Selective advantage stemming from sociolinguistic prestige of (the users of) competing variants has been shown to play a considerable role in change, both via competition between forms within the language community as well as borrowing from other languages (Labov 2011; Hernández-Campoy & Conde-Silvestre 2012). A foreign or novel variant may also be selected for by virtue of filling a lexical or morphosyntactic gap (McMahon 1994; Trask

1996). The form of a variant alone may convey a selective advantage. For example, it has been observed that, all other things being equal, speakers prefer shorter forms that take less effort to utter (Zipf 1949; Kanwal et al. 2017) and limited iconicity can be advantageous (Dingemanse et al. 2015). Various usage and acquisition properties have been shown to be predictors of success (Kershaw, Rowe & Stacey 2016; Calude, Miller & Pagel 2017; Grieve, Nini & Guo 2018; Monaghan & Roberts 2019). There is also evidence that certain phonetic changes are more likely than others, due to the articulatory and acoustic properties of human speech sounds (Ohala 1983; Baxter et al. 2006). In certain circumstances there may be even qualitative evidence of directed selection, such as knowledge of previous activities of some authoritative language planning body, prescriptive grammars, or other exogenous forces (Rubin et al. 1977; Anderwald 2012; Ghanbarnejad et al. 2014; Daoust 2017).

It is a reasonable hypothesis that, given adequately large and representative samples of language use over time (i.e., corpora), signatures of selection should be inferable from the usage data alone. This idea has recently been explored in a number of works (Hahn & Bentley 2003; Bentley 2008; Reali & Griffiths 2010; Blythe 2012; Sindi & Dale 2016; Amato et al. 2018), and has been also applied to domains of cumulative cultural evolution beyond language (Kandler, Wilder & Fortunato 2017; Kandler & Crema 2019). One of the more ambitious attempts is that of Newberry et al. (2017), who employ a standard method borrowed from the field of population genetics, which also deals with the inference of selection in a population and the assessment of drift in evolution. We will henceforth refer to this work as "Newberry et al." (an earlier version of the paper is Ahern et al. 2016). They use the Frequency Increment Test (Feder, Kryazhimskiy & Plotkin 2014), or FIT for short, and make an explicit connection with the Wright-Fisher model (Wright 1931; Ewens 2004) of neutral stochastic drift (not unlike a previous similar contribution, Sindi & Dale 2016).

Newberry et al. consider three grammatical changes in the English language. Their main focus is the (ir)regularization of past-tense verbs (e.g. the change from irregular *snuck* to regular *sneaked*), a topic that has been of some interest (Lieberman et al. 2007; Cuskley et al. 2014; Gray et al. 2018). They also investigate the change in periphrastic *do* (*say not that!* becoming *don't say that!*), the evolution of verbal negation (from the Old English pre-verbal to the Early Modern English post-verbal), and possible phonological neighborhood effects (which we will not discuss here). They use data from the Corpus of Historical American English (Davies 2010) and the Penn Parsed Corpora of Historical English (Kroch & Taylor 2000). Their method consists of calculating the relative frequencies of alternative forms in a corpus (e.g., the relative frequency of the irregular past tense form *snuck* against that of the regular *sneaked*), placing the count data into variable-length temporal bins, and running the FIT on the resulting time series. Ultimately, the test yields a p-value under the null hypothesis of change by drift alone. They also infer the "effective population size" of the verbs and show that the strength of drift (in a subset of verbs with a FIT $p > 0.2$) correlates inversely with corpus frequencies, echoing the analogous observation about small populations in genetics.

The FIT points towards selection being operative in some cases, while labelling others (in fact, most changes in past-tense forms) as changes stemming from drift. In this work, we replicate this analysis (using Newberry et al.'s original code; see the Data Availability section in the end). We highlight an important methodological issue that arises when applying the FIT to linguistic data and which should be taken into account in future applications of the FIT (and similar tests) to identify cases of selection from linguistic corpora. The key issue lies in the construction of the time series via binning counts (e.g.

from a corpus), and the application of the test in question to such time series, but we also draw attention to issues more specific to diachronic language data. While the FIT may be an appropriate test in some cases, we show that an incautious application of the FIT to linguistic data can end up incorrectly identifying cases of drift as cases of selection, and missing subjectively clear cases of selection.

While the approach of applying a test of selection to corpus-based time series shows promise as a method of linguistic analysis, we believe these issues deserve further investigation. We briefly explain the technical aspects of temporal binning and the FIT in the next subsections.

1.1 Linguistic corpora and data binning

In quantitative research on language dynamics, words and grammatical constructions are often equated with alleles (Reali & Griffiths 2010). This analogy is motivated by the observation that a given "underlying form" may have two or more (near-) synonymous actualizations or "surface forms" (e.g. as in the *sneaked–snuck* case which are both actualizations of *sneak*.PAST). Word variants are not quite like alleles though. Organisms inherit genetic material from their parents, and one can (in principle) test for the presence of a particular allele in each individual in the population over time. In the context of language use, the notions of parents, offspring and generations are more diffuse than they are in genetics. What is done in practice when analysing time series is to construct an artificial "generation" by collecting together all instances of the word variants under consideration that fall within a specific time window (or "bin"). Particularly troublesome is that fact that a given lexeme may not occur in a given corpus in a particular period of time, which means having to widen the bin to obtain a meaningful frequency. Such absences may occur simply because of the finite size of the sample: any corpus is in the end just a sample from a population of utterances. The smaller the corpus, the smaller the chance a lexeme has to occur. It may also be because people talked and wrote about other topics in that time window, which did not require the use of this particular sense. A corpus may be large, but not well balanced, in the sense that it does not cover all the relevant genres or topics of the time. Incidentally, this is a point of critique directed by Pechenick, Danforth & Dodds (2015) at another widely used diachronic corpus, the Google Books N-grams dataset.

To understand the issue of binning (or temporal segmentation) in more detail, let us consider for a moment a fictional corpus of a daily newspaper, spanning two centuries. Our goal is to count the occurrences of two competing spelling forms of a word and operationalise these as relative frequencies in a time series. The smallest possible temporal sample would consist of the text that makes up one daily issue of the paper (yielding a fine grained time series of about $n = 73000$ data points). One could also aggregate (bin) all the texts from one month ($n = 2400$), year ($n = 200$), decade ($n = 20$) or century ($n = 2$). However, there is no single ideal way to bin the data. A century, with only two data points, may be too large a chunk, as it may miss processes taking place in between — and it is difficult to infer anything about the dynamics of the change from two data points. A day is likely too small a sample, since the word (in either spelling) might not occur every day, unless it is a particularly commonly used one.

In corpus-based language research either years or decades therefore seem the most commonly used bins. Regardless, a decision has to be made regarding how to bin corpus data; our point here is to show that this decision (which potentially constitutes an additional researcher degree of freedom, since different binning decisions may yield different results) influences the outcome of analyses which use tests like the FIT to identify selection.

1.2 The Frequency Increment Test

The FIT (Feder, Kryazhimskiy & Plotkin 2014) belongs to a family of methods conceived to detect selection in time series genetic data, with intended application to population genetics experiments and historic DNA samples. All of them boil down to looking for certain patterns in time series of allele frequencies (Nishino 2013; Terhorst, Schlötterer & Song 2015; Schraiber, Evans & Slatkin 2016; Iranmehr et al. 2017; Taus, Futschik & Schlötterer 2017; Vlachos & Kofler 2018) (see Malaspinas 2016; Vlachos et al. 2019: for reviews). Such approaches rely on the presumption that a change driven by selection would look different, or leave different "signatures", from a change happening due to stochastic drift.

The FIT works as follows. Relative frequencies in the range (0, 1) are transformed into frequency increments *Y* according to

$$Y_i = (v_i - v_{i-1}) / \sqrt{2v_{i-1}(1 - v_{i-1})(t_i - t_{i-1})} \tag{1}$$

where v_i is the relative frequency of a variant at a measurement time t_i. The rationale behind this rescaling is that, under neutral evolution, the mean increment v_i–v_{i-1} (i.e. the change in frequency of v_i from time t_{i-1} to time t_i) is zero, and its variance is proportional to

$$v_{i-1}(1 - v_{i-1})(t_i - t_{i-1}), \tag{2}$$

i.e. the expected variance under drift is large when we are looking at the changes in frequency between two widely separated time points (i.e. t_{i-1} and t_i are far apart) or when values of v_i are close to 0.5 (i.e. changes in frequency driven by drift will tend to be small when the variant is very rare and v_i is close to 0, or very common and v_i is close to 1).

The FIT relies on the Gaussian approximation of the Wright-Fisher diffusion process. When the variant frequency v_i is not too close to either of the boundary values 0 or 1 and the time between successive measurements is sufficiently small, the random variables Y_i can be approximated as having a normal distribution with a mean of zero and a variance that is inversely proportional to an effective population size (which is taken to be constant over time). Thus a test under the null hypothesis of drift amounts to a test of how likely the transformed increments Y_i are under the assumption that they are drawn from a normal distribution with a mean of zero, as would be the case under drift: this can be evaluated using a one-sample *t*-test test under the assumption of normally-distributed increments with a zero mean and equal variance.

In this context, a failure to reject the null indicates a failure to reject the hypothesis of drift. On the other hand, if the null hypothesis is rejected, than the changes may be due to some non-neutral process. In this work, we check for the normality assumption using the Shapiro-Wilk test. Homoscedasticity (the assumption that the underlying distributions have equal variances) is less straightforward; we explore its relevance in the Supplementary Appendix.

The authors of the Frequency Increment Test (Feder, Kryazhimskiy & Plotkin 2014) note that its power increases with the number of sampled time points, but also that it has low power in cases of both very weak (near-drift) and very strong selection coefficients. The latter leads to a situation where fixation to a variant happens swiftly within the sampling interval (the range of the time series), making the rest of the time series uninformative. The frequencies should also be far from absorbing boundaries (i.e., situations where one variant is at (or near) 0% and the other at 100% of the population), which might pose a particular problem in corpus-based time series analysis: since linguistic change is (classically) believed to follow an S-shaped trajectory (Blythe & Croft 2012), a change which takes place near the start or end of a given corpus would throw off the test, since most of the length of the given time series would be (near-)stationary. Similarly, if a corpus (equivalent to the "sampling period" in a genetics experiment) is too "short", it might only chronicle a segment of a longer change process.

2 The FIT and binning decisions in linguistic corpora: A reanalysis of English past tense verb regularization

We focus here on the main result of Newberry et al. — the application of the FIT for assessing time series of verb form frequencies in order to determine if the observed patterns of change for 36 English verbs results from stochastic drift or selection. Technical data processing details described in this section are based on the Supplementary Information of Newberry et al., their code, and M. Newberry, p.c.

They construct a time series for each of 36 pre-selected verbs using 200 years of data in the Corpus of Historical American English (COHA), by counting how many times the regular past tense form occurs relative to the total number of instances of either the regular or irregular form. The yearly verb count series are then binned (grouped) into a number of variable-width quantile bins $n(b) = \lceil \ln(n(v)) \rceil$, where $n(v)$ is the sum of both (regular and irregular) past tense form tokens of the verb counted across the entire corpus. For example, *light*.PAST occurs $n(v) = 8869$ times in the corpus, resulting in $\lceil \ln(n(v)) \rceil = 10$ bins to group the years where the verb occurs. The first bin contains years 1810–1863 (and contains 897 tokens), the second 1864–1886 (890 tokens) and so on, up to the tenth (1994–2009, 884 tokens). Since the grouping is by years (years being the time resolution of the corpus), the bin size varies slightly in the exact number of tokens falling into each bin. More frequent verbs thus get more bins (up to 13), whereas less frequent verbs get fewer bins (down to 6). For each verb in each bin, the relative frequency of its regular past tense form in [0, 1] is calculated. Since the FIT assumes relative frequencies in (0, 1), Laplace $+1$ smoothing is applied to count values in bins where one of the variants has no occurrences at all in this section of the corpus.

As discussed above in the section on corpus binning, *some* temporal segmentation process is necessary. The binning procedure applied by Newberry et al. is somewhat different from the more common strategy of using fixed length bins such as years or decades. The advantage of their approach is that there is guaranteed to be data in every bin (whereas a low frequency lexeme might be entirely absent in a fixed-width bin), the bins are roughly the same size in terms of tokens, and the resulting increments tend (although are not guaranteed) to be normally distributed with equal variance. These properties are beneficial for the FIT, more likely yielding normally distributed increments with less sampling noise (Feder, Kryazhimskiy & Plotkin 2014). It should be noted though that the resulting bins differ quite widely in their temporal granularity — e.g. in the example above, the longest bin covers the earliest 53 years of the corpus, the shortest covers the most recent 15 years, and different verbs will use different time windows depending on their frequency in the corpus. Since the COHA is smaller on the early end (less tokens per year) and bigger on the more recent end, variable-width bins of the verb data are systematically longer in the early 1800s compared to the 20th century ones (cf. the Supplementary Appendix for more discussion).

The series of relative frequencies based on the resulting bins are fed into the Frequency Increment Test to assess whether one may reject the null hypothesis of drift and assert that a given trajectory is therefore probably a product of selection. Newberry et al. set the FIT $\alpha = 0.05$ but also report results for $\alpha = 0.2$. They conduct the Shapiro-Wilk normality test on the transformed frequency increments, as the FIT assumes the increments to be normally distributed.

We replicate their original results, using their code, and furthermore explore the consequences of manipulating the size of the bins, in two ways. We present results for both binning strategies. That is, variable-width bins, $n(b) = c\ln(n(v))$, where c is an additional arbitrary constant, and $c = 1$ recovers the Newberry et al. procedure; and fixed-width bins, each set to a fixed duration in years.

Importantly, the fixed-width binning approach necessitates the introduction of an additional parameter: since some bins may end up with no or few occurrences of either form of a verb, we set a threshold of minimum 10 total occurrences for a relative value to be calculated in a bin; otherwise the bin is excluded before applying the FIT (hence also reducing the number of bins that make up the time series). As the FIT assumes values in (0, 1), smoothing of boundary values is required. But if there is only a single occurrence of a verb in a bin (meaning the single present form would be at 100%, the other at 0), then the $+1$ smoothing would force the relative value to be 50–50, which is undesirable. Similar distortions would happen with small frequency values, hence the threshold of 10. See the Supplementary Appendix for more discussion on the differences between these approaches and how different minimal frequency thresholds affect the results. A more conservative threshold (such as 100) would yield more reliable bins (and less noisy time points), but given the size of COHA, most verbs don't have 100 occurrences per year (or some even in 5 years), which would preclude testing in shorter fixed bins.

Figure 1 shows the results of these various analyses, in terms of how many verbs (out of the 36) allow us to reject the null hypothesis of drift, given the thresholds mentioned in the original work, as well as taking into account the normality assumption of the FIT (see above). We use the Shapiro-Wilk normality test, following Newberry et al. (this test is of course subject to low power in small samples as well). Out of the 466 time series analyses summarised in Figure 1 (36 verbs times 13 binning choices, minus two series with not enough data points), 63% of the FIT p-values are eligible to be interpreted at Shapiro-Wilk $\alpha = 0.1$.

We find that binning strategy does have an effect on the results, both in variable and fixed binning. Importantly, in broad strokes, the picture presented by Newberry et al. holds. They found that 6 out of 36 verbs undergoing selection; since the majority of verbs do not give a positive signal for selection, they interpret this as indicating that language change is often primarily stochastic. Looking at a wider range of binnings, we find that in most cases, there are indeed 5 ± 2 verbs that get flagged as undergoing selection at FIT $\alpha = 0.05$, consistent with their conclusion. However, the specific verbs that are flagged as undergoing selection vary depending on the binning strategy. There are 4 verbs for which selection is detected in most binning choices — *light, smell, sneak, wake* (incidentally the ones with the strongest inferred selection coefficient, given the original binning, cf. EDT1 in Newberry et al.). There are also between 9 and 11 verbs (in variable-width binning; depending on how stringently the normality assumption is observed) which provide a robust absence of significant indications of selection, where the FIT p-value never drops below 0.2 regardless of binning. However, for the remaining verbs the decision as to whether or not they are undergoing selection depends on the binning choices. That being said, Newberry et al. do draw attention to the fact that results of applying the FIT come with a certain margin of error and report their false discovery estimates (30% for verbs with a FIT $\alpha = 0.05$, 45% at 0.2).

Given that binning leads to different sample sizes of increments for the underlying t-test, those in turn being based on differing distributions of the tokens, some variance in the p-values is to be expected (not unlike in a replication of an experiment). The interpretation of our results and the appropriate conclusion regarding the sensitivity of the FIT test to binning strategy ultimately depends on one's intention in carrying out a tests of selection in

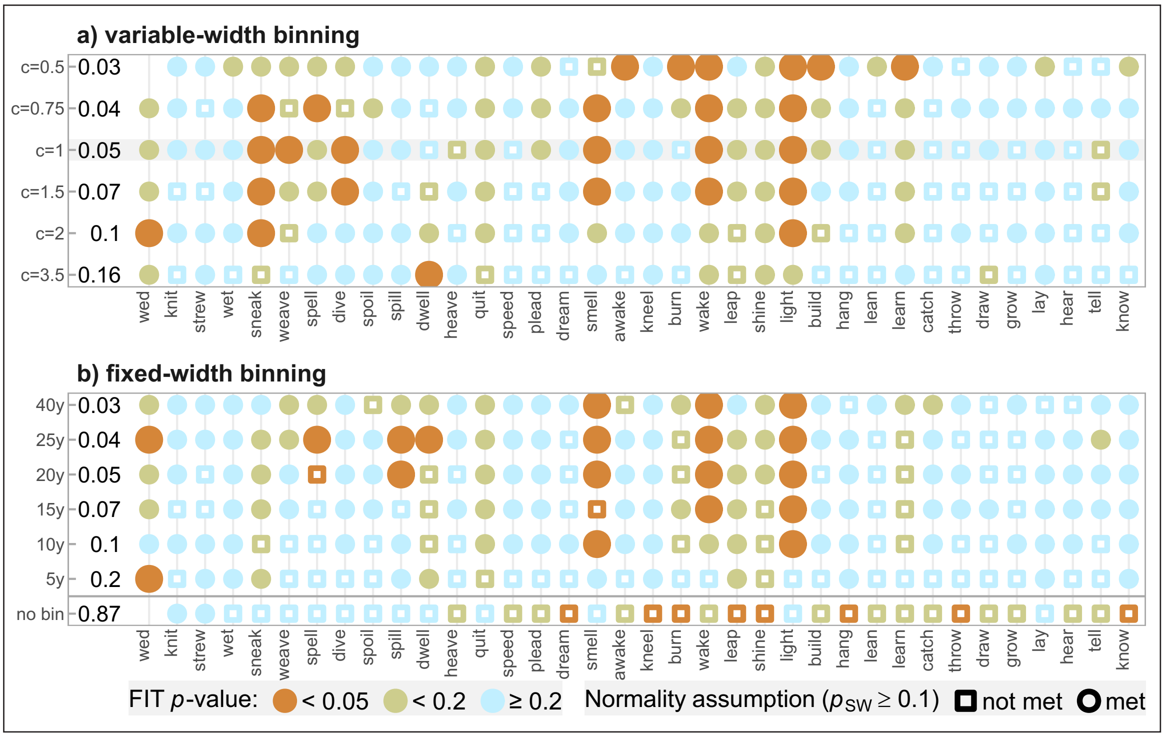


Figure 1: Results of applying the FIT to time series constructed based on 200 years of COHA frequency data. The verbs are ordered by overall frequency (low on the left). The constant c determines the number of variable length bins via $n(b) = c\ln(n(v))$. $c = 1$ corresponds to Newberry et al.'s original results. 10 years corresponds to fixed bin length of 10 years, etc; "no bin" refers to no additional binning on top of the default yearly bins in the corpus. The colour of each point corresponds to the result of the FIT test of a verb time series in each binning (orange: $p < 0.05$, gold: $0.05 \leq p < 0.2$, light blue: $p \geq 0.2$). The shape corresponds to the Shapiro-Wilk test result (filled circle: $p \geq 0.1$, hollow square: $p < 0.1$, likely not normal), with cases of selection meeting the normality assumption highlighted by a larger circle. The column of numbers on the left displays the (rounded) median of the bins to years ratio in the given binning strategy. Only years where the verb occurs are counted (exclusion of sparse bins also leads the median in the no-binning version to be below 1). The listed variable (panel a) and fixed-width strategies (b) yield comparable binning ratios, e.g. the "$c = 1$" version is comparable to 20-year fixed-width. In summary, the results presented here demonstrate that the FIT is sensitive to the strategy used for binning.

the first place. If the goal is to test a large set of series to determine general tendencies, as is the case for Newberry et al. then this approach may well be good enough — the qualitative result of Newberry et al. does broadly apply in most binning strategies.

However, most individual time series seem rather sensitive to binning, in the sense that the *p*-values fluctuate across conventional α levels between binnings. No verbs show an unambiguous signal of selection. For example, drift is not rejected in the time series of *wed* using the Newberry et al. binning, while it is when the number of variable-width bins is multiplied by 2. The verb *sneak* is significant at $\alpha = 0.05$ in almost all the variable-width binnings, but in none of the fixed length ones; *awake* is significant in only a single explored binning strategy (variable-width with $c = 0.5$) and there are 4 more such verbs particularly sensitive to binning (the 1-year bins notwithstanding).

The no-binning results (i.e., using the default 1-year bins of COHA without further binning) differ visibly from the rest, but the normality assumption is also mostly violated. Given the small and variable bin sizes (tokens per bin), the same is likely true for the homoskedasticity assumption (although how much that matters and how to set a threshold is not clear, cf. the Supplementary Appendix). Most importantly, using "default"

1-year bins leads to testing on series where the increments are often based on very small samples, which is not desirable for any statistical test.

These evaluations obviously depend on the choice of α thresholds for the FIT and the supporting normality test — for example, a more stringent FIT α would lead to more verbs being classified as unambiguous cases of drift. In any case, if the intention is to test a particular example of linguistic change for selection (something a linguist may well be interested in), things become difficult. The issue diminishes if there is sufficient data on the variants, but that does not seem to be the case for many of the verbs tested here, given the size of COHA.

All in all, these findings merit a further investigation into the inner workings of the Frequency Increment Test and its applicability to corpus-based time series, which we will conduct in the following two sections.

3 The behaviour of the Frequency Increment Test in artificial time series

We construct a number of artificial examples (Figure 2) to probe the behaviour of the FIT on time series of length and character similar to those investigated in the original paper (which contained between 6 and 13 time points). The FIT can be shown to yield robust results for a certain range of series (as already shown by the subset of binning-insensitive verbs in the previous section). Yet we also observe a number of scenarios — time series that could be plausibly derived from linguistic corpora — where the results of the FIT are perhaps not what one might expect, from a language science point of view. To put it another way, this is the section where we push the FIT and see if it breaks. The next section demonstrates scenarios where the results of the FIT remain robust.

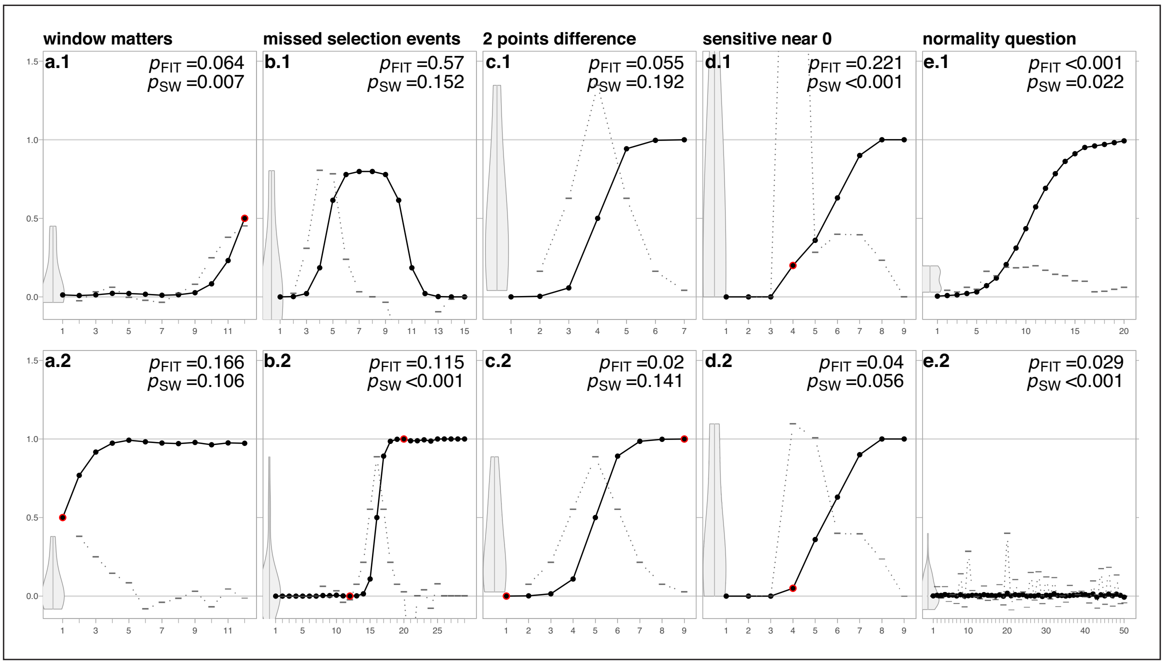


Figure 2: Artificially constructed time series of fictional variant relative frequencies (thick black lines, in (0, 1)); time on the x-axis. The rescaled increments (after adjusting for absorption) are shown as dotted grey lines with dash points, and their distribution is shown on the left side as a violin plot. Points of interest discussed in this section are highlighted with red on some panels. The FIT and Shapiro-Wilk test *p*-values are reported in the corners. This figure depicts a number of realistic scenarios where applying the FIT would yield unexpected results, due to either the range of the time series derived from the corpus **(a, b)**, a difference in the number of data points **(c)**, the sensitivity of FIT to near-zero values **(d, e)**, and how stringently the assumption of the normality of the distribution of increments is being observed (e). This figure illustrates reasons to exercise caution when applying a test like the FIT to linguistic time series.

Each series in Figure 2 may be interpreted as the percentage of a variant of some fictional linguistic element over time (after binning). We calculate the FIT *p*-value of each series, as well as the Shapiro-Wilk test *p*-values. Figure 2.a draws attention to how the temporal range of the time series (or that of the coverage of the corpus) can lead to quite different conclusions. Both 2.a.1 and 2.a.2 are different ends of the same series (the overlap highlighted with the red circle). The series, if analysed as a whole, would yield a $p_{FIT} = 0.02$, but neither end on its own holds sufficient data to reject drift (nor is the FIT technically applicable, if the assumption of normality is observed). This perspective may explain the case of the purportedly drift-driven regularization of the verbs *spill* and *burn*, which are brought up in Newberry et al. as examples where drift alone is sufficient to explain the change, but which are problematic because the regular forms were already highly frequent by the early 19th century where the COHA coverage starts. *spill* starts out with a share of 55% regular forms in the first bin given the variable-width binning strategy; *burn* is at 86% regular. Under fixed decade binning, *burn* is 36% regular in the first bin, increasing to 62% and then to 82%, indicating a sharp increase characteristic of strong selection rather than drift (but obscured by the variable binning approach).

This example also points to a case where different evolutionary domains (genetics, language) might have different expectations about what a reasonable time-series characteristic of selection should look like. The FIT assumes the Wright-Fisher as the underlying model (reasonably so in population genetics). The long tail of near-zero values followed by a sudden increase in 2.a.1 is something that is unlikely to be observed in a Wright-Fisher model with constant selection strength parameter. However, from a linguistic point of view, this is a very natural series: a recent innovation or borrowing will be represented in the corpus as an increase preceded by a period of zero frequencies as far back as the corpus goes; this pattern could be explained as a recent change in fitness (e.g. a change in the subjective sociolinguistic prestige of a word).

A similar case is presented in Figure 2.b.1: if the time series chronicles both strong selection for one variant, and subsequent selection for the competing variant, then a blind application of the FIT will invariably indicate drift. Using only (either) half of the series as input to the test would yield a *p*-value indicating selection. *knit* is a verb undergoing a somewhat similar process, with usage spiking towards the regular (observable under finer binnings), followed by mostly irregular usage. Figure 2.b.2 is an example of the behaviour of FIT if the corpus coverage is *too* wide. The S-curve in the middle would yield a FIT *p*-value of 0.02 — in fact, it is the exact same curve as in Figure 2.c.2 (highlighted by the red dots). Yet the S being surrounded by (near-)absorption values, the FIT would indicate drift (were the test to be used despite the possible non-normality of the distribution).

In the case of real data, the part of the time series depicting the long period of no change could in principle be clipped away. This is straightforward if the "tail" consists of zeroes, but less so given small near-boundary values. Similarly, only the part of the time series far enough from the boundaries could be analysed (keeping in mind the specifics of the FIT, see above). However, any such solutions would introduce yet another researcher degree of freedom (what part of the series to include in the analysis) (cf. Simmons, Nelson & Simonsohn 2011).

Figure 2.c further illustrates how the FIT result is affected by a change in the way the time series is operationalised (e.g., using a different number of bins). 2.c.1 and 2.c.2 are S-curves with identical parameters, differing only in length (by 2 data points). Yet their FIT *p*-values are notably different (see the next section for more on sensitivity to binning differences). As expected, the FIT is sensitive to small changes if the sample is small (being based on the *t*-test). This may explain to some extent the changes in FIT *p*-values of short time series, between similar binnings differing only by a few points in length (cf. Figure 1). However, fewer bins can also lead to a lower p, if it results in a less jagged time series (likely the case for e.g. *burn*; cf. Section 4 for the effects of binning on drift series).

The examples so far however have had more to do with particularities of pre-test data manipulation. Figure 2.d illustrates a property of the FIT, its sensitivity to changes near the boundaries. 2.d.1 and 2.d.1 differ only by the value of the fourth data point, but the resulting FIT p-value is quite different (and furthermore the Shapiro-Wilk test indicates departure from normality in the increment distribution due to the outlier). The issue of applicability of the FIT to series with increments departing from normality is further illustrated with the last pair of series. 2.e.1 is a typical S-curve often observed in language change, but the non-normal distribution of its increments would disallow the interpretation of the FIT p-value (that would otherwise indicate a clear case of selection).

We observe that in general, for longer series exhibiting monotonic increase (characteristic of strong selection), the distribution of the increments quickly veers into the non-normal (as indicated by the Shapiro-Wilk p-value; other normality tests behave similarly; see also the Supplementary appendix). Time series composed of random values drawn from a uniform or normal distribution (or log-normal with small σ) — i.e., the kind of series that should exhibit no selection — tend to have increments distributed approximately normally, as long as the series is away from the boundary values. However, the increments of S-shaped curves tend towards a bimodal distribution. Increment distributions of are severely skewed when a series is shaped like an S-curve but with a sharp "bend", a straight line (linear increase or decrease), and when a series include long periods of no change.

The assumption of normality could of course be relaxed. However, we observe that this would lead to at least one additional issue, in the form of false positives stemming from the sensitivity of the FIT to small near-boundary changes, illustrated by 2.e.2. Given a long enough series of random values (here sampled from a normal distribution) with a near-zero mean and small standard deviation, the FIT often yields a small p-value (the same applies to samples from the uniform and log-normal distributions; this effect is not observed when the mean is away from the boundaries). Such series would however usefully get flagged as having non-normal increment distributions.

This is also likely why the otherwise flat-lining series for *tell* in Newberry et al. ends up being included in the discussion as a possible case of selection (at FIT $p = 0.12$, with a red flag of Shapiro-Wilk $p = 0.001$). Among the 12 bins of its series (under the original variable-width quantile binning procedure), it has only a few once-per-bin occurrences of regular *telled* after the initial three bins — a total of 4 singleton occurrences spread out over the span of a century. The $+1$ absorption adjustment forces the zeroes for *telled* in the rest of the bins to be ones as well. The observed fluctuations (and resulting FIT p-value) in the series only reflect the slightly fluctuating token frequency of *tell*, which ranges between 9189 and 11940 in the variable-width bins. Keeping the relative frequency value constant after the third bin instead (at the value equal to the third bin to avoid bias) would result in a FIT $p = 0.21$.

These last four usages of the regular past form *telled* in COHA all occur in the fiction part of the corpus, all appearing to reflect the intention of the author to convey a particular kind of character (not used randomly as per a drift model). This would be an example of how an archaic variant can re-surface — quite possible in a language with a long written record, where speakers need not necessarily even directly "inherit" a variant from the previous generation. In that case, *telled* could be said to have been selected for, due to having increased fitness in a specific (stylistic) niche, and its usage is not due to random variation in the utterances of the speakers (or drift). However, as shown above, this possible (occasional) selection is not what the FIT is picking up on in this case, but rather simply the fluctuating frequency of *tell*.

Meaning change can also give rise to apparent re-emergence of variants. The occurrence of a form does not guarantee that it is being used in the same meaning or function that it had in another period or context (an implicit assumption in Newberry et al.). For example, the aforementioned *spill* in COHA quickly converges to the regular past tense *spilled*, but occasional usages of the irregular *spilt* still occur, yielding what appears to be a randomly fluctuating time series. On closer inspection, the latter appear to be mostly adjectival usages, not actual past tense verbs, and often turn up in the lexicalized (or "fossilized") phrase of *cry over spilt milk*. Examples like that of the time series of *telled* and *spilt*, or the series in Figure 2.a.2 and e.2. may possibly be seen as edge cases from the perspective of population genetics — the original domain of the Frequency Increment Test and related approaches. However, as highlighted here, they are examples of not particularly uncommon processes (lexicalization, stylistic usage of unusual variants) in the domain of language.

Finally, one might argue the examples in Figure 2 are not really counterexamples to the utility of the FIT, being representative of cases where the FIT is, strictly speaking, not designed to apply in the first place, such as series with not-quite-normal increments, long flat segments, and values near the boundaries. Excluding these however would mean excluding a fair share of language change scenarios easily observable in corpora, such as changes starting at zero as in cases of linguistic innovations, ongoing changes stretching beyond the bounds of a corpus, and many S-curves typical of language change (and series in general where the underlying selection coefficient is likely not constant). Yet dismissing these as invalid points of concern would also mean dismissing the FIT as a broadly applicable test of selection for the domain of language change.

In the next section, we turn to simulations to explore the behaviour of the FIT beyond that of a few specific series (Section 4), before finally trying to reconcile these conflicting viewpoints (Section 5).

4 The effect of binning frequency data for time series: A simulated example

Here we attempt to further explore the "parameter space" of applying the FIT to simulated data with known properties of selection strength and binning. (code to replicate these results: see the Data availability section in the end). We use the Wright-Fisher model (Ewens 2004) to simulate a large number of time series using the following parameters: population size $N = 1000$ (N here does not refer to the "population" of speakers, but is analogous to the sum of parallel variants in a corpus bin, e.g. the sum of the counts of *lit* and *lighted* in a given year); selection coefficients s *in* [0, 5]; 200 generations (the latter emulating COHA, where the minimal time resolution is 1 year, and there is 200 years of data). The update rule for this model is as follows. Given n_t "mutants" (e.g., regular past tense forms) in generation t, each individual in the next generation is a mutant individual with probability

(3) $$q = \frac{n_t(1+s)}{n_t(1+s)+(N-n_t)}$$

Otherwise, it is the wild type (e.g., irregular past tense forms). Where $s = 0$ we have random drift; higher values of s given an increasingly strong selective advantage to the mutant variant.

Each series (200 data points) is binned into a decreasing number of bins (i.e., [200, 4], of length [1, 50]), and the FIT is applied to every binned version. The simulation for each combination of selection strength and bin length is replicated 1000 times. In summary, in this section we vary the selection strength s and binning, while keeping N and the number of generations constant.

Importantly, we also apply binning to the series post-simulation the same way one would apply binning to corpus counts, as discussed above. The obvious difference from corpus-based time series is that the latter usually do not come from a population with a stable size (total lexeme frequency usually varies in addition to variation in its variants), and are often not continuous (gaps where a lexeme might be completely absent). Since our artificial series do not suffer from these problems, variable-width and fixed-length binning yield identical results, and we can simply use the latter.

We explore two scenarios, where the competing "mutant" variant starts out at 50% of the population and where it starts out at 5%. The former is useful for exploring the effects of binning at low *s* and false positive rates, the latter for exploring high *s* and false negatives. Obviously, any specific *s* thresholds and ranges discussed in this section apply to this specific experiment and would likely be somewhat different given series of different length and *N* (cf. the Supplementary Appendix for some further exploration).

4.1 Drift and low selection

Figure 3 depicts how the results of the FIT change depending on binning, given a time series with low selection ($s = 0.01$, bottom row) and no selection ($s = 0$, top row; corresponds to the leftmost column of pixels on the panels in Figure 4). At zero selection, the FIT has a reasonable false positive rate of around 5% at $\alpha = 0.05$. Binning such series into a smaller number of bins causes an increase in the share of *p*-values below 0.05 (presumably because noise is smoothed out). Binning appears to affect the $s = 0.01$ range even more (bottom row).

Figure 4 represents the entire parameter space explored in this experiment for the 50% start condition. Each pixel on the heat maps corresponds to a parameter combination of selection strength (horizontal axis) and number of bins (vertical axis). The vertical axis starts with 200 or no binning, corresponding to bin length 1 — and running up to 4 bins, with bin length 50, being the result of 200 data points squeezed into the 4 bins. Minimal binning — compressing 200 generations into 100 bins of length 2 — appears to make the clearest immediate difference: the share of $p < 0.05$ is consistently about 10%

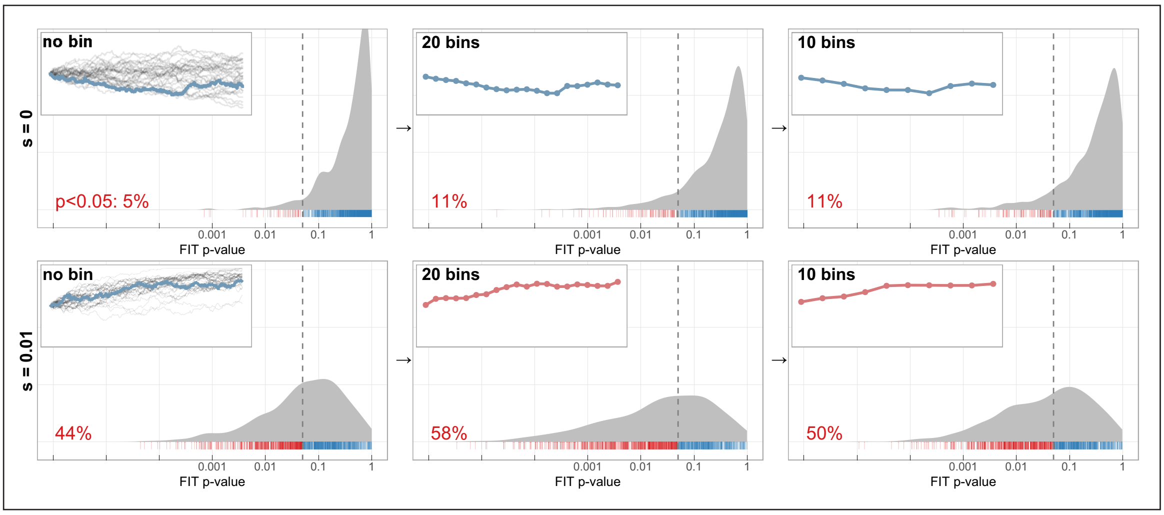


Figure 3: The distribution of FIT *p*-values given 1000 series from the Wright-Fisher model (200 generations, starting at 50%). The panels are arranged from left to right reflecting increased binning. The small inset panels display how binning affects a single example series. *p*-values below 0.05 are coloured red (left of the dashed line), above 0.05 in blue. Note the $\log_{10}$ x-axis. This figure illustrates that the false positive rate is susceptible to increasing when the series are binned (top row). At non-zero but low s, differences between binning and no binning can be more pronounced (bottom row). See Figure 4 for the full exploration of the parameter space.

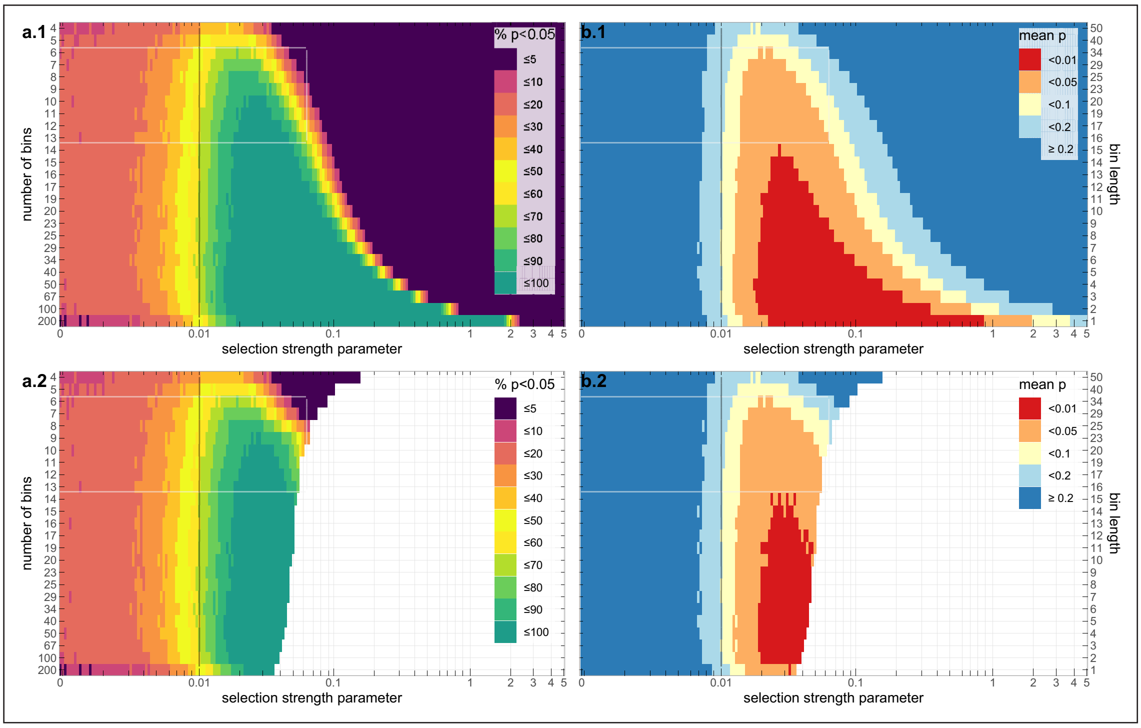


Figure 4: FIT *p*-values of time series generated using the Wright-Fisher model (with the "mutant" variant starting at 50%), across a range of selection coefficients (x-axis, note the log scale), binned into a decreasing number of bins (y-axis). Left in pink and green (a): % of time series with FIT $p < 0.05$, in 1000 replicates. Right in red and blue (b): mean FIT *p*-value. The bottom pair (a.2, b.2): the same data, but series with a Shapiro-Wilk $p < 0.1$ have been removed before calculating the percentages and means. The white rectangle: the range of *s* and binning explored in Newberry et al. The vertical black line highlights the *s* explored in Figure 3. A consistent colour across a column of pixels indicates robustness to binning choices under the corresponding *s*, while variable colouring indicates sensitivity to binning.

higher between the binned and non-binned series when *s* is low (observe the bottom two "shifted" looking pixel rows in Figure 4.a.2).

The 50% start is suitable for exploring low selection, as in the case of lower starting values, many such series hit absorption or "run into the ground", and the resulting mostly-zero series would violate the normality assumption (of its underlying Gaussian approximation of the diffusion process). However, the higher *s* range in Figure 4.a.2 could be interpreted as a model of the situation where a change is only partially chronicled by a corpus, e.g. Figure 2.a.2 in Section 3. Selection becomes understandably difficult to detect in very short series regardless of the underlying selection coefficient.

4.2 High selection

The 5% start is suitable for exploring high selection, as with higher starting values, many high-selection time series reach absorption fast, yielding series not meeting the increment normality assumption. Figure 5 depicts distributions of FIT *p*-values under different binnings, given time series with a moderately high *s* of 0.04, and the incoming variant starting out at 5%. This appears to be the subset of series where the FIT works very well and is most insensitive to binning choices.

Beyond that, things become more complicated. Our reanalysis of the 36 verb time series in Section 2 indicated that it is series exhibiting the strongest selection that would remain consistent in terms of their FIT result across the different binnings. However, as illustrated in Figure 6, it seems too high selection can have the inverse effect, as this is where false

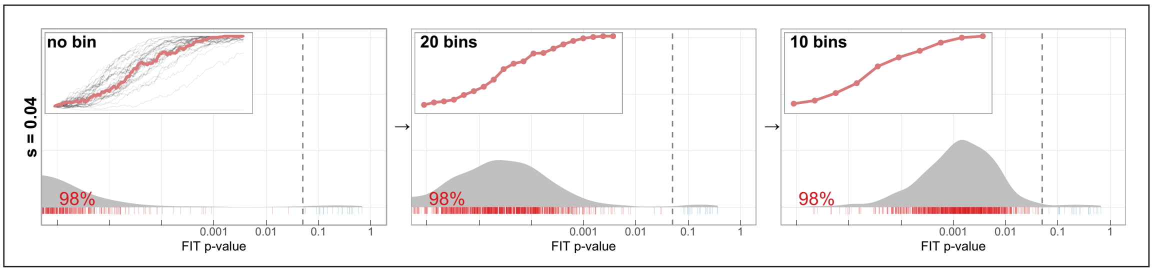


Figure 5: The distribution of FIT *p*-values given 1000 Wright-Fisher series with strong selection (200 generations, starting at 5%). The panels are arranged from left to right reflecting increased binning. The small inset panels display how binning affects a single example series. *p*-values below 0.05 are coloured red (left of the dashed line), above 0.05 in blue. Note the $\log_{10}$ x-axis. The red value in the bottom left corner shows the percentage of *p*-values below 0.05. This figure illustrates the *s* range where the FIT is most robust to binning, retaining a small and stable false negative rate (i.e. the inverse of the percentage value in the corner).

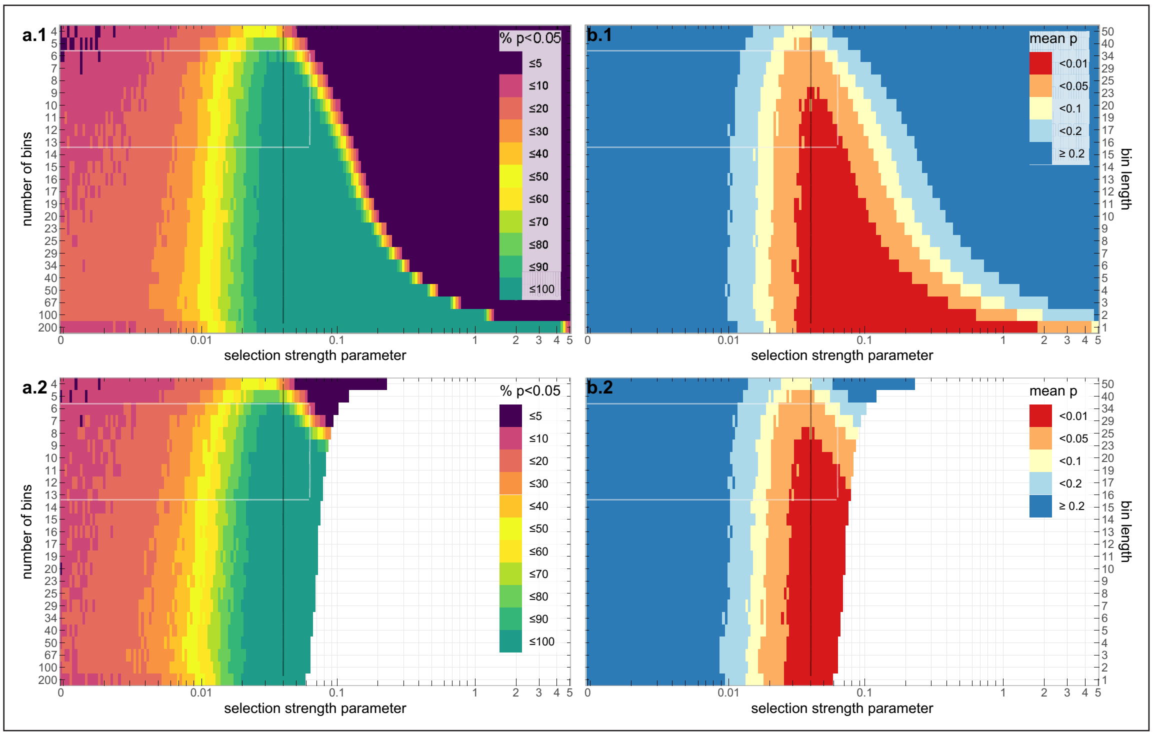


Figure 6: FIT *p*-values of time series generated using the Wright-Fisher model (with the "mutant" variant starting at 5%), across a range of selection coefficients (x-axis, note the log scale), binned into a decreasing number of bins (y-axis). Left in pink and green (a): % of time series with FIT $p < 0.05$, in 1000 replicates. Right in red and blue (b): mean FIT *p*-value. The bottom pair (a.2, b.2): the same data, but series with a Shapiro-Wilk $p < 0.1$ have been removed before calculating the percentages and means. The white rectangle: the range of s and binning explored in Newberry et al. The vertical black line highlights the s explored in Figure 5. A consistent colour across a column of pixels indicates robustness to binning choices under the corresponding s, while variable colouring indicates sensitivity to binning.

negatives begin to crop up under too much binning (e.g. with 10 bins, >10% at $s = 0.07$, >90% at $s = 0.1$). That is, if the increment normality assumption is being be strictly observed — if it is, then the results of the test are not valid any more at this range (cf. white area in Figure 6.a.2). This illustrates that the FIT has a maximum selection strength for which it is effective. At higher selection strengths, i.e. above 0.06.0.1 in our toy model, sensitivity to binning and violations of the normality assumption both become problematic, yielding results with a high false negative rate (if the assumption is relaxed; cf. Figure 6.a.1) or results which are invalid (if it is observed; 6.a.2). Incidentally, this also is the *s* range where S-curves characteristic of language change begin to form (cf. the Supplementary appendix).

In summary, these results indicate that if one is to take the same ensemble of language changes, with known selection strength, and apply different binning protocols, one could easily end up drawing very different conclusions depending on the bin length and the normality assumption threshold, if the conclusions are based solely on applying a test such as the FIT. However, if awareness of these limits is maintained, then the FIT works well on time series with moderately strong selection, and reasonably well (with the caveat of somewhat increased false positives rate under binning) on time series generated by a zero or low selection coefficient.

5 Discussion

We started out by focussing on the study of the (ir)regularisation of the past tense of 36 English verbs in Newberry et al. specifically their finding that drift cannot be rejected in most cases, leading to the claim of the "an underappreciated role for stochasticity in language evolution" (Newberry et al. 2017: 223). The conclusion of our reanalysis section — that their broad conclusion stands but that the FIT is sensitive in specific instances to the chosen binning strategy — prompted further investigation of the properties and range of potential applicability of the FIT. In the following sections, we demonstrated that the FIT yields reasonable results in a certain subset of possible time series, yet perhaps less expected results in others, when applied to a variety of series with different lengths, shapes and underlying selection coefficients.

The fundamental issue is that corpus data has to be operationalised one way or another if one is to apply a time series analysis that is based on variant frequencies. There is as yet no single best method to do so, and the additional researcher degree of freedom is practically unavoidable. Also, unlike microbial experimental data — for which the FIT was designed originally — the beginning and end of a corpus in terms of temporal coverage may not necessarily overlap with the beginning and end of a language change trajectory. The implications of these scenarios on the FIT approach were explored in Figures 2 and 4. Any test based on increment signatures is likely to miss a significant change, if it is recorded by very few data points. This could be either due to data sparsity or low number of bins, very high underlying selection, or the change happening in the middle of an otherwise long series. This could be remedied to an extent by only considering the bins of a corpus or the segments a time series where a change "looks like" it is taking place — but that introduces yet another parameter or researcher degree of freedom.

In what follows, we attempt to summarize our findings and distil them into actionable guidelines for applying tests of selection to linguistic corpus-derived time series.

5.1 Limitations for linguistic selection testing

Besides the fact that caution should be exercised when its statistical assumptions are not met (as with any statistical test), the following should be taken into account when applying the FIT or a similar test of selection to corpus data. *s* continues to refer to the selection coefficient driving the process of change (assuming an underlying Wright-Fisher like process; see Section 1.2 for related discussion). Obviously, a test of selection being carried out implies that *s* is actually unknown to the tester — the guidelines sketched here are meant to draw attention to situations where it might be beneficial to inspect the results more carefully. In terms of the input data quality, the results of a test can be misleading if the time series:

- chronicle only a part of a change (beginning or end);
- are too short (too few data points or bins);
- are too long (if covering multiple events, variable *s*);
- based on greatly variable bin sizes (avoidable with variable-width binning, which leads to variable bin lengths).

In terms of the types and shapes of possible series, binning can lead to unpredictable results in the case of FIT (and its assumption of increment normality is likely violated) in time series:

- which are S-curves (non-normal increments);
- where *s* may be suspected to vary over time (e.g. S-curves with long tails);
- where $s = 0$ (binning increases false positives);
- with a very high *s* (sharp changes, quick fixation);
- with tiny near-boundary fluctuations;
- where such values are introduced by smoothing (absorption adjustment).

The high *s* and absorption issue can be avoided by either excluding any series with a long span of zeroes or by making a choice to clip the post-absorption part of the series. That may leave a variable number of very few data points, and of course requires some consistent method of choosing the clipping point. The tiny fluctuations issue is typically caused by occasional occurrences of the less popular variant of a pair or set with a very high underlying total token frequency. Such series can be avoided by checking for the normality of increments.

As exemplified in this contribution, the way data is handled can in some cases drive the results of a test of selection. An application of such a test — particularly if it is borrowed from a different domain — should thus take into account the nature of the data. In the case of time series derived from diachronic corpora, a number of issues require attention. These include corpus size and normalisation (Gries 2010), quality of corpus tagging (cf. Supplementary appendix), genre (Szmrecsanyi 2016) and topic (Karjus et al. 2020) dynamics, representativeness and composition (Lijffijt, Säily & Nevalainen 2012; Pechenick, Danforth & Dodds 2015; Koplenig 2017). For example, imbalances in genre or register can easily lead to a drifty-looking series, if the usage of a variant differs between them. It is also not clear how the interplay of multiple, possibly opposing sources of selection (inherent properties of the variant, sociolinguistic prestige, top-down language planning, etc.) could be captured by a single test. Properties inherent to language can make a difference, such as the aforementioned re-use of archaic variants from the written record (Section 3), or meaning change, which may reasonably resolve competition between variants as they go on to inhabit different niches (automatic methods exist to detect the latter, cf. Dubossarsky et al. 2019). This relates to the issue of determining what variants do and which do not actually compete with one other for the same meaning or function, often referred to in sociolinguistics as the problem of the envelope of variation (cf. Walker 2010).

5.2 Opportunities for linguistic selection testing

On the bright side, despite these concerns, the Frequency Increment Test and presumably similar tests are likely reliably applicable to time series derived from linguistic corpus data when:

- the series covers the entire change (yet if possible also excludes near-boundary values);
- the assumptions of the test are checked for;
- the underlying *s* can be assumed to be constant;

- the interplay of *s* ranges and binning is taken into account (simulations help);
- the corpus is large, representative and consistently balanced over time for genre, style and topics;
- the target token count for each time bin is large (≳100, cf. the Appendix);
- the semantics of the pair (or set) of variants remain the same;
- the set of variants yielding the relative frequencies can be assumed to be competing, and the set contains all the competitors for a meaning or function.

Besides these rules of thumb, it would be beneficial in most cases to have some principled mechanisms to:

- evaluate multiple possible binning choices for the robustness of the test results;
- deal with the "leftover" flat part of the series before and after the change being analysed;
- distinguish drift and the effects of variable *s* over time.

Possible use cases in linguistics involving the FIT (or a similar test) presumably fall on a spectrum where on the one end the subject of a study would be a single change in the history of a language, and the aim would be to determine if that change has occurred due to drift or due to individuals consistently selecting for one of the variants, owing to its perceived higher fitness. On the other end would be the evaluation of a very large set of linguistic time series derived from a corpus, with the aim to reveal general patterns and dynamics of language change processes. The study of 36 English verbs by Newberry et al. falls closer towards this end of the spectrum.

When the subject of a study is a single change (or a few), and the result hinges on a single test result, then we would naturally advise to take the preceding concerns into careful consideration, from data sampling and preparation to the specifics of a given selection test, while being mindful of the involved researcher degrees of freedom. If a study veers toward the other end of the spectrum, involving a large set of series, then its design would largely come down to a choice between two approaches.

One could either take a "big-data" approach, feeding the test with a very large set of time series to explore the role of selection and drift in language change, checking for only the minimal statistical assumptions of the test. The upside is that, hopefully, despite the concerns specific to corpora and language, true patterns would emerge, given enough data. The downside is of course the danger of garbage in, garbage out.

Or alternatively, one could take the approach of also trying to check for the various linguistic assumptions in addition to the statistical ones, filtering out unsuitable series. This would hopefully lead to better language science. On the downside, this requires the meticulous introduction of a number of extra parameters, or researcher degrees of freedom. Furthermore, the results might not be representative of general language change dynamics in the end, if based on testing only a niche subset of series "suitable" for a given test — of which there might not be that many either. In other words, no free lunch.

5.3 Future prospects

The multitude of points listed above might sound like a lot of limitations. However, we would not by any means conclude that efforts to detect selection in linguistic data should be abandoned. The idea of detecting selection in diachronic linguistic data based on shapes or signatures is not new and remains an open challenge (Bentley 2008; Reali & Griffiths 2010; Blythe 2012; Sindi & Dale 2016; Amato et al. 2018). At the same time, methods for detecting selection continue being improved in the field of population genetics (Nishino

2013; Terhorst, Schlötterer & Song 2015; Schraiber, Evans & Slatkin 2016; Iranmehr et al. 2017; Taus, Futschik & Schlötterer 2017; Vlachos & Koer 2018).

Perhaps it would be useful to draw a distinction between exploratory and confirmatory findings. In essence, this strand of research (including Newberry et al.) has remained exploratory. Simulations with controlled properties allow for an evaluation of the performance of a test or model under various conditions and suspected confounds (cf. also Kauhanen 2017). However, to the best of our knowledge, there is currently no objective way to evaluate such methods or compare their accuracy against one another, in terms how well they reflect the actual selection biases operating on the level of the speaker, that may eventually give rise to a change in the consensus on the population level — a sample of which is (the only thing that is) eventually observable in a diachronic corpus. It would therefore be useful to distinguish between approaches that *test* for selection, and those that more accurately generate (albeit potentially interesting and worthwhile) hypotheses. The latter may be useful e.g. when positing causes of language change — be they linguistic, social, or cognitive in nature. If drift cannot be rejected, then theorising about possible "causes" of the change is unnecessary.

The difficulties with binning suggest that trying to manipulate the data to make it look more like the underlying Wright-Fisher model — i.e., coarse-graining individual instances of use to construct the continuously-varying variant frequencies that the model predicts — is not the way to go. An alternative procedure would be to include the process of sampling these instances of use to build the corpus as part of the model. For example, given some time series $x(t)$ generated by the Wright-Fisher model, then at an instant t this model says that we should expect to encounter one of the two word variants with probability $x(t)$. In an ideal world, one would then maximise the likelihood of the observed sequence of tokens with respect to the parameters of the Wright-Fisher model (i.e., the selection strength and effective population size). This procedure looks to be somewhat computationally demanding, and may prove intractable for large corpora. However, such a procedure could in principle be applied to token counts as they appear in a corpus, without the need for pre-processing (such as binning) and the researcher freedom associated with it.

Another domain besides language which has attracted similar genetics-inspired modelling approaches is that of archaeology, particularly datasets of (pre-)historical artefacts (Bentley & Shennan 2003). Similar concerns have followed: "time-averaged assemblages" of variants in cumulative cultural evolution (essentially binned data) can easily introduce bias in various tests (Premo 2014; Crema, Kandler & Shennan 2016). Diachronic datasets (e.g. those based on the archaeological record, but similarly, corpora) only provide sparse, aggregated frequency information, which may be the reflection of a variety of neutral or selective transmission processes at the individual level (Premo 2014; Crema, Kandler & Shennan 2016; Kandler, Wilder & Fortunato 2017; Kandler & Crema 2019). Since these underlying processes cannot be directly observed (particularly in prehistoric data), Kandler, Wilder & Fortunato (2017) suggest shifting the focus from identifying the single individual-level process that likely produced the observed data — to excluding those that likely did *not*. A corpus being a sample of individual utterances, this suggestion is worth consideration. Although the written record tends to have more metadata than the archaeological, the author of an utterance, along with their selective biases, is often unknown.

Detecting signatures of selection and drift in the evolution of language (and other domains of cumulative culture) remains an interesting prospect. It would be informative to see a comparison of the FIT-like selection detection methods that have been developed in population genetics or archaeology, applied to linguistic data, and systematically evaluated. If the issues listed in the sections above could be solved, then this would certainly improve possibilities for exciting linguistic inquiry, inviting answers to questions such as, do lexemes experience stronger drift than syntactic constructions? What is the relationship of selection and niche (Laland, Odling-Smee & Feldman 2001; Altmann, Pierrehumbert & Motter 2011) in language change? Are some parts of speech more susceptible to change via selection than others? (M. Newberry, p.c.) What is the role of drift in creole evolution? (Strimling, Jansson & Parkvall 2015) In semantic change? (Hamilton, Leskovec & Jurafsky 2016) Are some languages changing more due to drift than others? (and if that relates to community size; Atkinson, Kirby & Smith 2015; Reali, Chater & Christiansen 2018) Can different types of selection be distinguished, e.g. top-down planning, grassroots (Amato et al. 2018), momentum-driven (Stadler et al. 2016)?

6 Conclusions

We find ourselves witnessing an exciting time for linguistic research, where more and more data on actual language usage is becoming available, encompassing different languages, dialects, registers, modalities, but also centuries. At the same time computational means for analysing big data have become readily accessible, hand in hand with the development of methods providing new insight into how languages function, change and evolve over time. Alongside and perhaps interlinked with these developments, language as a domain of scientific investigation has attracted interest in recent decades from fields traditionally not engaged in linguistic research, such as physics and biology.

We evaluated the proposal of Newberry et al. (2017), consisting of the application of the Frequency Increment Test as a method for determining whether any time series constructed from corpus frequencies of competing variants is a case of selection or a case of change stemming from stochastic drift. We found that while some of the original results remain robust to binning choices, other do not. Based on constructed and simulated examples, we find that while the results of the FIT can be robust given a subset of suitable series, there are scenarios where they affected by the way the diachronic corpus data are binned.

We advocate that in the interest of reproducibility, binning, like any other data manipulation and operationalisation procedures, should be explicitly described in a contribution (as it is by Newberry et al.) — but additionally, if the results change given different choices, this should also be reported. Beyond data operationalisation, we drew attention to issues specific to linguistic data that should be taken into account to ensure quality of testing results, as well as to work in cultural evolution where it has been shown that the inference of individual transmission processes from population-level frequency aggregates is susceptible to error and should be handled with care.

To conclude, identifying the role and prevalence of stochastic drift in language change is an important goal, but our results suggest that great care should be exercised when applying such tests to linguistic data, in order for the results to not be biased by issues specific to the domain as well as properties of a particular test.

Data Accessibility Statement

The R code we used to replicate the results of the original paper is available at https://github.com/mnewberry/ldrift, and the corpus at https://corpus.byu.edu/coha. The code to run the simulations described is this paper is available at https://github.com/andreskarjus/wfsim_fit.

Additional File

The additional file for this article can be found as follows:

- **Supplementary file 1.** Supplementary appendix to "Challenges in detecting evolutionary forces in language change using diachronic corpora". DOI: https://doi.org/10.5334/gjgl.909.s1

Acknowledgements

The authors would like to thank Mitchell Newberry for discussion and comments on the paper that led to significant improvements and revisions, the anonymous reviewers for useful comments, and Alison Feder for providing an implementation of the FIT earlier when one was not yet publicly available.

Funding Information

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.

Competing Interests

The authors have no competing interests to declare.

Author Contributions

Andres Karjus carried out the research and wrote the paper. Richard A. Blythe, Simon Kirby and Kenny Smith provided revisions, comments and feedback on the design of the research and the paper.

References

Ahern, Christopher A., Mitchell G. Newberry, Robin Clark & Joshua B. Plotkin. 2016. Evolutionary forces in language change. *ArXiv e-prints*. (5 July, 2017).

Altmann, Eduardo G., Janet B. Pierrehumbert & Adilson E. Motter. 2011. Niche as a determinant of word fate in online groups. *PLOS ONE* 6(5). 1–12. DOI: https://doi.org/10.1371/journal.pone.0019009

Amato, Roberta, Lucas Lacasa, Albert Díaz-Guilera & 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. DOI: https://doi.org/10.1073/pnas.1721059115

Andersen, Henning. 1990. The structure of drift. In Henning Andersen & Konrad Koerner (eds.), *Historical Linguistics 1987. Papers from the 8th International Conference on Historical Linguistics*, 1–20. Amsterdam: Benjamins.

Anderwald, Lieselotte. 2012. Variable past-tense forms in nineteenth-century American English: Linking Normative Grammars and language change. *American Speech* 87(3). 257–293. (18 September, 2019). DOI: https://doi.org/10.1215/00031283-1958327

Atkinson, Mark, Simon Kirby & Kenny Smith. 2015. Speaker input variability does not explain why larger populations have simpler languages. *PLOS ONE* 10(6). 1–20. DOI: https://doi.org/10.1371/journal.pone.0129463

Baxter, G. J., R. A. Blythe, W. Croft & A. J. McKane. 2006. Utterance selection model of language change. *Physical Review E* 73(4). 046118. DOI: https://doi.org/10.1103/PhysRevE.73.046118

Bentley, R. Alexander. 2008. Random drift versus selection in academic vocabulary: an evolutionary analysis of published keywords. *PLOS ONE* 3(8). 1–7. DOI: https://doi.org/10.1371/journal.pone.0003057

Bentley, R. Alexander & Stephen J. Shennan. 2003. Cultural transmission and stochastic network growth. *American Antiquity* 68(3). 459–485. DOI: https://doi.org/10.2307/3557104

Blythe, Richard A. 2012. Neutral evolution: A null model for language dynamics. *Advances in complex systems* 15(3–4). DOI: https://doi.org/10.1142/S0219525911003414

Blythe, Richard A. & William Croft. 2012. S-curves and the mechanisms of propagation in language change. *Language* 88(2). 269–304. (5 July, 2017). DOI: https://doi.org/10.1353/lan.2012.0027

Calude, Andreea S., Steven D. Miller & Mark Pagel. 2017. Modelling loanword success a sociolinguistic quantitative study of Māori loanwords in New Zealand English. *Corpus Linguistics and Linguistic Theory*, 1–38. DOI: https://doi.org/10.1515/cllt-2017-0010

Crema, Enrico R., Anne Kandler & Stephen Shennan. 2016. Revealing patterns of cultural transmission from frequency data: Equilibrium and nonequilibrium assumptions. *Scientific reports* 6. 39122. DOI: https://doi.org/10.1038/srep39122

Croft, W. 2000. *Explaining language change: An evolutionary approach*. Longman.

Cuskley, Christine F., Martina Pugliese, Claudio Castellano, Francesca Colaiori, Vittorio Loreto & 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. DOI: https://doi.org/10.1371/journal.pone.0102882

Daoust, Demise. 2017. Language planning and language reform. In *The handbook of sociolinguistics*, 436–452. Wiley-Blackwell. DOI: https://doi.org/10.1002/9781405166256.ch27

Davies, Mark. 2010. *The Corpus of Historical American English (COHA): 400 million words, 1810–2009*. Available online at https://www.englishcorpora.org/coha.

Dingemanse, Mark, Damián E. Blasi, Gary Lupyan, Morten H. Christiansen & Padraic Monaghan. 2015. Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences* 19(10). 603–615. DOI: https://doi.org/10.1016/j.tics.2015.07.013

Dubossarsky, Haim, Simon Hengchen, Nina Tahmasebi & Dominik Schlechtweg. 2019. Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 457–470. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/P19-1044

Enfield, N. J. 2014. Transmission biases in the cultural evolution of language: towards an explanatory framework. In Daniel Dor, Chris Knight & Jerome Lewis (eds.), *The social origins of language*. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199665327.003.0023

Ewens, Warren J. 2004. *Mathematical population genetics 1: Theoretical introduction* (Interdisciplinary Applied Mathematics). New York: Springer. DOI: https://doi.org/10.1007/978-0-387-21822-9

Fay, Nicolas, Simon Garrod, Leo Roberts & Nik Swoboda. 2010. The interactive evolution of human communication systems. *Cognitive science* 34(3). 351–386. DOI: https://doi.org/10.1111/j.1551-6709.2009.01090.x

Feder, Alison F., Sergey Kryazhimskiy & Joshua B. Plotkin. 2014. Identifying signatures of selection in genetic time series. *Genetics* 196(2). 509–522. (5 July, 2017). DOI: https://doi.org/10.1534/genetics.113.158220

Ghanbarnejad, Fakhteh, Martin Gerlach, José M. Miotto & Eduardo G. Altmann. 2014. Extracting information from S-Curves of language change. *Journal of The Royal Society Interface* 11(101). DOI: https://doi.org/10.1098/rsif.2014.1044

Gray, Tyler J., Andrew J. Reagan, Peter Sheridan Dodds & Christopher M. Danforth. 2018. English verb regularization in books and tweets. *ArXiv e-prints*. DOI: https://doi.org/10.1371/journal.pone.0209651

Gries, Stefan Th. 2010. Useful statistics for corpus linguistics. *A mosaic of corpus linguistics: Selected approaches* 66. 269–291.

Grieve, Jack, Andrea Nini & Diansheng Guo. 2018. Mapping lexical innovation on American social media. *Journal of English Linguistics* 46(4). 293–319. DOI: https://doi.org/10.1177/0075424218793191

Hahn, Matthew W. & 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. S120–S123. DOI: https://doi.org/10.1098/rsbl.2003.0045

Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing 2016*. 2116–2121. DOI: https://doi.org/10.18653/v1/D16-1229

Haspelmath, Martin. 1999. Optimality and diachronic adaptation. *Zeitschrift für Sprachwissenschaft* 18(2). 180–205. DOI: https://doi.org/10.1515/zfsw.1999.18.2.180

Hernández-Campoy, Juan Manuel & Juan Camilo Conde-Silvestre. 2012. *The handbook of historical sociolinguistics*. Wiley-Blackwell. DOI: https://doi.org/10.1002/9781118257227

Iranmehr, Arya, Ali Akbari, Christian Schlötterer & Vineet Bafna. 2017. CLEAR: Composition of likelihoods for evolve and resequence experiments. *Genetics* 206(2). 1011–1023. (5 July, 2017). DOI: https://doi.org/10.1534/genetics.116.197566

Jespersen, Otto. 1922. *Language, its nature, development, and origin*. H. Holt.

Kandler, Anne & Enrico R. Crema. 2019. Analysing cultural frequency data: Neutral theory and beyond. In Anna Marie Prentiss (ed.), *Handbook of evolutionary research in archaeology*, 83–108. Cham: Springer International Publishing. DOI: https://doi.org/10.1007/978-3-030-11117-5

Kandler, Anne, Bryan Wilder & Laura Fortunato. 2017. Inferring individuallevel processes from population-level patterns in cultural evolution. *Royal Society Open Science* 4(9). DOI: https://doi.org/10.1098/rsos.170949

Kanwal, Jasmeen, Kenny Smith, Jennifer Culbertson & 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. DOI: https://doi.org/10.1016/j.cognition.2017.05.001

Karjus, Andres, Richard A. Blythe, Simon Kirby & Kenny Smith. 2020. Quantifying the dynamics of topical fluctuations in language. *Language Dynamics and Change*, 1–40. DOI: https://doi.org/10.1163/22105832-01001200

Kauhanen, Henri. 2017. Neutral change. *Journal of Linguistics* 53(2). 327–358. DOI: https://doi.org/10.1017/S0022226716000141

Kershaw, Daniel, Matthew Rowe & Patrick Stacey. 2016. Towards modelling language innovation acceptance in online social networks. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM '16)*. 553–562. ACM. DOI: https://doi.org/10.1145/2835776.2835784

Kirby, Simon, Hannah Cornish & 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. DOI: https://doi.org/10.1073/pnas.0707835105

Koplenig, Alexander. 2017. 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. DOI: https://doi.org/10.1093/llc/fqv037

Kroch, Anthony & Ann Taylor. 2000. *The Penn-Helsinki Parsed Corpus of Middle English (PPCME2)*. Department of Linguistics, University of Pennsylvania.

Labov, W. 2011. *Principles of linguistic change, volume 3: Cognitive and cultural factors* (Language in Society). Wiley-Blackwell. DOI: https://doi.org/10.1002/9781444327496

Laland, K. N., J. Odling-Smee & M. W. Feldman. 2001. Cultural niche construction and human evolution. *Journal of Evolutionary Biology* 14(1). 22–33. DOI: https://doi.org/10.1046/j.1420-9101.2001.00262.x

Lieberman, Erez, Jean-Baptiste Michel, Joe Jackson, Tina Tang & Martin A. Nowak. 2007. Quantifying the evolutionary dynamics of language. *Nature* 449(7163). 713–716. DOI: https://doi.org/10.1038/nature06137

Lijffijt, Jefrey, Tanja Säily & Terttu Nevalainen. 2012. CEECing the baseline: lexical stability and significant change in a historical corpus. In Jukka Tyrkkö, Matti Kilpiö, Terttu Nevalainen, Matti Rissanen (ed.), *Outposts of historical corpus linguistics: From the Helsinki Corpus to a proliferation of resources* (Studies in Variation, Contacts and Change in English 10). Helsinki: Research Unit for Variation, Contacts and Change in English (VARIENG).

Malaspinas, Anna-Sapfo. 2016. Methods to characterize selective sweeps using time serial samples: An ancient DNA perspective. *Molecular Ecology* 25(1). 24–41. DOI: https://doi.org/10.1111/mec.13492

McMahon, April M. S. 1994. *Understanding language change*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139166591

Monaghan, Padraic & Seán G. Roberts. 2019. Cognitive inuences in language evolution: Psycholinguistic predictors of loan word borrowing. *Cognition* 186. 147–158. DOI: https://doi.org/10.1016/j.cognition.2019.02.007

Newberry, Mitchell G., Christopher A. Ahern, Robin Clark & Joshua B. Plotkin. 2017. Detecting evolutionary forces in language change. *Nature* 551(7679). 223–226. DOI: https://doi.org/10.1038/nature24455

Nishino, Jo. 2013. Detecting selection using time-series data of allele frequencies with multiple independent reference loci. *G3: Genes, Genomes, Genetics* 3(12). 2151–2161. DOI: https://doi.org/10.1534/g3.113.008276

Ohala, John J. 1983. The origin of sound patterns in vocal tract constraints. In *The production of speech*, 189–216. New York, NY: Springer. DOI: https://doi.org/10.1007/978-1-4613-8202-7_9

Pechenick, Eitan Adam, Christopher M. Danforth & Peter Sheridan Dodds. 2015. Characterizing the Google Books Corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLoS ONE* 10(10). e0137041. DOI: https://doi.org/10.1371/journal.pone.0137041

Premo, L. S. 2014. Cultural transmission and diversity in time-averaged assemblages. *Current Anthropology* 55(1). 105–114. DOI: https://doi.org/10.1086/674873

Reali, Florencia, Nick Chater & Morten H. Christiansen. 2018. Simpler grammar, larger vocabulary: How population size affects language. *Proceedings of the Royal Society of London B: Biological Sciences* 285(1871). DOI: https://doi.org/10.1098/rspb.2017.2586

Reali, Florencia & Thomas 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. (8 July, 2017). DOI: https://doi.org/10.1098/rspb.2009.1513

Rubin, Joan, Björn H. Jernudd, Jyotirindra DasGupta, Joshua A. Fishman & Charles A. Ferguson. 1977. *Language planning processes* (Contributions to the Sociology of Language). Mouton. DOI: https://doi.org/10.1515/9783110806199

Sapir, Edward. 1921. *Language. An introduction to the study of speech*. New York: Harcourt, Brace and Company.

Schraiber, Joshua G., Steven N. Evans & Montgomery Slatkin. 2016. Bayesian inference of natural selection from allele frequency time series. *Genetics*. DOI: https://doi.org/10.1534/genetics.116.187278

Simmons, Joseph P., Leif D. Nelson & Uri Simonsohn. 2011. False-positive psychology: Undisclosed exibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22(11). 1359–1366. DOI: https://doi.org/10.1177/0956797611417632

Sindi, Suzanne S. & 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. DOI: https://doi.org/10.1016/j.jtbi.2015.12.012

Smith, Kenny, Monica Tamariz & Simon Kirby. 2013. Linguistic structure is an evolutionary trade-off between simplicity and expressivity. In Markus Knauff, Michael Pauen, Natalie Sebanz & Ipke Wachsmuth (eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, 1348–1353. Cognitive Science Society.

Stadler, Kevin, Richard A. Blythe, Kenny Smith & Simon Kirby. 2016. Momentum in language change: A model of self-actuating S-shaped curves. *Language Dynamics and Change* 6(2). 171–198. (5 July, 2017). DOI: https://doi.org/10.1163/22105832-00602005

Steels, Luc & Eörs Szathmáry. 2018. The evolutionary dynamics of language. Biosystems 164. 128–137. DOI: https://doi.org/10.1016/j.biosystems.2017.11.003

Strimling, Pontus, Fredrik Jansson & Mikael Parkvall. 2015. Modeling the evolution of creoles. *Language Dynamics and Change* 5(1). 1–51. (5 July, 2017). DOI: https://doi.org/10.1163/22105832-00501005

Szmrecsanyi, Benedikt. 2016. About text frequencies in historical linguistics: Disentangling environmental and grammatical change. *Corpus Linguistics and Linguistic Theory* 12(1). 153–171. DOI: https://doi.org/10.1515/cllt-2015-0068

Tamariz, Monica, T. Mark Ellison, Dale J. Barr & Nicolas Fay. 2014. Cultural selection drives the evolution of human communication systems. *Proceedings of the Royal Society B: Biological Sciences* 281(1788). 20140488. DOI: https://doi.org/10.1098/rspb.2014.0488

Taus, Thomas, Andreas Futschik & Christian Schlötterer. 2017. Quantifying Selection with Pool-Seq Time Series Data. *Molecular Biology and Evolution* 34(11). 3023–3034. DOI: https://doi.org/10.1093/molbev/msx225

Terhorst, Jonathan, Christian Schlötterer & Yun S. Song. 2015. Multi-locus analysis of genomic time series data from experimental evolution. *PLoS genetics* 11(4). e1005069. DOI: https://doi.org/10.1371/journal.pgen.1005069

Trask, Robert Lawrence. 1996. Historical linguistics. London: Arnold.

Van de Velde, Freek. 2014. Degeneracy: The maintenance of constructional networks. In *The extending scope of construction grammar* 54. 141–179. Berlin/Boston: Walter De Gruyter GmbH.

Vlachos, Christos, Claire Burny, Marta Pelizzola, Rui Borges, Andreas Futschik, Robert Koer & Christian Schlötterer. 2019. Benchmarking software tools for detecting and quantifying selection in evolve and resequencing studies. *Genome Biology* 20(1). 169. DOI: https://doi.org/10.1186/s13059-019-1770-8

Vlachos, Christos & Robert Kofler. 2018. MimicrEE2: Genome-wide forward simulations of Evolve and Resequencing studies. *PLOS Computational Biology* 14(8). 1–10. DOI: https://doi.org/10.1371/journal.pcbi.1006413

Walker, James A. 2010. *Variation in linguistic systems*. New York: Routledge.

Wright, Sewall. 1931. Evolution in Mendelian populations. *Genetics* 16(2). 97–159.

Zipf, George Kingsley. 1949. *Human behavior and the principle of least effort: An introduction to human ecology*. Reading, MA: Addison-Wesley Press.

How to cite this article: Karjus, Andres, Richard A. Blythe, Simon Kirby and Kenny Smith. 2020. Challenges in detecting evolutionary forces in language change using diachronic corpora. *Glossa: a journal of general linguistics* 5(1): 45. 1–25. DOI: https://doi.org/10.5334/gjgl.909

Submitted: 31 January 2019 **Accepted:** 31 October 2019 **Published:** 07 May 2020

]u[*Glossa: a journal of general linguistics* is a peer-reviewed open access journal published by Ubiquity Press. OPEN ACCESS

Supplementary appendix to:
Challenges in detecting evolutionary forces in language change using diachronic corpora

Andres Karjus[1], Richard A. Blythe[1,2], Simon Kirby[1], Kenny Smith[1]
[1] Centre for Language Evolution, School of Philosophy, Psychology and Language Sciences,
University of Edinburgh; [2]School of Physics and Astronomy, University of Edinburgh

This appendix expands on the main text, providing additional information, technical details, and further exploration of the parameter spaces of the models.

A note on corpus annotation quality

While not discussed at length in the main text, the quality of corpus annotation such as lemmatization and part-of-speech tagging plays an equally important role in addition to other corpora-related issues mentioned in the Discussion. Studying the large-scale usage of any linguistic elements of interest relies on the identification of relevant targets in a corpus. Too many erroneously extracted examples can mislead the results. Among the 36 verbs in the sample of Newberry et al, this is especially pertinent for homonymous words like *wet* and *wed*. We already discussed the adjectival usage of *spilt* above. We also found that, for example, 44% of the extracted examples of *wet*.PAST in the first bin (1812-1875 in COHA, under the variable-width binning procedure) were cases of erroneous tagging — being instead other non-past forms of *wet* and occurrences of the adjective *wet*. The same issue applies to *wed*, in addition to being confused with the abbreviation for Wednesday.

Results based on a different minimal frequency threshold

Figure S1 is intended to complement Figure 1, where we applied a minimal frequency threshold of 10 in each bin (this is mostly relevant for fixed-width binning, as variable-width ensures largely similar bin sizes). Since this is an arbitrary threshold, we also tried a more conservative value of minimal 100 occurrences per bin (for a bin to be included in the time series), with the results reflected in Figure S1. In summary, the higher threshold does not change the results for variable-width binning, besides some lower-frequency verbs being excluded (the empty lower left corner). In fixed-width binning, some results change, e.g. *spill* is now always flagged as drift, while *burn*, *dive* and *quit* get flagged as selection.

Results of no binning (i.e. using default COHA 1-year bins) should still be taken with a pinch of salt, even when the normality assumption is now met (circles instead of squares) — removing bins with less than 100 tokens leaves even medium-frequency verbs with only a few bins (e.g., 5 in the case of *light*, spread uneven across 200 years; observe also the median bins-to-years ratio of 0.55).

The fact that the minimal threshold affects fixed binning more is not surprising, as the frequencies vary more. This makes variable-width binning a more attractive solution, but

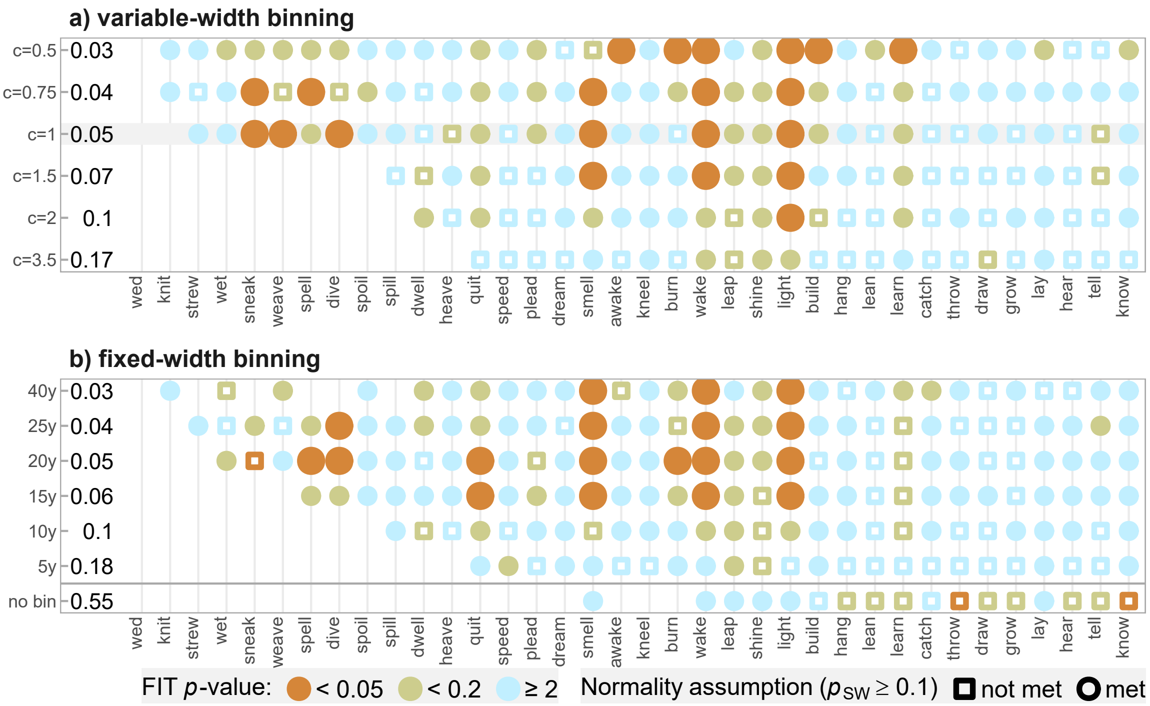


Figure S1: Results of applying the FIT to time series constructed based on 200 years of COHA frequency data. The interpretation of this figure is the same as that of Figure 1 the only difference being the increased minimal within-bin frequency threshold of 100. The constant c determines the number of variable length bins via $n(b) = c\ln(n(v))$. Thus "$c = 1$" corresponds to Newberry et al.'s original results (highlighted with the horizontal grey line). 10y corresponds to fixed bin length of 10 years, etc; 'no bin' refers to no additional binning on top of the default yearly bins in the corpus

its different behaviour should also be taken into consideration. Should the overall frequency of a pair (or set) of variants change over the course of the corpus, it will end up with more bins over the more frequent end of the time scale. As COHA is not uniform in size across time, having considerably less data per year in the first few decades, time series based on variable-width binning of COHA data systematically have longer segments in the beginning and shorter ones towards the end. The "long bins" allow for drawing time series over more sparse corpus segments, where fixed binning would yield unreliably small or empty bins. At the same time, variable-width may by nature gloss over some fluctuations (characteristic of drift) while making a series look more smooth (more characteristic of selection).

Results based on series of different lengths

This figure is intended to complement the simulation section in the main text, which focused on the results of binning a 200-length series into shorter series. Here, no binning is being applied. Figure S2 shows that the FIT produces somewhat different results with the same s given series of different lengths, as expected: when the selection signal is strong enough to be detected (above ~ 0.02), then it is easier to detect it in longer series with more data points than in shorter series. Regardless of series length (at least up to the 200), the false

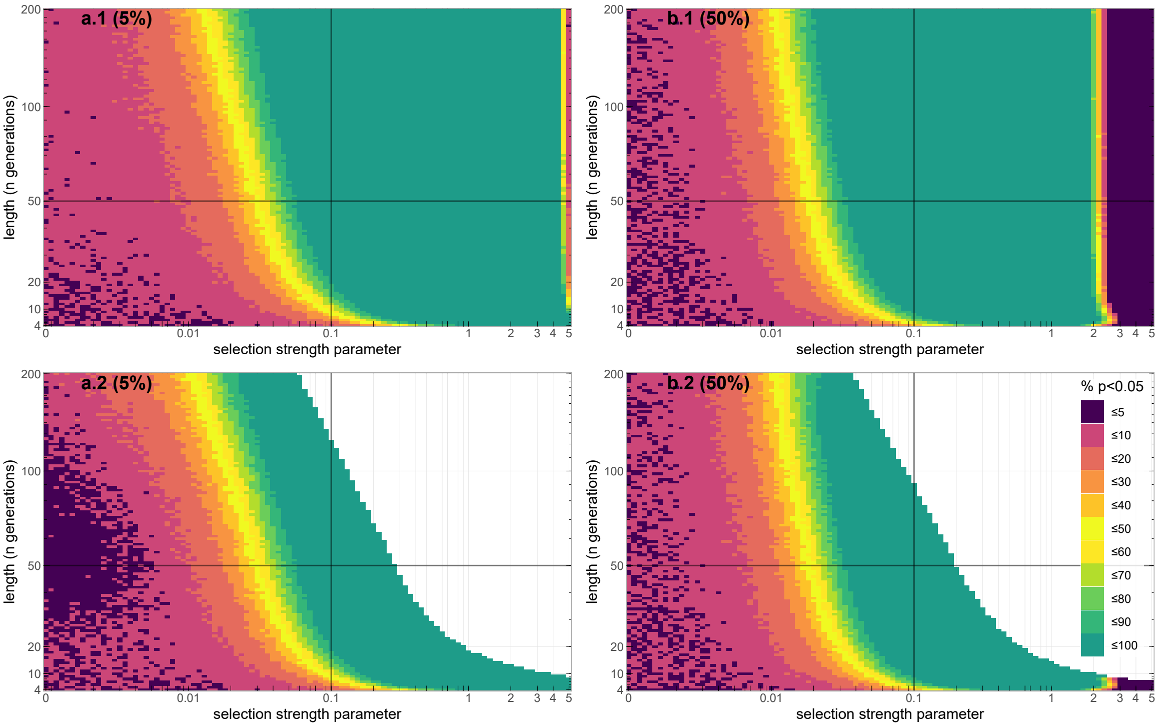


Figure S2: The effect of the interplay of time series length and selection strength s on the results of the FIT. The percentage of FIT $p < 0.05$ (out of 1000 replications for each combination) is reported for a range of time series lengths (y-axis, $[4, 200]$, note the log scale) and the same range of s as above. The left side pair (a) illustrates the case of the time series starting out at 5%, with the 50% condition on the right (b). In the bottom panels (a.2, b.2), series with a Shapiro-Wilk $p < 0.1$ are removed before calculating the percentage. This figure further illustrates the interplay of series length and s that affect the results of FIT

positive rate stays in $[0.03, 0.07]$ (Figure S2.b). This also shows that the higher false positive rate under binning shown in Section 4 does originate in the binning process (which smooths out small fluctuations) rather than simply length difference (binning naturally also making a series shorter).

More examples of the selection coefficient

Figure S3 is intended to complement Figures 3 and 5, where some example Wright-Fisher series where plotted. The s range in our experiments consisted of 200 equidistant values from a log scale between 0.001 and 5, with the addition of 0 in the beginning to be able to explore pure drift.

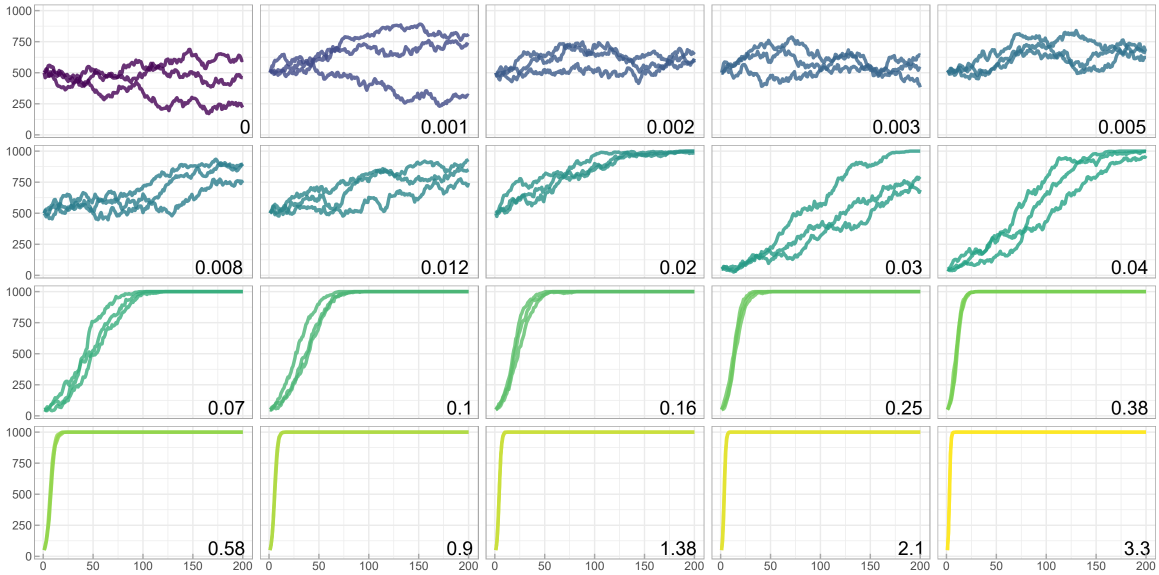


Figure S3: A visualization of the range of selection strength s values explored in the simulation section of this study (shown in the corner of each panel). The horizontal axis corresponds to population size (of the 'mutant' individuals), with time on the horizontal axis. Higher levels of s lead to the mutants taking over the population at faster rates

The increment normality assumption

The interpretation of the results of the FIT depends how stringently its assumption of the normality of the increments distribution is observed, particularly when s is high. In Figures 4 and 6 we used a Shapiro-Wilk test with a cut-off theshold of 0.1. We conducted additional simulations to see if a lower would yield qualitatively different results, and found it makes very little difference. We also tried using the Lilliefors-Kolmogorov-Smirnov test and the Anderson-Darling test and found all of them to be broadly in agreement: depending on the series starting point and chosen α, the increment normality assumption becomes violated as series (of length 200) approach the s range of 0.05..0.1, with the breaking point being somewhat lower on the s scale in non-binned series and higher in series binned into 10-15 bins (i.e. it is easier to meet the normality assumption if the series is binned).

Increment heteroskedasticity and the Fitness Increment Test

In this additional section, we shed some light on another mathematical aspect of the FIT, the homoskedasticity assumption, as the FIT is, in its core, a one-sample t-test for a zero mean under the assumption of normally-distributed increments with equal variance. For reference, this is the increment transformation process (cf. Section 1.2):

$$Y_i = \frac{v_i - v_{i-1}}{\sqrt{2v_{i-1}(1 - v_{i-1})(t_i - t_{i-1})}} \tag{1}$$

where v_i is the relative frequency of a variant in $(0, 1)$ at time t_i. The rationale behind this rescaling is that, under neutral evolution, the mean increment $v_i - v_{i-1}$ is zero, and its variance is proportional to

$$v_{i-1}(1-v_{i-1})(t_i-t_{i-1}) \tag{2}$$

However, here we are dealing with estimates of v_{i-1} and v_i obtained from finite samples of size M_{i-1} and M_i, respectively. This leads to additional contributions to the variance of the increment $v_i - v_{i-1}$, arising from the variance of the binomial distribution $v(1-v)/M$, where v is the mean value and M is the sample size. To a first approximation, the total variance of the increment is obtained by summing the three contributions. That is, $(v_i - v_{i-1})$ has a variance of

$$\frac{v_{i-1}(1-v_{i-1})(t_i-t_{i-1})}{N} + \frac{v_{i-1}(1-v_{i-1})}{M_{i-1}} + \frac{v_i(1-v_i)}{M_i} \ . \tag{3}$$

where N stands for effective population size. The transform divides all of this by $v_{i-1}(1-v_{i-1})(t_i-t_{i-1})$ which leads to a variance of each *transformed* increment Y_i of approximately

$$\frac{1}{N} + \frac{1}{M_{i-1}(t_i-t_{i-1})} + \frac{v_i(1-v_{i-1})}{M_i(t_i-t_{i-1})v_{i-1}(1-v_{i-1})} \ . \tag{4}$$

The FIT can be expected to perform as intended when this variance is constant. This is the case when $1/M \ll (t_i - t_{i-1})/N$ or $M(t_i - t_{i-1}) \gg N$ (assuming N can be inferred, which is not trivial, but cf. Newberry et al.). The transformed increments based on corpus data basically never have perfectly equal variance once sample size is taken into account, but will be roughly constant when the sample sizes are large (relative to N). The variable-width binning, as employed by Newberry et al., assures that the variances are more or less equal, as each bin has roughly the same number of tokens. The worry with fixed-width binning — including the default data binning of one year in COHA as well further binning of the years into decades and so on — is that the variance is not going to be equal, as bins may or may not cover a similar number of tokens.

We calculated these values for the English verb data (with the simplification of excluding N, which is not trivial to infer). Variable-width binning consistently yields small increment variances with a very small standard deviation (depending in turn on the variance in the bin sizes in the original data). Using the data without further binning (i.e. 1-year bins from COHA) yields multiple magnitudes higher values for both, as does fixed binning into short bins. But starting at decade-length bins (for higher-frequency verbs like *light*) and 20-year bins (for lower-frequency verbs like *spell*), as bin sizes approach 100 tokens, the picture becomes quite similar to variable-width binning.

It is not clear, however, how much heteroskedasticity is bad enough to lead to spurious results. For example, is it invalid to interpret the results of the FIT based on 1-year or 5-year bins at all, given typical sample sizes in a corpus like COHA? While this would benefit from more through future investigation, we attempt to shed some light on this by conducting more Wright-Fisher simulations where we manipulate the size of M in each generation, and the standard deviation of sample sizes (σ_S), as well as apply different binning strategies to the resulting time series. In Figure S4, the series length is 200 as in the previous simulations, $s = 0$ (as we are interested in the false positives rate), and we explore two N sizes, 10000 (left side column in Figure S4) and 1000 (right side). Each pixel represents the share of FIT $p < 0.05$ in 1000 replications with the given parameter combination.

For each replication in a combination, we run a Wright-Fisher simulation, but to construct the time series, take a random sample of individuals M at each of the 200 generations. The sample sizes are in turn generated by sampling values from a log-normal distribution with a mean of $\ln(M)$ (y-axis in Figure S4) and $\sigma_S \in \ln([1, 2])$ (corresponding to the x-axis in Figure S4). The log-normal distribution excludes 0, but with a high standard deviation, some of the generated M values can exceed that of the population size, so when reconstructing the time series, they are truncated by taking $\min(M, N)$. After this manipulation however, where M and V are both high, the resulting actual standard deviation across the 200 M sample sizes would not correspond to the predetermined parameter of σ_S, therefore, such replications are filtered out (along with series where the normality assumption is violated, at Shapiro-Wilk $p < 0.1$). When less than 10% of the replications for a combination are valid, it is excluded from plotting (the white areas in Figure S4).

The leftmost column of pixels on each panel corresponds to no variance in sample sizes, i.e., the samples are of equal size, corresponding exactly to the value on the y-axis. The top left pixel therefore represents the baseline Wright-Fisher simulation result with no downsampling (the false positive rate being around 5% in both N at $\alpha = 0.05$). Each top panel shows the results without binning, with the lower ones showing results when the series are binned into a smaller number of bins (after the aforementioned downsampling procedure).

In Figure S4, where cold blueish colours represent percentages of FIT $p < 0.05$. Ideally, as $s = 0$, all of the panels should be devoid of any warm colours. Looking at any single panel, the columns of pixels right of the no-variance leftmost column are not any more yellow than the leftmost column. This demonstrates that variance in sample sizes does not make any discernable difference — it does not make the already borderline false positive rate any worse. This observation holds between binning choices. Binning itself does increase the false positive rate, as already determined in Section 4 (panels below the top ones exhibit more yellow). If anything, it would seem series based on samples of size $M < N$ and increased σ_S have an improved (i.e. smaller) false positive rate, an effect particularly pronounced when the series are binned. This is however an expected result stemming from the added sampling noise (making any series look more "random" to the test).

The heteroskedasticity question remains somewhat unresolved, but based on these results we can say that at least the false positive rate of FIT is unlikely to be considerably affected by differing bin sizes. In terms practical guidelines, to be safe, if applicable variable-width binning should be used with FIT as proposed by Newberry et al. If fixed-width binning is used, then bins should consist of 100 occurrences or more. In the end this is not only a variance problem, but a small sample size problem. A large number of bins consisting each of only tens of occurrences has considerable sampling noise. Given the same corpus, a small number of bins consisting each of hundreds of occurrences can gloss over the true trajectory of change, but also any statistical test based on too few data points is unreliable. In other words, there's no data like more data.

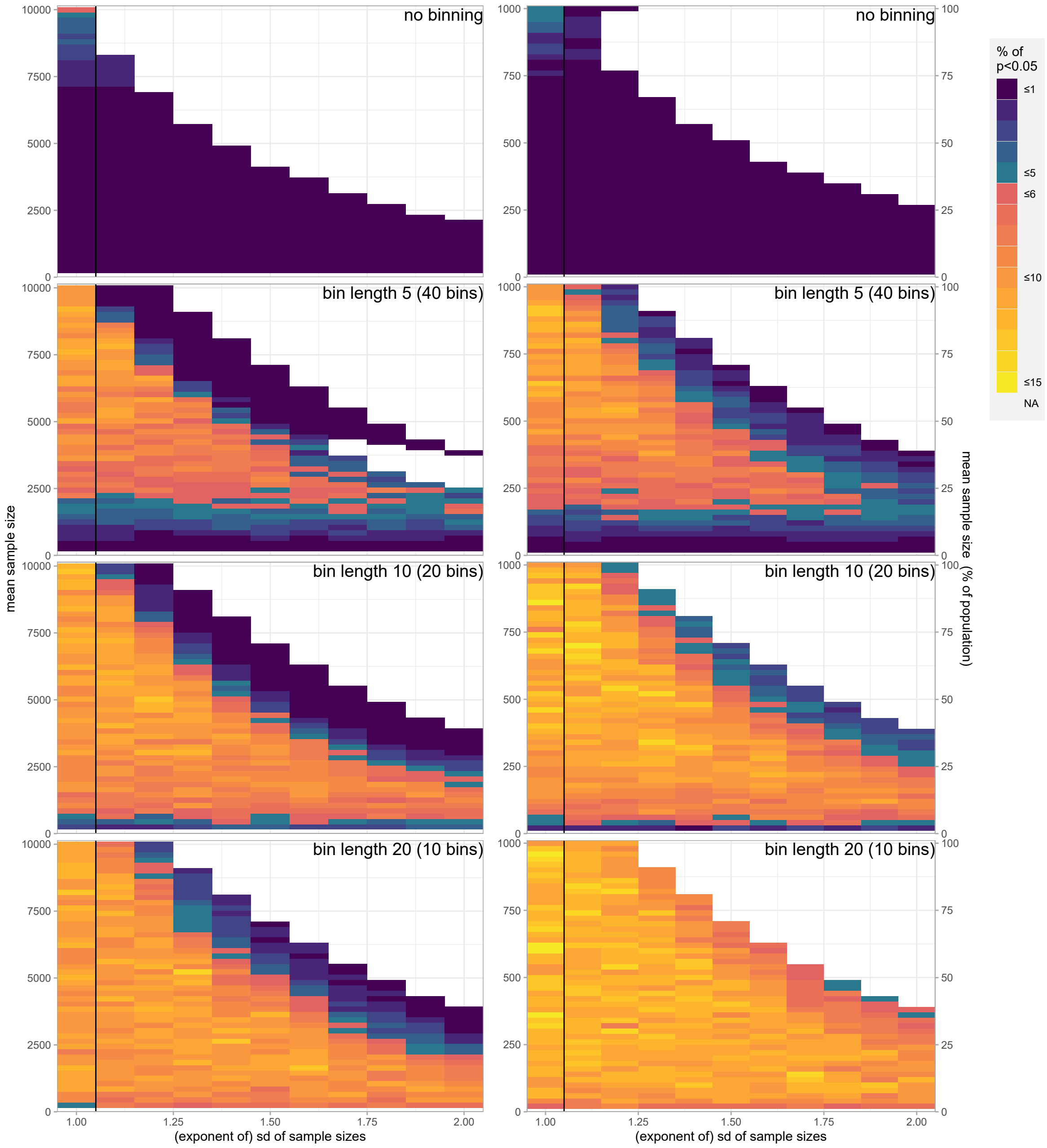


Figure S3: False positive rates of the FIT based on Wright-Fisher simulations with downsampled populations. Column of panels on the left: $N = 10000$. Panels on the right: $N = 1000$. The cool colours correspond to percentages of $p < 0.05$ below 5%, warm colours indicate higher percentages. This figure illustrates that while binning tends to introduce more false positives, in any given binning strategy, added variance in the underlying occurrence counts (and thus bin sizes) does not

2.3 Conclusions

The Frequency Increment Test, as discussed in this Chapter, is undoubtedly a promising approach and part of a larger family of potentially useful methods, but from the extensive evaluation carried out above, appears not quite mature yet to be widely applied to linguistics research. This study was nevertheless useful for informing technical decisions in later chapters, including those involving corpus binning, minimal term frequency in models, and drove the idea of evaluating the models by means of simulations (see Chapters 3, 4).

Finally, in the period between submitting this thesis and its defense, a paper (Karsdorp et al. 2020) came out that in turn replicates our work (Karjus et al. 2020a), and attempts to solve the issues we raised, by approaching the task as one of time series classification, and employing a neural network architecture to do so. While a longer review of this work is out of scope for this post-viva addition here, it does seem like a major step forward in terms of testing for selection in linguistic time series.

Chapter 3

Quantifying the dynamics of topical fluctuations in language

In this chapter, I review previous corpus-based research on frequency change in natural language, and go on to develop and evaluate a simple but effective computational model to quantify fluctuations in topics of conversation in diachronic corpora, with particular focus on the lexicon. The approach, termed the topical advection model, is based on word co-occurrence statistics, and captures these fluctuations as a weighted average of changes in related word frequencies.

I argue that the advection model can be used as a baseline in any predictive models concerned with usage frequency changes of linguistic elements (not just words). Changes is language usage concerns a variety of linguistic disciplines, including historical linguistics but also applied fields like lexicography. The advection model can also be used to adjust for the effect of topical prevalence in linguistic time series (the topic of Chapter 2). Finally, it can be considered as a proxy for changing communicative needs (see Chapter 4), the motivation being that increases and decreases in the prevalence of topics reflect the changing priorities and interests of speakers' communities over time. Using a sample of lexical innovations from American English, I show that new words are statistically more likely to be introduced when their associated topics, and therefore presumably the associated topical communicative needs, are on the rise.

I test two implementations for inferring the advection value of a given word, one based on a weighted list of most associated context words, and the other on Latent Dirichlet Allocation, a popular topic model. I find these to perform comparably, while the former has the advantage of simplicity and easier interpretability. I also evaluate the methodology by simulating plausible scenarios of language change using synthetic corpora, as well as by randomization tests, and conclude that the advection model reliably quantifies a meaningful aspect of language. The inherent downsides of binning corpora were discussed in Chapter 2, but the implementations of the advection model however do still rely on comparison between discrete subcorpora or bins as a practical simplification (variations of operationalizing the binning are explored in the Appendix of the paper below). Future research could look into developing an alternative that would be applicable to continuous time series. For now, I will continue making use of the current version of the advection model in Chapters 4 and 5.

3.1 Author contributions

The following paper has been published in *Language Dynamics and Change*. The reproduction of this online version, over the subsequent pages, is in accordance with the publication licence. I carried out the analysis, wrote the paper, created the figures, and handled the submission process. Kenny Smith, Richard A. Blythe and Simon Kirby provided advice on the design of the study and the analysis, as well as edits and comments on the paper. Note that the reference "Karjus et al. 2018a" refers to the preprint of the paper that forms Chapter 2 in this thesis, which was accepted but not published yet at the time this paper was being typeset.

3.2 Karjus et al. (2020): Quantifying the dynamics of topical fluctuations in language

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

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

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

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, Frermann 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).

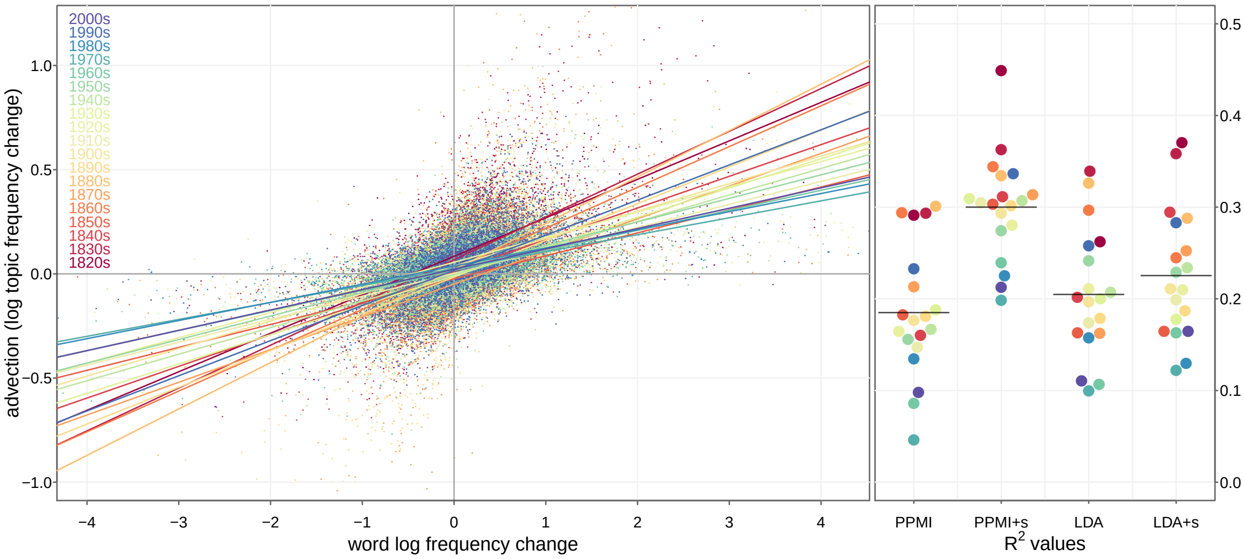


FIGURE 1 Left panel: log frequency changes of nouns and their corresponding topical advection (log topic change) values from two centuries of language change (from the PPMI-based model with topic smoothing). Each of the 107475 dots indicates the frequency change and advection value of one of the 10060 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

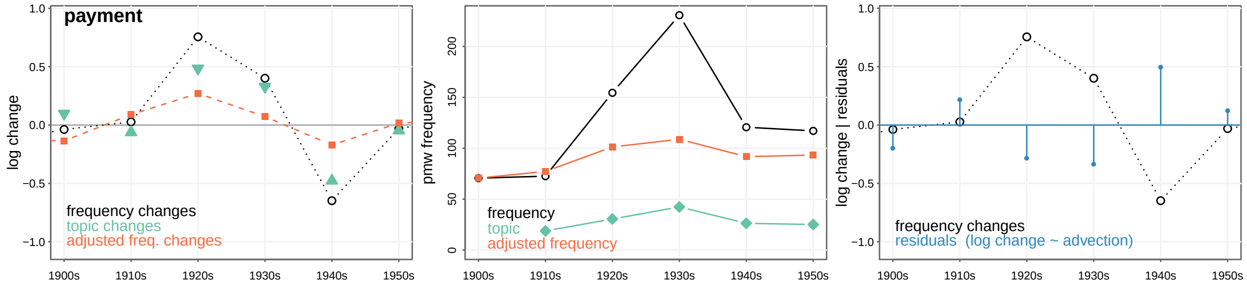


FIGURE 2 Time series of *payment* in the first half of the 20th century. Usage of the word increases considerably in the 1930s, but so does its topic. Black circles: log frequency change values (dotted line), actual frequency (solid line). Green triangles: topic frequency; change values on the left panel, 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).

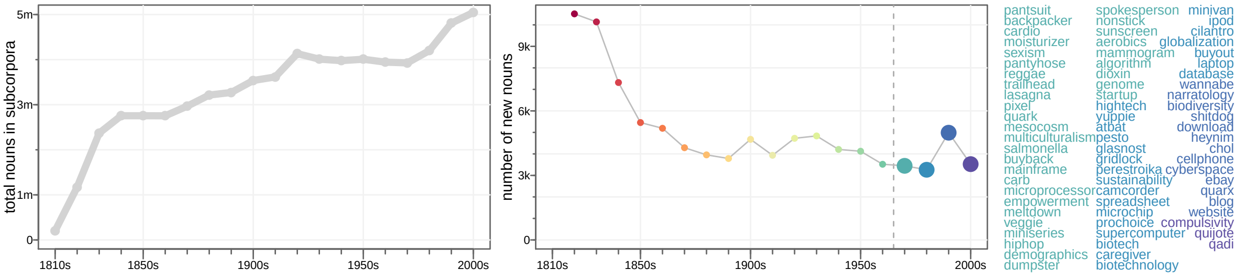


FIGURE 3 Token frequencies of nouns (left) and type frequencies of new nouns (middle panel) in the (preparsed) COHA corpus across period subcorpora. The vertical dashed line on the middle panel indicates the last four decades used to determine the test set of new words in this section; these words are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

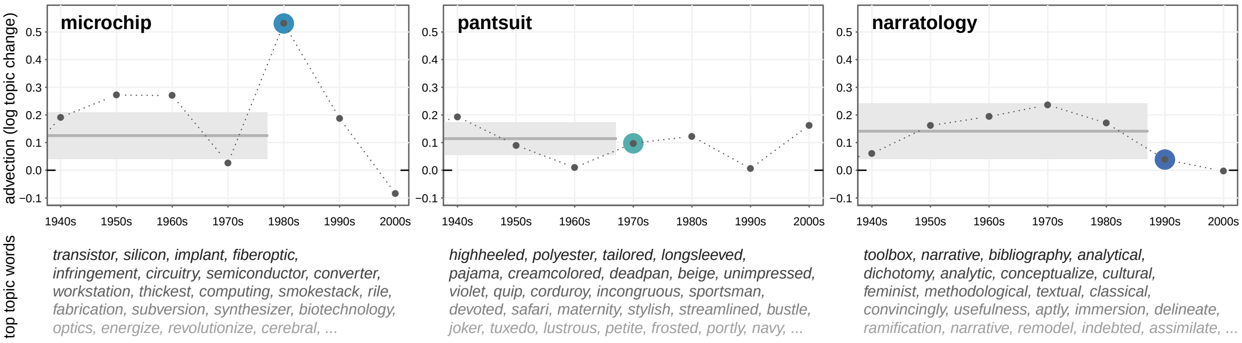


FIGURE 4 Three example novel words. The dashed and dotted dark gray line: the advection (log change) values of the topic of the word; above 0 indicates an increase, below 0 a decrease in the topic (note that this is not the frequency of the word, but the mean log changes in the topic). The brightly colored 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).

Frermann, 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: $\mathrm{logChange}(\omega; t) = \ln[f(\omega; t) + s] - \ln[f(\omega; t - 1) + s]$ 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., $\text{logTopicChange}(\tau; \omega, \text{t}) = \sum_{\omega' \neq \omega} p(\omega'|\tau) \text{logChange}(\omega'; t) p(\tau|\omega') / [1 - p(\omega|\tau)]$.
Then, $\text{advection}(\omega; t) = \sum_{\tau} \text{logTopicChange}(\tau; \omega, \text{t})\ p(\tau|\omega)$. 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

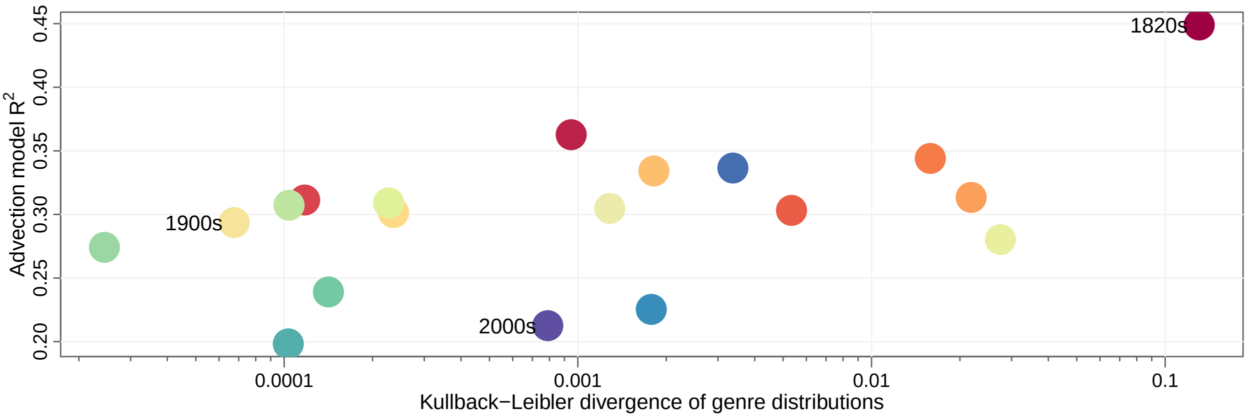


FIGURE 5 Divergence of genre distributions and the descriptive power of the advection measure (in the PPMI-based model, with smoothing). Each dot stands for one decade pair comparison, e.g. the dark purple dot marks the comparison of the 2000s to the preceding 1990s. The colors correspond to the colors 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.

3.3 Conclusions

In this Chapter, I developed a model for quantifying topical fluctuations in diachronic linguistic data, and demonstrated its applicability to predicting frequency changes in words, to adjusting time series for the topical elements, and its potential usage as a model of changes in communicative need. As such, it forms a methodological foundation for the following two Chapters (4 and 5). The advection model, meticulously evaluated and tested here, will be used to describe variance in lexical competition and colexification dynamics. The lexical innovation section of the paper that forms this Chapter was, in hindsight, fairly simplistic in its approach to defining new successful words — this methodology will also be further refined in Chapter 4, to weed out false positive cases of innovations stemming from uneven corpus composition.

Chapter 4

Communicative need modulates competition in language change

Chapter 2 approached language change from the interrogative angle, looking for ways to determine *if* selection is taking place. In this Chapter, I shift to the *why* question — given two or more functionally similar elements in language, why is it that in some cases one of them undergoes strong enough selection to kill the other(s) off, while in other cases similar elements coexist?

Chapter 3 laid out an approach to corpus data that I will continue to take in the remainder of the thesis. Here, the focus remains on the lexicon, as I use the topical advection model as a proxy measure of changes in communicative need, and develop a novel methodology to estimate the extent of competition between similar linguistic elements, inferring it directly from diachronic corpus data. I argue that the former is predictive of the latter. Elevated communicative needs allow similar words to co-exist in language without competing, as their minute differences are presumably seen as useful by speakers to express themselves in the topic that is currently relevant in their social and cultural environment. In terms of overarching general needs or evolutionary pressures (see Chapter 1), that of expressivity is allowed to overrule the need for efficiency (to reduce complexity by optimizing the lexicon to be smaller). In contrast, lower communicative needs around a given topic are more likely to lead to competition between near synonyms, as the general need for efficiency in language kicks in, and overshadows that of expressivity.

This Chapter (and the next) also take a more conservative approach to what counts as cases of innovation and spread in a corpus, using lessons learned in Chapter 3, to filter out words which would at first appear to have undergone a significant frequency change, but on closer look reveal to be artefacts of uneven corpus composition, e.g. terms from a single long book appearing frequent in a decade subcorpus simply due to being very frequent in a given book. In addition to that, a number of lexicostatistical control variables are introduced in statistical modelling to account for potential confounds.

4.1 Author contributions

The following paper has been submitted to *Language*. The version reproduced here is the submitted version, which is also posted as a preprint on Arxiv. I carried out the analysis, wrote the paper, created the figures, and handled the submission process. Kenny Smith, Richard A. Blythe and Simon Kirby provided advice on the design of the study and the analysis, as well as edits and comments on the paper. Note that the reference "Karjus et al. 2020a" refers to the paper that forms Chapter 3 in this thesis, which was in the online-only Advance Article stage at *Language Dynamics and Change* at the time this paper was submitted; "Karjus et al. 2020b" refers to the paper that forms Chapter 2.

4.2 Karjus et al. (2020): Communicative need modulates competition in language change

Communicative need modulates competition in language change

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

[1] Centre for Language Evolution, School of Philosophy, Psychology and Language Sciences, University of Edinburgh;

[2]School of Physics and Astronomy, University of Edinburgh

a.karjus@sms.ed.ac.uk, {r.a.blythe, simon.kirby, kenny.smith}@ed.ac.uk

Abstract

All living languages change over time. The causes for this are many, one being the emergence and borrowing of new linguistic elements. Competition between the new elements and older ones with a similar semantic or grammatical function may lead to speakers preferring one of them, and leaving the other to go out of use. We introduce a general method for quantifying competition between linguistic elements in diachronic corpora which does not require language-specific resources other than a sufficiently large corpus. This approach is readily applicable to a wide range of languages and linguistic subsystems. Here, we apply it to lexical data in five corpora differing in language, type, genre, and time span. We find that changes in communicative need are consistently predictive of lexical competition dynamics. Near-synonymous words are more likely to directly compete if they belong to a topic of conversation whose importance to language users is constant over time, possibly leading to the extinction of one of the competing words. By contrast, in topics which are increasing in importance for language users, near-synonymous words tend not to compete directly and can coexist. This suggests that, in addition to direct competition between words, language change can be driven by competition between topics or semantic subspaces.

1 Introduction

The literature on language change is full of examples of new elements, such as borrowings or morphological alternatives, replacing previous variants with similar functions. In English for example, past tense regular forms have been replacing irregular ones and vice versa (Pinker and Ullman 2002), a number of speech sounds were swapped out with other ones during the the Great Vowel Shift (Lass 1992), and the Norman Conquest led to the replacement of a large number of Middle English words with French alternatives (Durkin 2014). This kind of competition and replacement is core to the study of borrowing and innovation in historical linguistics and sociolinguistics (cf. McMahon 1994; Labov 2011; Mufwene 2002; Croft 2000), to discussions of linguistic selection and drift (Baxter et al. 2009; Cuskley et al. 2014; Sindi and Dale 2016; Newberry et al. 2017; Turney and Mohammad 2019; Pagel et al. 2019), S-shaped curves in language change (Blythe and Croft 2012; Ghanbarnejad et al. 2014; Stadler et al. 2016; Feltgen et al. 2017), and to studies of lexical growth and competition in big data computational linguistics (cf. Altmann et al. 2011; Stewart and Eisenstein 2018).[1] These population-level approaches, which track competition between variants over potentially substantial time spans in

[1] Orthogonal to this is competition between entire languages or varieties (Abrams and Strogatz 2003; Castelló et al. 2013; Zhang and Gong 2013; Karjus and Ehala 2018).

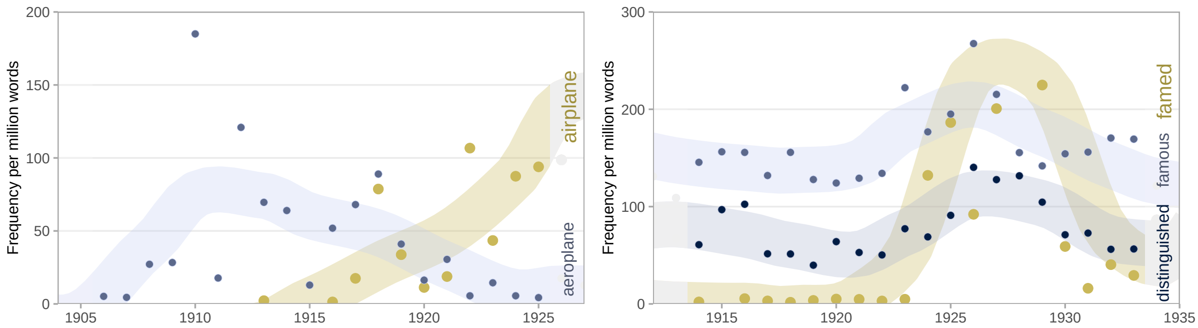


Figure 1: Example time series from the Corpus of Historical American English (COHA). Two decades after the invention of heavier-than-air powered aircraft, *airplane* replaced the initial term *aeroplane* (left side panel; the points are normalized yearly frequencies, with the lines representing smoothed averages for visual aid). Around the same time, *famed* appears to be increasing in usage. Yet it does not replace any semantically close words — *famed*, *famous* and *distinguished* all increase in tandem. Why do some successful new words replace their near-synonyms when their usage spreads, yet some do not, instead enriching their immediate semantic space?

populations of many individuals, are complemented by the psycholinguistic literature studying competition between representations within individual brains (cf. MacWhinney 1989; Brouwer et al. 2012; Mickan et al. 2020). The choices of individual speakers are of course what constitute synchronic variation, which in turn may accumulate as changes observable on the level of the larger community consensus over time.

We make three contributions in this paper. First, we propose a quantitative model of competition between linguistic elements in large-scale diachronic corpus data. Then, we use this to demonstrate that competition dynamics are modulated by communicative needs of speakers. Finally, we argue that not all competition takes place between individual elements like words but rather collections of elements (topics of conversation). As illustrated in Figure 1, while some linguistic innovations lead to direct competition between synonymous variants (like *aeroplane* and *airplane*), potentially resulting in the eventual decline and replacement of all but one of the competing forms, many cases of innovation do not. Figure 1 gives as example the non-competition between *famed* and *famous*, near-synonyms that increased in frequency in lock-step in the 1920s, at around the same time that *airplane* was replacing *aeroplane*. In short, there is variation in the presence or nature of competition — some words like *airplane* which enter a language or spread beyond niche usage compete with and replace a similar word, and may end up being replaced themselves in the future, whereas some words like *famed* seem to exist companionably alongside other closely-related words.

Our hypothesis is that communicative need affects the nature (specifically the “directness”) of the competition between individual words. We regard communicative need as a property of a topic of conversation, i.e. a subject consisting of related themes and ideas, encompassing a subset of co-occurring vocabulary. When communicative need within a topic is constant — its importance to a language community is not changing rapidly – then any newly introduced words must compete with words with similar semantic functions that are already present in the language. This is the case for the topic of early aviation which *airplane* belonged to in the first decades of the 20th century (cf. Figure 5 in Section 2.4). In contrast, where communicative need is on the rise — a topic is increasing in importance for language users — there is less need for competition between words, with multiple words able to co-exist and ride the wave of the users' communicative needs (like *famed* and *famous* do). We use computational methods (see Section 2) to quantify the notions of topic, communicative need, and directness of competition. While our focus here is on the lexicon, we believe that given the general nature of our proposed approach to quantifying competition dynamics, it could be applied to other areas of language like syntax or phonology, provided a sufficient quantity of suitably annotated data is available.

Our hypothesis, as stated above, is informed by prior work arguing that the shape of the lexicons and grammars of natural languages reflect the communicative needs and preferences of users of a given language. This idea has a long tradition, going back to Boas (1911: 26), Sapir (1921: 228), and Martinet (1952: 2). These needs are unlikely to be uniform across languages and time, or as Lupyan and Dale (2016) put it, "aspects of language that promote its learning and effective use are likely to spread, but what is optimal for one environment may be suboptimal for another". This view is widely shared by authors discussing communicative needs as a possible driving force in language change (Givón 1982: 117; Arends and Bruyn 1994: 118; Tomasello 1999: 74; Hopper and Traugott 2003: 37; Frajzyngier and Shay 2003: 286; van Trijp 2012; Mufwene 2013; Dor 2015; Kemp et al. 2018; Winters et al. 2015; Winters et al. 2018; Altmann et al. 2011). For example, languages are known to vary in the number of colours they lexify and how elaborate their kinship vocabulary systems are, which has been argued to reflect differences in communicative needs of linguistic communities (Gibson et al. 2017; Zaslavsky et al. 2019; Kemp and Regier 2012), and languages in warm climates are more likely to have a single word for both *ice* and *snow*, while these are lexified as individual words in colder climates (Regier et al. 2016). Similar environment-driven effects have been shown to operate in number marking systems (Haspelmath and Karjus 2017), and proposed as a possible driver of colexification dynamics besides conceptual similarity (Xu et al. 2020).

Perhaps the most obvious locus where (semantic) communicative need can lead to change in any given language is a lexical gap — a semantic subspace lacking an expression (Trask and Trask 1993: 157; Blank 1999: 79) or occupied by a word that has lost its expressive force (McMahon 1994: 201; Tamariz et al. 2014). This gap may be filled with a suitable word or construction, either innovated within a language, or borrowed from another language, often from a socially more prestigious one (cf. Hernández-Campoy and Conde-Silvestre 2012; Monaghan and Roberts 2019; Calude et al. 2017).

In our case, it is this more local and transient sense of communicative need we seek to measure and explore, specific to a given language and a given culture in a given population at a particular time in its history. This is in contrast to the broader sense of communicative need of languages being required to meet certain general criteria to be both learnable and useful as tools of communication (e.g. Zipf 1949; Labov 1982; Christiansen and Chater 2008; Kirby et al. 2015; Dingemanse et al. 2015; Auer and Hinskens 2005).

The motivation for our argument on competition between collections of linguistic elements stems from these more general communicative needs. Given the pressure on languages to be learnable and efficient systems of communication, it would be reasonable to expect that the increase of complexity in one part of the lexicon (i.e. the entry of a new lexical item) would require a compensating simplification elsewhere. This could be either in the same lexical subspace, i.e. the new word replaces a semantically similar word — or elsewhere, i.e. the incoming word is associated with a topic experiencing high communicative need, driving out words associated with other topics. In contrast to word-level competition, to our knowledge these dynamics have been given little attention in language change literature. Karjus et al. (2020a) demonstrate that individual word frequencies tend to follow the fluctuations of topics over time, as observable in population-level aggregate data such as corpora[2] (but presumably also in the differential salience of topics in the minds of the individual). For example, in times of war people talk about war-related things using more detailed vocabulary than they would otherwise; around major sports events like the Olympic Games various sports-related terms occur more frequently. These topical fluctuations can be taken to reflect the changing communicative needs of language users, reflecting the things that language users want and need to use language to communicate about.

We combine four distinct computational components in order to test our main hypothesis that communicative need affects lexical change and measure the directness of competition, ranging from word-level to topic-level. The first step is to collect samples of words which we will refer to as "targets" (Section 2.2). This is the test set for the model, words which have increased considerably in usage frequency over some period of time, and possibly replaced some other words. Our focus on the lexicon is partially driven by technical challenges: words are by far the most straightforward to model using the lexico-statistical machinery we employ to infer meaning from data, given the current state of available tools and datasets.

Word similarity is operationalized by training a distributional semantics model where distances between all words can be measured (Section 2.3). Words similar to the targets will be referred to as "(semantic) neighbors", referring to proximity in semantic space. This is a more suitable term than "synonyms", as our unsupervised machine learning approach conflates various possible semantic relations — such as synonymy, antonymy, hyperonymy, associativity — into a single similarity metric.

Previous research has pointed at the difficulties of capturing competitive dynamics and its effect on word growth (Grieve 2018: 155; Stewart and Eisenstein 2018: 4368). We propose an approach supported by machine learning to solve this (Section 2.3) We identify the locus of competition by summing up frequency changes in the ordered set of words similar to a target word, inferred from our model of semantics. In cases of competition between related words, the increase in frequency in the target word will be balanced by a decrease in frequency in a close semantic neighbor, whereas in more indirect inter-topic competition the frequency change will be balanced by decreases in distant, even unrelated, words.

As the final step, communicative need is inferred by proxy (as proposed in Karjus et al. 2020a) using a simple information-theoretic topic model (Section 2.4). Both this and our measure of word similarity rely on different operationalizations of word co-occurrence statistics — we take care to ensure that these measures do not cause autocorrelation in the final explanatory model. Our approach does involve a number of parameters and technical choices when it comes to training the machine learning models and operationalizing the corpus data. However, we find the results of our approach to be fairly robust within reasonable parametrizations. We apply these techniques to corpora spanning 5 language varieties and 3 centuries. We find a small but significant effect repeating across all five datasets, supporting our hypothesis that communicative need modulates competition. Words that are introduced into language or disseminate beyond occasional niche usage are more likely to take over the semantic functions and cause a decrease in the usage of neighboring words, if communicative need in their topics remains

[2] Hofmann et al. (2020) find a similar correlation between morphological family size and topical dynamics.

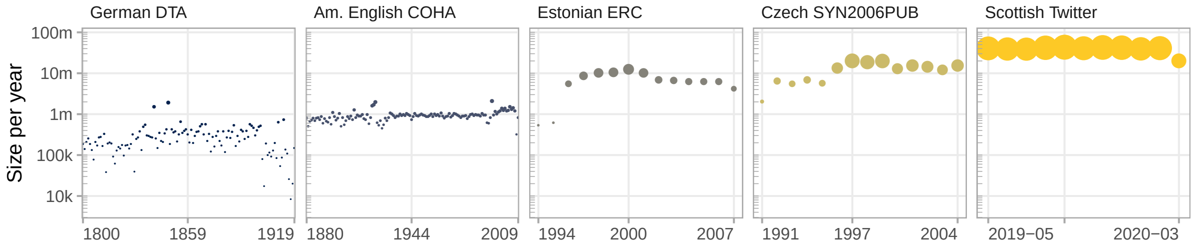


Figure 2: An illustration of the variability of the corpora as used in this study. The points reflect token counts by year (after filtering out stopwords; also note the $\log_{10}$ scale). The Twitter corpus of 315 days has over 3 times more data than 130 years of COHA put together (the Twitter panel shows monthly counts instead of yearly).

stable. On the other had, if communicative need is elevated, they instead enrich the semantic space without causing their semantic neighbors to go out of use.

2 Methods and materials

2.1 The corpora

We test our hypothesis on data from five corpora, mostly selected by availability, but intended to cover a variety of corpus types, languages and time periods, as illustrated in Figure 2. The Corpus of Historical American English (COHA; Davies 2010) spans the years 1810-2009 and includes 400 million words, balanced between four genres (newspapers, magazines, fiction and non-fiction). We use only data from 1880 onward as this part of the corpus is better balanced and more homogeneous.The Deutsche Textarchiv (DTA) is a corpus of German spanning 1600-1919, comparable in size and genre composition to the COHA. We use data from 1800 onward for the same reasons of genre balance and homogeneity as given for COHA. The Estonian Reference Corpus (ERC; Kaalep et al. 2010) is a corpus of modern Estonian; we use the media and fiction parts between 1994-2007, most of the data consisting of daily newspapers. The SYN2006PUB (Čermák et al. 2006) is a corpus of Czech newspapers between 1989-2004; we omit the first two years which have little data.

To diversify our test sets, we also mined Twitter for 315 days between 2019-2020 for all tweets posted in Scotland, and compiled it into a lemmatized corpus of 431 million words. Despite the idiosyncratic nature of Twitter communication, previous research has shown it to be a useful resource for studying language variation and change (cf. Grieve et al. 2018; Goel et al. 2016; Kershaw et al. 2016). The timespan of this corpus is obviously much shorter than the traditional diachronic corpora, but it provides magnitudes more data per unit of time, metadata on the source of each utterance, and yields insight into how widely the predicted relationship between competition and communicative need holds.

COHA, DTA, ERC and SYN2006PUB were conveniently already tagged and lemmatized, and underwent similar preprocessing. Since we are interested primarily in the content word lexicon, we filtered out stopwords (function words, numbers, punctuation, etc.), using custom stopword lists and part-of-speech tags as available in the corpora themselves. We lowercased all texts and excluded proper nouns (using POS tags), as what they refer to can vary arbitrarily both diachronically and synchronically (e.g. a *Bill* can be a president or a man on the street). We noticed some proper nouns still seeped into the test sets due to tagging errors, but did not filter any of them out post-hoc. For our Twitter corpus, we excluded duplicates and retweets, lowercased and lemmatized the texts (using spaCy; Honnibal and Montani 2017), filtered out stopwords and @-tags (usernames, being essentially proper nouns) and further homogenized the data by removing the # from hashtags (*#brexit* presumably means the same as *Brexit*, while multi-word hashtags like *#borisjohnsonlies* retain the meaning with or without the #).

2.2 Target words and time series

Historical corpora are noisy population-level aggregate samples of utterances produced over time. Instead of attempting to model the dynamics of entire time series of all lexical items in a corpus — inevitably mostly based on small noisy samples — we opt to select a small set of examples of significant usage frequency change between well-defined time spans. This ensures that what we are modelling is not corpus compilation sampling noise. The large size of our corpora provides the luxury of collecting a sufficient number of such cases. This means precluding potentially interesting competition dynamics at low frequencies, but our unsupervised machine learning approach to meaning requires substantial amounts of data to remain reliable (cf. Wendlandt et al. 2018; Dubossarsky et al. 2017).

We search each corpus for words that fit the criteria set out below. These criteria make reference to "time spans" and "units of time" which vary between the corpora, due to their different temporal resolution. We take the minimal time resolution within the corpora (days for Twitter, years for the others) to define the unit of time. The time span is a period over which a word's frequency is measured: each time span comprises multiple time units. For each potential target word w at each unit of time in a corpus we evaluate the frequency change between (normalized) total frequency f in the preceding time span (t_1) and the following one (t_2). For example, we use 10-year spans for COHA, so if the year considered by the search algorithm is 1916, then $t_1 = [1906, 1915]$ and $t_2 = [1916, 1925]$. In the *airplane* example (cf. Figure 3), the log (per millions words) frequency difference between $[1906, 1915]$ and $[1916, 1925]$ is $\ln(48/0.2) = 5.4$, which is the largest change for *airplane* between any two 10-year spans in COHA. We used 10-year spans in DTA; ERC and SYN2006PUB both span just over a decade, but contain much more data per year, so we used 5-year spans for those, and 30-day spans for the year-long Scottish Twitter corpus (see the Appendix for a longer discussion on the necessity of aggregating or "binning" corpus data).

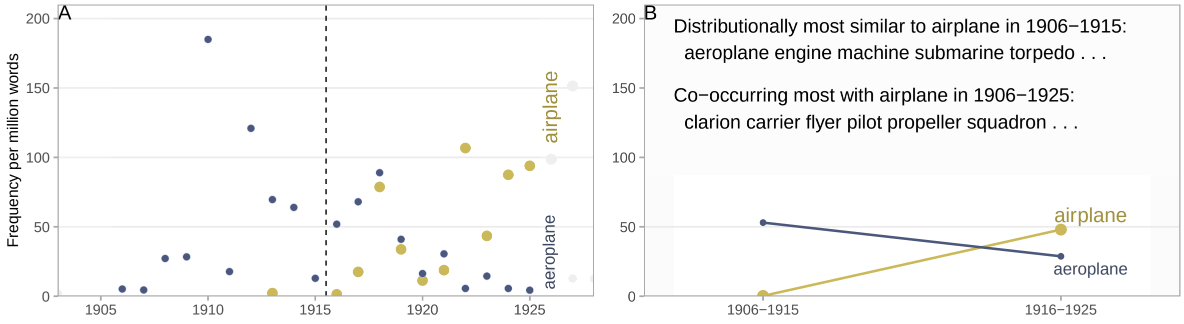


Figure 3: The *airplane* data: yearly frequencies on the left and the binned competition model input on the right. Counts are aggregated into two 10-year bins. The dashed vertical line highlights the border between these 10-year spans. The total increase of *airplane* is almost matched by the decrease of its nearest semantic neighbor, *aeroplane*, the remainder by decreases in *engine* and *machine*). The list of distributionally similar words on panel B include those that could be used instead of *airplane* in a similar context (but not necessarily the same syntactic function or exact meaning; see Section 2.3). The list of co-occurring words include those that are used near *airplane*, e.g. in a phrase like *airplane pilot*, and form a basis for the model for inferring changes in communicative need (see Section 2.4).

The selection criteria for a target word w are as follows:

1. Most importantly, the token frequency change of a potential target should be stark enough to cut through sampling noise ($\ln(f_{t_2}/f_{t_1}) \geq 2$).
2. Its absolute frequency should be high enough for related statistics and distributional semantics inferences to be reliable ($f_{t_2} \geq 200$).
3. In the Twitter corpus, where more metadata is available, we also require w to be spread out across the user base (see Section 2.5 for details).
4. w should also be used throughout t_2 (in at least 80% of units of time within the span of t_2)
5. The frequency increase of w should be consistent, the time series should not include outlying peaks (see Section 2.5 for technical details).

The last two criteria avoid cases where an apparent word frequency increase (simply based on comparing t_1 and t_2) stems from a word being a frequent term in some specialized corner of language or for a very short time period, while seeing little to no use in common language. If multiple stretches along the time series of a word meet these criteria, we simply use the pair of time spans with the greatest frequency change between them; therefore each word only occurs once in the resulting dataset used for statistical analysis (Section 3).

This filtering procedure yields on average 270 target words per corpus (COHA: 240, DTA: 489, ERC: 274, SYN2006PUB: 257, Twitter: 97). Each target is associated with two time spans, between which the target word increased considerably in frequency, like the two decades of *airplane* in Figure 3 (see Sections 3 and 2.3 for more examples of target words), and a number of lexicostatistical variables as described in the next sections. Further technical details on these parameters, and other implementation decisions, are discussed in the Appendix.

2.3 Modelling competition

Our measure of competition derives from a simple notion about frequencies. After normalization, frequencies of words in a corpus sum to 1. Let us consider two possible sub-corpora from a historical corpus, e.g. for the 1990s and 2000s, each normalized separately. If a target word of interest increases between these periods, then it follows that some other word(s) must by definition decrease in frequency — because in the end, everything sums to 1. In that sense, change in frequency always entails competition — the increase of one word is matched by or "equalized" by the sum of decrease(s) of some other word(s). Note that we use values multiplied by 1 million (per-million frequency) for more interpretable figures.

The distance to the target of the words whose decreases make up for the increase of the target can be taken as an indicator of the locus of competition. An increase that is directly compensated by an equivalent decrease in a semantically similar word (a near-synonym) indicates competition between two words that represent the same meaning. When no similar words decrease, and the increase in the target can only be matched by looking further away in the semantic space, then this indicates competition is less direct, and more likely to be between topics rather than within topics. We will refer to this distance — where the frequency increase in a target is equalized by the cumulation of decreases in other words — as the "equalization range".

Importantly, this measure becomes less informative as its value increases. Large equalization range values should be interpreted as indicating that there are no direct (losing) competitors to be found, rather than considering the last equalizing word as a competitor. These are rather cases of what we suspect to be competition between topics — which is of course much more indirect and hard to capture than competition between words with similar meaning. As the model searches for decreasing words further and further from the target in semantic space, the ones it finds may be quite unrelated to the target word (like *son* in Figure 4D). In contrast, at short equalization ranges, the competitors are usually clearly semantically related words (like *aeroplane* in Figure 4A).

Obviously, this approach requires some way of obtaining the similarity between all words in the corpus. This could be done using a dictionary (cf. Ramiro et al. 2018) or a lexical database such as a Wordnet (cf. Turney and Mohammad 2019), or using machine learning (Xu and Kemp 2015; Hamilton et al. 2016; Rosenfeld and Erk 2018; Fremann and Lapata 2016). We opt for the latter approach, as this can be readily applied to any sufficiently large corpus, without the need for external language-specific resources. We use Latent Semantic Analysis (LSA; cf. Bullinaria and Levy 2007), an application of Singular Value Decomposition. Like all distributional semantics models, it relies on word co-occurrence statistics. Words that are used together with the same words — i.e. have similar distributions of co-occurrences across the lexicon — end up with cosine-similar vectors in the resulting high-dimensional vector space. This acts as the computational approximation of a lexico-semantic space of a language, but with all semantic associations (cf. Section 1) collapsed into a simplified proximity metric.

We measure the target equalization range using a normalized version of cosine distance. We observe that vector spaces trained on different corpora or even segments of the same corpus can have quite variable densities. Therefore it makes sense to normalize the distance, which we do by dividing by the distance value of the closest neighbor. Distance values of 0 in the results (cf. Figure 6) therefore refer to cases where the increase in frequency of a target word is completely matched by a compensating decrease in the nearest neighbor.

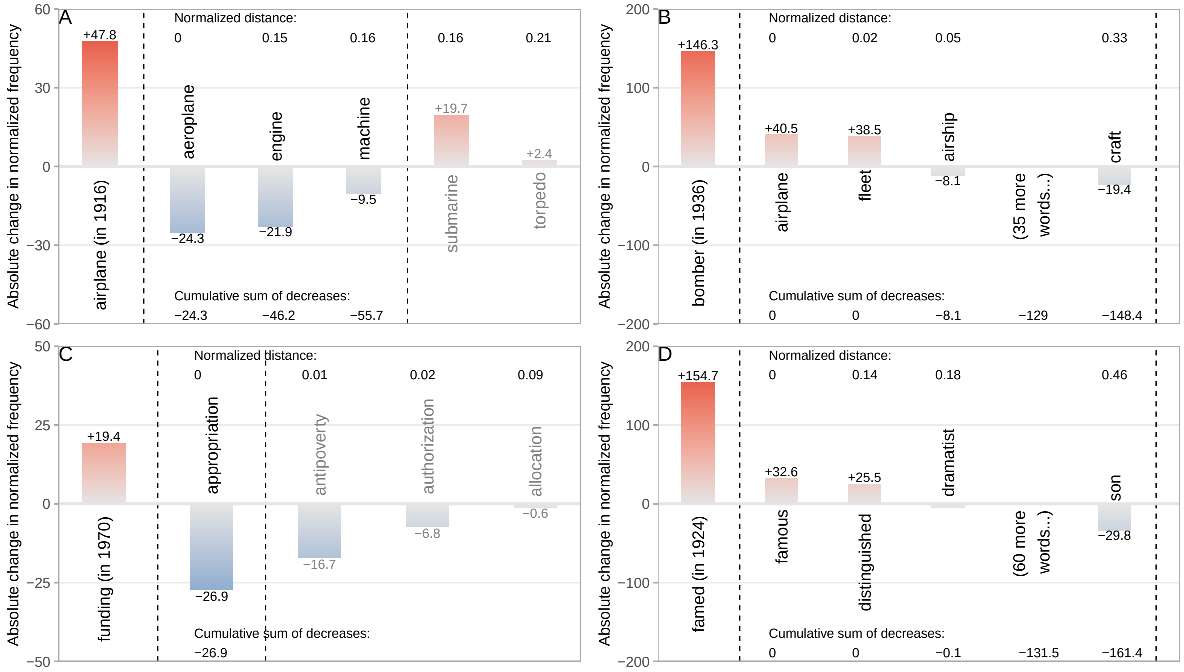


Figure 4: Examples of the competition model in action. The dashed lines indicate the equalization range. Panels A and C (the earlier *airplane* example and *funding* from the 1970s) are fairly clear cases of competition — their increase is compensated by decrease in a near neighbor. For *airplane* the equalization range is 0.16 (the normalized distance to *machine*), as its increase of +47.8 per million words is matched by the sum of *aeroplane*, *engine* and *machine*. The right side panels B and D illustrate lack of direct competition. *bomber* came to prominence two decades after *airplane*, when the topic of aerial warfare was too important to have only one word for all things with wings; note that *bomber* emphatically does not compete with its nearest neighbor, *airplane*, since both increase in frequency. For *famed*, the equalization range is 0.46, and the decreasing words that compensate its increase are spread out between 60-odd words (note that at ranges this large, the measure simply indicates "no direct competition" rather than reliably identifying actual competitors).

The LSA model is based on aligned co-occurrence data from the two time spans associated with a target word. The target is assigned a meaning vector using data from the second span where it is (thanks to the initial filtering) frequent enough, and the rest of the words in the lexicon are assigned meaning based on the first time span. There are two reasons for this. Since we require targets to undergo notable frequency change, most targets in the test sets have little to no presence before this increase, so it would not be possible to reliable infer their semantics. It is also not impossible that the increase of a target would change its immediate semantic landscape, forcing semantic change in related words (cf. McMahon 1994: 178). Our approach ensures the resulting semantic neighbors in the model are those that reside in the semantic space near the target just before its usage started increasing (see Figure 4, and a more technical description in the Appendix).

The ease of operating with co-occurrence vectors and LSA in this manner is one reason to use this approach instead of a more recent model like word2vec combined with vector space alignment (Mikolov et al. 2013; Hamilton et al. 2016; Yao et al. 2018). Our approach is analogous to the one described in Dubossarsky et al. (2019) and Sagi et al. (2011), using common context words to model semantics over time. A context-sensitive model (Devlin et al. 2019; Hu et al. 2019) could potentially provide better meaning estimates, but would make comparing words between diachronic subcorpora less straightforward; this could be explored in future research. Judging both by qualitative evaluation and testing against a gold standard test set (Hill et al. 2015), we found LSA to perform reasonably well despite the small size of time period subcorpora (distributional semantics models are usually trained on corpora of tens of billions of tokens, not mere tens of millions).

Our general approach is similar to Turney and Mohammad (2019) who also investigate competition between words, but rely on the dictionary-like Wordnet data for determining similarity. Obviously inferring meaning using machine learning instead of using an expert-crafted lexicological resource has the downside of introducing additional noise. The upside is that since we infer meanings for words directly from respective time period sub-corpora, our approach does not require additional language-specific resources (such as a Wordnet), but also accounts for older and changed meanings (which a synchronic Wordnet does not). Furthermore, instead of modeling competition within predefined sets of synonyms (the "synsets" of a Wordnet), our approach takes into account the entire lexicon with explicit similarity values, and allows us to account for indirect (topic-level) competition.

2.4 Modeling communicative need

Determining the communicative needs of the largely invisible speakers whose texts ended up in a historical corpus is by no means a trivial task. We estimate changes in communicative needs by assuming the following relatively simple model linking the observed corpus data and the presumed underlying process (see also Kemp et al. 2018: 120).

A diachronic corpus such as the COHA is essentially a large sample of utterances by numerous speakers (or more specifically, writers, in a written language corpus) expressing themselves across a variety of contexts and genres. If a topic of conversation is gaining importance for speakers, it would hopefully be reflected in the language, and therefore be observable as frequency changes in a representative corpus (assuming of course the apparent changes do not stem from sampling noise in an unbalanced corpus; cf. Pechenick et al. 2015). If the prevalence of a topic differs between two sub-corpora — such as two decades — then this can be taken to indicate differing communicative need within this topic. If a topic is of socio-cultural importance to speakers — the associated communicative need is elevated — then it is reasonable to expect that speakers use the relevant vocabulary more, and use more detailed semantics for references in the discourse to successfully communicate more fine-grained distinctions, which may in turn result in the coining or borrowing of new words or repurposing old ones. For example, the topic of *bomber* (Figure 4B), relating to aerial warfare, naturally became more prevalent during World War 2 — which is reflected in widespread increases in frequency not only in *bomber* itself but also in words it would co-occur (i.e., form a topic) with, such as *squadron* or *air force*, as well as the introduction of new ones such as *blitz*.

We make use of the topical advection model from Karjus et al. (2020a) to estimate changes in communicative need through quantifying the shifts in latent topics between time period subcorpora. "Advection" is a term borrowed from physics, referring to the transport of a substance by the bulk motion of a fluid — the analogy being words swept along by prevalence fluctuations of associated topics. Karjus et al. (2020a) show that this measure is a fair baseline predictor

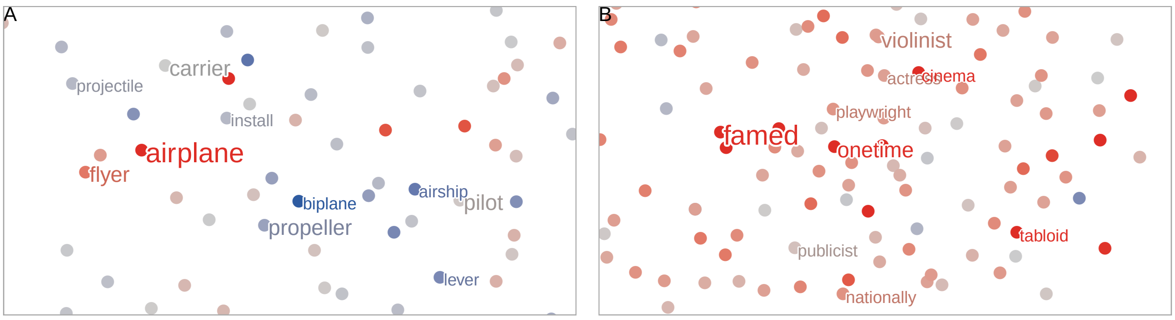


Figure 5: Topic landscapes for *airplane* (in 1906-1925) and *famed* (1914-1933). This is a dimension-reduced rough projection of the co-occurrence matrices of the subcorpora of these periods, with proximity between any two words approximately corresponding to not their semantic similarity but simply the extent which they occur together (more specifically, their PPMI association scores). A topic may be thought of as a group of words that are used together in similar contexts. Blue colors indicate words with decreasing frequency over the relevant time span, reds indicate increasing words. The advection value (weighted mean log topic change) for *airplane* is close to neutral at 0.13, while it's strongly positive for *famed* (0.82), reflected by being surrounded by plenty of red.

for word frequency changes — it is possible to make a reasonable prediction about how much a word's frequency will change by looking at how well its related topic is doing. It is of course more informative for words that drift along with the flow of topics (such as *famed* at the rise of cinema and celebrity culture in the interwar period; cf. Figure 5) rather than those which compete with and are selected for (or against) by speakers, such as *aeroplane*, which simply replaced a similar word with a similar spelling.

The topical advection model measures the change in topic frequencies between time periods (or sub-corpora more generally), not the prevalence of a topic at a given point in time. We infer the "topic" of each target word as a list of top k context words which co-occur in the same context as the target (in a wider window of ±10 words), scored by their Positive Point-Wise Mutual Information (PPMI; see Appendix), as illustrated in Figure 5. Change in topic frequency is then measured as the weighted mean (log) frequency change of these topic words. Corpus data from both of the target's associated time spans t_1 and t_2 are concatenated, as this approach was shown in Karjus et al. (2020a) to improve the model's performance.

At the heart of our approach are then essentially two non-overlapping lists of words: the list of (top 75 PPMI-scored) topic words — and the list of semantic neighbors, ordered by similarity, spanning the entire lexicon (minus the topic words, to avoid autocorrelation). Both lists are based on corpus co-occurrence statistics: topics consist of words occurring in the same context as the target, and semantic neighbors are essentially words which have similar context words. Sometimes a few of those may overlap: for example, besides *aeroplane*, *aircraft*, *balloon* and *propeller* all also have high similarity scores to *airplane* — but feature among its top topic words as well (cf. Figure 3B), indicating co-occurrence in common contexts with *airplane*. It is crucial to avoid autocorrelation between the two measures — which we do by filtering out such overlapping topic words (such as *balloon*) from the list of neighbours when determining the equalization range.[3] Leaving out topic words from the neighbours list unavoidably limits

[3] The ease of decorrelating the measures this way is one reason to use a simple topic model based on discrete

the descriptive power of the competition model: word(s) that sometimes occur in the same context with the target may also be among the ones that the target is actually in the process of replacing.

2.5 Controlling for other lexico-statistical variables

We include a number of lexicostatistical measures as controls in the statistical model used to test the relationship between competition and communicative need. This is to exclude other possible explanations for variance in directness of competition, at least ones that can be inferred from a corpus (this unfortunately does not include possibly also relevant sociolinguistic variables). Frequency change in the target (difference in per-million frequency values; cf. Figure 4) is an important potential predictor: bigger increases could perhaps lead lower frequency neighbors to go out of use, or the opposite, bigger increases might require a larger equalization range. We also control for maximum (z-scored) peak value in the time series across the two time spans of each target (e.g. in COHA, yearly frequencies); the time point associated with the start of the increase in a target's time series (as a numeric value), and the length of the target word (long words might have different dynamics than short ones).

As for variables relating to the immediate semantic space, we control for minimum (Damerau-Levenshtein) edit distance of closest neighbors (is the target competing with a similarly spelled word?), cosine distance to nearest neighbor (does the target actually have close synonyms?), and the maximum percentage change among the nearest neighbors (does the target cause an extinction?). The last one differentiates cases of direct competition which lead to near-100% decrease in a neighbor — if the equalization range is short — from changes which just lead to either a relatively small decrease in a high-frequency neighbor, or small decreases spread out between multiple neighbors.

We also include a variable for leftover frequency mass (e.g. in Figure 4, for *funding* it would be $26.9 - 19.4 = 7.5$ units of per-million frequency, or 39% of the +19.4 increase of *funding*). If the decrease of the final equalizing neighbor is considerably larger than the increase in the target, then presumably either the model is not doing a good job capturing the semantics, or there is something more complex than just one-to-one competition going on (we also filter out targets where the leftover is actually larger than the increase value of the target). Additionally, in Twitter data, we control for the (median of daily) user to frequency ratio (see Section 2.2).

For the Twitter corpus, we further make use of the available user metadata and, as mentioned in Section 2.2, only consider targets which are reasonably widely used. Some words or hashtags may look very frequent at first, but a closer look often reveals (possibly automated) lone accounts or small groups that post the same or similar message hundreds of times a day e.g. to promote their views or products. This is of course not representative of common language use. We therefore excluded candidate targets with an account-to-frequency ratio[4] of < 0.75, and also include this as a control variable in the statistical model for the Twitter dataset (see Section 3.

words here rather than something like LDA (Blei et al. 2003) which models topics as distributions; they were shown to perform comparably in Karjus et al. (2020a). It is only feasible to do it this way around: the neighbours list spans the entire lexicon, while there are only 75 words in the topic list.

[4] I.e., the number of accounts who used a given term, divided by the total frequency of the term, yielding a value between $(0, 1]$. A "1" means every occurrence is associated with a unique account; if 50 accounts each tweet a term twice (100 total), then it's 0.5; a single account tweeting a term 100 times yields 0.01. The filtering threshold uses the median of these daily values.

We did not make use of like and retweet counts, as they apply to entire tweets and not individual words — although some averaged measure could potentially be considered in future research.

3 Results: communicative need predicts lexical competition dynamics

Figure 6 illustrates the results of applying the model to the target words extracted from our five corpora. We model the variables in a straightforward linear regression model, one for each data set. In all models, advection is a significant predictor ($p < 0.05$) for the response variable of equalization range. The R^2 values quoted in Figure 6 refer to the amount of variance accounted for by the communicative need variable, on top of all the lexicostatistical controls described in Section 2.5 (adjusted R^2, based on comparing the full model to the reduced controls-only model, cf. Anderson-Sprecher 1994). As apparent in Figure 6, the model behaves comparably

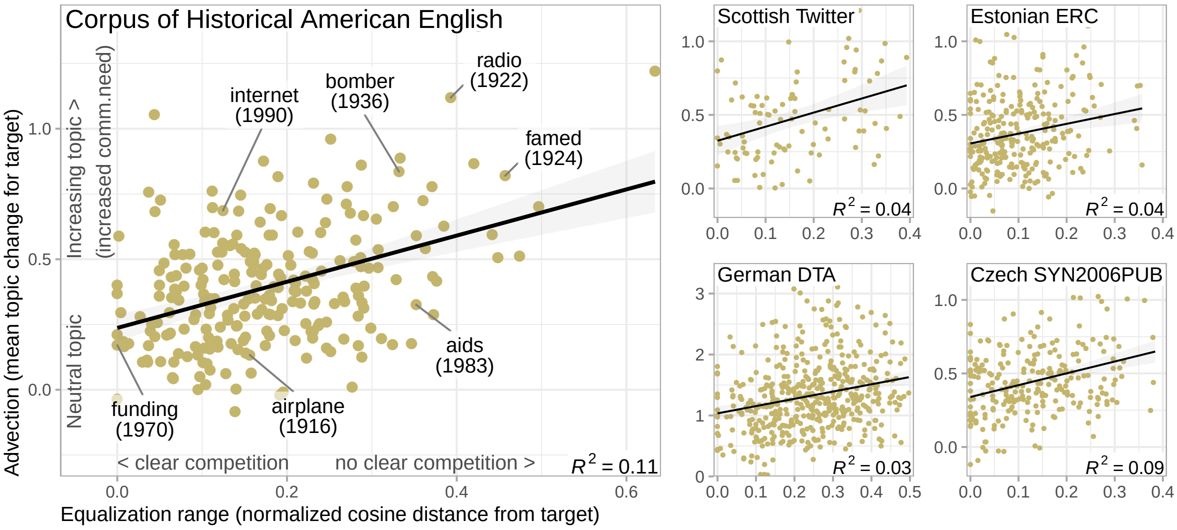


Figure 6: The extent of competition between words correlates with changes in communicative need, as operationalized by the topical advection model. The x-axis shows the equalization range — where the frequency decreases of semantic neighbors match the increase in the target (cf. Section 2.3). Points on the left of the plot therefore represent words which compete and replace their immediate neighbor(s), whereas words on the right do not. The y-axis represents the mean topic change for the target word: advection values around 0 (lower on the plot) represent words in a topic which is relatively stable, whereas points higher up in the graph represent words whose topics are increasing in frequency in the corpus, which we assume reflects elevated communicative need around that topic. There is a clear correlation between these two quantities (as reflected in points lying roughly on the plotted diagonal): words clustered in the bottom left corner are those in a stable topic and which compete and replace their immediate neighbor(s); words towards the top right corner increase in tandem with their neighbors, without a clear signal of immediate competition. The main panel shows this plot for the COHA corpus; the smaller side panels show that the effect of communicative need on competition dynamics persists with similar magnitude across data sets of target words based on corpora that differ in languages, time period and type of media.

across the data sets, describing a moderate amount of variance (up to 11%) in competition dynamics. The German data set turns out to be somewhat of an outlier, with much higher absolute advection values, meaning that the topical composition of the corpus must fluctuate considerably over time. However, the correlation between advection and lexical replacement is still present. The full models have R^2 values between 0.17 and 0.25, as the lexicostatistical controls such as frequency change magnitude account for some additional variance.

If the usage frequency increase of a word — or appearance in a language in the case of novel words — does not coincide with a rising topic of conversation, then the word is more likely to take over the semantic functions of a similar word. For example, in American English, *funding* encroached the semantic field of *appropriation* in financial contexts in the 1970s, and *boy scout* partially replaced the functions of *cadet* in the 1910s after the founding of the namesake organization. In Estonian newspapers, it appears the term *respublikaan* pretty much replaced *koonderakondlane* from 2002 onward — the former meaning 'member of Res Publica', a center-right political party that became active in 2002, and the latter meaning 'member of the Coalition Party', a center-right political party disbanded in 2002. This pair is of course not an example of synonymy, but reflects our model capturing terms used in very similar contexts to refer to similar political actors. In the Twitter corpus, *movember*[5] starts trending towards the end of October 2019, replacing another charity-related term, *greatscottishrun*; the rest of the increase is compensated by a slight decrease in the more frequent general term *charity*.

In contrast, a word that increases in usage and belongs to a topic experiencing elevated communicative need is more likely to co-exist with synonymous or similar words. This would be the earlier *famed* and *bomber* examples, or *radio* in the 1920s. In the Twitter corpus, the term *corona* occasionally pops up throughout the year referring to the beverage, but in the sense of the virus starts trending in January-February 2020 — the pandemic of that year constituting a new high advection topic consisting of terms like *virus*, *spreading*, *#coronavirusupdate* but also the toilet paper emoji, all increasing in tandem with *corona*.

4 Discussion

4.1 Technical limitations and possible improvements

We have shown that changes in the communicative needs of speakers contribute to lexical change and competition dynamics. However, we believe the real effect may well be larger than detected by our model. In addition to the peculiarities of written language corpora as discussed below (Section 4.2), the models used here rely on statistical machine learning — meaning, similarity and topics are all inferred from co-occurrence data. In other words, we rely on statistical approximations to communicative need and conventions, based on another proxy (corpora) to actual usage. Noise is unavoidable. The model is further weakened by the necessary purging of the semantic neighbors lists of often high similarity words to avoid autocorrelation with the topic model (cf. Section 2.4). Yet we find the effect persists.

A reasonable worry would be that the small correlation between communicative need and competition dynamics we observe is a spurious one, an artifact of our statistical machinery, or some

[5] An organization and annual event involving the growing of mustaches during November to raise awareness of men's health issues.

aspect of corpus composition. We do not have reason to believe so, based on carrying out simulations with randomized data on the competition model (see the Appendix), the advection model having undergone similar validation (cf. Karjus et al. 2020a), having controlled for a slew of other lexico-statistical variables, and having tested the model on a variety of different corpora.

There are several avenues of technical improvement that could be explored to build on the current contribution. These include using more sophisticated word embeddings (see Section 2.3), bigger corpora as they become available, and exploring the effects of different model parameterizations. In terms of corpora, investigating the role of communicative need in selection and competition in creole and new variety formation would be particularly interesting (cf. Baxter et al. 2009; Strimling et al. 2015; Winford 2017). Our essentially correlational results could be improved with causal analysis, and the methodology could potentially be extended to work with continuous time series (cf. Koplenig 2017). The current competition measure identifies cases of direct competition, but becomes less informative as the equalization range increases. This calls for a method for more accurately inferring topic-level competition. Connecting the competition model with tests for selection and drift could be explored (cf. Newberry et al. 2017; Karjus et al. 2020b; Kauhanen 2017). Communicative need could perhaps be operationalized in ways that better approximate real world usage situations, possibly also by estimating diachronic developments via synchronic data (cf. Regier et al. 2016; Karjus 2015).

4.2 Scarcity of direct competition

We note that there are numerous examples among the target sets (cf. Section 3) where the equalization range consists of only a single neighbor. Yet examples of competition where the increase of a target word would lead to the complete disappearance of a neighboring one, at least within the timespan of a generation, are almost non-existent. It seems once a word has already entered conventional usage, it takes a while for it to completely disappear, even if it is on a clear downward path. Even though airplane (beside just plane) is the preferred variant in American English, *aeroplane* keeps popping up in the corpus throughout the 20th century, albeit at low frequencies, as does for example *larboard* (the archaic nautical term for the left side of a ship) and *cumbrous* (cumbersome). This echoes findings in previous research: while the entry of new linguistic material into language is often claimed to follow an S-shaped curve (Blythe and Croft 2012), extinction has been argued to follow a decelerated trajectory (Nini et al. 2017).

The unwillingness of words to die makes more sense if one considers the nature of written language corpora — which may well include texts referring to historical events and objects, texts from more archaic varieties of a language (as British is to American English), and texts written (or edited) by older speakers for whom using older variants of modern terms comes naturally. It has also been pointed out that the shape of the lexicon may not always reflect the current cultural interests and communicative needs of a community, with terms in semantic subspaces of waning relevance nevertheless surviving generations of speakers (Malt and Majid 2013: 591). Finally, there is also a further explanatory variable that we do not control for in our model: our approach to competition is based on usage frequencies, but there is also the possibility that a word losing out in competition might change meaning and continue to survive in another function (see the Appendix for details). Our simple model of semantics also treats each form as having a single (vector of) meaning, and competition may also resolve thought the loss or gain of semantic functions in polysemous words.

4.3 Different kinds of competition

Our findings point to language change being driven by yet another kind of competition in addition to those discussed in Section 1. This is the competition between topics of conversation — in turn presumably reflecting the events and state of the changing world. Word frequencies loosely follow topical fluctuations over time (Karjus et al. 2020a), and our findings further illustrate that indeed many words that get introduced to language (or spread beyond previous niche usage) do not do so directly at the expense of older synonyms — the nearest words that can be found decreasing in frequency are often semantically unrelated to the target. Instead, they follow the fluctuations of topics.

Inevitably, when some topics of conversation increase in prevalence, others must diminish (there are only so many hours in a day). And in less relevant topics, semantic spaces will become sparse, as multiple words with slightly different shades of meaning become redundant, due to lowered communicative needs in the area. Historical and sociolinguistics often focuses on isolated examples of lexical replacement by borrowing or competition between language-internal variants. We believe competition between topics or semantic subspaces is something that deserves further investigation. Furthermore, while grammatical complexity is widely studied and shown to correlate with population size and structure (Atkinson et al. 2015; Bentz and Winter 2014; Reali et al. 2018), linguistic topical complexity — not just vocabulary size — remains virtually unexplored.

4.4 Using experimentation to further understanding of linguistic change

Human language is a unique system seen nowhere else in nature. Understanding how and why it works requires understanding how it changes, change being one of the few absolutely universal properties of living languages. This in turn requires understanding both individual and population level dynamics. On the one hand, behaviour of linguistic communities is not necessarily indicative of the biases or choices of individual language learners and users, and different biases may lead to similar outcomes; on the other hand, constraints at the population level may arise from weak individual biases that may be hard to detect in isolation (Smith and Wonnacott 2010; Smith et al. 2017; Kandler et al. 2017).

While the exact histories of the sociolinguistic environments where changes take place cannot be reconstructed, corpora, though imperfect lenses, provide a way to systematically observe wider changes in populations over time, like the growth and decline of elements of the lexicon. The correlation we have observed calls for further investigation into the role of communicative need and fluctuations of topics in language change, also from the perspective of individual learning and communication biases. Unlike historical dynamics, this is something that can be studied in controlled experimental settings, either using natural (Lev-Ari and Peperkamp 2014) or artificial languages (cf. Kirby et al. 2008; Winters et al. 2015; Scott-Phillips and Kirby 2010).

5 Conclusions

Previous research using experimental approaches and synchronic data has shown how languages adapt to the communicative needs of their speakers. We have shown how to model these processes and correspondences using data that reflects changes in language communities over longer time spans. Our methods do not require language-specific resources other than a sufficiently large diachronic corpus, and produce comparable results across corpora of different languages, types, genres, and time spans. In particular, we have described a language-agnostic approach to quantifying competition between elements of language, here on the example of lexical items. We found that these dynamics correlate with changes in communicative need, as operationalised by the topical advection model. In summary, we find support for the idea that languages keep changing in ways that are useful for their speakers. All other things being equal, multiple similar words can co-exist in a lexicon as long as the finer shades of meaning they provide are useful in discourse — while new words will eventually replace old ones if a single word will do in the given semantic subspace.

Acknowledgments

We thank Steven Piantadosi for a discussion that led to the operationalization of the competition model (cf. Section 2.3) and Jennifer Culbertson for useful questions and comments. 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.

Data and code availability

Some of the corpora are publicly available (see respective references below). The code to run the models described in this paper is available at `https://github.com/andreskarjus/competition-langchange`

References

Abrams, Daniel M. and Steven H. Strogatz (2003). “Modelling the Dynamics of Language Death”. In: *Nature* 424, p. 900.

Altmann, Eduardo G., Janet B. Pierrehumbert, and Adilson E. Motter (2011). “Niche as a Determinant of Word Fate in Online Groups”. In: *PLOS ONE* 6.5, pp. 1–12.

Anderson-Sprecher, Richard (1994). “Model Comparisons and R2”. In: *The American Statistician* 48.2, pp. 113–117.

Arends, Jacques and Adrienne Bruyn (1994). “Gradualist and Developmental Hypotheses”. In: *Pidgins and Creoles: An Introduction.* John Benjamins Publishing, pp. 111–120.

Atkinson, Mark, Simon Kirby, and Kenny Smith (2015). “Speaker Input Variability Does Not Explain Why Larger Populations Have Simpler Languages”. In: *PLOS ONE* 10.6, pp. 1–20.

Auer, Peter and Frans Hinskens (2005). “The Role of Interpersonal Accommodation in a Theory of Language Change”. In: *Dialect Change: Convergence and Divergence in European Languages.* Ed. by Peter Auer, Frans Hinskens, and PaulEditors Kerswill. Cambridge University Press, pp. 335–357.

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”. In: *Language Variation and Change* 21.02, pp. 257–296.

Bentz, Christian and Bodo Winter (2014). “Languages with More Second Language Learners Tend to Lose Nominal Case”. In: *Quantifying Language Dynamics.* Brill, pp. 96–124.

Blank, Andreas (1999). “Why Do New Meanings Occur? A Cognitive Typology of the Motivations for Lexical Semantic Change”. In: Berlin, Boston: De Gruyter Mouton, pp. 61–90.

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

Blythe, Richard A. and William Croft (2012). “S-Curves and the Mechanisms of Propagation in Language Change”. In: *Language* 88.2, pp. 269–304.

Boas, Franz (1911). *The Mind of Primitive Man.* Macmillan. 302 pp.

Brouwer, Susanne, Holger Mitterer, and Falk Huettig (2012). “Can Hearing Puter Activate Pupil? Phonological Competition and the Processing of Reduced Spoken Words in Spontaneous Conversations:” in: *Quarterly Journal of Experimental Psychology.*

Bullinaria, John A. and Joseph P. Levy (2007). “Extracting Semantic Representations from Word Co-Occurrence Statistics: A Computational Study”. In: *Behavior Research Methods* 39.3, pp. 510–526.

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

Casler, Stephen D (2015). “Why Growth Rates? Which Growth Rate? Specification and Measurement Issues in Estimating Elasticity Values”. In: *The American Economist* 60.2, pp. 142–161.

Castelló, Xavier, Lucía Loureiro-Porto, and Maxi San Miguel (2013). “Agent-Based Models of Language Competition”. In: *International journal of the sociology of language* 2013.221, pp. 21–51.

Čermák, František, Jaroslava Hlaváčová, Milena Hnátková, Tomáš Jelínek, Jan Kocek, Marie Kopřivová, Michal Křen, Renata Novotná, Vladimír Petkevič, Věra Schmiedtová, et al. (2006). *SYN2006PUB: Corpus of Czech Newspapers.* Faculty of Arts, Institute of the Czech National Corpus, Charles University.

Christiansen, Morten H. and Nick Chater (2008). “Language as Shaped by the Brain”. In: *Behavioral and Brain Sciences* 31.5, pp. 489–509.

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

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”. In: *PLOS ONE* 9.8, pp. 1–7.

Davies, Mark (2010). *The Corpus of Historical American English (COHA): 400 Million Words, 1810-2009.* Available online at https://www.english-corpora.org/coha.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). NAACL-HLT 2019. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186.

Dingemanse, Mark, Damián E. Blasi, Gary Lupyan, Morten H. Christiansen, and Padraic Monaghan (2015). "Arbitrariness, Iconicity, and Systematicity in Language". In: *Trends in Cognitive Sciences* 19.10, pp. 603–615.

Dor, Daniel (2015). *The Instruction of Imagination: Language as a Social Communication Technology*. Foundations of Human Interaction. Oxford University Press.

DTA (2019). *Deutsches Textarchiv. Version Vom 6. Februar 2019: DTA-Kernkorpus Und Ergänzungstexte. http://www.deutschestextarchiv.de.*

Dubossarsky, Haim, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg (2019). "Time-out: Temporal Referencing for Robust Modeling of Lexical Semantic Change". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 457–470.

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*, pp. 1147–1156.

Durkin, Philip (2014). *Borrowed Words: A History of Loanwords in English*. Oxford: Oxford University Press.

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". In: *Open Science* 4.11.

Frajzyngier, Zygmunt and Erin Shay (2003). *Explaining Language Structure through Systems Interaction*. John Benjamins Publishing. 329 pp.

Frermann, Lea and Mirella Lapata (2016). "A Bayesian Model of Diachronic Meaning Change". In: *Transactions of the Association for Computational Linguistics* 4, pp. 31–45.

Ghanbarnejad, Fakhteh, Martin Gerlach, José M. Miotto, and Eduardo G. Altmann (2014). "Extracting Information from S-Curves of Language Change". In: *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". In: *Proceedings of the National Academy of Sciences* 114 (40), pp. 10785–10790.

Givón, Thomas (1982). "Tense-Aspect-Modality: The Creole Prototype and Beyond". In: *Tense-aspect: Between semantics and pragmatics*, pp. 115–163.

Goel, Rahul, Sandeep Soni, Naman Goyal, John Paparrizos, Hanna Wallach, Fernando Diaz, and Jacob Eisenstein (2016). "The Social Dynamics of Language Change in Online Networks". In: *Social Informatics*. Ed. by Emma Spiro and Yong-Yeol Ahn. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 41–57.

Grieve, Jack (2018). "Natural Selection in the Modern English Lexicon". In: *The Evolution of Language: Proceedings of the 12th International Conference on the Evolution of Language*. Ed. by C. Cuskley, M. Flaherty, H. Little, Luke McCrohon, A. Ravignani, and T. Verhoef. NCU Press.

Grieve, Jack, Andrea Nini, and Diansheng Guo (2018). "Mapping Lexical Innovation on American Social Media". In: *Journal of English Linguistics* 46.4, pp. 293–319.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky (2016). "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*, pp. 1489–1501.

Haspelmath, Martin and Andres Karjus (2017). “Explaining Asymmetries in Number Marking: Singulatives, Pluratives, and Usage Frequency”. In: *Linguistics* 55.6, pp. 1213–1235.

Hernández-Campoy, Juan Manuel and Juan Camilo Conde-Silvestre (2012). *The Handbook of Historical Sociolinguistics.* Wiley-Blackwell.

Hill, Felix, Roi Reichart, and Anna Korhonen (2015). “SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation”. In: *Computational Linguistics* 41.4, pp. 665–695.

Hofmann, V., J.B. Pierrehumbert, and H. Schuetze (2020). “Predicting the Growth of Morphological Families from Social and Linguistic Factors”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Seattle WA, July 5 - July 10.* Association for Computational Linguistics.

Honnibal, Matthew and Ines Montani (2017). “spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing”.

Hopper, Paul J. and Elizabeth Closs Traugott (2003). *Grammaticalization.* Cambridge University Press. 300 pp.

Hu, Renfen, Shen Li, and Shichen Liang (2019). “Diachronic Sense Modeling with Deep Contextualized Word Embeddings: An Ecological View”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* ACL 2019. Florence, Italy: Association for Computational Linguistics, pp. 3899–3908.

Kaalep, Heiki-Jaan, Kadri Muischnek, Kristel Uiboaed, and Kaarel Veskis (2010). “The Estonian Reference Corpus: Its Composition and Morphology-Aware User Interface”. In: *Proceedings of the 2010 Conference on Human Language Technologies – The Baltic Perspective: Proceedings of the Fourth International Conference Baltic HLT 2010.* NLD: IOS Press, pp. 143–146.

Kandler, Anne, Bryan Wilder, and Laura Fortunato (2017). “Inferring Individual-Level Processes from Population-Level Patterns in Cultural Evolution”. In: *Royal Society Open Science* 4.9.

Karjus, A., R. A. Blythe, S. Kirby, and K. Smith (2020a). “Quantifying the Dynamics of Topical Fluctuations in Language”. In: *Language Dynamics and Change.*

Karjus, Andres (2015). “Through the Spyglass of Synchrony: Grammaticalization of the Exterior Space in the Eastern Circum-Baltic”. In: *New Trends in Nordic and General Linguistics.*

Karjus, Andres, Richard A. Blythe, Simon Kirby, and Kenny Smith (2020b). “Challenges in Detecting Evolutionary Forces in Language Change Using Diachronic Corpora”. In: *Glossa: a journal of general linguistics* 5.1, p. 45.

Karjus, Andres and Martin Ehala (2018). “Testing an Agent-Based Model of Language Choice on Sociolinguistic Survey Data”. In: *Language Dynamics and Change* 8.2, pp. 219–252.

Kauhanen, Henri (2017). “Neutral Change”. In: *Journal of Linguistics* 53.2, pp. 327–358.

Kemp, Charles and Terry Regier (2012). “Kinship Categories across Languages Reflect General Communicative Principles”. In: *Science (New York, N.Y.)* 336.6084, pp. 1049–1054.

Kemp, Charles, Yang Xu, and Terry Regier (2018). “Semantic Typology and Efficient Communication”. In: *Annual Review of Linguistics* 4.1, pp. 109–128.

Kershaw, Daniel, Matthew Rowe, and Patrick Stacey (2016). “Towards Modelling Language Innovation Acceptance in Online Social Networks”. In: *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* (San Francisco, California, USA). WSDM ’16. ACM, pp. 553–562.

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”. In: *Proceedings of the National Academy of Sciences* 105.31, pp. 10681–10686.

Kirby, Simon, Monica Tamariz, Hannah Cornish, and Kenny Smith (2015). “Compression and Communication in the Cultural Evolution of Linguistic Structure”. In: *Cognition* 141, pp. 87–102.

Koplenig, Alexander (2017). “A Data-Driven Method to Identify (Correlated) Changes in Chronological Corpora”. In: *Journal of Quantitative Linguistics* 24.4, pp. 289–318.

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

Labov, William (1982). “Building on Empirical Foundations”. In: *Perspectives on Historical Linguistics*. Ed. by W. Lehmann and Y. Malkiel. Vol. 24. Amsterdam and Philadelphia: Benjamins, pp. 17–92.

Lass, Roger (1992). “What, If Anything, Was the Great Vowel Shift”. In: *History of Englishes: New Methods and Interpretations in Historical Linguistics*. Ed. by Matti Rissanen et al. Berlin: Mouton de Gruyter, pp. 144–155.

Lev-Ari, Shiri and Sharon Peperkamp (2014). “An Experimental Study of the Role of Social Factors in Language Change: The Case of Loanword Adaptations”. In: *Laboratory Phonology* 5.3, pp. 379–401.

Levy, Omer, Yoav Goldberg, and Ido Dagan (2015). “Improving Distributional Similarity with Lessons Learned from Word Embeddings”. In: *Transactions of the Association for Computational Linguistics* 3, pp. 211–225.

Lupyan, Gary and Rick Dale (2016). “Why Are There Different Languages? The Role of Adaptation in Linguistic Diversity”. In: *Trends in Cognitive Sciences* 20.9, pp. 649–660.

MacWhinney, Brian (1989). “Competition and Lexical Categorization”. In: *Linguistic categorization* 61, pp. 195–241.

Malt, Barbara C. and Asifa Majid (2013). “How Thought Is Mapped into Words”. In: *WIREs Cognitive Science* 4.6, pp. 583–597.

Martinet, André (1952). “Function, Structure, and Sound Change”. In: *WORD* 8.1, pp. 1–32.

McMahon, April M.S. (1994). *Understanding Language Change*. Cambridge University Press.

Mickan, Anne, James M. McQueen, and Kristin Lemhöfer (2020). “Between-Language Competition as a Driving Force in Foreign Language Attrition”. In: *Cognition* 198, p. 104218.

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

Monaghan, Padraic and Seán G. Roberts (2019). “Cognitive Influences in Language Evolution: Psycholinguistic Predictors of Loan Word Borrowing”. In: *Cognition* 186, pp. 147–158.

Mufwene, Salikoko S (2013). “Language as Technology Some Questions That Evolutionary”. In: *In search of universal grammar: From old Norse to Zoque* 202, p. 327.

Mufwene, Salikoko S. (2002). “Competition and Selection in Language Evolution”. In: *Selection* 3.1, pp. 45–56.

Newberry, Mitchell G., Christopher A. Ahern, Robin Clark, and Joshua B. Plotkin (2017). “Detecting Evolutionary Forces in Language Change”. In: *Nature* 551.7679, pp. 223–226.

Nini, Andrea, Carlo Corradini, Diansheng Guo, and Jack Grieve (2017). “The Application of Growth Curve Modeling for the Analysis of Diachronic Corpora”. In: *Language Dynamics and Change* 7.1, pp. 102–125.

Pagel, Mark, Mark Beaumont, Andrew Meade, Annemarie Verkerk, and Andreea Calude (2019). “Dominant Words Rise to the Top by Positive Frequency-Dependent Selection”. In: *Proceedings of the National Academy of Sciences* 116.15, pp. 7397–7402.

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”. In: *PLoS ONE* 10.10. Ed. by Alain Barrat, e0137041.

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

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”. In: *Scientific Reports* 2, 313 (2012).

Pinker, Steven and Michael T Ullman (2002). “The Past and Future of the Past Tense”. In: *Trends in cognitive sciences* 6.11, pp. 456–463.

Ramiro, Christian, Mahesh Srinivasan, Barbara C. Malt, and Yang Xu (2018). “Algorithms in the Historical Emergence of Word Senses”. In: *Proceedings of the National Academy of Sciences* 115.10, pp. 2323–2328.

Reali, Florencia, Nick Chater, and Morten H. Christiansen (2018). “Simpler Grammar, Larger Vocabulary: How Population Size Affects Language”. In: *Proceedings of the Royal Society of London B: Biological Sciences* 285.1871.

Regier, Terry, Alexandra Carstensen, and Charles Kemp (2016). “Languages Support Efficient Communication about the Environment: Words for Snow Revisited”. In: *PLOS ONE* 11.4, pp. 1–17.

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)*. Vol. 1, pp. 474–484.

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

Santus, Enrico, Emmanuele Chersoni, Alessandro Lenci, Chu-Ren Huang, and Philippe Blache (2016). “Testing APSyn against Vector Cosine on Similarity Estimation”. In: *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers* (Seoul, South Korea), pp. 229–238.

Sapir, Edward (1921). *Language. An Introduction to the Study of Speech.* New York: Harcourt, Brace and Company.

Schlechtweg, Dominik, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde (2019). “A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 732–746.

Scott-Phillips, Thomas C. and Simon Kirby (2010). “Language Evolution in the Laboratory”. In: *Trends in Cognitive Sciences* 14.9, pp. 411–417.

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

Smith, Kenny, Amy Perfors, Olga Fehér, Anna Samara, Kate Swoboda, and Elizabeth Wonnacott (2017). “Language Learning, Language Use and the Evolution of Linguistic Variation”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 372.1711, p. 20160051.

Smith, Kenny and Elizabeth Wonnacott (2010). “Eliminating Unpredictable Variation through Iterated Learning”. In: *Cognition* 116.3, pp. 444–449.

Stadler, Kevin, Richard A. Blythe, Kenny Smith, and Simon Kirby (2016). “Momentum in Language Change: A Model of Self-Actuating S-Shaped Curves”. In: *Language Dynamics and Change* 6.2, pp. 171–198.

Stewart, Ian and Jacob Eisenstein (2018). “Making “Fetch” Happen: The Influence of Social and Linguistic Context on Nonstandard Word Growth and Decline”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels, Belgium). Association for Computational Linguistics, pp. 4360–4370.

Strimling, Pontus, Fredrik Jansson, and Mikael Parkvall (2015). “Modeling the Evolution of Creoles”. In: *Language Dynamics and Change* 5.1, pp. 1–51.

Tamariz, Monica, T. Mark Ellison, Dale J. Barr, and Nicolas Fay (2014). “Cultural Selection Drives the Evolution of Human Communication Systems”. In: *Proceedings of the Royal Society B: Biological Sciences* 281.1788, p. 20140488.

Tomasello, Michael (1999). *The Cultural Origins of Human Cognition*. Harvard University Press. 257 pp.

Törnqvist, Leo, Pentti Vartia, and Yrjö O. Vartia (1985). “How Should Relative Changes Be Measured?” In: *The American Statistician* 39.1, pp. 43–46.

Trask, R.L. and R.L. Trask (1993). *A Dictionary of Grammatical Terms in Linguistics*. Linguistics - Routledge. Routledge.

Turney, Peter D. and Saif M. Mohammad (2019). “The Natural Selection of Words: Finding the Features of Fitness”. In: *PLOS ONE* 14.1, pp. 1–20.

Van Trijp, Remi (2012). “Self-Assessing Agents for Explaining Language Change: A Case Study in German”. In: *Proceedings of the 20th European Conference on Artificial Intelligence*. ECAI’12. Montpellier, France: IOS Press, pp. 798–803.

Wendlandt, Laura, Jonathan K. Kummerfeld, and Rada Mihalcea (2018). “Factors Influencing the Surprising Instability of Word Embeddings”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. NAACL-HLT 2018. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2092–2102.

Wetherell, Charles (1986). “The Log Percent (L%): An Absolute Measure of Relative Change”. In: *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 19.1, pp. 25–26.

Winford, Donald (2017). “The Ecology of Language and the New Englishes: Toward an Integrative Framework”. In: *Changing English: Global and Local Perspectives: Markku Filppula, Juhani Klemola, Anna Mauranen, Svetlana Vetchinnikova,* 92, p. 25.

Winters, James, Simon Kirby, and Kenny Smith (2015). “Languages Adapt to Their Contextual Niche”. In: *Language and Cognition* 7.3, pp. 415–449.

Winters, James, Simon Kirby, and Kenny Smith (2018). “Contextual Predictability Shapes Signal Autonomy”. In: *Cognition* 176, pp. 15–30.

Xu, Yang, Khang Duong, Barbara C. Malt, Serena Jiang, and Mahesh Srinivasan (2020). “Conceptual Relations Predict Colexification across Languages”. In: *Cognition* 201, p. 104280.

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

Yao, Zijun, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong (2018). “Dynamic Word Embeddings for Evolving Semantic Discovery”. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (Marina Del Rey, CA, USA). WSDM ’18. ACM, pp. 673–681.

Zaslavsky, Noga, Charles Kemp, Naftali Tishby, and Terry Regier (2019). “Color Naming Reflects Both Perceptual Structure and Communicative Need”. In: *Topics in Cognitive Science* 11.1, pp. 207–219.

Zhang, Menghan and Tao Gong (2013). “Principles of Parametric Estimation in Modeling Language Competition”. In: *Proceedings of the National Academy of Sciences* 110.24, pp. 9698–9703.

Zipf, George Kingsley (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology.* Reading, MA: Addison-Wesley Press.

Appendix

5.1 Using corpus data requires some form of aggregation

The minimal time resolution in most of the corpora we used is one year. However, there is not enough data in most diachronic corpora per year for word embedding models (which we use to estimate word semantics) to work as intended. For COHA and DTA we used time spans or “bins” of 10 years. ERC and SYN2006PUB both span just over a decade, but contain much more data per year, so we used 5-year spans for those. For the year-long Scottish Twitter corpus we used 30-day spans. The limitation of comparing pairs of discrete time spans is of technical nature: the version of the topical advection model (Section 2.4) that we use as a proxy to communicative need is not readily applicable to continuous time series.

Instead of simply using fixed calendric spans e.g. decades or months, we carry out binning for each word separately, depending on where a word starts increasing. Although a choice like splitting a centuries-spanning corpus into 10-year spans or a year into 30-days spans might feel intuitive, all these choices really are quite arbitrary. Binning has indeed also been shown to affect statistical models based on corpus time series (Karjus et al. 2020b). While we remain reasonably confident in the results produced in this paper, values of parameters like this — and the ones discussed in Section 2.2 and further below in the Appendix — is something that should be critically evaluated in future research.

Parameters and alternative setups

The complexity of our approach to lexical competition, necessitated by the complexity of the linguistic processes and the challenges of estimating these from diachronic data, entails a number of relatively arbitrary parameters and design choices. We describe the results of what we consider an intuitively reasonable set of choices, but further research could further explore the parameter space. We also explored slightly longer time spans in COHA where this is possible (20 years, yielding similar results), and flipping the model to quantify the “losers” of competition instead: in the main text, we focus on words increasing in frequency, which provide a clearer case for competition, as discussed in Section 4.2. The model works the other way around as well, with targets being words decreasing in frequency, above some chosen change threshold —

in the COHA data, we find a significant but even smaller correlation between equalization range and advection for words going out of usage.

There is no doubt that some changes in language take more than our chosen time spans (e.g., 10 + 10 years). There is no technical reason why we could not use longer time spans like 50 years, or compare a decade to another decade 100 years earlier — but the worry is that multiple processes may well take place within longer periods, which our competition model (Section 2.3) is not tailored to handle. Figure 7 illustrates this, how very similar words can go through periods of co-existence and competition, and how competition can play out over longer time scales. *Moslem* occurs at low frequencies in COHA throughout the 19th century, *Muslim* appearing in the beginning of the 20th. These spelling variants then co-exist for half a century still at low frequencies (with a few exceptions like that of 1921, the year of the Malabar Rebellion), the former slightly more frequent than the latter, until *Muslim* starts increasing — but it still takes another 30 or so years until *Moslem* starts decreasing.

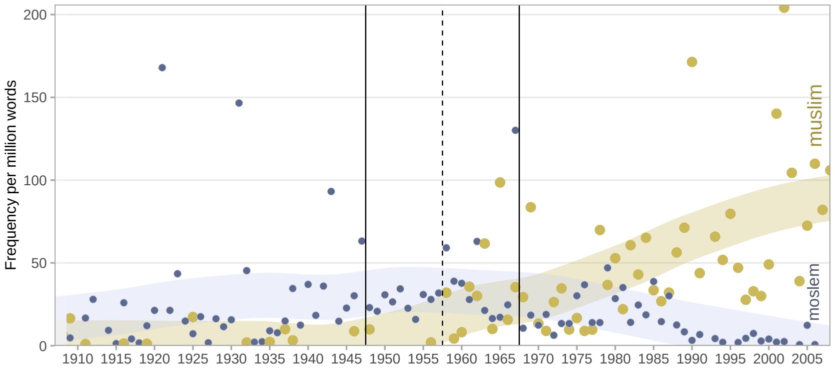


Figure 7: More example time series from COHA; points are normalized yearly frequencies, lines reflect smoothed averages. The black vertical lines indicate the period captured by our model as the largest (log) change between any 10-year spans in the time series of *Muslim*. The outliers tend to coincide with related historical events. This graph further illustrates the complexities of lexical change over longer periods.

PPMI and log change

We employ two useful metrics in multiple stages of our model, namely Positive Pointwise Mutual Information (1) and log frequency change (2). Both are used in the advection model (see Section 2.4), PPMI is used to detect multi-word units (see below), and log change is used to filter target words that have increased considerably between two time spans t_1 and t_2

$$\mathrm{PPMI}(\omega, c) := \max\left\{\log_2 \frac{P(\omega, c)}{P(\omega)P(c)}, 0\right\} \tag{1}$$

$$\mathrm{logChange}(\omega; t) := \begin{cases} 0 & \text{if } f(\omega; t_1) = 0 \text{ and } f(\omega; t_2) = 0 \\ \ln[f(\omega; t_2) + s_{t_2}] - \ln[f(\omega; t_1) + s_{t_1}] & \text{otherwise} \end{cases} \tag{2}$$

ω is a word and c is a context word it may co-occur with. $f(\omega; t)$ is the (in our case, normalized) frequency of ω in the corpus during time period t. We use Laplace smoothing offset to avoid ln(0), setting the values of s to the equivalent of 1 occurrence in t in after normalizing to per-million counts; s is set to 0 if the frequency $f(\omega; t) > 0$. If both frequencies are 0, then change is set to be 0. While we refer to (2) as "log change" (cf. also Altmann et al. 2011; Petersen et al. 2012), it is also called log percent (Törnqvist et al. 1985; Wetherell 1986) or logarithmic growth rate (Casler 2015). Like absolute change $(f(\omega; t) - f(\omega; t-1))$, but unlike percentage change, log change is symmetric and additive. However, on the absolute scale, the biggest frequency changes are those of fluctuating already high-frequency words, while log change highlights sharp and therefore hopefully meaningful changes at lower frequency bands. This, in combination with chosen minimum increase threshold, results in target sets that mostly start out at 0 or low frequency (in COHA, the median across targets in t_1 is 1.9 per million words, or about 19 occurrences in an average decade of 10m words) and reach frequencies reflecting widespread usage by t_2 (median 30.1 pmw in COHA).

Spelling smoothing and multi-word units

We homogenize spelling by removing all punctuation (including hyphens) from lemmas in all corpora, and in COHA, concatenate the most common multi-word units, which we detect as follows. The latter motivated by the fact that the spelling of compounds in English varies both diachronically and synchronically (e.g. *long term*, *long-term*, *longterm*). This was done with the aim of improving our co-occurrence based measures of synonymy and topicality: in compounds (or phrases, collocations) such as *social worker* or *death row*, the words on their own often have different or at least more general meaning. Lexical innovations such as *website* (also occurring as *web site*) often go through multiple variants, making it harder to track their spread. In this contribution we focus on (homogenized) lemmas, leaving competition between low-frequency spelling variants like that for future research (one that would likely need larger corpora in terms of data per unit of time).

In multi-word unit detection, we only consider two-word units to keep things simple. As the first pass, the 200-year COHA is split into 20-year subcorpora and in each one, common multi-word units are determined using PPMI as the collocation metric (with a threshold of 7). We do not do this using the entire corpus at once, as collocation statistics may well change over time. The union of these 10 sets yields a total of 501 units, mostly compounds such as *post office*, some phrases like *absolutely necessary*, and a few proper nouns like *Gulf War*. On the second pass, when parsing and cleaning the corpus for the analysis proper, these multi-word units are concatenated when encountered (e.g., to become *postoffice*), and treated as single words in the subsequent frequency counts, semantic and topic models. However, we find that this operation only marginally improves the power of the statistical model based on COHA data at the end of the pipeline (see Section 3).

Notes on Twitter

Our Twitter corpus is slightly different from the others in that it covers communication on the platform by all users from a given geographical region, in a short time span. In contrast to written language corpora, this should reflect more "natural", unedited, and relatively homogenous language use — but then again Twitter is also only a narrow, situational slice of language.

Its user base and demographics are not necessarily representative of the actual population, and the utterances expressed on Twitter are (hopefully) still only a subset of the daily utterances produced by its users.

Looking back at Figure 6, the number of targets in the Twitter dataset is relatively small. This is due to our stringent selection criteria for targets, one of which is consistent usage over the given time span. Looking at the data, many time series appear “spiky” instead (see below for more on peak detection). A word or hashtag occurs rarely except for a day or two where its usage then skyrockets, often referring to some event, a piece of news, or a TV show. This of course which makes sense given the nature of Twitter, and we naturally do not expect to see considerable language change in the span of a year, but rather the topic-type competition discussed above.

Notes on control variables

We control for edit distance between a target and its nearest neighbors to account for words which may potentially be competing with their spelling variants, such as *airplane* (see Section 2.5). This necessitates arbitrarily defining how near “near” is, and we pick a range of 20 words. Maximum decrease percentage among neighbors also involves an implicit range parameter, but this is just set to be the same as the equalization range (Section 2.3), and as such varies from target to target.

Since we work with aggregated (binned) frequencies, we also account for differences in the time series within the aggregates by quantifying their maximum peak value — seeing how some words increase steadily, while for some words, an apparent large increase in aggregated (e.g. decade) frequency stems from a single high-frequency peak on closer inspection. Each frequency value in an examined time series (e.g. a 20-year span in COHA) is z-scored, using the mean and standard deviation of the rest of the series, i.e. excluding the value itself. We record the maximum of these z-scores, and during the target search phase (Section 2.2) also exclude candidate series where the maximum is > 10, indicating a series with a large outlying peak (10 standard deviations away from the mean). Such peaks can stem from sampling noise (a yearly subcorpus may for example include a book where some certain term is highly frequent) or real-world events which get a lot of coverage in the short term but do not affect the lexicon in the long term (as is very common in the Twitter corpus).

Details of the semantics model

The LSA model is trained on a PPMI-weighted co-occurrence matrix based on corpus data from the first of the two time spans associated with each target word (t_1, cf. Section 2.2), reflecting the semantic space of the language before the usage of the target started increasing. We use a window of ± 2 words (cf. Levy et al. 2015), $k = 100$ for LSA dimensionality, and a minimal occurrence threshold of 100 tokens. Most targets in the test sets have little to no presence in t_1, which would hinder reliable semantic inference. We collect the lexicon-length co-occurrence vector for the target from the second time span (t_2) subcorpus where its usage is by definition widespread and frequent, align it to the lexicon of t_1, and then fit this into the t_1-trained LSA. This way, the resulting semantic neighbors are those that reside in the semantic space near the target just before its usage started increasing (cf. Figure 4).

We remove words from these neighbors lists which do no occur widely in t_1 (threshold to occur at least in half of a time span). This filters out words that appear prevalent in a decade but only because of high frequency in a single year, often a single document like one book. This does not reflect widespread usage and is likely sampling noise.

Evaluating the approach using randomized data

Since our approach to modeling competition relies heavily on machine learning, the natural worry is that the results may result from some unknown property or artifact of the underlying complex models (cf. Dubossarsky et al. 2017; Wendlandt et al. 2018), or be driven by some other lexicostatistical confounds such as frequency. We therefore include a number of plausible control variables in our statistical analysis (see Section 2.5), set a frequency threshold to exclude low-frequency and therefore unreliable words (Section 2.2), make sure our two co-occurrence-based measures do not overlap (Section 2.4), but also evaluate the competition model by feeding it randomized data.

The competition model relies on an ordered list of similarity-scored words, the closest of which could be considered near-synonyms of the target (see Section 2.3). We carry out a randomization test by giving each target word arbitrary semantic neighbors with arbitrary similarity scores (drawn from the distribution of actual similarities), but calculating the equalization range as usual (see Figure 4). Under this randomization the closest neighbours of *airplane* will be usually be unrelated words, for instance *chocolate* or *rabbit*, instead of *aeroplane*. If advection (or proxy for communicative need) still correlated with the equalization range based on the assumption that *airplanes* may be considered synonymous and competing with *rabbits* then this would be reason for concern about the validity of the approach. However, we find that advection is a significant predictor ($p < 0.05$) in less than 5% of 1000 permutations of the model, as tested on the COHA dataset (i.e., as expected, given an α of 0.05), indicating that this is (hopefully) not the case.

Polysemy

We operationalized two further control variable which we omitted from the main text, semantic change and polysemy. Both are somewhat complex and difficult to parametrize. Polysemy (and homonymy) constitutes a commonly acknowledged weakness of type-based vector semantic models like LSA or word2vec, which collapse the possible multiple meanings of a word form into a single vector. We sought to estimate polysemy of target words and include it as a control variable, implementing the measure of “dissemination” proposed by Stewart and Eisenstein (2018), which is used to model a proxy to polysemy using a linear regression model predicting the (log of the) number of words a word co-occurs with (in a window of ± 2 words) by its (log) frequency, with positive residuals indicating polysemy. We found a simple linear regression to yield an inadequate fit, improved by using a second-order polynomial. However, the initial results based on COHA data were not particularly intuitive, and as a control variable it did not turn up significant in the statistical model at the end of the pipeline, so we omit this from the analysis and the main text.

Semantic change

The semantic change measure derives from our model of synonymy, which has diachronicity and context alignment already built in (see Section 2.3). Semantic change is simply a measure of the (inverse of) the similarity between the (context-aligned) vectors of words in the two time spans. Semantic change in targets cannot be estimated, as most are very low frequency in the first time span. Measuring change in nearest neighbors requires a similar range parameter as the edit distance variable, but only the semantic change of neighbors that occur frequently enough in both time spans can be estimated.

Looking at the distribution of change values which indicate most words as slightly changing between decades, we suspect there is also likely some noise in the measure, possibly due to the relatively small size of the time period subcorpora (in machine learning terms anyway). We carried out simulation experiments to probe a possible correlation between frequency difference and semantic similarity, as a proxy to frequency change possibly causing what would look like semantic change (which would be highly undesirable). We did this by taking the last decade of the COHA, making a copy, and randomly relabelling some occurrences of a sample of words from various frequency bands as *word'* in the copy (similarly to the evaluation approach in Karjus et al. 2020a). This has the effect that nothing else except the frequency of the target words changes — so if a measure of similarity between *word* and *word'* changes, then the given method of inferring semantic similarity (and change) must be frequency-biased, as it is in reality the exact same word. We measured both cosine similarity and the fact of *word'* remaining the top closest neighbour of *word*, for a range of simulated frequency differences between -10% to -99.9%, and did this with a few different count-based vector semantics models — LSA, but also full-length PPMI vectors, APSyn (Santus et al. 2016) and GloVe (Pennington et al. 2014). We find that all those are to some extent frequency-biased (echoing findings on word2vec by Wendlandt et al. 2018), at least given data of the size and composition of a COHA decade, but also that the results of LSA did remain relatively stable as long as the downsampled frequency did not fall below 100-200 (hence our choice of frequency thresholds for the context and target words).

Frameworks have been proposed to evaluate semantic change metrics (cf. Dubossarsky et al. 2019; Schlechtweg et al. 2019), but given the complexities listed above and in order to keep the main text focused on the central question, we decided to omit modelling semantic change in this contribution.

4.3 An alternative measure of communicative need

This short Section, included here for the sake of completeness, reviews an alternative to estimating changes in communicative need that I came across in the brief time period between submitting the paper above and submitting this thesis.

In a paper focused on predicting the occurrence of lexical neologisms, Ryskina et al. (2020) discuss language change as being driven by pressures of "demand" and "supply". While they do not make explicit reference to the literature, the former is conceptually highly similar to what has been referred to as communicative need (cf. Regier et al. 2016; Kemp et al. 2018; Karjus et al. 2020c), while "supply" resembles what has been referred to as the pressure for informativeness or expressivity (cf. Smith et al. 2013; Kemp et al. 2018; Carr et al. 2018). In this Chapter, I used topical advection as a proxy for communicative need, based on a simple topic modelling approach. In other words, the advection value of a target word is estimated based on the frequency change values of words it co-occurs with in the same contexts (i.e., first-order similarity). Ryskina et al. (2020) estimate demand or communicative need using the frequency change values of words that are similar to the target in a semantic space (i.e. near-synonyms in the same semantic subspace, second-order similarity).

The advantage of using topical advection in the context of modelling competition is that it is easy to de-correlate these measures — the advection model, as used here, uses a list of most associated (i.e. co-occurring) context words, while the competition estimate, equalization range, is computed based on a list of nearest neighbours in the semantic space. While these lists have some overlaps, they are mostly orthogonal to each other. Still, some potentially relevant competitors may end up being disregarded, due to also occurring in list of topic words (cf. the *airplane* example above, where *aircraft* is excluded despite being a close synonym and potential competitor). Another arguable advantage is that a topic captures words a target co-occurs with, while the subspace contains words that may be derivations or variants of the target word itself, compounds containing it, or its other case forms. The latter is less of an issue in English given its limited morphological complexity, but more so in non-isolating languages, if a corpus is not lemmatized (all those used here were) or if the lemmatization has failed to reduce some forms to their dictionary form (possibly relevant in the case of the Estonian, Czech and German corpora used here).

To compare the results and predictability of competition outcomes with a subspace-based communicative need estimate (à la Ryskina et al. 2020), some adjustment is required, as both measures would now be based on the same subset of words, the nearest neighbours in a semantic space. Since the equalization range approach, for targets increasing in frequency (the focus of this Chapter) effectively only takes into account neighbours that are decreasing in frequency, one way to do this is to calculate subspace advection based only on words that have non-negative change values, which avoid overlap. This means the resulting advection value is by definition always zero or positive, but the magnitude of the value should still be informative. The hypothesis remains the same: high advection (high communicative need) should predict less competition, low advection makes competition between neighbours more likely.

The downside is that it may end up misrepresenting the average change value of the immediate subspace (if all closest neighbours are decreasing in frequency and thus excluded). The upside is that the competition measure itself is potentially somewhat more accurate, as there is no need to filter it (in contrast to the approach taken above). In other words, neither measure is perfect.

To make the subspace-based advection measure comparable to the topical advection one (which used top 75 topic words), for delineating a subspace, I used a cosine similarity threshold c that on average also yields 75 neighbours in a given vector space (i.e. calculated individually for each LSA model, as they can slightly differ in average density). I set a threshold of minimal 10 words for a subspace advection value to be calculated (as some subspaces might consist only of decreasing words or be very sparse; but this led to only a few targets being excluded). In short, I calculated subspace advection for each target word as the weighted mean of the positive frequency changes among the target's nearest neighbours (with similarity > c), weighted by their cosine similarity scores to the target.

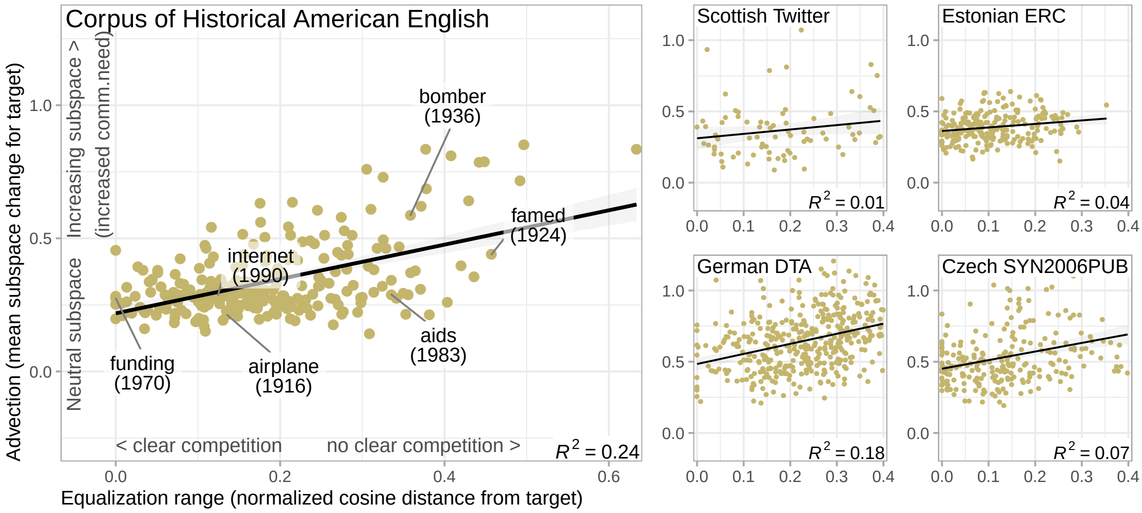


Figure 7: This is a companion graph to Figure 6 in Karjus et al. (2020b), displaying the results of applying the alternative, subspace-based advection measure for estimating changes in communicative need. The x-axis shows the equalization range — where the frequency decreases of semantic neighbors match the increase in the target. Points on the left of the plot represent target words which compete and replace their immediate neighbor(s), whereas words on the right do not. The y-axis represents the mean positive subspace change for the target word. Advection values close to 0 (lower on the plot) represent words in a subspace where words are either stable or possibly decreasing in frequency (negative change values are necessarily excluded). Points higher up in the graph represent words in subspaces that consists of neighbours increasing in frequency, which we assume reflects elevated communicative need around that subspace. As in the original figure, the R^2 values refer to the amount of variance accounted for by the communicative need variable, on top of all the lexicostatistical controls.

Figure 7, corresponding to Figure 6 in this Chapter Karjus et al. (2020b), displays the results. The differences vary by corpora — subspace advection does a slightly better job at predicting equalization range in the COHA and German DTA datasets, but worse in the others (notably in the Twitter dataset: in the full model with controls, $p = 0.18$ for subspace-based advection). The two measures are moderately correlated with one another (mean $r = 0.37$, see Figure 8), making broadly similar predictions about equalization range, but disagreeing when it comes to some individual target words. As mentioned above though, the alternative measure is somewhat handicapped here, as the subspace neighbour lists need to be filtered to avoid overlap with the competition (equalization range) measure, as discussed above.

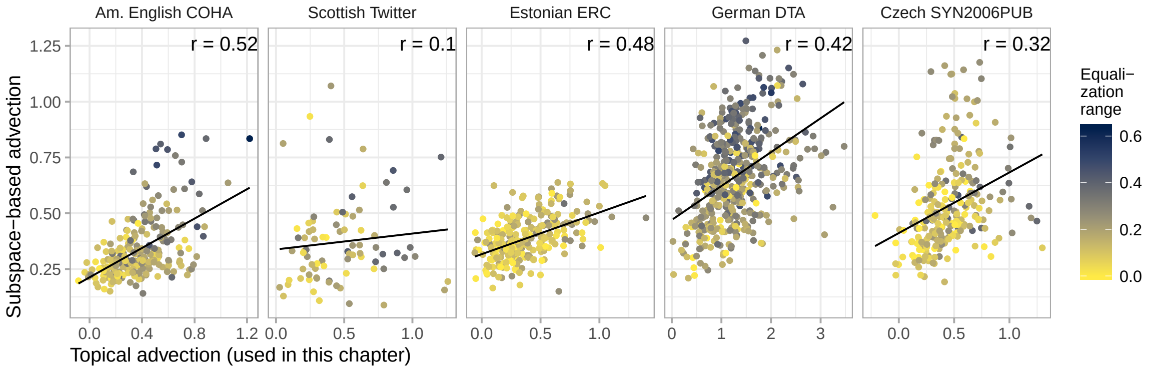


Figure 8: Comparisons of the topical and subspace-based advection measures. Each point is a target word in the corresponding dataset. Their colour corresponds to the equalization range, in particular, the unfiltered one, used in the model with the subspace-based advection (being arguably more objective; topical advection requires filtering the neighbours list for topic words, affecting the calculation of the equalization range). With the exception of Twitter data, the two approaches to advection (and communicative need) roughly agree. Low equalization range words, those competing with their neighbours, tend to have lower advection values (yellow shades, bottom left corner on the panels). High equalization range, i.e. lack of competition, is associated with higher advection values in both approaches (darker blue shades, top right corners).

The Twitter dataset again displays the greatest disagreement. As a conjecture, one notable difference between the Twitter target word set and those based on the other corpora is that, besides the vastly different time frame, it includes more proper names and proper-name-like hashtags than the other ones. This may have an effect here, as the semantics and subspaces of proper nouns may reasonably be expected to be qualitatively different from those of other words.

The divergence between these two measures is also somewhat reminiscent of the findings of Hamilton et al. (2016b) who discussed two different measures of semantic change. They showed that the one based on semantic neighbours is more sensitive to linguistic change, while the other, based on global co-occurrence statistics, is more sensitive to “cultural" shifts. It is not impossible that the two proxies to communicative need differ along similar axes. This all calls for further investigation.

Importantly though, the application of this alternative estimate of communicative need does roughly replicate the results of Karjus et al. (2020b). This provides some additional confidence that such measures do capture some real and tangible aspect of language dynamics. It is hoped that the applicability, predictive differences, and possible improvements to these various operationalizations of communicative need can be further compared and probed in future research.

4.4 Conclusions

In this Chapter, I looked into lexical competition and how variance in competition can be explained by differences in communicative need, in turn operationalized by the advection model from Chapter 3. The competition model, as proposed above, is a rather general statistical approach to change in variant frequencies over time, requiring only counts of elements in two time periods, and some operationalization of co-occurrence, in a large enough dataset, to infer similarity. No other language or culture specific resources are needed. As such, it — like the topical advection model — could potentially be applied in fields beyond (diachronic) linguistics, dealing with any products of cumulative culture, not just language.

However, all these processes can be reliably observed and inferred only to an extent, in the messy population aggregate data that are corpora and similar datasets. Therefore, Chapter 5 will turn to controlled human experiments to explore the effect of differing situational communicative needs on the lexification decisions of individual speakers.

Chapter 5

Conceptual similarity predicts colexification patterns, unless blocked by communicative need

Chapter 4 focused on the issue of competition between lexical elements. Here, I look at the other side of the coin — colexification, when a number of meanings are expressed by a single word, possibly one that competed with another and made it disappear, leading to a more sparse semantic subspace. This is in contrast to a situation where each shade of meaning in subspace is expressed by its own word (which do not compete against one another for space), leading to a more expressive language, but one that is also more complex. Here, I make use of two methodologies. I start out with an artificial language experiment, based on a fairly well established paradigm in evolutionary linguistics, in turn based on preceding work in experimental psychology and semiotics. I use it to study differences in speaker behaviour given two communicative situations differing in the crucial detail of focused communicative need. I then go on to develop a approximate measure of colexification for corpus data, and show that the diachronic results — communicative need (again inferred using the advection model from Chapter 3) describing variance in semantic subspace density — reflect the individual-level lexical choice dynamics inferred from the experiments.

5.1 Author contributions

This chapter is written with the intention to be submitted as a paper to a journal. Like the other papers in this thesis, it is multi-authored (hence the plural pronouns in the rest of the text), with the following author contributions: I designed and carried out the experiments, conducted the analysis, wrote the text, and created the figures; Kenny Smith, Richard A. Blythe and Simon Kirby provided advice on the design of the experiment, data analysis, and the corpus study, as well as edits and comments on the text.[1] Note that the references Karjus et al. (2020c) and Karjus et al. (2020b) refer to papers that respectively form Chapter 3 and Chapter 4 in this thesis.

[1] We would also like to thank Yang Xu, Barbara C. Malt and Mahesh Srinivasan for useful discussions, and Jonas Nölle for help with the initial experimental design.

5.2 Introduction

Colexification refers to the phenomenon of multiple meanings sharing one word in a language, i.e. when two or more functionally distinct senses are associated with a single lexical form (François 2008). Recognising that a word lexifies more than one sense however requires determining what the minimal units of meaning are in the first place. This would be difficult to do based on just one language, but can be done utilizing systematic cross-linguistic comparison. For example, English has individual monomorphemic words for *hand* and *arm*, while many other languages have a single word for the whole thing, e.g. Kazakh (*qol*), Swahili (*mkono*) (for more examples, see the CLICS database; Rzymski et al. 2020). This does not imply that it would be impossible to refer to these concepts in these languages, but that it may involve more complex compounds or expressions to describe them. It does imply that ARM and HAND could be considered as being the comparable, minimally distinct senses — which some languages colexify, while others do not.

In a recent study, Xu et al. (2020) demonstrate, using a large sample of languages, that similar and associated senses (like FIRE and FLAME) are more frequently colexified than unrelated or weakly associated meanings (like FIRE and SALT), suggesting that this provides an important constraint on the evolution of lexicons. This work follows a line of research on the variability of lexification patterns across languages of the world (e.g. Malt et al. 1999; François 2008; List et al. 2013; Majid et al. 2015; Srinivasan and Rabagliati 2015; Thompson et al. 2018). Studying colexification involves two questions: why a language might simplify its vocabulary by colexifying some meanings, and why it might introduce complexity by dedicating individual lexemes to some other set of meanings. It has been argued that the cross-linguistic variability in the complexity of a number of lexical subsystems — e.g. expressions of colour (Lindsey and Brown 2002; Gibson et al. 2017; Zaslavsky et al. 2019a), kinship (Kemp and Regier 2012), numeral systems (Xu and Regier 2014) and natural phenomena (Berlin 1992; Regier et al. 2016) — can often be explained by differences in communicative needs arising from differences in the cultural and natural environments of linguistic communities (cf. also Lupyan and Dale 2016). Similar contextual adaption effects have been found in experimental settings with artificial languages (cf. Winters et al. 2015; Nölle et al. 2018; Tinits et al. 2017).

Consequently, Xu et al. (2020) also propose that beyond the tendency of colexification of similar senses, language and culture specific communicative needs should be expected to affect the likelihood of colexification of similar concepts — such as SISTER and BROTHER, or ICE and SNOW — if it is necessary for efficient communication to distinguish them. A successful language needs to be efficient and learnable (Christiansen and Chater 2008; Smith et al. 2013), but also meet the speakers' requirements for expressive communication — also known as the simplicity vs informativeness or cognitive vs communicative cost trade-off (Kemp and Regier 2012; Carr et al. 2020). This perpetual optimisation process can be expected to be affected to an extent by communicative needs (Kemp et al. 2018).

Extrapolating this argument beyond the lexical subsystems mentioned above to the scale of entire languages, we would expect semantic spaces of languages to be mostly uniform in density — how many words are used to express shades of any given concept or meaning — but differ where culture-specific communicative needs of the time either require more detail or the opposite, where fewer words will suffice (analogous to uniform information density on the level of utterances; cf. Levy 2018).

In this contribution, we aim to first investigate how the typological predictions by Xu et al. (2020) play out on the grassroots level of discourse, and then test the hypothesis that communicative needs of speakers modulate colexification dynamics beyond conceptual similarity. We employ an artificial language experimental setup to probe lexification decisions by individual speakers and how they may give rise to language change — as well as a machine learning driven historical corpus approach, to survey changes in lexical densities over decades.

5.3 The experiment

Our experiment has two conditions (see Section 5.3.2 below), one where we replicate the typological findings of Xu et al. (2020) and show that, everything else being equal, speakers do indeed favour colexifying similar concepts (such as TRIP and JOURNEY) when faced with a task where the available signal space is limited. In the second condition, we manipulate local communicative need, creating a situation where colexifying similar concepts would instead hinder the accurate exchange of messages — in this case, we hypothesize speakers would resign to colexify dissimilar concepts rather than sacrifice communicative success.

5.3.1 Participants

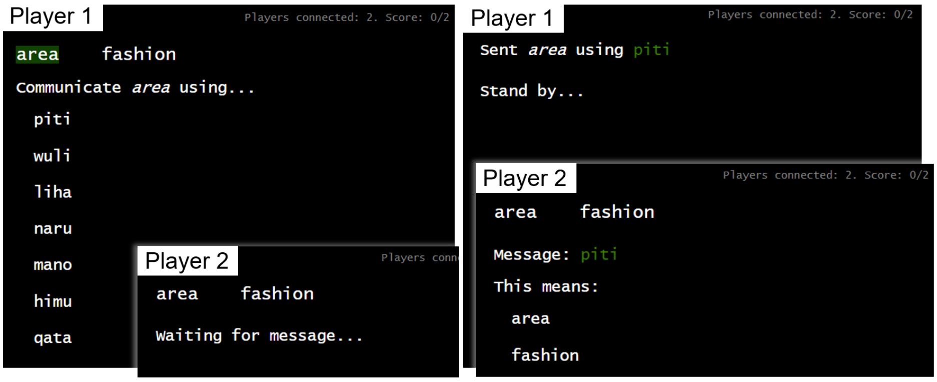
Our pool of participants consists of students of the University of Edinburgh, recruited though the university's CareerHub portal and departmental mailing lists. All participants identified as native or near-native speakers of English. The experiment was approved by the Ethics Committee of the School of Philosophy, Psychology and Language Sciences of the University of Edinburgh. All participants provided informed consent prior to participating and were compensated monetarily for their time. We collected 20 dyads per condition (80 participants total). Data from an additional 7 dyads was excluded either because they had communicative accuracy close to random guessing (we set a quality threshold of at least 70% accuracy) or did not finish the experiment.

5.3.2 Experimental procedure

Our experiment simulates language evolution using a dyadic computer-mediated communication game setup (cf. Scott-Phillips and Kirby 2010; Galantucci et al. 2012; Winters et al. 2015; Kirby et al. 2015). Participants attempt to communicate concepts using a small set of artificial words — in order to successfully communicate, the participants must negotiate the meanings for the signals through trial and error. The task is framed as an “espionage" game where the usage of the “secret codes" is justified as keeping the messages hidden from “the enemy".

The meaning space in each game consists of 10 English common nouns (see Section 5.3.3). The participants never see more than two of the meanings (nouns) on the screen at any time. In each game, there are 3 target pairs of high-similarity nouns, such as *motor* and *engine*. The rest of the nouns serve as distractors that have low similarity scores to all other nouns, including the targets (see below for details). The signal space consists of 7 artificial words such as *wewi* or *nufo*. Since there are fewer signals than meanings, participants must colexify some meanings.

The experiment consists of 135 rounds (trials). At each round, the participants take turns being the “sender" and the “receiver" of a message. The sender of the round is shown two meanings, represented by English words. The sender is instructed to signal one of them, using any word from the artificial lexicon. The receiver is then shown the same pair of meanings (in random order) and the communicated signal. The receiver is instructed to guess which of the two meanings the signal represents. After taking a guess, both participants are shown an identical feedback screen which informs them whether the receiver guessed correctly. Their roles are then switched for the next round (see Figure 1). We calculate the communicative accuracy of a dyad as the percentage of correct guesses (see Section 5.3.4). We assume it takes a while to establish stable meaning correspondences, so we consider the first 1/3 of the rounds as a “burn-in" phase, and only take into account the data after that.



```
Players connected: 2. Score: 0/2
area     fashion
Communicate area using...
  piti
  wuli
  liha
  naru
  mano
  himu
  qata
```

```
Players conn
area     fashion
Waiting for message...
```

```
Players connected: 2. Score: 0/2
Sent area using piti
Stand by...
```

```
Players connected: 2. Score: 0/2
area     fashion
Message: piti
This means:
  area
  fashion
```

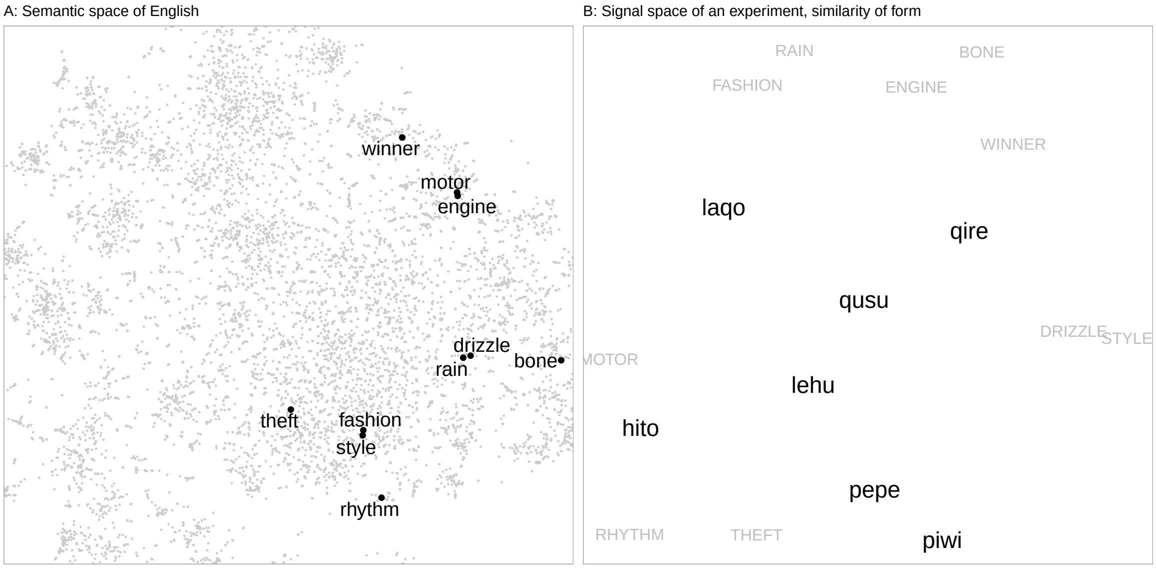
Figure 1: An illustration of one round of the game. The left side screenshots show what the players see on their screens as one of them, the sender of the round, selects and sends a message. The screens on the right show what they see while the other player, the guesser of the round, is taking a guess. After this, the feedback screen is shown, and the player roles are switched for the next round.

5.3.3 Stimuli

5.3.3.1 The meaning space

We populate the meaning space of the experiment with English nouns drawn from the Simlex999 dataset (Hill et al. 2015) which consists of pairs of words and their crowd-sourced similarity judgements. We use Simlex999, as it was built for evaluating models of meaning with the explicit goal of distinguishing genuine similarity (synonymy) from simple associativity. We use a subset of the common nouns in the dataset that are 3 to 7 characters in length. The target pairs are required to have a Simlex similarity score of at least 8 out of 10, but an association score below 1 out of 10. This should yield meanings which are near-synonymous and not simply contextually associated (e.g. a *beaver* is not synonymous with *dam* but highly associated with one).

Since Simlex does not have scores for all possible word pairs in its lexicon, we also used publicly available pre-trained word embeddings (fasttext trained on the Wikipedia dump, cf. Bojanowski et al. 2017) to obtain additional computational measures of similarity. We use these to ensure low similarity across the board in the distractor set, which was sampled so that no two distractors and no distractor-target pair would have vector cosine similarity above 0.2 (out of 1). Furthermore, no two nouns would be allowed be substrings of each other, nor otherwise similar in form (we used a Damerau-Levenshtein edit distance threshold of 3), and targets were not allowed to share the same first letter. This leaves 13 target pairs with similar meaning such as fashion-style, motor-engine and drizzle-rain. For each dyad we selected 3 target pairs and an additional 4 distractors; Figure 2.A illustrates the stimuli of one game to the backdrop of the rest of the English meaning space (here obtained from the same fasttext model, and dimension-reduced using t-Sne; van der Maaten and Hinton 2008).



A: Semantic space of English

B: Signal space of an experiment, similarity of form

Figure 2: An illustration of the meanings and signals used in one game. The left side A panel shows a semantic space of present-day English. Every dot is a word; the words used for the meaning space are highlighted. The meaning space for the game consists of 3 pairs of target meanings, all highly similar within the pairs (close to each other), but semantically distant from all other pairs as well as the 4 distractor meanings, scattered around the semantic space by design. On B panel, proximity corresponds to similarity in form, instead that of meaning (approximately; this is a dimension-reduced projection of an edit distance matrix). The signals (black) are all different in form from each other, and different from all the English words used for the meaning space (grey), none of which is very similar to each other either.

5.3.3.2 The signal space

The artificial language for the signal space of the experiment was created algorithmically as follows. Each “word" would have a length of 4 characters, randomly generated from CV syllables, in turn constructed from a set of consonants $\{qwtpsfhnmrl\}$ and vowels $\{aeoui\}$. We further constrained the artificial language so that the initial letters of the words would not overlap with any initial letters of the English nouns in a given stimulus set. We used a large English wordlist to exclude any actual English words, and made sure all artificial words were at least 3 edits distant from each other as well as from the English nouns in the same game (see Figure 2.B).

5.3.4 Experimental manipulation

In the baseline condition, the distribution of meaning pairs (e.g. DRIZZLE-RAIN, STYLE-FASHION, PAYMENT-BULL, RAIN-PAYMENT, RAIN-FASHION) is uniform — each possible combination is shown to the participants exactly 3 times. Pairs are displayed in a randomized order, under the constraint that the distributions of pairs are roughly the same in the burn-in period and the rest of the game. In this condition, we expect participants to colexify similar meanings.

In the target condition, all possible meaning pair combinations occur in the game, but we manipulate the frequencies of the pairs so that the target (related) pairs occur together more often than the distractor pairs. The target pairs (e.g. DRIZZLE-RAIN, STYLE-FASHION) are shown 11 times each, and the pairs consisting of distractor meanings 5 times (e.g., PAYMENT-BULL). Pairs consisting of a meaning from a target pair and another meaning are shown 2 times (e.g. RAIN-PAYMENT or RAIN-FASHION). This means that participants are required to select signals which allow their partner to differentiate between DRIZZLE and RAIN 11 times, but are only required to differentiate between RAIN and PAYMENT 2 times.

We manipulate the pair frequency distribution in this way to make sure all individual meanings all occur exactly the same number of times (27). It is necessary to control for individual frequencies, as simply making target pairs more frequent would also mean making the meanings in those pairs more frequent than the distractor ones. This would introduce a confound: another reasonable hypothesis could be that it is occurrence frequency that drives colexification (i.e. colexifying frequent meanings is preferred, or avoided; cf. Xu et al. 2020), over and above communicative need or word similarity.

The increased co-occurrence of similar meanings in the target condition simulates communicative need. If a pair of similar meanings never or seldom needs to be distinguished, then it is efficient to colexify them, both from a learning and communication perspective. In contrast, if the communicative context requires often disambiguating between two similar meanings or referents — such as RAIN and DRIZZLE in a culture obsessed with talking about the weather — then colexifying them as *rain* or melding them into something like *rainzzle* would obviously be detrimental to communicative success. We expect this to be reflected in the outcomes of the target condition.

It should be noted that this setup puts a possibly heavier cognitive load on the participants in the target condition. In the baseline condition, participants can colexify similar meanings (like RAIN and DRIZZLE), which is presumably easier to remember, without paying an additional communicative cost. In the target condition, the similar meanings keep occurring together, so to maintain successful communication, the participants must colexify meanings which are maximally dissimilar by design (e.g., DENTIST and FASHION).[2]

In that sense, we aim to test the strong version of the communicative need hypothesis — we predict that given high enough communicative need, speakers would rather colexify unrelated meanings than sacrifice communicative efficiency by colexifying similar meanings. The "weaker" alternative is discussed in Section 5.5.

5.3.5 Experimental results: communicative need predicts colexification patterns

Although we expected the target condition would be more difficult for the participants, differences in communicative accuracy turn out to be negligible (estimated probability of making a correct guess being 0.86 in the target condition compared to 0.88 in the baseline, $p = 0.34$).[3] There was no difference in average game length either (29 minutes in both conditions).

Our experimental setup logically lends itself to two approaches in terms of analysis, so we show results from both of them. One is to aggregate the lexicalization choices of the participants in a dyad over the course of a game and compare the colexification rates of target pairs between the two conditions. This treats the choices of the participants across a game as (hopefully) converging on a set of stable signal-meaning associations — which most of them do, judging by the communicative accuracy scores. The results of this approach are easy to interpret, but may gloss over changes in individual representations over the course of the game.

The alternative is to model the choices of every participant, trial by trial, keeping a record of their past choices, and measuring the likelihood that they re-use the same signal for multiple meanings. This requires a more complex modelling solution, but accounts for the individual behavior of the participants at greater detail.

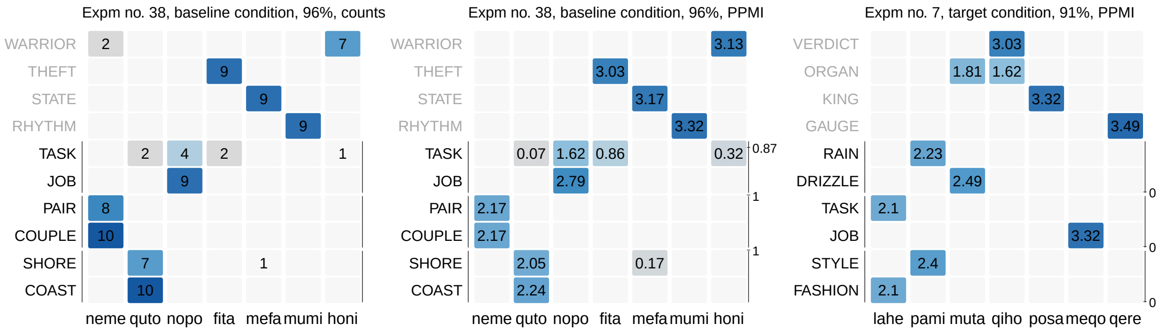
5.3.5.1 Results based on the aggregation measure

In this approach, we quantify the extent the artificial language signals are used to colexify the meanings using a simple information-theoretic measure. The summary results of each game,

[2] However, note that since the guess is always made just between two options, the baseline accuracy is 50%. This means it is still possible to achieve a reasonably high communicative accuracy score even if the dyad decides to colexify all target meaning pairs — as some dyads indeed do — and when a target pair comes up, would just take a guess. Assuming 50% random guesses hit the mark, and otherwise perfect 100% performance on all other non-target pairs (this is however never observed), it would be technically possible to achieve 82% accuracy in the post-burn-in part of the game, using this strategy.

[3] Based on a mixed effects logistic model, predicting correctness of guess by condition, with random slopes and intercepts for meaning and dyad; only taking into account the post-burn-in part of each game, and excluding the few games with overall accuracy below 70%, as described above

excluding the burn-in period, are converted into a signal-meaning matrix, where the values represent how many times each signal was used to attempt to communicate a meaning. We keep both lexifications that led to a correct guess and those that did not, but remove hapaxes, i.e. cases where a signal was used only once in the entire game, as presumable noise. An argument could be made about also filtering out all one-time-only associations between a signal and a meaning, as each "1" count (see Figure 3) represents a single (possibly accidental) attempt by only one of the participants to use a signal for a meaning, and thus does not represent a converged association by the dyad. Clearing these out would lead to cleaner measures; we present results both with and without this filter.



Expm no. 38, baseline condition, 96%, counts

	neme	quto	nopo	fita	mefa	mumi	honi
WARRIOR	2						7
THEFT				9			
STATE					9		
RHYTHM						9	
TASK		2	4	2			1
JOB			9				
PAIR	8						
COUPLE	10						
SHORE		7			1		
COAST		10					

Expm no. 38, baseline condition, 96%, PPMI

	neme	quto	nopo	fita	mefa	mumi	honi	
WARRIOR							3.13	
THEFT				3.03				
STATE					3.17			
RHYTHM						3.32		
TASK		0.07	1.62	0.86			0.32	0.87
JOB			2.79					
PAIR	2.17							1
COUPLE	2.17							
SHORE		2.05			0.17			1
COAST		2.24						

Expm no. 7, target condition, 91%, PPMI

	lahe	pami	muta	qiho	posa	meqo	qere	
VERDICT				3.03				
ORGAN			1.81	1.62				
KING					3.32			
GAUGE							3.49	
RAIN		2.23						0
DRIZZLE			2.49					
TASK	2.1							0
JOB						3.32		
STYLE		2.4						0
FASHION	2.1							

Figure 3: Signal-meaning matrices from a baseline condition (left) and a target condition game (right). The vertical black bars highlight the similar-meaning pairs, with their colexification values displayed on the two PPMI panels (most of them here are either 0 or 1). Count values in the cells on the leftmost panel indicate how many times each signal was used to communicate a given meaning. The middle panel illustrates how the counts translate to PPMI association scores. In this baseline condition game, the players have chosen to colexify similar meanings such as SHORE and COAST, as expected. In the target condition, similar meanings often need to be distinguished from one another. The players in game no. 7 (right) have figured this out and colexified RAIN with STYLE and DRIZZLE with ORGAN instead.

Since we are after association between meaning (m) and signal (s), we convert these counts into Positive Point-Wise Mutual Information scores (see Equation 1), which gives a more useful picture than just signal-meaning counts, discounting signals that are used indiscriminately across the board.

$$\mathrm{PPMI}(m,s) := \max\left\{\log_2 \frac{P(m,s)}{P(m)P(s)}, 0\right\} \tag{1}$$

The left and middle panel of Figure 3 illustrates this conversion process. Each meaning ends up with has a vector of PPMI scores (the "rows" on Figure 3). The colexification score for each target meaning pair is just the cosine similarity between these vectors — like SHORE and COAST in Experiment number 38, the middle panel, which have a colexification score of almost 1 (0.9966 to be more precise); while TASK and JOB get a score of 0.87, as they are sometimes lexified using different signals. In Experiment number 7 (left side panel), all target pairs get a colexification score of 0, as they are consistently lexified using different signals. The same values but across all dyads are plotted in Figure 4.

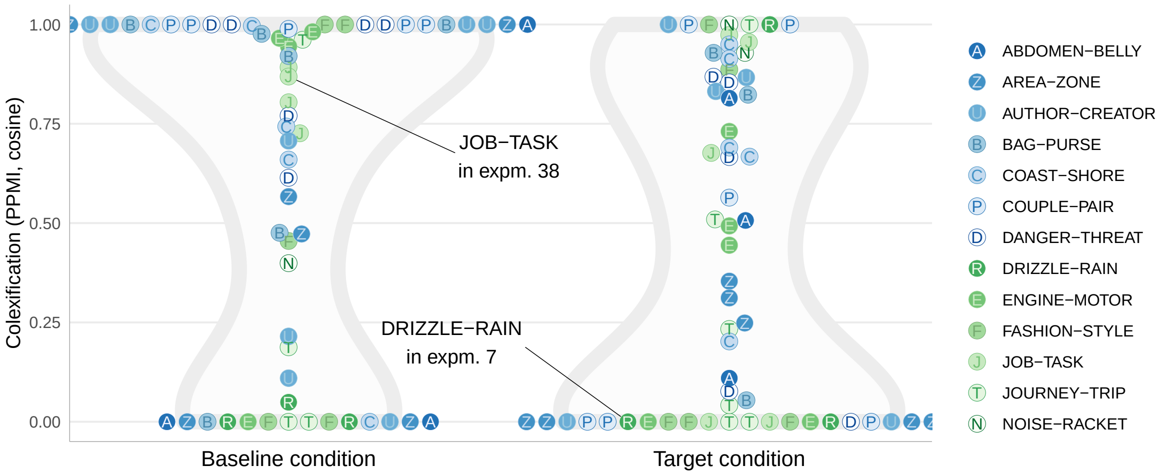


Figure 4: Results based on the aggregation approach (without filtering out single associations). Each point is a meaning pair, with the smoothed density violins (grey) illustrating their distributions. In the baseline condition (left side violin), participants are more likely to colexify similar meanings, but occasionally some pairs e.g. JOURNEY-TRIP still get assigned separate signals. In the target condition (right), participants are more likely to avoid colexifying target pairs, but not always either.

We test the difference between conditions using a mixed-effects quasi-binomial model with a logit link function in a generalized additive regression framework (using the mgcv package in R; Wood 2011), predicting the colexification value (vertical axis in Figure 4) by condition.[4] The estimate of the fixed effect of condition is negative (with the baseline condition as the reference level, indicating lower colexification in the target condition), and of similar magnitude regardless of the filtering choice on the one-time-only associations, while the p-value varies around the conventional $\alpha = 0.05$ threshold. In the model without the filter, $\beta = -0.83$, $SE = 0.48$, $p = 0.088$, i.e. in the target condition, the estimated probability of colexification decreases to 0.16, compared to 0.63 ($\beta_0 = 0.54$) in the baseline. When one-time-only associations are filtered out, then $\beta = -0.86$, $SE = 0.42$, $p = 0.045$.

5.3.5.2 Results using the trial by trial measure

In this approach, we obtain the colexification variable by iterating through all the (target meaning) messages sent in each game, and checking if the most recent usage of a given signal by the same player lexified another meaning, and if so, was that other meaning related to the current one (i.e., in the same target pair). Messages containing a distractor meaning are excluded, as they do not have any semantically related counterparts in the meaning space by design. Cases where the same signal was used to lexify the same meaning again are excluded as well (as there

[4] With a random slope for condition by meaning pair, and a random intercept by dyad, to take into account the repeated measurements from dyads and meaning pairs. The response variable varies (and is censored) between [0, 1], with most of the values at or near the boundaries; as such, common models like linear, binomial or beta regression would not be suitable.

is no colexification to account for). This allows us to compare the conditions in terms of the likelihood of target pairs being colexified.

To exemplify: given a player signals RAIN using *pami*, we check the most recent usage of *pami* by the same player. If they previously used *pami* to signal DRIZZLE, then that counts as colexification between related target meanings, and a case gets added to the new dataset, with the colexification variable assigned as "yes". If *pami* was last used to signal an unrelated meaning, e.g. *dentist*, then that yields a "no", i.e. a colexification happened, but not between target meanings. If *pami* was last used to signal RAIN, then nothing is added to the dataset.

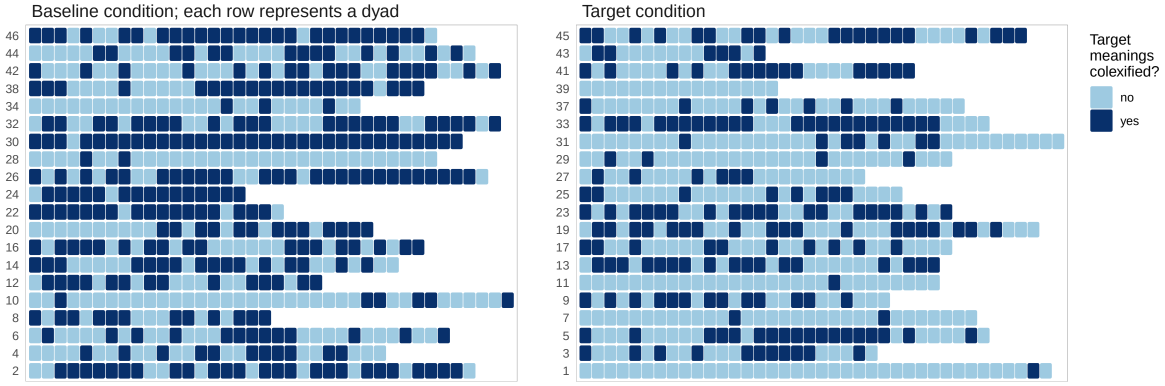


Figure 5: All the data resulting from the trial by trial measure. Each tile corresponds to a message sent by a player that contains a target meaning, which is being colexified with another meaning using the same signal (dark blue if with a related meaning, e.g. RAIN-DRIZZLE, light blue if with an unrelated meaning). Trials are shown in order, but those involving distractor meanings, or where the given signal was last used to lexify the same meaning, are excluded. Some dyads therefore yield fewer data points than others, if they assigned unique signals to target meanings, resulting in less data points involving colexification. The difference between conditions is visually apparent: the baseline condition on the left has more dark blue, indicating colexification of similar meanings, while the target condition has more light blue tiles.

This procedure, repeated for all players across all 40 games, yields a dataset of 1183 cases, a median of 30 per dyad; see Figure 5.[5] We then fit a mixed effects logistic regression model (using the lme4 package in R; Bates et al. 2015) with the binomial colexification variable as the response, predicted by the interaction between condition (baseline or target) and trial number, to account for possible changes over the course of the game. We fit random intercepts for meaning and sender (the latter nested in dyad), and a random slope for condition by meaning (a full random effects structure would be desirable but could not be included due to model convergence issues).

To simplify interpretation, trial number is centred to the middle of the main part of the game (i.e. after the burn-in). In the model described in Table 1, the intercept value of 0.1 therefore

[5] As in the aggregation approach above, we exclude the data from the burn-in period from the statistical model, but here do take the burn-in into account when checking for most recent lexification for signals, as the players do not start from a clean sheet after the end of the burn-in, but at least to some extent already have a system in place.

stands for the log odds of target meanings being colexified in the baseline condition, mid-game (i.e. a 0.52 probability). By mid-game, the difference in colexification probability between the conditions is only marginally significant ($p = 0.08$). The model indicates each passing trial does increase the probability of colexification in the baseline condition. Importantly, the interaction between condition and trial is in the opposite direction ($\beta = -0.02$, $p = 0.002$), indicating participants were less likely to colexify related meanings in the target condition (at the end of a game, the pooled probability estimate of that is only 0.29, compared to 0.69 in the baseline condition).

	Estimate	SE	*p*
Intercept	0.1	0.37	0.79
condition = target	−0.89	0.51	0.08
trial number	0.02	<0.01	<0.01
condition × trial	−0.02	<0.01	<0.01

Table 1: The fixed effects of the regression model used to analyse the trial-by-trial measure.

In summary, having applied the two alternative methods of analysis to our experimental data, we find converging evidence indicating that, when provided a limited signal space, participants are prone to colexifying similar meanings (confirming the typological find by Xu et al. 2020). However, when faced with a situation where there is elevated communicative need to distinguish related meanings, participants are more likely to colexify some pairs or clusters of meanings which have low inter-similarity, to maintain communicative efficiency. This confirms the prediction made by Xu et al. (2020), and is in line with previous research on communicative need in general (cf. Section 5.2). We now turn to a historical corpus with the aim to test a corresponding diachronic hypothesis.

5.4 The corpus study

The aim of our corpus study is to replicate the assumed mechanisms behind our experimental findings on data from historical time scales. The corpus-based Chapter 4 (Karjus et al. 2020b) explored the relationship of lexical competition and communicative need, with the conclusion that the latter modulates the former, with elevated communicative need allowing for larger number similar words to co-exist in a semantic subspace instead of competing with each other. The colexification angle is the other side of the same coin: Chapter 4 focused on modelling competition between similar words, i.e. members of the same subspace or cluster, as operationalized below — while here we are interested in how these dense subspaces or tight clusters of words emerge in language. Communicative need is predicted to play a role in both processes. This section builds directly on the word embeddings-supported corpus approach developed in Chapter 4 (Karjus et al. 2020b); the technical details of that will be revisited below in the context of the topic of this section. As before, we make sure to also control for a number of other relevant lexicostatistical variables (e.g. similarity of form).

We use a large diachronic English dataset (the Corpus of Historical American English, COHA;

Davies 2010), infer semantic spaces using word embeddings, and correlate changes in subspace density to changes in population-level communicative need, approximated by another computational measure. Figure 6 illustrates this approach; this is a temporally aligned word embedding model of semantic spaces based on COHA data (approximate 2D t-Sne projections of the high-dimensional vector space models; cf. van der Maaten and Hinton 2008), showing how most semantic subspaces remain relatively stable over time, while some gain and lose words.

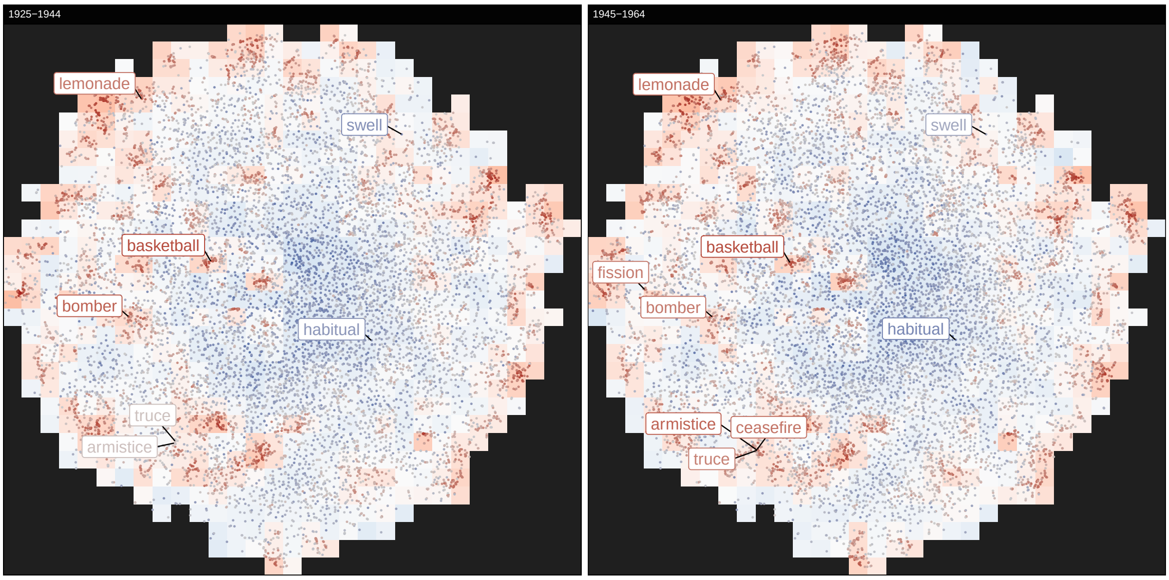


Figure 6: Two aligned semantic spaces, based on data from the Corpus of Historical American English, the years 1945-1964 (left) and 1925-1944 (right). Each dot is a word; colour indicates density (as average proximity to top 10 nearest neighbours), centred on the mean, so words in average-density spaces are light grey, words with few and distant synonyms are blue, while words with many similar terms or synonyms are red. Note that some blue dots may still appear close together, as this is only a rough approximation of the high-dimensional vector space. The maps are further divided into 30x30 grids for visual aid, each cell showing the mean of the values therein (in slightly lighter shades; empty areas are black). Most subspaces stay the same in terms of density: for example, having detailed basketball terms remained relevant in common discourse both before and after World War 2, while the term *habitual* continued not needing a close synonym. Some areas change: the subspace of *truce*, *armistice* (bottom left) gains a number of terms at the end of war.

The experimental evidence (Section 5.3.5) supports the idea that speakers adapt their languages to meet the communicative needs of the situation. It follows that the compounding of such smaller changes may eventually lead to a change in the population consensus regarding the given expression. While an experimental setup allows us to look closely at interaction as it happens, it is not straightforward to predict how these choice patterns in (very) short timescales in an artificial communication scenario would translate into language change dynamics on historical timescales. We therefore conduct a corpus study with the aim to link individual-level choices with population-level changes. While this link cannot be demonstrated explicitly, finding a similar correlation of colexification dynamics and communicative need would support the idea that languages are shaped, both in the moment and over centuries, to adapt to the needs of the speakers.

We introduce a novel measure of colexification applicable to corpus data, based on detecting tight clusters of words in the semantic space of a language. A dense cluster (see the red areas in Figure 6) indicates that a concept is likely *not* colexified, but instead lexified by multiple words. A sparse area of a space (blue in Figure 6) can be taken to mean that the concepts therein *are* colexified. We focus on clusters that are tighter than expected based on the structure of the rest of the space, and track their formation, survival and dissolution — any of which should say something interesting about the communicative needs of the time. If multiple words are used to refer to similar concepts, then presumably communicative needs are elevated (comparable to the target condition in our experiment). If concepts are colexified, then it is probably because there is less need to distinguish them (comparable to our baseline condition).

We make use of a corpus-based measure of communicative need from previous research, the topical advection model (Karjus et al. 2020c; Karjus et al. 2020b). Using data from the Corpus of Historical American English (COHA; Davies 2010), we demonstrate that changes in communicative need are predictive of changes in colexification, i.e. how many words in common language use are being employed to describe a concept (i.e., a subspace in the semantic space). Our approach relies on machine learning and does not require any language-specific resources other than a sufficiently large diachronic corpus. While we use English as an example here, it should be in principle readily applicable to any other language.

5.4.1 Quantifying colexification

Colexification is defined as multiple meanings sharing one word form. While it is easy to estimate the number of words in a language using a large corpus, it would be difficult to estimate the number of meanings (that may or may not be colexified). Typologists often approach similar issues of estimating the total set of meanings or grammatical functions by manually surveying grammars or dictionaries of a large number of languages for the lexifications therein (cf. Haspelmath 2003; François 2008; Rzymski et al. 2020). However using already compiled cross-linguistic lexical databases (as done by Xu et al. 2020) would still entail the highly laborious manual task of matching the meanings to words in a historical corpus.

We take a probabilistic approach, with the following rationale, to determine which subspaces of a semantic space are colexified and which subspaces (meanings) are expressed with more than one word. Like in Chapter 4 (Karjus et al. 2020b), we infer the semantic spaces from corpus data using a distributional approach (Section 5.4.2 goes over the details of that), yielding a high-dimensional vector space, where each word is represented by a vector, and their similarity calculated as the cosine between these vectors. Partitioning a high-dimensional space would be difficult enough, even more so when the number of partitions (meanings) is unknown. We solve this problem by flipping the question: instead of asking which meanings are being colexified, we can ask which meanings are *not*, or more specifically, are but to a lesser degree.

We make an assumption that it is in principle possible to split every meaning into infinitely finer shades of meaning — that meaning is functionally infinite. Natural languages just stop at some point: for a language to be learnable and an efficient tool of communication, its lexicon must be of a size the average speaker can remember. It is useful to differentiate *apples* from *oranges*, and maybe even different types or sorts of *apple* from one another (have a cluster of various *apple*-words). However assigning unique words to every possible instance of apples differing in minute detail would likely be inefficient both in terms of communication and learnability. This is the complexity-informativeness trade-off discussed in Section 5.2. It follows from this assumption that every word in a corpus can be treated as somewhat polysemous, colexifying multiple meanings by default. If there are more words than expected within some specified range from one another, then this would indicate that this subspace is colexified to a lesser degree. In contrast, a lone word distant from all others can be assumed to colexify its subspace (cf. the blue areas in Figure 6).

With this in mind, the first goal is identifying dense subspaces of tightly clustered similar words (e.g. {orange, mandarin, tangerine}. Data on these clusters will be aggregated, amended with relevant lexicostatistical variables, and their formation (change in type frequency, i.e. cluster size) will be correlated with an estimate of changes in population-level communicative need.

To do that, we need a measure of what counts as a cluster in a continuous space. We use cosine similarity (of the vectors of) of words — if the inter-similarities of two or more words are above a chosen threshold, then they can be considered a cluster (for a similar density-based approach to neologism prediction, cf. Ryskina et al. 2020). We obtain the threshold to delineate subspaces by taking the cosine similarity of each word to its nearest semantic neighbour in a given temporal subcorpus model (e.g. 1945-1964 in Figure 6), ranking these values, and obtaining the cosine similarity value *c* at the 95th quantile — the range at which in 95% of the cases there would be no other word (see Figure 7). All words with a similarity to their nearest neighbour at $\leq c$ — i.e. left of the black line in Figure 7 — are considered colexifying the (shades of) meaning in their local subspace. All words with a nearest neighbour at $> c$ exist in subspaces with higher than expected density, i.e., the infinite meanings of the given subspace are still colexified, but to a lesser degree. There may be more than two words clustered in a subspace, counted as such if all their inter-similarities are $> c$.

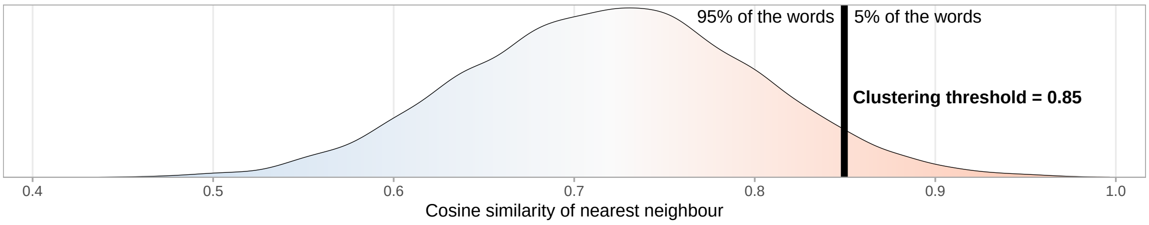


Figure 7: A typical distribution of nearest neighbour similarities in an LSA model trained on a 20-year subcorpus of COHA. The inferred colexification threshold *c* is marked with the bold vertical line. The shading corresponds to the colour scale of the background cells in Figure 6.

These are the subspaces or clusters we are interested in. When such a dense cluster forms over the course of the history of a language, then, given the communicative need hypothesis explored in our experiment, we would expect the given subspace to be experiencing elevated communicative need. This is what we are going to test: we will obtain a dataset of clusters and their lexicostatistical properties from the diachronic corpus, and correlate type frequency changes in these clusters to changes in communicative need in the subspace (inferred using the topical advection model, cf. Section 5.4.3). However, there are some technical details that need to be discussed first.

5.4.2 Detecting near synonym clusters

The dataset is compiled in the following manner. We run a sliding window of 20 years in length (in 5-year steps) over the COHA data, train a temporally-aligned distributional vector space model for pairs of 20+20 years periods (see below), and identify any clusters of words in close proximity (all cosine similarities $> c$ in our model of semantics; recall the dark red areas in Figure 6). COHA spans 200 years, but we only use the 20th century data, as it appears to be better balanced than most 19th century decades.

We require words to be sufficiently frequent for reliable statistical inference — to occur in at least 15 out of the 20 years, and at least 100 times total within the 20-year window window (a typical year of COHA consists of about 1 million words after stopword cleaning). This threshold pulls double duty: it makes sure we only consider words that are frequent enough for reliable inference (hence the raw frequency threshold, not a normalized one), but also words that are widespread, part of common language usage. When talking about words going out of use, we are referring to them falling below this threshold of common usage. A potential shortcoming of using a threshold like this is that, in the semantic spaces based on the data, it could make words that just shift slightly around the 100 threshold look either as new words or words going extinct. We therefore exclude all such clusters where type frequency change stems from near-boundary token frequency change.

There are logically two ways to analyse the history of word clusters: either to look at their past, i.e. how they formed, or their future, i.e. what happens to them once formed. We will refer to these as the “lookback" and “lookforward" model respectively, and use the same length of 20 years for both the “past" and “future" time span. Figure 8 illustrates this on the example of the lookback model. The resulting datasets of clusters and their attributes, prepared for statistical modelling, feature no repeated measures within a given model type, as we make sure the clusters do not overlap (within their 20+20 time spans). In the case of shared words between clusters within a model type, we keep the tightest one (highest mean inter-similarity between members), and in the case of overlapping words between clusters over time, we keep the biggest one (breaking ties again by inter-similarity). This filtering by maximizing cluster size also results in the feature that in the lookback model, the change values are either 0 or increasing, and in the lookforward model, they are all 0 or decreasing.

We infer the semantic space and diachronic changes therein using the setup developed in Karjus et al. (2020b), training Latent Semantic Analysis (LSA; cf. Bullinaria and Levy 2007) models

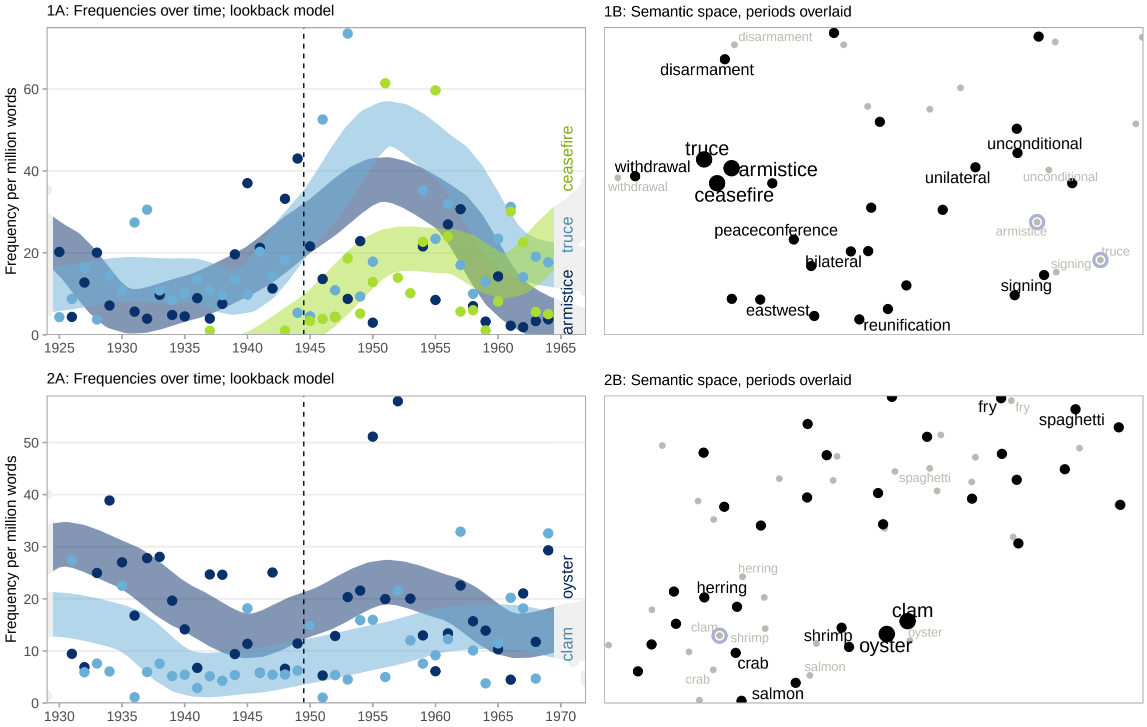


Figure 8: Examples of clusters, with their time series on the left (with the vertical dashed line indicating the border between the 20-year spans) and their corresponding local semantic spaces (again dimension-reduced using t-SNE) on the right. Black dots therein represent words in the second time span (right of the dashed line on the time series panel), grey dots their "old" meanings (first time span). Towards the end of WW2, *truce* and *armistice* (top row) increase in frequency in discourse, joined by a new term, *ceasefire*. The semantic space (1B) shows that the former two shifted somewhat in contextual meaning, moving closer to *withdrawal*, *disarmament*, and the new term, *peace conference*. Bottom row: *clam* and *oyster* are used fairly consistently throughout the middle part of the 20th century, but the meaning of *clam* becomes less herring'y and more synonymous with oysters, reflecting presumable changes in its usage context.

for each span of data where a dense word cluster is found. As each cluster is associated with two 20-year time spans (both in the lookback and lookforward model), the LSA model is trained on a joint aligned (PPMI-weighted) co-occurrence matrix based on the two spans. Analogously to the approach proposed by Dubossarsky et al. (2019), each word is suffixed with the time span label (first or second), which allows us to estimate semantic change (cosine distance between $word_t1$ and $word_t2$), and measure changes in subspace density, i.e., the type frequency within each cluster. For example, in the lookback model, upon finding a cluster like {*truce, armistice, ceasefire*} (see Figure 8), we can check how many words existed in the same subspace (i.e. within the range of the colexification threshold) — which in the latter case is 0 words: the forming of this cluster involved semantic shift in two old words and the introduction of a new one.

Similarly to the small token frequency shifts exclusion criterion (see above), we exclude clusters where changes in type frequency are caused by small shifts in the semantic space (above the colexification threshold), which might as well be sampling noise in the data underlying the distributional model. Naturally, homonymy is a concern for any distributional approach, particularly in a model where each form is represented by a single vector (while time period specific training and alignment is considerably easier than it would be with a contextualized model e.g. Devlin et al. 2019). We hope that the size of our resulting datasets somewhat alleviates this unavoidable source of noise.

Since spelling of compounds in English varies, we homogenize the COHA data by removing all punctuation (including hyphens) from lemmas, and concatenate the most common multiword units, using the same setup as Karjus et al. (2020b) (looking up strongly associated units in the first pass over the raw corpus, in 20-year steps, and then concatenating these units in the second pass while cleaning the corpus). The latter step is motivated by the fact that the spelling of compounds in English varies, involving spaces, hyphens, and concatenation. Since we expect lexical innovation to be among the mayor sources in lexical density change, detecting such units, with all their initially likely variable spellings, is of interest (e.g., *web site, web-site, website*).

A distributional semantics approach can conflate similarity relations other than synonymy, so we expect there to be some noise and false positives in our measure. However, we survey a sample of the word clusters produced by our approach (709 unique clusters in total across all the 4 setups, see Section 5.4.4) and find the results adequate to proceed with statistical analysis. Most clusters make sense: such as spelling variants e.g. {*instalment, installment*}, near synonyms (like {*birth control, contraception*}, {*amazement, astonishment*}, {*sitting room, living room*}, and things that could be labelled with a single general word like {*bourbon, gin, scotch*} or {*protestant, catholic*}. Some clusters are somewhat less intuitive, such as those containing similar concepts which could be considered antonyms like {*east, west*}, tangentially related concepts like {*tennis, golf*}, things that just often occur together e.g. {*cheese, bread, butter*}, or derivations of the same root ({*constitution, constitutional*} — however the fact of derivation itself could be seen as indicating elevated communicative need around the topic).

5.4.3 Quantifying communicative need and control variables

Another technical aspect of our corpus study involves the quantification of changes in population-level communicative need. Analogous to Karjus et al. (2020b), we use the topical advection model from Karjus et al. (2020c) to estimate the shifts in latent topics between time period sub-corpora.[6] This version of the advection model operationalizes a topic for each word, as an association-weighted list of words that most commonly co-occur with the target word in context. The association-weighted mean log frequency change of the context words is the advection value.

Topical shifts can be taken taken as an approximation of changes in what the speakers need to

[6] Chapter 4 also explored an alternative measure of communicative need, based on the mean frequency change of semantically similar words in the neighbouring subspace (cf. Ryskina et al. 2020). As the results were roughly comparable to the topical advection measure there, only the latter is used here. Other measures could be explored and compared in the future.

communicate about, and as such, the communicative need related to the topic. If a topic is of importance to speakers, then it is reasonable to expect that speakers use the related vocabulary more, and in more detail when it comes to the semantics in the discourse to successfully communicate more fine-grained distinctions, which may in turn result in the coining or borrowing of new words or repurposing old ones. Conversely, if a topic is decreasing in prevalence, then presumably it is of lesser importance, and it can be expected less detailed semantics will do. This was essentially the main finding of Karjus et al. (2020b), which focused on the interword competition aspect. The advection model assigns a topic to each word; since we are dealing with clusters, we simply use the mean value of the words within a given cluster. Since both the topic model and semantics model rely on co-occurrence statistics, we must make sure to avoid any autocorrelation between the measures. We do this by excluding words belonging to the same cluster from the topic word lists, before calculating the advection value.

To test the hypothesis that communicative need is predictive of changes in colexification, we correlate the advection (topic frequency change) value of a cluster with the change in its type frequency, i.e. how many words form the cluster (Section 5.4.4). We also control for a number of other lexico-statistical variables: the mean log token frequency of the words in the cluster (as more widespread words may behave differently than less common words; and may suffer from estimation issues, cf. Faruqui et al. 2016), mean semantic change (to differentiate clusters which form due to small shifts in meaning from those including words which meaning has been categorically changed), as well as mean of the edit distances between the cluster members (as spelling variant clusters may behave differently from near synonym clusters).

5.4.4 Results of the corpus study

Figure 9 illustrates the results of our analysis. Since the response variable of change in type frequency in cluster is best modelled as an ordinal one, we use a proportional odds logistic regression model. The R^2 values reported in Figure 9 are calculated based on log likelihood, as $1 - \text{logLik}(\text{full model})/\text{logLik}(\text{null model})$, and the improvement provided by the advection variable compared to the controls-only model, as well as average baseline-adjusted accuracy (kappa) scores based on 10-fold cross-validation (and the respective improvement).

We test both the lookback and lookforward version of the model, on two datasets: one with all eligible clusters as described in Section 5.4.2, and one with only clusters where a change involves the introduction or extinction of a word. The former includes cases where the change in or a formation of a cluster is just caused by words shifting in the semantic space. The latter also includes eligible stable (i.e. zero-change) clusters for comparison.

After accounting for the lexicostatistical control variables (Section 5.4.3, the explanatory variable of topical advection remains as a small but significant effect in all models except the full lookforward model (Figure 9.2A; apparently communicative need, or at least given our estimation of it and our chosen time spans, is less informative when it comes to the dissolving of lexical clusters).[7] In summary, changes in our estimate of communicative need correlate with changes in lexical density: rising need tends to lead to higher-than-expected density in the semantic space (Figure 9.1A, 1B; comparable to the target condition of our experiment) — while lower communicative need facilitates colexification (at least when words go out of use, not just shift in meaning, i.e. 9.2B).

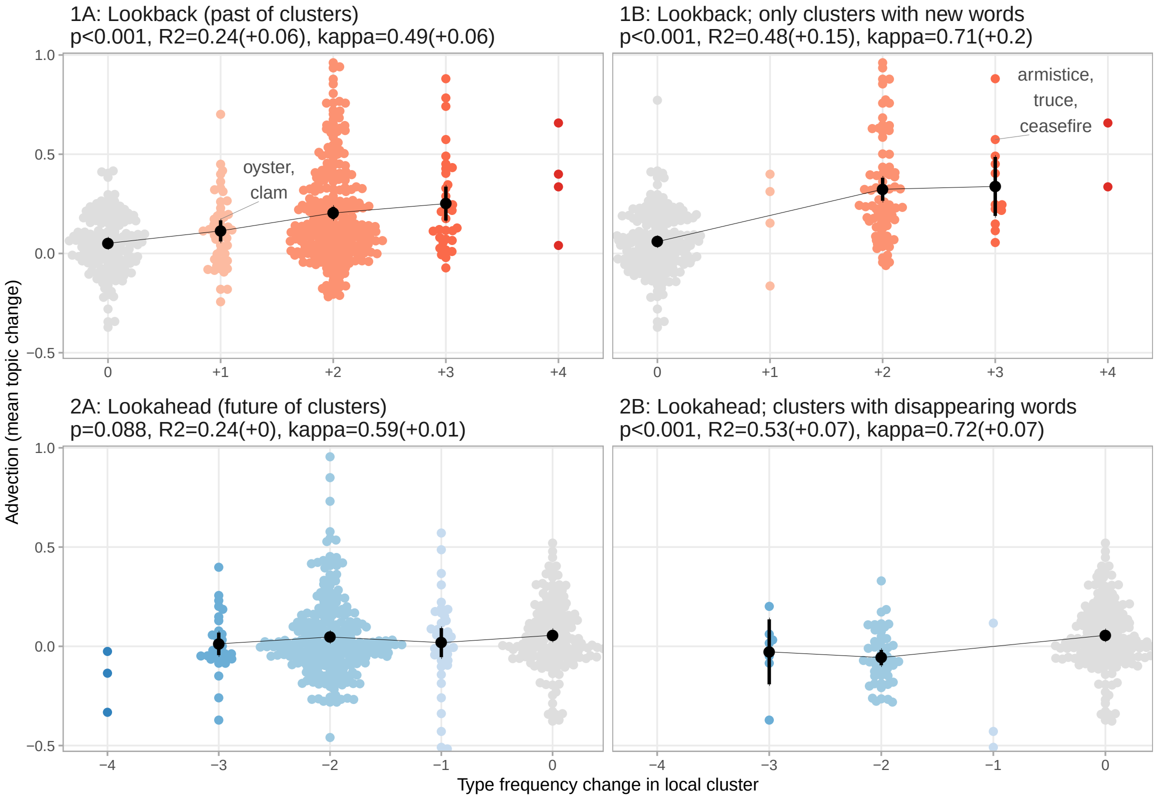


Figure 9: Results of the model correlating changes in communicative need (y-axis) and changes in the number of words in a cluster (x-axis). Each dot represents a near-synonym cluster like *clam, oyster*. Means across each (larger group of) change value are displayed as black dots, connected by lines for visual reference. Top row: the lookback model, quantifying the changes that led up to clusters forming; bottom row: the lookforward model, observing the changes in clusters found in the data. Left column: all eligible clusters; right: only clusters where a new word or a word going out of use was involved in the change in density. The reported (model log-likelihood based) R^2 values describe the full model (with the control variables), with the added descriptive power of the advection variable shown in brackets; same for the kappa (baseline-adjusted accuracy) scores; p-values correspond to likelihood ratio tests comparing the reduced controls-only model to the full model, reflecting the significance of the advection variable.

[7] It also appears changes of ±1 are not very common in our models, apprent in Figure 9. A cluster requires at least two adjacent words by definition, while a change of ±1 would require there to have been at least one word in the subspace already (lookback) or a cluster of at least size 2 that loses a single word (lookforward) — which seems to be less common cases, at least in our operationalization of colexification.

5.5 Discussion

In our experimental approach and corpus study, we find evidence supporting the idea that speakers' communicative needs drive or at least modulate aspects of language change. While neither results from an unnatural communicative scenario like a short gamified dyadic experiment, nor highly noisy correlates extracted from historical corpora using machine learning based rough estimates — would perhaps be particularly conclusive on their own, finding similar effects on both fronts provides us with some confidence in our findings.

However, our experimental and corpus approach explored only one a small subset of possible relationships between these forces and processes. There are a number plausible alternatives to our chosen experimental setup that could be probed in future research. One is to test the “weaker" version of the hypothesis mentioned in Section 5.3.4: our design essentially pushes participants to colexify meanings which are highly dissimilar; they appear to struggle with that, understandably, and subsequently the results are somewhat noisy. It may be worth exploring an alternative, more “natural" target condition where the distractors also form high-similarity pairs, which would be easier to colexify, while assigning individual signals to the target meanings (the ones that need to be distinguished more often).

We chose direct similarity or synonymy as the semantic relationship to explore. Xu et al. (2020) show that conceptual associativity (i.e. the *beaver-dam* type) also correlates with cross-linguistic colexification patterns. It would be interesting to see if, when given a choice, the relation more preferable in terms of colexification for speakers would be associativity or similarity. Differences between more fine-grained relationships like register-varying synonymy (*abdomen, belly*) and hyponymy (*bag, purse*) could also be probed, as well as how these preferences may correlate with historical patterns of sense formation and expansion (cf. Ramiro et al. 2018). Xu et al. (2020) also discuss a potential role of frequency — if more commonly referred to senses may be more likely colexified. We control for frequency in our experiment by making sure the occurrence distribution of meanings is uniform in any given game. Future experimental research could let frequency vary systematically to determine its importance. As another extension, community size and network effects could be explored in using a larger-scale experiment (cf. Raviv et al. 2019). Instead of fixing communicative need by condition as we did, changing pressures could be introduced over the course of a longer experiment instead.

Similarly, the corpus methodology could be built upon in a number of ways. The technical aspects of and possible improvements to the distributional semantics model and the topical advection model that we re-used here are discussed in Chapter 4 (Karjus et al. 2020b). The approach does not require any linguistic resources other than a large corpus, and as such could be readily applied to other languages beyond English. Lexical density derived from vector spaces is of course only a rough approximation of the actual semantic space of a language, and not without issues (cf. Faruqui et al. 2016). The estimate of lexical density introduced here, as a proxy to colexification, also requires an arbitrary threshold parameter (we use the 95th quantile). Alternative measures could be explored, and polysemy or colexification measures from historical lexical databases (which would presumably be more accurate; cf. Ramiro et al.

2018; Xu et al. 2017) could be aligned to historical corpus data, and compared to the results of the automatic measure. Finally, while our application was diachronic, synchronic semantic space density in a larger sample of languages (cf. Thompson et al. 2018; Thompson and Lupyan 2018; Rabinovich et al. 2020) could be compared to some synchronic proxy of communicative need (cf. Regier et al. 2016).

5.6 Conclusions

We replicated a typological finding from earlier research on colexification tendencies, and tested the claim that colexification dynamics may be driven not only by similarity between concepts but also the communicative needs of linguistic communities. We investigated this interaction using an artificial language experiment, to probe how meaning to signal ratios are shaped in discourse, and found speakers readily colexify similar concepts as expected — unless distinguishing them is vital for successful communication, in which case they do not. We also conducted a diachronic corpus study, where our population-level findings mirrored the observed individual-level tendencies in the experiments: rising communicative need in a semantic subspace correlates with higher subspace density, i.e. more words to tease apart finer shades of meaning.

Language change is driven by a multitude of interacting forces, ranging from random drift to sociolinguistic pressures to institutional language planning, to selection for more efficient and expressive forms. Our work supports the argument that speakers' communicative needs — a factor balancing and modulating the relative importance of the higher-level pressures for efficiency and informativity — should be considered as one of them. In more general terms, living languages will never stop changing because (among other things) the needs of their speakers are dynamic, existing in a state of constant flux.

Chapter 6

Conclusions

In this thesis, I studied dynamics of language change, with particular focus on words, how their usage ebbs and flows over time, their usage changes, and how their functions are taken over by new words. I evaluated a proposal to categorise the rise and fall of linguistic elements as being driven by drift or selection (Chapter 2), showed that the fluctuations of word frequencies are predictable to a degree (Chapter 3), and that the outcomes of historical competition between lexical items and changes in lexical density in the semantic space are modulated by communicative need (Chapters 4, 5) — which can also be observed driving lexification choices on the level of discourse, as evidenced by an experimental investigation (Chapter 5).

6.1 Future directions

The sections below further detail some technical and theoretical contributions of this thesis. There is yet much to be done on the front of understanding the interplay of communicative needs, other communicative pressures, and language change. I hope that this thesis provides some useful methods and starting points for doing so, both from a corpus-based and an experimental angle.

The computational approaches developed here for inferring changes in needs, competition, and colexification provide rough estimates about linguistic processes. Chapter 4 charted a number of potential methodological improvements that could be explored in the future. The experiment conducted in Chapter 5 to investigate the relationship of conceptual similarity, communicative need and lexification choices covers one type of similarity and only in a dyadic interaction setting; further research should aim to explore other conceptual relations (Xu et al. 2020) and investigate possible larger-scale communication effects (cf. Raviv et al. 2019; Segovia-Martín et al. 2020).

6.2 Wider implications of this work

6.2.1 Historical and corpus linguistics

Many approaches and subfields in the language sciences investigate specific linguistic elements in specific languages. This thesis took a largely language-agnostic, statistical approach to words, text and context (not unlike “distant reading" as it is referred to in digital humanities;

cf. Moretti 2005), focusing on the wider dynamics of change, rather than individual changes.

The natural worry is that results from such an approach may be subject to possible hidden biases or artefacts in the statistical and machine learning pipelines (cf. Dubossarsky et al. 2017). I explored ways of evaluating these methods against artificial, controlled language change scenarios, emulated using simulations of change trajectories (Chapter 2) and synthetic modification of natural corpora (Chapter 3), as well as against randomized baselines (Chapters 3, 4). Chapter 2 also probed the implications of binning corpora into temporal subcorpus chunks — a widespread and often unavoidable practice, where the researcher degree of freedom regarding bin size can either be mostly inconsequential, or end up considerably altering the results.

I hope to have shown that evaluation techniques based on simulations, controlled modification experiments and randomisation tests can be useful in diachronic language research, particularly if the object of study is larger than what could be analysed by hand.

6.2.2 Expressivity, simplicity, and language evolution

There is a growing body of work on the interplay between the orthogonal pressures of simplicity and informativeness. The former relates to compressibility and ease of learning, ease of articulation, and having low cognitive cost. Informativeness or expressivity refers to having high communicative accuracy, being low in communicative cost i.e. having low information loss, and relates to ease of decoding (the exact terms and foci vary between authors and disciplines; cf. Kemp and Regier 2012; Kirby et al. 2015; Winters et al. 2015; Carstensen et al. 2015; Beckner et al. 2017; Bentz et al. 2017; Nölle et al. 2018; Zaslavsky et al. 2019b; Carr et al. 2020; Smith 2020; Steinert-Threlkeld and Szymanik 2020; Uegaki in prep; Haspelmath to appear). These studies represent converging evidence that languages that are learned and used in communication — the real-world ones, artificial ones grown in the lab, as well as those evolved by computational agents — all aspire to balance these two pressures, ending up somewhere along the optimal frontier.

Kemp et al. (2018) discuss how the location of a language on that frontier may be modulated by culture and environment specific communicative needs. The results of the corpus-based studies and experiments in this thesis provide concrete support for this argument: the pressure for simplicity in lexicons can be relaxed in favour of more expressivity, given high enough communicative need, while informativeness can give way to simplicity when a less expressive lexical subspace does the job.

A possible next step in studying the evolution of lexicons on the scale of entire languages (as opposed to isolated domains like kinship or colour) would be to combine explicitly quantified measures of simplicity and expressivity (cf. Piantadosi et al. 2011; Bentz et al. 2017; Zaslavsky et al. 2019b; Steinert-Threlkeld and Szymanik 2020), some estimate of communicative need (Chapter 3 provides a starting point), and either a joint semantic model of multiple languages allowing for direct comparison of lexical densities and colexification (e.g. Chen and Cardie 2018; Thompson et al. 2018; Rabinovich et al. 2020), or language-specific diachronic semantic spaces, ideally continuous or more fine-grained (cf. Rosenfeld and Erk 2018; Dubossarsky et al. 2019; Ryskina et al. 2020) than the discrete subcorpora comparisons utilised in this thesis.

6.2.3 Sociolinguistics

This thesis has mostly focused on language-internal dynamics, using a slew of lexicostatistical variables (as a proxy to language-external forces) to predict other lexicostatistical variables. In the experimental part I tried my best to exclude all language-external effects that may play a role in real world language use: the participants did not see or hear each other, and the forms of the artificial languages were generated in a way to be as alien as possible to English speaking participants. Yet the central idea of this thesis — that communicative needs drive language change — refers to the (social, cultural, natural) environments of language communities, where social structures, cultural preferences and dynamics of multilingualism may all play a role in shaping language use. Combining the study of these processes with that of situational communicative need as explored in this thesis, and that of the pressures for informativeness and simplicity, would likely yield a more complete understanding of language evolution, variation and change.

6.2.4 The sociology of language and language planning

It is probably fair to say that there are plenty of people in most (at least Western) societies who vehemently believe that “slang" and loanwords is something that must be fought against, new registers of language like texting are leading to low literacy, or that some demographic group (usually a minority or just young people) are straight up ruining the language and must be stopped (cf. Crystal 2009). These are views often held not only by laypersons but also governmental bodies tasked with language planning and policy, in countries where these exist (the *Académie Française* being a commonly cited example). This is of course not to say language planning and directed terminology cultivation does not have its place; the lack of it can lead to a situation of the kind already bemoaned by Leibniz, who in 1697 complained that the German language “experience[s] the worst insufficiency in words referring to morality, passion of the mind, social intercourse, governmental matters and all sorts of affairs of civil and public conduct" (Coulmas 1989).

However, my findings do indicate that when it comes to the big picture, the widely held view of “change equals bad" might not be the case. Not only is language change natural and universal, but often enough likely serves some purpose, which may well be beneficial for its efficiency (like the shortening of words), expressivity (like the borrowing of new words with slightly different connotations or shades of meaning), or metalinguistic functions as discussed in Chapter 1. In short, maybe all these young people are not ruining language after all.

Bibliography

Abrams, Daniel M. and Steven H. Strogatz (2003). “Modelling the Dynamics of Language Death”. *Nature* 424, p. 900.

Ahern, Christopher A., Mitchell G. Newberry, Robin Clark, and Joshua B. Plotkin (2016). *Evolutionary Forces in Language Change*. URL: https://arxiv.org/abs/1608.00938 (visited on 07/05/2017).

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, pp. 1–12. DOI: 10.1371/journal.pone.0019009.

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, pp. 8260–8265. DOI: 10.1073/pnas.1721059115.

Andersen, Henning (1990). “The Structure of Drift”. *Historical Linguistics 1987. Papers from the 8th International Conference on Historical Linguistics*. Ed. by Henning Andersen and Konrad Koerner. Amsterdam: Benjamins, pp. 1–20.

Anderson-Sprecher, Richard (1994). “Model Comparisons and R2”. *The American Statistician* 48.2, pp. 113–117. DOI: 10.1080/00031305.1994.10476036.

Anderwald, Lieselotte (2012). “Variable Past-Tense Forms in Nineteenth-Century American English: Linking Normative Grammars and Language Change”. *American Speech* 87.3, pp. 257–293. DOI: 10.1215/00031283-1958327.

Arends, Jacques and Adrienne Bruyn (1994). “Gradualist and Developmental Hypotheses”. *Pidgins and Creoles: An Introduction*. John Benjamins Publishing, pp. 111–120.

Atkinson, Mark, Simon Kirby, and Kenny Smith (2015). “Speaker Input Variability Does Not Explain Why Larger Populations Have Simpler Languages”. *PLOS ONE* 10.6, pp. 1–20. DOI: 10.1371/journal.pone.0129463.

Auer, Peter and Frans Hinskens (2005). “The Role of Interpersonal Accommodation in a Theory of Language Change”. *Dialect Change: Convergence and Divergence in European Languages*. Ed. by Peter Auer, Frans Hinskens, and PaulEditors Kerswill. Cambridge University Press, pp. 335–357. DOI: 10.1017/CBO9780511486623.015.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker (2015). “Fitting Linear Mixed-Effects Models Using Lme4”. *Journal of Statistical Software* 67.1, pp. 1–48. DOI: 10.18637/jss.v067.i01.

Bates, Douglas and Martin Maechler (2018). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.2-17.

Baxter, G. J., R. A. Blythe, W. Croft, and A. J. McKane (2006). “Utterance Selection Model of Language Change”. *Physical Review E* 73.4, p. 046118. DOI: 10.1103/PhysRevE.73.046118.

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, pp. 257–296. DOI: 10.1017/S095439450999010X.

Beckner, Clay, Richard Blythe, Joan Bybee, Morten H. Christiansen, William Croft, Nick C. Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman, and Tom Schoenemann (2009). "Language Is a Complex Adaptive System: Position Paper". *Language Learning* 59.s1, pp. 1–26. DOI: 10.1111/j.1467-9922.2009.00533.x.

Beckner, Clay, Janet B Pierrehumbert, and Jennifer Hay (2017). "The Emergence of Linguistic Structure in an Online Iterated Learning Task". *Journal of Language Evolution* 2.2, pp. 160–176.

Bentley, R. Alexander (2008). "Random Drift versus Selection in Academic Vocabulary: An Evolutionary Analysis of Published Keywords". *PLOS ONE* 3.8, pp. 1–7. DOI: 10.1371/journal.pone.0003057.

Bentley, R. Alexander, Alberto Acerbi, Paul Ormerod, and Vasileios Lampos (2014). "Books Average Previous Decade of Economic Misery". *PLoS ONE* 9.1. Ed. by Matjaž Perc, e83147. DOI: 10.1371/journal.pone.0083147.

Bentley, R. Alexander and Stephen J. Shennan (2003). "Cultural Transmission and Stochastic Network Growth". *American Antiquity* 68.3, pp. 459–485.

Bentz, Christian, Dimitrios Alikaniotis, Michael Cysouw, and Ramon Ferrer-i-Cancho (2017). "The Entropy of Words—Learnability and Expressivity across More than 1000 Languages". *Entropy* 19.6, p. 275.

Bentz, Christian and Bodo Winter (2014). "Languages with More Second Language Learners Tend to Lose Nominal Case". *Quantifying Language Dynamics*. Brill, pp. 96–124.

Berlin, Brent (1992). *Ethnobiological Classification: Principles of Categorization of Plants and Animals in Traditional Societies*. Princeton University Press.

Blank, Andreas (1999). "Why Do New Meanings Occur? A Cognitive Typology of the Motivations for Lexical Semantic Change". Berlin, Boston: De Gruyter Mouton, pp. 61–90.

Blei, David M. and John D. Lafferty (2006). "Dynamic Topic Models". *Proceedings of the 23rd International Conference on Machine Learning* (Pittsburgh, Pennsylvania, USA). ICML '06. ACM, pp. 113–120. DOI: 10.1145/1143844.1143859.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). "Latent Dirichlet Allocation". *Journal of Machine Learning Research* 3, pp. 993–1022.

Blythe, Richard A. (2012). "Neutral Evolution: A Null Model for Language Dynamics". *Advances in complex systems* 15.3-4. DOI: 10.1142/S0219525911003414.

Blythe, Richard A. and William Croft (2012). "S-Curves and the Mechanisms of Propagation in Language Change". *Language* 88.2, pp. 269–304. DOI: 10.1353/lan.2012.0027.

Boas, Franz (1911). *The Mind of Primitive Man*. The Macmillan Company.

Bochkarev, V., V. Solovyev, and S. Wichmann (2014). "Universals versus Historical Contingencies in Lexical Evolution". *Journal of The Royal Society Interface* 11.101. DOI: 10.1098/rsif.2014.0841.

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, pp. 153–175.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017). “Enriching Word Vectors with Subword Information”. *Transactions of the Association for Computational Linguistics* 5, pp. 135–146.

Brouwer, Susanne, Holger Mitterer, and Falk Huettig (2012). “Can Hearing Puter Activate Pupil? Phonological Competition and the Processing of Reduced Spoken Words in Spontaneous Conversations:” *Quarterly Journal of Experimental Psychology*.

Bullinaria, John A. and Joseph P. Levy (2007). “Extracting Semantic Representations from Word Co-Occurrence Statistics: A Computational Study”. *Behavior Research Methods* 39.3, pp. 510–526. DOI: 10.3758/BF03193020.

Bybee, Joan (2002). “Lexical Diffusion in Regular Sound Change”. *Sounds and Systems, Trends in Linguistics Studies and Monographs* 141, pp. 59–74.

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*, pp. 1–38. DOI: 10.1515/cllt-2017-0010.

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, pp. 892–923. DOI: 10.1111/cogs.12371.

Carr, Jon W., Kenny Smith, Jennifer Culbertson, and Simon Kirby (2020). “Simplicity and Informativeness in Semantic Category Systems”. *Cognition* 202, p. 104289. DOI: 10.1016/j.cognition.2020.104289.

Carr, Jon W, Kenny Smith, Jennifer Culbertson, and Simon Kirby (2018). “Simplicity and Informativeness in Semantic Category Systems”. DOI: 10.31234/osf.io/jkfyx.

Carstensen, Alexandra, Jing Xu, Cameron T Smith, and Terry Regier (2015). “Language Evolution in the Lab Tends toward Informative Communication”. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. Ed. by D. Noelle et al. Austin, TX: Cognitive Science Society, p. 6.

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, pp. 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, pp. 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, pp. 21–51.

Čermák, František, Jaroslava Hlaváčová, Milena Hnátková, Tomáš Jelínek, Jan Kocek, Marie Kopřivová, Michal Křen, Renata Novotná, Vladimír Petkevič, Věra Schmiedtová, et al. (2006). *SYN2006PUB: Corpus of Czech Newspapers*. Faculty of Arts, Institute of the Czech National Corpus, Charles University.

Chelsey, Paula and Harald R. Baayen (2010). “Predicting New Words from Newer Words: Lexical Borrowings in French”. *Linguistics* 48.6, pp. 1343–1374.

Chen, Xilun and Claire Cardie (2018). "Unsupervised Multilingual Word Embeddings". *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 261–270. doi: 10.18653/v1/D18-1024.

Christensen, Peer, Riccardo Fusaroli, and Kristian Tylén (2016). "Environmental Constraints Shaping Constituent Order in Emerging Communication Systems: Structural Iconicity, Interactive Alignment and Conventionalization". *Cognition* 146, pp. 67–80. doi: 10.1016/j.cognition.2015.09.004.

Christiansen, Morten H. and Nick Chater (2008). "Language as Shaped by the Brain". *Behavioral and Brain Sciences* 31.5, pp. 489–509. doi: 10.1017/S0140525X08004998.

Church, Kenneth Ward and Patrick Hanks (1990). "Word Association Norms, Mutual Information, and Lexicography". *Computational linguistics* 16.1, pp. 22–29.

Coulmas, Florian (1989). "Language Adaptation". *Language Adaptation*. Ed. by Florian Coulmas. Cambridge University Press, pp. 1–25.

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). doi: 10.1038/srep39122.

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

Crystal, D. (2009). *Txtng: The Gr8 Db8*. OUP Oxford.

Culbertson, Jennifer and Simon Kirby (2016). "Simplicity and Specificity in Language: Domain-General Biases Have Domain-Specific Effects". *Frontiers in Psychology* 6: 1964. doi: 10.3389/fpsyg.2015.01964.

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, pp. 1–7. doi: 10.1371/journal.pone.0102882.

Daoust, Demise (2017). "Language Planning and Language Reform". *The Handbook of Sociolinguistics*. Wiley-Blackwell, pp. 436–452.

Dautriche, Isabelle, Kyle Mahowald, Edward Gibson, Anne Christophe, and Steven T. Piantadosi (2017). "Words Cluster Phonetically beyond Phonotactic Regularities". *Cognition* 163, pp. 128–145. doi: 10.1016/j.cognition.2017.02.001.

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, pp. 2149–2169. doi: 10.1111/cogs.12453.

Davies, Mark (2008). *The Corpus of Contemporary American English (COCA): 450 Million Words, 1990-2012*. Available online at https://www.english-corpora.org/coca.

Davies, Mark (2010). *The Corpus of Historical American English (COHA): 400 Million Words, 1810-2009*. Available online at https://www.english-corpora.org/coha.

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, p. 391.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding". *Proceedings* *of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. NAACL-HLT 2019. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

Dingemanse, Mark, Damián E. Blasi, Gary Lupyan, Morten H. Christiansen, and Padraic Monaghan (2015). "Arbitrariness, Iconicity, and Systematicity in Language". *Trends in Cognitive Sciences* 19.10, pp. 603–615. doi: 10.1016/j.tics.2015.07.013.

Dor, Daniel (2015). *The Instruction of Imagination: Language as a Social Communication Technology*. Foundations of Human Interaction. Oxford University Press.

DTA (2019). *Deutsches Textarchiv*. Version Vom 6. Februar 2019: DTA-Kernkorpus Und Ergänzungstexte. http://www.deutschestextarchiv.de.

Dubossarsky, Haim, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg (2019). "Time-out: Temporal Referencing for Robust Modeling of Lexical Semantic Change". *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 457–470. doi: 10.18653/v1/P19-1044.

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*, pp. 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, pp. 7–28.

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

Durkin, Philip (2014). *Borrowed Words: A History of Loanwords in English*. Oxford: Oxford University Press.

Enfield, N. J. (2014). "Transmission Biases in the Cultural Evolution of Language: Towards an Explanatory Framework". *The Social Origins of Language*. Ed. by Daniel Dor, Chris Knight, and Jerome Lewis. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199665327.003.0023.

Ewens, Warren J. (2004). *Mathematical Population Genetics 1: Theoretical Introduction*. Interdisciplinary Applied Mathematics. Springer New York.

Faruqui, Manaal, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer (2016). "Problems with Evaluation of Word Embeddings Using Word Similarity Tasks". *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. Berlin, Germany: Association for Computational Linguistics, pp. 30–35. doi: 10.18653/v1/W16-2506.

Fay, Nicolas, Simon Garrod, Leo Roberts, and Nik Swoboda (2010). "The Interactive Evolution of Human Communication Systems". *Cognitive science* 34.3, pp. 351–386.

Feder, Alison F., Sergey Kryazhimskiy, and Joshua B. Plotkin (2014). "Identifying Signatures of Selection in Genetic Time Series". *Genetics* 196.2, pp. 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. doi: 10.1098/rsos.170830.

Frajzyngier, Zygmunt and Erin Shay (2003). *Explaining Language Structure through Systems Interaction*. John Benjamins Publishing. 329 pp.

François, Alexandre (2008). "Semantic Maps and the Typology of Colexification: Intertwining Polysemous Networks across Languages". *Studies in Language Companion Series*. Ed. by Martine Vanhove. Vol. 106. Amsterdam: John Benjamins Publishing Company, pp. 163–215. DOI: 10.1075/slcs.106.09fra.

Frermann, Lea and Mirella Lapata (2016). "A Bayesian Model of Diachronic Meaning Change". *Transactions of the Association for Computational Linguistics* 4, pp. 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, pp. 6591–6594. DOI: 10.1073/pnas.1500355112.

Galantucci, Bruno, Simon Garrod, and Gareth Roberts (2012). "Experimental Semiotics." *Language and Linguistics Compass* 6.8, pp. 477–493. DOI: 10.1002/lnc3.351.

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. DOI: 10.1098/rsif.2014.1044.

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), pp. 10785–10790. DOI: 10.1073/pnas.1619666114.

Gibson, Edward, Richard Futrell, Steven P. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy (2019). "How Efficiency Shapes Human Language". *Trends in Cognitive Sciences* 23.5, pp. 389–407. DOI: 10.1016/j.tics.2019.02.003.

Givón, Thomas (1982). "Tense-Aspect-Modality: The Creole Prototype and Beyond". *Tense-aspect: Between semantics and pragmatics*, pp. 115–163.

Goel, Rahul, Sandeep Soni, Naman Goyal, John Paparrizos, Hanna Wallach, Fernando Diaz, and Jacob Eisenstein (2016). "The Social Dynamics of Language Change in Online Networks". *Social Informatics*. Ed. by Emma Spiro and Yong-Yeol Ahn. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 41–57. DOI: 10.1007/978-3-319-47880-7_3.

Gray, Tyler J., Andrew J. Reagan, Peter Sheridan Dodds, and Christopher M. Danforth (2018). "English Verb Regularization in Books and Tweets". *PLOS ONE* 13.12, pp. 1–17. DOI: 10.1371/journal.pone.0209651.

Gries, Stefan Th. (2010). "Useful Statistics for Corpus Linguistics". *A mosaic of corpus linguistics: Selected approaches* 66, pp. 269–291.

Grieve, Jack (2018). "Natural Selection in the Modern English Lexicon". *The Evolution of Language: Proceedings of the 12th International Conference on the Evolution of Language*. Ed. by C. Cuskley, M. Flaherty, H. Little, Luke McCrohon, A. Ravignani, and T. Verhoef. NCU Press. DOI: 10.12775/3991-1.037.

Grieve, Jack, Andrea Nini, and Diansheng Guo (2018). "Mapping Lexical Innovation on American Social Media". *Journal of English Linguistics* 46.4, pp. 293–319. DOI: 10.1177/0075424218793191.

Gulordava, Kristina and Marco Baroni (2011). "A Distributional Similarity Approach to the Detection of Semantic Change in the Google Books Ngram Corpus". *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*. Association for Computational Linguistics, pp. 67–71.

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". *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, pp. 1489–1501. DOI: 10.18653/v1/P16-1141.

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, pp. 2116–2121.

Haspelmath, Martin (to appear). "Explaining Grammatical Coding Asymmetries: Form-Frequency Correspondences and Predictability". *Journal of Linguistics*.

Haspelmath, Martin (1999). "Optimality and Diachronic Adaptation". *Zeitschrift für Sprachwissenschaft* 18.2, pp. 180–205.

Haspelmath, Martin (2003). "The Geometry of Grammatical Meaning: Semantic Maps and Cross-Linguistic Comparison". *New Psychology of Language* 2, pp. 211–242.

Haspelmath, Martin and Andres Karjus (2017). "Explaining Asymmetries in Number Marking: Singulatives, Pluratives, and Usage Frequency". *Linguistics* 55.6, pp. 1213–1235. DOI: 10.1515/ling-2017-0026.

Hernández-Campoy, Juan Manuel and Juan Camilo Conde-Silvestre (2012). *The Handbook of Historical Sociolinguistics*. Wiley-Blackwell.

Hill, Felix, Roi Reichart, and Anna Korhonen (2015). "SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation". *Computational Linguistics* 41.4, pp. 665–695. DOI: 10.1162/COLI_a_00237.

Hinrichs, Lars, Benedikt Szmrecsanyi, and Axel Bohmann (2015). "Which-Hunting and the Standard English Relative Clause". *Language* 91.4, pp. 806–836.

Hofmann, V., J.B. Pierrehumbert, and H. Schuetze (2020). "Predicting the Growth of Morphological Families from Social and Linguistic Factors". *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Seattle WA, July 5 - July 10*. Association for Computational Linguistics.

Honnibal, Matthew and Ines Montani (2017). "spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing".

Hopper, Paul J. and Elizabeth Closs Traugott (2003). *Grammaticalization*. Cambridge University Press. 300 pp.

Hu, Renfen, Shen Li, and Shichen Liang (2019). "Diachronic Sense Modeling with Deep Contextualized Word Embeddings: An Ecological View". *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL 2019. Florence, Italy: Association for Computational Linguistics, pp. 3899–3908. DOI: 10.18653/v1/P19-1379.

Iranmehr, Arya, Ali Akbari, Christian Schlötterer, and Vineet Bafna (2017). “CLEAR: Composition of Likelihoods for Evolve and Resequence Experiments”. *Genetics* 206.2, pp. 1011–1023. DOI: 10.1534/genetics.116.197566.

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

Joseph, J.E. (2004). *Language and Identity: National, Ethnic, Religious*. Palgrave Macmillan.

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.

Kaalep, Heiki-Jaan, Kadri Muischnek, Kristel Uiboaed, and Kaarel Veskis (2010). “The Estonian Reference Corpus: Its Composition and Morphology-Aware User Interface”. *Proceedings of the 2010 Conference on Human Language Technologies – The Baltic Perspective: Proceedings of the Fourth International Conference Baltic HLT 2010*. NLD: IOS Press, pp. 143–146.

Kandler, Anne and Enrico R. Crema (2019). “Analysing Cultural Frequency Data: Neutral Theory and Beyond”. *Handbook of Evolutionary Research in Archaeology*. Ed. by Anna Marie Prentiss. Cham: Springer International Publishing, pp. 83–108. DOI: 10.1007/978-3-030-11117-5_5.

Kandler, Anne and Adam Powell (2018). “Generative Inference for Cultural Evolution”. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 373.1743. DOI: 10.1098/rstb.2017.0056.

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. DOI: 10.1098/rsos.170949.

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, pp. 45–52. DOI: 10.1016/j.cognition.2017.05.001.

Karjus, Andres (2015). “Through the Spyglass of Synchrony: Grammaticalization of the Exterior Space in the Eastern Circum-Baltic”. *New Trends in Nordic and General Linguistics*. DOI: 10.1515/9783110346978.267.

Karjus, Andres, Richard A. Blythe, Simon Kirby, and Kenny Smith (2018a). “Challenges in Detecting Evolutionary Forces in Language Change Using Diachronic Corpora”. *arXiv preprint* 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, pp. 186–188. DOI: 10.7275/R5RR1WFX.

Karjus, Andres, Richard A. Blythe, Simon Kirby, and Kenny Smith (2020a). “Challenges in Detecting Evolutionary Forces in Language Change Using Diachronic Corpora”. *Glossa: a journal of general linguistics* 5.1, p. 45. DOI: 10.5334/gjgl.909.

Karjus, Andres, Richard A. Blythe, Simon Kirby, and Kenny Smith (2020b). “Communicative Need Modulates Competition in Language Change”. *arXiv preprint* arXiv:2006.09277.

Karjus, Andres, Richard A. Blythe, Simon Kirby, and Kenny Smith (2020c). “Quantifying the Dynamics of Topical Fluctuations in Language”. *Language Dynamics and Change* 10.1, pp. 86–125. DOI: 10.1163/22105832-01001200.

Karjus, Andres and Martin Ehala (2018). “Testing an Agent-Based Model of Language Choice on Sociolinguistic Survey Data”. *Language Dynamics and Change* 8.2, pp. 219–252. DOI: 10.1163/22105832-00802004.

Karsdorp, Folgert, Enrique Manjavacas, Lauren Fonteyn, and Mike Kestemont (2020). “Classifying Evolutionary Forces in Language Change Using Neural Networks”. *Evolutionary Human Sciences*, pp. 1–40. DOI: 10.1017/ehs.2020.52.

Kauhanen, Henri (2017). “Neutral Change”. *Journal of Linguistics* 53.2, pp. 327–358. DOI: 10.1017/S0022226716000141.

Keller, Daniela Barbara and Jörg Schultz (2013). “Connectivity, Not Frequency, Determines the Fate of a Morpheme”. *PLOS ONE* 8.7, pp. 1–8. DOI: 10.1371/journal.pone.0069945.

Keller, Daniela Barbara and Jörg Schultz (2014). “Word Formation Is Aware of Morpheme Family Size”. *PLOS ONE* 9.4, pp. 1–6. DOI: 10.1371/journal.pone.0093978.

Kemp, Charles and Terry Regier (2012). “Kinship Categories across Languages Reflect General Communicative Principles”. *Science (New York, N.Y.)* 336.6084, pp. 1049–1054. DOI: 10.1126/science.1218811.

Kemp, Charles, Yang Xu, and Terry Regier (2018). “Semantic Typology and Efficient Communication”. *Annual Review of Linguistics* 4.1, pp. 109–128. DOI: 10.1146/annurev-linguistics-011817-045406.

Kershaw, Daniel, Matthew Rowe, and Patrick Stacey (2016). “Towards Modelling Language Innovation Acceptance in Online Social Networks”. *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* (San Francisco, California, USA). WSDM ’16. ACM, pp. 553–562. DOI: 10.1145/2835776.2835784.

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

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

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, pp. 10681–10686. DOI: 10.1073/pnas.0707835105.

Kirby, Simon, Monica Tamariz, Hannah Cornish, and Kenny Smith (2015). “Compression and Communication in the Cultural Evolution of Linguistic Structure”. *Cognition* 141, pp. 87–102. DOI: 10.1016/j.cognition.2015.03.016.

Koplenig, Alexander (2017a). “A Data-Driven Method to Identify (Correlated) Changes in Chronological Corpora”. *Journal of Quantitative Linguistics* 24.4, pp. 289–318. DOI: 10.1080/09296174.2017.1311447.

Koplenig, Alexander (2017b). “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, pp. 169–188. DOI: 10.1093/llc/fqv037.

Koplenig, Alexander (2017c). “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, pp. 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. Ed. by Karen Lidzba, e0150771. DOI: 10.1371/journal.pone.0150771.

Kroch, Anthony and Ann Taylor (2000). *The Penn-Helsinki Parsed Corpus of Middle English (PPCME2)*. Department of Linguistics, University of Pennsylvania.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena (2015). “Statistically Significant Detection of Linguistic Change”. *Proceedings of the 24th International Conference on World Wide Web* (Republic and Canton of Geneva, Switzerland). WWW ’15. International World Wide Web Conferences Steering Committee, pp. 625–635. DOI: 10.1145/2736277.2741627.

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

Labov, William (1982). “Building on Empirical Foundations”. *Perspectives on Historical Linguistics*. Ed. by W. Lehmann and Y. Malkiel. Vol. 24. Amsterdam and Philadelphia: Benjamins, pp. 17–92.

Laland, K. N., J. Odling-Smee, and M. W. Feldman (2001). “Cultural Niche Construction and Human Evolution”. *Journal of Evolutionary Biology* 14.1, pp. 22–33. DOI: 10.1046/j.1420-9101.2001.00262.x.

Lass, Roger (1992). “What, If Anything, Was the Great Vowel Shift”. *History of Englishes: New Methods and Interpretations in Historical Linguistics*. Ed. by Matti Rissanen et al. Berlin: Mouton de Gruyter, pp. 144–155.

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, pp. 379–401.

Levy, Omer, Yoav Goldberg, and Ido Dagan (2015). “Improving Distributional Similarity with Lessons Learned from Word Embeddings”. *Transactions of the Association for Computational Linguistics* 3, pp. 211–225. DOI: 10.1162/tacl_a_00134.

Levy, Roger P. (2018). “Communicative Efficiency, Uniform Information Density, and the Rational Speech Act Theory”. *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, pp. 684–689.

Lieberman, Erez, Jean-Baptiste Michel, Joe Jackson, Tina Tang, and Martin A. Nowak (2007). “Quantifying the Evolutionary Dynamics of Language”. *Nature* 449.7163, pp. 713–716.

Lijffijt, Jefrey, Tanja Säily, and Terttu Nevalainen (2012). “CEECing the Baseline: Lexical Stability and Significant Change in a Historical Corpus”. *Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources*. Ed. by Jukka Tyrkkö, Matti Kilpiö, Terttu Nevalainen, Matti Rissanen. Studies in Variation, Contacts and Change in English 10. Helsinki: Research Unit for Variation, Contacts and Change in English (VARIENG).

Lindsey, Delwin T. and Angela M. Brown (2002). “Color Naming and the Phototoxic Effects of Sunlight on the Eye”. *Psychological Science* 13.6, pp. 506–512. DOI: 10.1111/1467-9280.00489.

List, Johann-Mattis, Anselm Terhalle, and Matthias Urban (2013). “Using Network Approaches to Enhance the Analysis of Cross-Linguistic Polysemies”. *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Short Papers*. Potsdam, Germany: Association for Computational Linguistics, pp. 347–353.

Lupyan, Gary and Rick Dale (2010). “Language Structure Is Partly Determined by Social Structure”. *PLOS ONE* 5.1, pp. 1–10. DOI: 10.1371/journal.pone.0008559.

Lupyan, Gary and Rick Dale (2016). “Why Are There Different Languages? The Role of Adaptation in Linguistic Diversity”. *Trends in Cognitive Sciences* 20.9, pp. 649–660. DOI: 10.1016/j.tics.2016.07.005.

MacWhinney, Brian (1989). “Competition and Lexical Categorization”. *Linguistic categorization* 61, pp. 195–241.

Majid, Asifa, Fiona Jordan, and Michael Dunn (2015). “Semantic Systems in Closely Related Languages”. *Language Sciences*. Semantic Systems in Closely Related Languages 49, pp. 1–18. DOI: 10.1016/j.langsci.2014.11.002.

Malaspinas, Anna-Sapfo (2016). “Methods to Characterize Selective Sweeps Using Time Serial Samples: An Ancient DNA Perspective”. *Molecular Ecology* 25.1, pp. 24–41. DOI: 10.1111/mec.13492.

Malt, Barbara C. and Asifa Majid (2013). “How Thought Is Mapped into Words”. *WIREs Cognitive Science* 4.6, pp. 583–597. DOI: 10.1002/wcs.1251.

Malt, Barbara C., Steven A. Sloman, Silvia Gennari, Meiyi Shi, and Yuan Wang (1999). “Knowing versus Naming: Similarity and the Linguistic Categorization of Artifacts”. *Journal of Memory and Language* 40.2, pp. 230–262. DOI: 10.1006/jmla.1998.2593.

Martinet, André (1952). “Function, Structure, and Sound Change”. *WORD* 8.1, pp. 1–32. DOI: 10.1080/00437956.1952.11659416.

McMahon, April M.S. (1994). *Understanding Language Change*. Cambridge University Press.

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, pp. 176–182. DOI: 10.1126/science.1199644.

Mickan, Anne, James M. McQueen, and Kristin Lemhöfer (2020). “Between-Language Competition as a Driving Force in Foreign Language Attrition”. *Cognition* 198, p. 104218. DOI: 10.1016/j.cognition.2020.104218.

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

Monaghan, Padraic and Seán G. Roberts (2019). “Cognitive Influences in Language Evolution: Psycholinguistic Predictors of Loan Word Borrowing”. *Cognition* 186, pp. 147–158. DOI: 10.1016/j.cognition.2019.02.007.

Moretti, Franco (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso.

Mufwene, Salikoko S (2013). “Language as Technology Some Questions That Evolutionary”. *In search of universal grammar: From old Norse to Zoque* 202, p. 327.

Mufwene, Salikoko S. (2002). “Competition and Selection in Language Evolution”. *Selection* 3.1, pp. 45–56. DOI: 10.1556/Select.3.2002.1.5.

Newberry, Mitchell G., Christopher A. Ahern, Robin Clark, and Joshua B. Plotkin (2017). “Detecting Evolutionary Forces in Language Change”. *Nature* 551.7679, pp. 223–226. DOI: 10.1038/nature24455.

Nini, Andrea, Carlo Corradini, Diansheng Guo, and Jack Grieve (2017). “The Application of Growth Curve Modeling for the Analysis of Diachronic Corpora”. *Language Dynamics and Change* 7.1, pp. 102–125. DOI: 10.1163/22105832-00701001.

Nishino, Jo (2013). “Detecting Selection Using Time-Series Data of Allele Frequencies with Multiple Independent Reference Loci”. *G3: Genes, Genomes, Genetics* 3.12, pp. 2151–2161.

Nölle, Jonas, Marlene Staib, Riccardo Fusaroli, and Kristian Tylén (2018). “The Emergence of Systematicity: How Environmental and Communicative Factors Shape a Novel Communication System”. *Cognition* 181, pp. 93–104.

Ohala, John J (1983). “The Origin of Sound Patterns in Vocal Tract Constraints”. *The Production of Speech*. New York, NY: Springer, pp. 189–216.

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, pp. 717–720.

Pagel, Mark, Quentin D. Atkinson, Andreea S. Calude, and Andrew Meade (2013). “Ultraconserved Words Point to Deep Language Ancestry across Eurasia”. *Proceedings of the National Academy of Sciences* 110.21, pp. 8471–8476. DOI: 10.1073/pnas.1218726110.

Pagel, Mark, Mark Beaumont, Andrew Meade, Annemarie Verkerk, and Andreea Calude (2019). “Dominant Words Rise to the Top by Positive Frequency-Dependent Selection”. *Proceedings of the National Academy of Sciences* 116.15, pp. 7397–7402. DOI: 10.1073/pnas.1816994116.

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. Ed. by Alain Barrat, e0137041. DOI: 10.1371/journal.pone.0137041.

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

Perek, Florent (2016). “Using Distributional Semantics to Study Syntactic Productivity in Diachrony: A Case Study”. *Linguistics* 54.1, pp. 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). DOI: 10.1038/srep00313.

Piantadosi, Steven T., Harry Tily, and Edward Gibson (2011). “Word Lengths Are Optimized for Efficient Communication”. *Proceedings of the National Academy of Sciences* 108.9, pp. 3526–3529. DOI: 10.1073/pnas.1012551108.

Pierrehumbert, Janet B., Forrest Stonedahl, and Robert Daland (2014). *A Model of Grassroots Changes in Linguistic Systems*. URL: https://arxiv.org/abs/1408.1985 (visited on 07/08/2017).

Pinker, Steven and Michael T Ullman (2002). "The Past and Future of the Past Tense". *Trends in cognitive sciences* 6.11, pp. 456–463.

Premo, L. S. (2014). "Cultural Transmission and Diversity in Time-Averaged Assemblages". *Current Anthropology* 55.1, pp. 105–114. DOI: 10.1086/674873.

R Core Team (2016–2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Rabinovich, Ella, Yang Xu, and Suzanne Stevenson (2020). "The Typology of Polysemy: A Multilingual Distributional Framework". *arXiv preprint* arXiv:2006.01966.

Ramiro, Christian, Mahesh Srinivasan, Barbara C. Malt, and Yang Xu (2018). "Algorithms in the Historical Emergence of Word Senses". *Proceedings of the National Academy of Sciences* 115.10, pp. 2323–2328. DOI: 10.1073/pnas.1714730115.

Raviv, Limor, Antje Meyer, and Shiri Lev-Ari (2019). "Larger Communities Create More Systematic Languages". *Proceedings of the Royal Society B: Biological Sciences* 286.1907, p. 20191262. DOI: 10.1098/rspb.2019.1262.

Reali, Florencia, Nick Chater, and Morten H. Christiansen (2018). "Simpler Grammar, Larger Vocabulary: How Population Size Affects Language". *Proceedings of the Royal Society of London B: Biological Sciences* 285.1871. DOI: 10.1098/rspb.2017.2586.

Reali, Florencia and Thomas 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, pp. 429–436. DOI: 10.1098/rspb.2009.1513.

Regier, Terry, Alexandra Carstensen, and Charles Kemp (2016). "Languages Support Efficient Communication about the Environment: Words for Snow Revisited". *PLOS ONE* 11.4, pp. 1–17. DOI: 10.1371/journal.pone.0151138.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Edoardo M Airoldi, et al. (2013). "The Structural Topic Model and Applied Social Science". *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, pp. 11–24.

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

Rubin, Joan, Björn H. Jernudd, Jyotirindra DasGupta, Joshua A. Fishman, and Charles A. Ferguson (1977). *Language Planning Processes*. Contributions to the Sociology of Language. Mouton.

Ryskina, Maria, Ella Rabinovich, Taylor Berg-Kirkpatrick, David Mortensen, and Yulia Tsvetkov (2020). "Where New Words Are Born: Distributional Semantic Analysis of Neologisms and Their Semantic Neighborhoods". *Proceedings of the Society for Computation in Linguistics* 3.1, pp. 43–52. DOI: 10.7275/1jra-8m83.

Rzymski, Christoph, Tiago Tresoldi, Simon J. Greenhill, Mei-Shin Wu, Nathanael E. Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A. Bodt, Abbie Hantgan, Gereon A. Kaiping, Sophie Chang, Yunfan Lai, Natalia Morozova, Heini Arjava, Nataliia Hübler, Ezequiel Koile, Steve Pepper, Mariann Proos, Briana Van Epps, Ingrid Blanco, Carolin Hundt, Sergei Monakhov, Kristina Pianykh, Sallona Ramesh, Russell D. Gray, Robert Forkel, and Johann-Mattis List (2020). "The Database of Cross-Linguistic Colexifications, Reproducible Analysis of Cross-Linguistic Polysemies". *Scientific Data* 7.1, pp. 1–12. DOI: 10.1038/s41597-019-0341-x.

Sagi, Eyal, Stefan Kaufmann, and Brady Clark (2011). "Tracing Semantic Change with Latent Semantic Analysis". *Current methods in historical semantics*, pp. 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, pp. 85–114.

Santus, Enrico, Emmanuele Chersoni, Alessandro Lenci, Chu-Ren Huang, and Philippe Blache (2016). "Testing APSyn against Vector Cosine on Similarity Estimation". *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers* (Seoul, South Korea), pp. 229–238.

Sapir, Edward (1921). *Language. An Introduction to the Study of Speech*. New York: Harcourt, Brace and Company.

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

Schlechtweg, Dominik, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde (2019). "A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains". *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 732–746. DOI: 10.18653/v1/P19-1072.

Schraiber, Joshua G., Steven N. Evans, and Montgomery Slatkin (2016). "Bayesian Inference of Natural Selection from Allele Frequency Time Series". *Genetics*. DOI: 10.1534/genetics.116.187278.

Scott-Phillips, Thomas C. and Simon Kirby (2010). "Language Evolution in the Laboratory". *Trends in Cognitive Sciences* 14.9, pp. 411–417. DOI: 10.1016/j.tics.2010.06.006.

Searle, J.R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Segovia-Martín, José, Bradley Walker, Nicolas Fay, and Monica Tamariz (2020). "Network Connectivity Dynamics, Cognitive Biases, and the Evolution of Cultural Diversity in Round-Robin Interactive Micro-Societies". *Cognitive Science* 44.7, e12852. DOI: 10.1111/cogs.12852.

Selivanov, Dmitriy and Qing Wang (2018). *Text2vec: Modern Text Mining Framework for R*. Vol. R package version 0.5.1.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn (2011). "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant". *Psychological Science* 22.11, pp. 1359–1366. DOI: 10.1177/0956797611417632.

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, pp. 140–149. doi: http://dx.doi.org/10.1016/j.jtbi.2015.12.012.

Smith, Kenny (2020). "How Culture and Biology Interact to Shape Language and the Language Faculty". *Topics in Cognitive Science* 12.2, pp. 690–712. doi: 10.1111/tops.12377.

Smith, Kenny, Amy Perfors, Olga Fehér, Anna Samara, Kate Swoboda, and Elizabeth Wonnacott (2017). "Language Learning, Language Use and the Evolution of Linguistic Variation". *Philosophical Transactions of the Royal Society B: Biological Sciences* 372.1711, p. 20160051. doi: 10.1098/rstb.2016.0051.

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

Smith, Kenny and Elizabeth Wonnacott (2010). "Eliminating Unpredictable Variation through Iterated Learning". *Cognition* 116.3, pp. 444–449. doi: 10.1016/j.cognition.2010.06.004.

Srinivasan, Mahesh and Hugh Rabagliati (2015). "How Concepts and Conventions Structure the Lexicon: Cross-Linguistic Evidence from Polysemy". *Lingua* 157, pp. 124–152. doi: 10.1016/j.lingua.2014.12.004.

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, pp. 171–198. doi: 10.1163/22105832-00602005.

Steels, Luc and Eörs Szathmáry (2018). "The Evolutionary Dynamics of Language". *Biosystems* 164, pp. 128–137. doi: 10.1016/j.biosystems.2017.11.003.

Steinert-Threlkeld, Shane and Jakub Szymanik (2020). "Ease of Learning Explains Semantic Universals". *Cognition* 195, p. 104076. doi: 10.1016/j.cognition.2019.104076.

Stewart, Ian and Jacob Eisenstein (2018). "Making "Fetch" Happen: The Influence of Social and Linguistic Context on Nonstandard Word Growth and Decline". *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels, Belgium). Association for Computational Linguistics, pp. 4360–4370. doi: 10.18653/v1/D18-1467.

Strang, B.M.H. (2015). *A History of English*. Routledge Library Editions: The English Language. Taylor & Francis.

Strimling, Pontus, Fredrik Jansson, and Mikael Parkvall (2015). "Modeling the Evolution of Creoles". *Language Dynamics and Change* 5.1, pp. 1–51. doi: 10.1163/22105832-00501005.

Szmrecsanyi, Benedikt (2016). "About Text Frequencies in Historical Linguistics: Disentangling Environmental and Grammatical Change". *Corpus Linguistics and Linguistic Theory* 12.1, pp. 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". *Late Modern English Syntax*. Ed. by M Hundt. Studies in English Language. Cambridge University Press, pp. 133–152.

Tamariz, Monica, T. Mark Ellison, Dale J. Barr, and Nicolas Fay (2014). "Cultural Selection Drives the Evolution of Human Communication Systems". *Proceedings of the Royal Society B: Biological Sciences* 281.1788, p. 20140488. doi: 10.1098/rspb.2014.0488.

Taus, Thomas, Andreas Futschik, and Christian Schlötterer (2017). “Quantifying Selection with Pool-Seq Time Series Data”. *Molecular Biology and Evolution* 34.11, pp. 3023–3034. DOI: 10.1093/molbev/msx225.

Terhorst, Jonathan, Christian Schlötterer, and Yun S Song (2015). “Multi-Locus Analysis of Genomic Time Series Data from Experimental Evolution”. *PLoS genetics* 11.4, e1005069.

Thompson, Bill and Gary Lupyan (2018). “Automatic Estimation of Lexical Concreteness in 77 Languages”. *The 40th Annual Conference of the Cognitive Science Society (CogSci 2018)*. Cognitive Science Society, pp. 1122–1127.

Thompson, Bill, Sean Roberts, and Gary Lupyan (2018). “Quantifying Semantic Similarity across Languages”. *Proceedings of the 40th Annual Conference of the Cognitive Science Society (CogSci 2018)*.

Tinits, Peeter, Jonas Nölle, and Stefan Hartmann (2017). “Usage Context Influences the Evolution of Overspecification in Iterated Learning”. *Journal of Language Evolution* 2.2, pp. 148–159. DOI: 10.1093/jole/lzx011.

Tomasello, Michael (1999). *The Cultural Origins of Human Cognition*. Harvard University Press. 257 pp.

Törnqvist, Leo, Pentti Vartia, and Yrjö O. Vartia (1985). “How Should Relative Changes Be Measured?” *The American Statistician* 39.1, pp. 43–46.

Trask, R.L. and R.L. Trask (1993). *A Dictionary of Grammatical Terms in Linguistics*. Linguistics - Routledge. Routledge.

Trask, Robert Lawrence (1996). *Historical Linguistics*. London: Arnold.

Traugott, E.C. and R.B. Dasher (2001). *Regularity in Semantic Change*. Cambridge Studies in Linguistics. Cambridge University Press.

Turney, Peter D. and Saif M. Mohammad (2019). “The Natural Selection of Words: Finding the Features of Fitness”. *PLOS ONE* 14.1, pp. 1–20. DOI: 10.1371/journal.pone.0211512.

Uegaki, Wataru (in prep). **NAND and the Communicative Efficiency Model*. Unpublished manuscript, https://semanticsarchive.net/Archive/2M0YTUzN.

Van de Velde, Freek (2014). “Degeneracy: The Maintenance of Constructional Networks”. *The Extending Scope of Construction Grammar*. Vol. 54. Berlin/Boston: Walter De Gruyter GmbH, pp. 141–179.

Van der Loo, M.P.J. (2014). “The Stringdist Package for Approximate String Matching”. *The R Journal* 6.1, pp. 111–122.

Van der Maaten, L.J.P. and G.E. Hinton (2008). “Visualizing High-Dimensional Data Using t-SNE”. *Journal of Machine Learning Research* 9 (nov), pp. 2579–2605.

Van Gelderen, E., ed. (2009). *Cyclical Change*. Linguistik Aktuell/Linguistics Today. John Benjamins Publishing Company.

Van Trijp, Remi (2012). “Self-Assessing Agents for Explaining Language Change: A Case Study in German”. *Proceedings of the 20th European Conference on Artificial Intelligence*. ECAI’12. Montpellier, France: IOS Press, pp. 798–803.

Vlachos, Christos, Claire Burny, Marta Pelizzola, Rui Borges, Andreas Futschik, Robert Kofler, and Christian Schlötterer (2019). “Benchmarking Software Tools for Detecting and Quantifying Selection in Evolve and Resequencing Studies”. *Genome Biology* 20.1, p. 169. DOI: 10.1186/s13059-019-1770-8.

Vlachos, Christos and Robert Kofler (2018). "MimicrEE2: Genome-Wide Forward Simulations of Evolve and Resequencing Studies". *PLOS Computational Biology* 14.8, pp. 1–10. DOI: 10.1371/journal.pcbi.1006413.

Walker, James A. (2010). *Variation in Linguistic Systems*. New York: Routledge.

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

Wendlandt, Laura, Jonathan K. Kummerfeld, and Rada Mihalcea (2018). "Factors Influencing the Surprising Instability of Word Embeddings". *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. NAACL-HLT 2018. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2092–2102. DOI: 10.18653/v1/N18-1190.

Wetherell, Charles (1986). "The Log Percent (L%): An Absolute Measure of Relative Change". *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 19.1, pp. 25–26.

Wichmann, Søren (2008). "The Emerging Field of Language Dynamics". *Language and Linguistics Compass* 2.3, pp. 442–455. DOI: 10.1111/j.1749-818X.2008.00062.x.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani (2019). "Welcome to the tidyverse". *Journal of Open Source Software* 4.43, p. 1686. DOI: 10.21105/joss.01686.

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

Winford, Donald (2017). "The Ecology of Language and the New Englishes: Toward an Integrative Framework". *Changing English: Global and Local Perspectives: Markku Filppula, Juhani Klemola, Anna Mauranen, Svetlana Vetchinnikova,* 92, p. 25.

Winters, James, Simon Kirby, and Kenny Smith (2015). "Languages Adapt to Their Contextual Niche". *Language and Cognition* 7.3, pp. 415–449. DOI: 10.1017/langcog.2014.35.

Winters, James, Simon Kirby, and Kenny Smith (2018). "Contextual Predictability Shapes Signal Autonomy". *Cognition* 176, pp. 15–30. DOI: 10.1016/j.cognition.2018.03.002.

Wood, S. N. (2011). "Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models". *Journal of the Royal Statistical Society (B)* 73.1, pp. 3–36.

Wright, Sewall (1931). "Evolution in Mendelian Populations". *Genetics* 16.2, pp. 97–159.

Xu, Yang, Khang Duong, Barbara C. Malt, Serena Jiang, and Mahesh Srinivasan (2020). "Conceptual Relations Predict Colexification across Languages". *Cognition* 201, p. 104280. DOI: 10.1016/j.cognition.2020.104280.

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

Xu, Yang, Barbara C. Malt, and Mahesh Srinivasan (2017). “Evolution of Word Meanings through Metaphorical Mapping: Systematicity over the Past Millennium”. *Cognitive Psychology* 96, pp. 41–53. DOI: 10.1016/j.cogpsych.2017.05.005.

Xu, Yang and Terry Regier (2014). “Numeral Systems across Languages Support Efficient Communication: From Approximate Numerosity to Recursion”. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society, CogSci 2014, Quebec City, Canada, July 23-26, 2014.* Ed. by Paul Bello, Marcello Guarini, Marjorie McShane, and Brian Scassellati. cognitivesciencesociety.org.

Yao, Zijun, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong (2018). “Dynamic Word Embeddings for Evolving Semantic Discovery”. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (Marina Del Rey, CA, USA). WSDM ’18. ACM, pp. 673–681. DOI: 10.1145/3159652.3159703.

Zaslavsky, Noga, Charles Kemp, Naftali Tishby, and Terry Regier (2019a). “Color Naming Reflects Both Perceptual Structure and Communicative Need”. *Topics in Cognitive Science* 11.1, pp. 207–219. DOI: 10.1111/tops.12395.

Zaslavsky, Noga, Terry Regier, Naftali Tishby, and Charles Kemp (2019b). “Semantic Categories of Artifacts and Animals Reflect Efficient Coding”. *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, pp. 1254–1260.

Zhang, Menghan and Tao Gong (2013). “Principles of Parametric Estimation in Modeling Language Competition”. *Proceedings of the National Academy of Sciences* 110.24, pp. 9698–9703.

Zipf, George Kingsley (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Reading, MA: Addison-Wesley Press.