Embedding Regression:
Models for Context-Specific Description and Inference*

Pedro L. Rodriguez† Arthur Spirling‡ Brandon M. Stewart§

Abstract

Social scientists commonly seek to make statements about how a word's use varies over circumstances—whether that be time, partisan identity, or some other document-level covariate. A promising avenue is the use of domain-specific word embeddings, that simultaneously allow for statements of uncertainty and statistical inference. We introduce the à la Carte on Text (`conText`) embedding regression model for this purpose. We extend and validate a simple linear method of refitting pre-trained embeddings to local contexts that requires minimal input data. It outperforms well-known competitors for studying changes in meaning across groups and time. Our approach allows us to speak descriptively of systematic differences across covariates in the context in which words appear, and to comment on whether a particular use is statistically significantly different to another. We provide evidence of excellent relative performance of the model, and show how it might be used in substantive research.

*First draft: July 2020. This draft: June 25, 2021. We thank Clark Bernier for discussions on finite sample bias and Hauke Licht for suggestions on the sprawling embeddings literature. We are grateful for comments from an audience at the Political Methodology society including Walter Mebane, John Londregan, Molly Roberts and especially our discussant, Max Goplerud.

†Postdoctoral Fellow, Data Science Institute (joint with Political Science), Vanderbilt University and Instituto de Estudios Superiores de Administración (`pedro.rodriguez@Vanderbilt.Edu`)

‡Professor of Politics and Data Science, New York University (`arthur.spirling@nyu.edu`)

§Assistant Professor of Sociology and Arthur H. Scribner Bicentennial Preceptor at Princeton University (`bms4@princeton.edu`)

1 Introduction

All human communication requires common understandings of meaning. This is nowhere more true than political and social life, where the success of an appeal—rhetorical or otherwise—relies on an audience perceiving a message in the particular way that a speaker seeks to deliver it. Scholars have therefore spent much effort exploring the meanings of terms, how those meanings are manipulated, and how they change over time and space. Historically, this work has been qualitative (e.g. Austin, 1962; Skinner, 1969; Geertz, 1973). But in recent times, quantitative analysts have turned to modeling and measuring "context" directly from natural language (e.g. Hopkins, 2018; Aslett et al., 2020; Park, Greene and Colaresi, 2020).

A promising avenue for such investigations has been the use of "word embeddings"—a family of techniques that conceive of meaning as emerging from the distribution of words that surround a term in text (e.g. Mikolov et al., 2013). By representing each word as a vector of real numbers, and examining the relationships between vectors for the vocabulary of a corpus, scholars have uncovered new facts about language and the people that produce it (e.g. Caliskan, Bryson and Narayanan, 2017). This is also true in the study of politics, society and culture (Garg et al., 2018; Kozlowski, Taddy and Evans, 2019; Rodman, 2019; Rheault and Cochrane, 2019; Wu et al., 2019).

While borrowing existing techniques has certainly produced insights, for social scientists two problems remain. First, traditional approaches generally require a lot of data to produce high quality representations—that is, to produce embeddings that make sense and connote meaning of terms correctly. The issue is less that our typical corpora are small—though they are compared to those on the web-scale collections often used in computer science—and more that terms for which we would like to estimate contexts are subject-specific and thus typically quite *rare*. As an example, there are fewer than twenty parliamentary mentions of the "special relationship" between the US and the UK in some years of the 1980s—despite this arguably being the high watermark of elite closeness between the two countries. The second problem is

one of inference. While representations themselves are helpful, social scientists want to make statements about the statistical properties and relationships between embeddings. That is, they want to speak meaningfully of whether language is used differently across subcorpora and whether those apparent differences are larger than we would expect by chance. Neither of these problems are well-addressed by current techniques. While there have been efforts to address inference in embeddings (see, e.g, Kulkarni et al., 2015; Han et al., 2018; Lauretig, 2019), they are typically data intensive, computationally intensive or run into issues due to lack of identification in the underlying space.

We tackle these two problems together in what follows. We provide both a statistical framework for making statements about covariate effects on embeddings, and one that performs particularly well in cases of rare words or small corpora. Specifically, we innovate on Khodak et al. (2018) which introduced *à la carte embeddings* (ALC). In a nutshell, the method takes embeddings which have been pre-trained on large corpora (e.g. `word2vec` or `GloVe` embeddings readily available online), combined with a small sample of example uses for a focal word, and then induces a new context-specific embedding for the focal word. This requires only a simple linear transformation of the averaged embeddings for words within the context of the focal word.

We place ALC in a *regression* setting that allows for fast solutions to queries of the type "do authors with these covariate values use these terms in a different way than authors with different covariate values? If yes, how do they differ?" We provide two proofs of concept. First, we demonstrate the strength of our approach by comparing its performance to the "industry standard" as laid out by Rodman (2019) in a study of a New York Times corpus, where slow changes over long periods are the norm. Second, we show that our method can also identify drastic switches in meaning over short time periods—specifically in our case, for the term `Trump` before and after the 2016 election.

We study three substantive cases to show how the technique may be put to work. First, we explore partisan differences in Congressional speech—a topic of long-standing interest

in political science (see, e.g., Monroe, Colaresi and Quinn, 2008). Below, we show that `immigration` is, perhaps unsurprisingly, one of the most differently expressed terms for contemporary Democrats and Republicans. Our second substantive case is historical in nature: we compare across polities (and corpora) to show how elites in the UK and US expressed `empire` in the post-war period, how that usage diverged and converged, and when. Our third case builds on earlier work on open-ended responses in surveys (Roberts et al., 2014), to study how liberals and conservatives differ as regards pressing issues in the United States.

These innovations allow for social scientists to go beyond general meanings of words to capture the situation-specific usage of a particular term. This is possible without substantial computation and, in contrast to other approaches, requires only the text immediately around the word of interest.

We proceed as follows: in Section 2 we provide some context for what social scientists mean by 'context'. We then introduce the ALC algorithm, and provide a proof of concept. Subsequently, we extend ALC to a regression framework, and then present results from several substantive use-cases. We discuss limitations and future directions before concluding.

2 Context in Context

> ... they are casting their problems on society and who is society? There is no such thing!
>
> —Margaret Thatcher, interview with *Woman's Own* (1987).

Paraphrased as "there is no such thing as society", Thatcher's quote has produced lively debate in the study and practice of UK politics. Critics—especially from the left—argued that this was primarily an endorsement of individual selfishness and greed. But more sympathetic accounts have argued that the quote must be seen in its full *context* to be understood. The implication is that reading the line in its original surroundings changes the meaning:

rather than embracing egotism, it emphasizes the importance of citizens' obligations to each other above and beyond what the state requires.

Beyond this specific example, the measurement and modeling of "context" is obviously a general problem. In a basic sense, context is vital: we literally cannot understand what is meant by a speaker or author without it. This is partly due to polysemy—the word "society" might mean many different things. But the issue is broader than this and is at the core of human communication. Unsurprisingly then, the study of context has been a long-standing endeavor in social science. Its centrality has been emphasized in everything from the history of ideas (Skinner, 1969) through the lens of "speech acts" (Austin, 1962); to describing cultural practices via "thick description" (Geertz, 1973); to understanding "political culture" (Verba and Almond, 1963); and to the psychology of decision making (Tversky and Kahneman, 1981). This latter meaning of context, as a "frame" through which humans view situations and make decisions, has motivated a large empirical literature in political science (see Chong and Druckman, 2007), with experimental approaches providing consistent evidence of its importance (e.g. Mutz, 2011).

2.1 Approaches to Studying Context

While the framing literature has demonstrated how one might use context as a treatment, it has generally not provided tools for *detecting* or *describing* context in observational data. For this second task, social science has turned to text approaches—with topic models being popular (see Grimmer, 2010; Quinn et al., 2010; Roberts et al., 2014). Topic models provide a way to understand the allocation of attention across groupings of words. While such models have a built-in notion of polysemy (a single word can be allocated to different topics depending on the words that occur with it), they are rarely used as a mechanism for studying how individual words are used to convey different ideas (Grimmer and Stewart, 2013). In other words, the common unit of analysis in the social science use of the topic model is the document, whereas in the questions we discuss the interest is in the contextual use of a

specific word.[1]

Social scientists have turned more recently to *word embeddings* (e.g. Rheault and Cochrane, 2019; Rodman, 2019; Rodriguez and Spirling, 2021). These methods predict a focal word as a function of the other words that appear within a small window of the focal word in the corpus (or the reverse, predict the neighboring words from the focal word). Using the central insight of the *distributional hypothesis* (Firth, 1957) we can talk about the "context" of a word in a very literal sense: it is the tokens that appear near it in text, on average. For example, Caliskan, Bryson and Narayanan (2017) and Garg et al. (2018) have explored relationships between words captured by embeddings to describe problematic gender and ethnic stereotypes in society at large. In these settings the embeddings tell us that words such as "man" or "woman" appear nearby different types of words which may have more negative or positive connotations.

Existing word embedding models work tolerably well for description in the social sciences. But we want more—we want inference, which requires statements about uncertainty. Suppose we wish to compare the context of "society" as conveyed by British Prime Ministers with that of US Presidents. Do they differ in a statistically significant way? To judge this, we need some notion of a null hypothesis, some understanding of the variance of our estimates, and a test statistic. While there have been efforts to compare embeddings across groups (Rudolph et al., 2017), and to give frameworks for such conditional relationships (Han et al., 2018), these are non-trivial to implement. Perhaps more problematically for most social science cases, they rely on underlying embedding models that struggle to produce "good" representations—that make sense, and correctly capture how that word is actually used—when we have few instances of a term of interest. This matters because we are typically far short of the word numbers which standard models require for optimal performance and

[1]Some models, such as the Structural Topic Model (Roberts, Stewart and Airoldi, 2016), allow for systematic variation in the use of a word across topics by different pieces of observed metadata. However, these techniques are extremely computationally intensive (especially relative to the approaches we present in this paper) and perhaps tellingly are more often used as a way to marginalize out differences than to study them directly (Lucas et al., 2015).

terms (like "society") may be used in ways that are idiosyncratic to a particular document or author.

In the next section, we will explain how we build on earlier insights from ALC embeddings (Khodak et al., 2018) to solve these problems in a fast, simple and sample-efficient regression framework. Before doing so, we note three substantive use cases that both motivate the methodological work we do, and show its power as a tool for social scientists. The exercise in all cases is linguistic *discovery* insofar as our priors are not especially sharp, and we want to see what we can learn from the texts in question. Nonetheless, in using the specific approach we outline in this paper, we will be able to make inferences with attendant statements about uncertainty. In that sense, our examples are intended to be illuminating for other scholars comparing corpora or comparing authors within a corpus.

Use-case I: Partisan Differences in word usage. A common problem in Americanist political science is to estimate partisan differences in the usage of a given term. Put literally: do Republicans and Democrats mean something different when they use otherwise identical words like `abortion`, `immigration` and `marriage`? While there have been efforts to understand differential word *rate of use* within topics pertaining to these terms (e.g. Monroe, Colaresi and Quinn, 2008), there has been relatively little work on whether the *same* words appear in different contexts. Below, we use the *Congressional Record* (Sessions 111–114) as our corpus for this study (Gentzkow, Shapiro and Taddy, 2018). This requires that we compare embeddings as a function of party (and other covariates).

Use-case II: Changing UK-US Understandings of 'Empire'. The United Kingdom's relative decline as a Great Power in the post-war period has been well-documented (e.g. Hennessy, 1992; Sanders and Houghton, 2016). One way that we might investigate the timing of US dominance (over the UK, at least) is to study the changing understanding of the term "`Empire`" in both places. That is, beyond any attitudinal or sentiment shift, did American and British policy-makers alter the way they used empire as the century wore on? If they did, when did this occur? And did the elites of these countries converge or

diverge in terms of their understandings and associations of the term? To answer these questions, we will statistically compare the embedding for the term "`Empire`" for the UK House of Commons (via *Hansard*) versus the US Congress (via the *Congressional Record* from 1935–2010.

Use-case III: Open-Ended Responses in the ANES. Open-ended questions—in which respondents may expand at length given a prompt—have long been of interest to scholars of public opinion (Krosnick, 1999). This is especially true given recent advances in analyzing such responses with topic models (Roberts et al., 2014) and other text analysis techniques (Hobbs, 2019). Like differential rates of word use, these approaches capture the frequency with which something is discussed, but does not explore the way in which it is discussed. Below, we study whether respondents of different ideological stripes have a fundamentally different understanding of issue areas. In particular, we embed responses to the American National Election Study (2016) question "what do you think is the most important political problem facing the United States today?", and 'regress' these on conservative or liberal identity.

3 Using ALC Embeddings To Measure Meaning

Our methodological goal in what follows is a regression-like hypothesis-testing framework for embeddings. This requires three related operations:

1. an efficient and transparent way to embed words, such that we can produce high quality representations even when a given word is rare. As we explain in this section, ALC is a promising avenue for that task.

2. given (1), a demonstration that in real problems, a *single* instance of a word's use is enough to produce a good embedding. This allows us to set up the hypothesis-testing problem as a multivariate regression, and is the subject of Section 4.1.

3. given (1) and (2), a method for making claims about the statistical significance of differences in embeddings, based on covariate profiles. We tackle that in Section 4.3.

We want a framework for comparing the contexts of words across groups or time. Ideally, that framework will deliver good representations of meaning even in cases where we have very few incidences of the words in question. ALC embeddings (Khodak et al., 2018) promise exactly this. We now give some background and intuition on that technique. We then replicate Rodman (2019) to demonstrate ALC's efficiency and quality.

3.1 Word Embeddings Measure Meaning Through Word Co-Occurence

Word embeddings techniques give every word a *distributed representation*—that is, a vector. The length or dimension (D) of this vector is—by convention—between 100 and 500. When the inner product between two different words (two different vectors) is high, we infer that they are likely to co-occur in similar contexts. The distributional hypothesis (Firth, 1957)—that we shall know a word by the company it keeps—then allows us to infer that those two words are similar in *meaning*. While such techniques are not new conceptually (e.g. Hinton et al., 1986; Bengio et al., 2003), recent methodological advances (Mikolov et al., 2013; Pennington, Socher and Manning, 2014) allow them to be estimated much more quickly than in the past. More substantively, word embeddings have been shown to be surprisingly useful, both as inputs to supervised learning problems and for understanding language directly. For example, embedding representations can be used to solve analogy reasoning tasks, implying the vectors do indeed capture relational meaning between words (e.g. Mikolov et al., 2013; Arora et al., 2018).

Understanding exactly *why* word embeddings work is non-trivial. Roughly speaking the most popular procedures are an approximate matrix factorization (Levy and Goldberg, 2014) of a reweighted word co-occurence matrix, and there is now a large literature proposing variants of the original techniques (e.g. Pennington, Socher and Manning, 2014; Faruqui et al., 2014; Lauretig, 2019). Some of these are geared specifically to social science applica-

tions where the general interest is in measuring changes in meanings, especially via "nearest neighbors" of specific words.

While the learned embeddings provide a rough sense of what a word means, it is difficult to use them to answer questions of the sort we posed in our case studies. Consider our interest in how Republicans and Democrats use the same word (e.g. `immigration`) differently. If we train a set of word embeddings on the entire *Congressional Record* we only have a single meaning of the word. We could instead train a separate set of embeddings—one for Republicans and one for Democrats. Unfortunately, lack of identification in the space means that these two embeddings are not comparable. Different approaches exist for aligning the two spaces together, but doing so requires that we have substantial amounts of data for each group. We now discuss a way to proceed that is considerably easier.

3.2 A Random Walk Theoretical Framework and ALC Embeddings

The core of our approach are ALC embeddings. The theory behind that approach is given by Arora et al. (2016) and Arora et al. (2018). Those papers conceive of documents being a 'random walk' in a discourse space, where words are more likely to follow other words if they are closer to them in an embedding space. Crucially for ALC, Arora et al. (2018) also proves that under this random walk model, a particular relationship will follow for the embedding of a word and the embeddings of the words that appear in the contexts *around it*.

To fix ideas, consider the following toy example. Our corpus is the memoirs of a politician, and we observe two entries, both mentioning the word 'bill':

1. *The debate lasted hours, but finally we* [voted on the bill and it passed] *with a large majority.*

2. *At the restaurant we ran up* [a huge wine bill to be paid] *by our host.*

As one can gather from the context—here, the three words either side of the instance of

'bill' in square brackets—the politician is using the term in two different (but grammatically correct) ways.

The key result from Arora et al. (2018) shows the following: if the random walk model holds, the researcher can obtain an embedding for word w (e.g. 'bill') by taking the average of the embeddings of the words around w ($\mathbf{u}_w$) and multiplying them by a particular square matrix $\mathbf{A}$. Put otherwise, if we can take averages of some vectors of words that surround w (based on some pre-existing set of embeddings) and if we can find a way to obtain $\mathbf{A}$ (which we will see is also straightforward), we can provide new embeddings for even very rare words. And we can do this almost instantaneously.

Returning to our toy example, consider the first, legislative, use of 'bill' and the words around it. Suppose we have embedding vectors for those words from some other larger corpus, like Wikipedia. To keep things compact, we will suppose those embeddings are all of three dimensions (such that $D = 3$), and take the following values:

$$\underbrace{\begin{bmatrix} -1.22 \\ 1.33 \\ 0.53 \end{bmatrix}}_{\text{voted}} \underbrace{\begin{bmatrix} 1.83 \\ 0.56 \\ -0.81 \end{bmatrix}}_{\text{on}} \underbrace{\begin{bmatrix} -0.06 \\ -0.73 \\ 0.82 \end{bmatrix}}_{\text{the}} \quad \text{bill} \quad \underbrace{\begin{bmatrix} 1.81 \\ 1.86 \\ 1.57 \end{bmatrix}}_{\text{and}} \underbrace{\begin{bmatrix} -1.50 \\ -1.65 \\ 0.48 \end{bmatrix}}_{\text{it}} \underbrace{\begin{bmatrix} -0.12 \\ 1.63 \\ -0.17 \end{bmatrix}}_{\text{passed}}$$

Obtaining $\mathbf{u}_w$ for 'bill' simply requires averaging these vectors and thus

$$\mathbf{u}_{\text{bill}_1} = \begin{bmatrix} 0.12 \\ 0.50 \\ 0.40 \end{bmatrix},$$

with the subscript denoting the first use case. We can do the same for the second case—the

restaurant sense of 'bill'—from the vectors of `a`, `huge`, `wine`, `to`, `be` and `paid`. We obtain

$$\mathbf{u}_{\text{bill}_2} = \begin{bmatrix} 0.35 \\ -0.38 \\ -0.24 \end{bmatrix},$$

which differs from the average for the first meaning. A reasonable instinct is that these two vectors should be enough to give us an embedding for 'bill' in the two senses. Unfortunately, they will not—this is shown empirically in Khodak et al. (2018) and in our Trump/trump example below. The intuition is that averaging embeddings over-exaggerates common components of frequent words (like stop words). So we will need the $\mathbf{A}$ matrix too: it downweights these directions so they don't overwhelm the induced embedding.

Khodak et al. (2018) show how to put this logic into practice. The idea is that a large corpus (generally the corpus the embeddings were originally trained on, such as Wikipedia) can be used to estimate the transformation matrix $\mathbf{A}$. This is a one time cost after which each new word embedding can be computed *à la carte* (hence the name), rather than needing to retrain an entire corpus just to get the embedding for a single word. As a practical matter, the estimator for $\mathbf{A}$ can be learned efficiently with a lightly modified linear regression model which reweights the words by a non-decreasing function $\alpha(\cdot)$ of the total instances of each word (n_w) in the corpus. This reweighting addresses the fact that words which appear more frequently have embeddings which are measured with greater certainty. Thus we learn the transformation matrix as,

$$\hat{\mathbf{A}} = \arg\min_{\mathbf{A}} \sum_{w=1}^{W} \alpha(n_w) \|\mathbf{v}_w - \mathbf{A}\mathbf{u}_w\|_2^2 \tag{1}$$

The natural log is a simple choice for $\alpha(\cdot)$, and it works well. Given $\hat{\mathbf{A}}$, we can introduce new embeddings for any word by averaging the existing embeddings for all words in its context to create $\mathbf{u}_w$ and then applying the transformation such that $\hat{\mathbf{v}}_w = \hat{\mathbf{A}}\mathbf{u}_w$. The transformation

matrix is not particularly hard to learn (it is a linear regression problem) and each subsequent induced word embedding is a single matrix multiply.

Returning to our toy example, suppose that we estimate $\hat{\mathbf{A}}$ from a large corpus like *Hansard* or the *Congressional Record* or wherever we obtained the embeddings for the words that surround 'bill.' Suppose that we estimate

$$\hat{\mathbf{A}} = \begin{bmatrix} 0.81 & 3.96 & 2.86 \\ 2.02 & 4.81 & 1.93 \\ 3.14 & 3.81 & 1.13 \end{bmatrix}.$$

Taking inner products, we have

$$\mathbf{v}_{\text{bill}_1} = \hat{\mathbf{A}} \cdot \mathbf{u}_{\text{bill}_1} = \begin{bmatrix} 3.22 \\ 3.42 \\ 2.73 \end{bmatrix} \quad \text{and} \quad \mathbf{v}_{\text{bill}_2} = \hat{\mathbf{A}} \cdot \mathbf{u}_{\text{bill}_2} = \begin{bmatrix} -1.91 \\ -1.58 \\ -0.62 \end{bmatrix}.$$

These two transformed embeddings vectors are more different than they were—a result of downweighting the commonly appearing words around them—but that is not the point *per se*. Rather, we expect them to be informative about the word sense by, for example, comparing them to other (pre-estimated) embeddings in terms of distance. Thus we might find that the nearest neighbors of $\mathbf{v}_{\text{bill}_1}$ are

$$\text{legislation} = \begin{bmatrix} 0.11 \\ 0.52 \\ 0.38 \end{bmatrix} \quad \text{and} \quad \text{amendment} = \begin{bmatrix} 0.15 \\ 0.47 \\ 0.42 \end{bmatrix}$$

while the nearest neighbors of $\mathbf{v}_{\text{bill}_2}$ are

$$\text{dollars} = \begin{bmatrix} -1.92 \\ -1.54 \\ -0.60 \end{bmatrix} \quad \text{and} \quad \text{cost} = \begin{bmatrix} -1.95 \\ -1.61 \\ -0.63 \end{bmatrix}.$$

This makes sense, given how we would typically read the politician's lines above. The key here is that the ALC method allowed us to infer the meaning of words that occurred rarely in a small corpus (the memoirs) without having to build embeddings for those rare words in that small corpus: we could 'borrow' and transform the embeddings from another source. Well beyond this toy example, Khodak et al. (2018) finds empirically that the learned $\hat{\mathbf{A}}$ in a large corpus recovers the original word vectors with high accuracy (greater than .9 cosine similarity). They also demonstrate that this strategy achieves state-of-the-art and near state-of-the-art performance on a wide variety of natural language processing tasks (e.g. learning the embedding of a word using only its definition, learning meaningful n-grams, classification tasks etc.) at a fraction of the computational cost of the alternatives.

The ALC framework has three major advantages for our setting: transparency, computational ease, and efficiency. First, compared to many other embedding strategies for calculating conditional embeddings (e.g., words over time) the information used in ALC is transparent. The embeddings are derived directly from the additive information of the words in the context window around the focal word, there is no additional smoothing or complex interactions across different words. Furthermore, the embedding space itself does not change, it remains fixed to the space defined by the pre-trained embeddings. Second, this same transparency leads to computational ease. The transformation matrix $\mathbf{A}$ only has to be estimated once and then each subsequent induction of a new word is a single matrix multiply and thus effectively instantaneous. Later we will be able to exploit this speed to allow bootstrapping and permutation procedures that would be unthinkable if there was an expensive model fitting procedure for each word. Finally, ALC is efficient in the use

of information. Once the transformation matrix is estimated, it is only necessary that $\mathbf{u}_w$ converges—in other words, we only need to estimate a D-dimensional mean from a set of samples. In the case of a 6-word symmetric context window there are twelve words total within the context window; thus, for each instance of the focal word we get 12 samples from which to estimate the mean.

While Khodak et al. (2018) focused on using the ALC framework to induce embeddings for rare words and phrases, we will apply this technique to embed words used in different partitions of a single corpus or to compare across corpora. This allows us to capture differences in embeddings over time or by speaker, even when we have only a few instances within each sample. Importantly, unlike other methods, we don't need an entirely new corpus to learn embeddings for select focal words, we can select particular words and calculate (only) their embeddings using only the contexts around those particular words.[2] We now demonstrate this power of ALC by replicating Rodman (2019).[3]

3.3 Proof of Concept for ALC in Small Political Science Corpora: Reanalyzing Rodman (2019)

The task in Rodman (2019) is to understand changes in the meaning of `equality` over the period 1855–2016 in a corpus consisting of the headlines and other summaries of news articles. As a gold standard, a subset of the articles is hand-coded into fifteen topic word categories—of which five are ultimately used in the analysis—and the remaining articles are coded using a supervised topic model with the hand-coded data as input. Four embeddings techniques are used to approximate trends in coverage of those categories, via the (cosine) distance between the embedding for the word `equality` and the embeddings for the category

[2]For context, other approaches in computer science have used anchoring words (Yin, Sachidananda and Prabhakar, 2018), incremental training (Kim et al., 2014*a*), atemporal vectors (Di Carlo, Bianchi and Palmonari, 2019), and vector space alignment (Hamilton, Leskovec and Jurafsky, 2016). Gonen et al. (2020) provide another approach that like ours, emphasizes stability and simplicity.

[3]Many papers in computer science have studied semantic change (see Kutuzov et al., 2018, for a survey). Rodman (2019) provides the state of the art in political science.

labels. This is technically challenging, because the corpus is small—the first 25 year slice of data has only 80 documents—and in almost 30% of the word-era combinations there are fewer than 10 observations.[4]

Rodman (2019) tests four different methods by comparing results to the gold standard; ultimately, the chronologically trained model (Kim et al., 2014*b*) is the best performer. In each era (of 25 years), the model is fit several times on a bootstrap resampled collection of documents and then averaged over the resulting solutions (Antoniak and Mimno, 2018). Importantly, the model in period t is initialized with period $t-1$ embeddings, while the first period is initialized with vectors trained on the full corpus. Even for a relatively small corpus this process is computationally expensive, and our replication took about five hours of compute time on an 8-core machine.

The ALC approach to the problem is simple. For each period we use ALC to induce a period-specific embedding for `equality` as well as each of the five category words: `gender`, `treaty`, `german`, `race` and `african_american`. We use `GloVe` pre-trained embeddings and the corresponding transformation matrix estimated by Khodak et al. (2018)—in other words, we make use of no corpus-specific information in the initial embeddings and require as inputs *only the context window around each category word*. Following Rodman, we compute the cosine similarity between `equality` and each of the five category words, for each period. We then standardize (make into z-scores) those similarities. The entire process is transparent, involves no parameters to set, and takes only a few milliseconds (the embeddings themselves involve six matrix multiplies).

How does ALC do? Figure 1 is the equivalent of Figure 3 in Rodman (2019). It displays the normalized cosine similarities for the chronological model (CHR, taken from Rodman (2019)) and ALC, along with the gold standard (GS). We observe that ALC tracks approximately as well as Rodman's chronological model on its own terms. Where ALC clearly does better is on each model's nearest neighbors (Tables 1 and 2): it produces more

[4]We provide more information on the sample constraints in Supporting Information A.

semantically interpretable and conceptually precise nearest neighbors than the chronological model. This is partly a result of the ALC model being able to produce nearest neighbors beyond those in the original corpus, borrowing from semantic information stored in the pre-trained embeddings.

We emphasize that in the 1855 corpus, four of the five category words (all except `african_american`) are estimated using *five or fewer instances*. While the chronological model is sharing information across time periods, ALC is treating each slice separately, meaning that our analysis could be conducted effectively with even fewer time periods.

african_american		gender		treaty		german		race		equality	
CHR	ALC	CHR	ALC	CHR	ALC	CHR	ALC	CHR	ALC	CHR	ALC
equality	suffrage	will	legislatures	britain	equality	reich	visit	enfranchisement	enfranchisement	of	enacment
the	emancipation	performing	missourians	extradition	toleration	berlin	france	marriage	equality	the	abolition
and	fairness	give	suffrage	interpolation	speech	arms	eugenia	newmarket	interrelation	and	enacting
of	guaranteeing	blackwell	disestablish	minister	championing	hitler	bilateral	louise	expounder	in	effecting
whites	slavery	american	constitutions	rouher	extradition	von	relations	need	abrogation	to	abolishment

Table 1: Nearest neighbors for the 1855 corpus.

african_american		gender		treaty		german		race		equality	
CHR	ALC	CHR	ALC	CHR	ALC	CHR	ALC	CHR	ALC	CHR	ALC
crandall's	nonwhites	equality	equality	narrow	equality	maintains	universities	universe	equality	the	gender
costs	asians	the	inequalities	designed	affirms	hinge	colleges	1950s	segregation	for	gays
unraveling	cubans	for	inequity	missed	reaffirms	holstein's	campuses	warriors	inequalities	of	lesbians
treats	suburbanites	of	inequality	assure	affirming	equality's	striving	posits	discrimination	and	transgender
congresswoman	championing	and	lesbians	trade	upholds	kiel	decades	purdyâ\u0080\u0099s	affirmative	to	lgbt

Table 2: Nearest neighbors for the 2005 corpus.

Collectively, these results suggest that ALC is competitive with the current state of the art within the kind of small corpora that arise in social science settings. We now turn to providing a hypothesis testing framework that will allow us to answer the types of questions we introduced above.

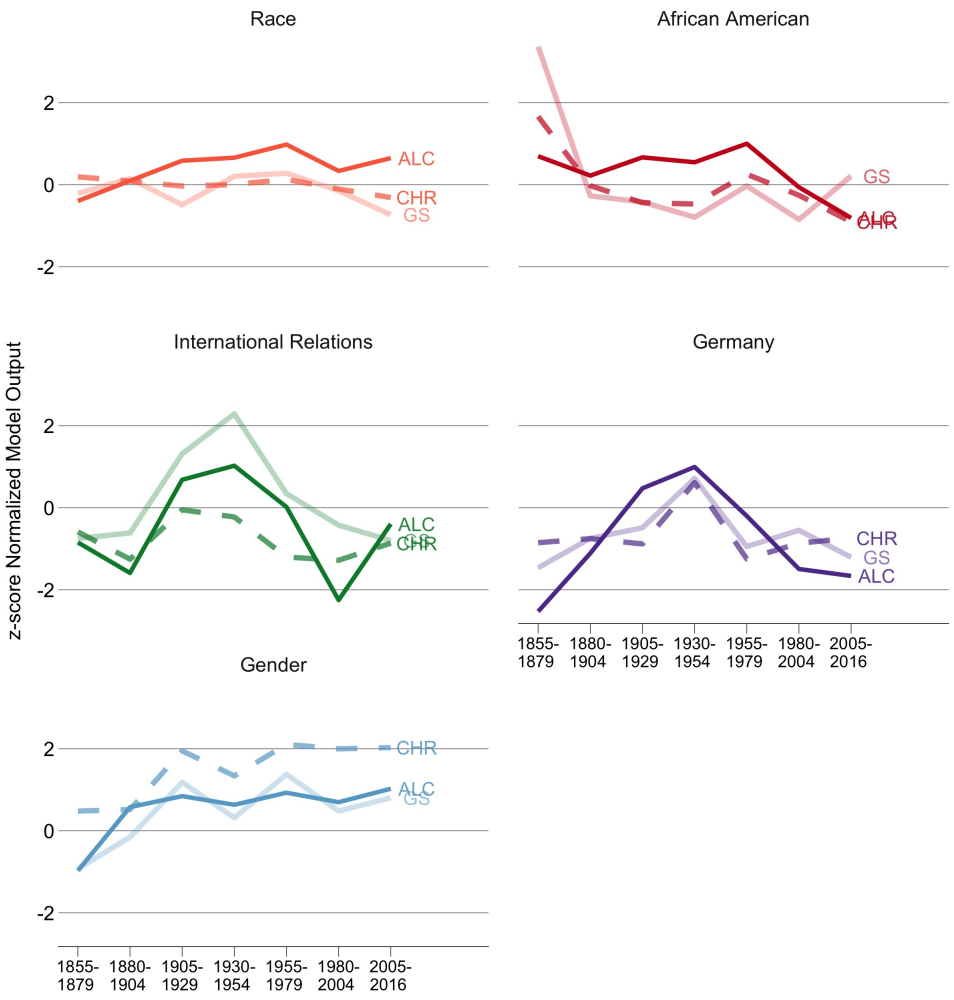


Figure 1: Replication of Figure 3 in Rodman (2019) adding ALC results. ALC = ALC model; CHR = chronological model and GS = gold standard.

4 Testing Hypotheses about Embeddings

Ultimately we want to speak of the way that embeddings differ systematically across levels of covariates. To do this, we will set up a regression-like framework, where each 'observation' is the embedding of a single word. ALC will assist us, but first we show that it can learn meaningful embeddings from *one* example use.

4.1 ALC Can Distinguish Word Meanings From One Example Use

Above we explained that ALC averaged pre-trained embeddings and then applied a linear transformation. This new embedding vector has, say, 300 dimensions, and we might reasonably be concerned that it is too noisy to be useful. To evaluate this, we need a ground truth. So we study a recent *New York Times* corpus; based on lead paragraphs, we show that we can reliably distinguish `Trump` the person (2017–2020) from other sense of `trump` as a verb or noun (1990–2020).

For each sense of the word (based on capitalization) we take a random sample of 100 realizations from our New York Times corpus and embed them using ALC. We apply k-means clustering with two clusters to the set of embedded instances and evaluate whether the clusters partition the two senses. If ALC works, we should obtain two separate clouds of points that are internally consistent (in terms of the senses of the term). This is approximately what we see. Figure 2 provides a visualization of the 300-dimensional space projected to two dimensions with PCA and identifying the two clusters by their dominant word sense. We explicitly mark misclassifications with an x.

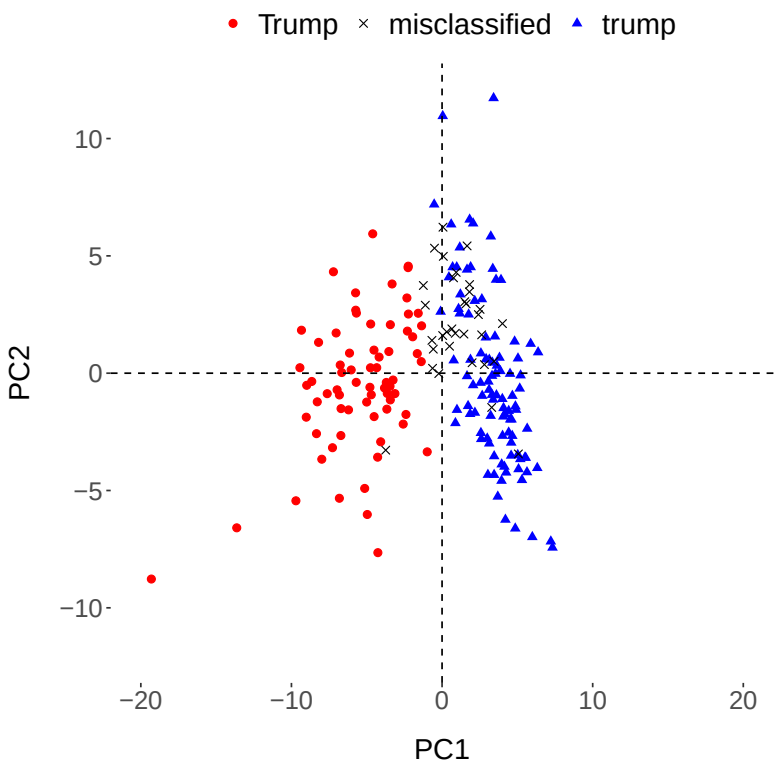


Figure 2: Each observation represents a single realization of a target word, either of `trump` or `Trump`. Misclassified instances refer to instances of either target word that were assigned the majority cluster of the opposite target word.

To provide a quantitative measure of performance we compute the average cluster homogeneity: the difference in proportions between the majority and minority classes in each cluster. This value ranges between 0—both clusters have equal numbers of both context types—and 1—each cluster consists entirely of a single context type. By way of comparison, we do the same exercise using other popular methods of computing word vectors for each target realization including: simple averaging of the corresponding pre-trained embeddings (ALC without transformation by $\mathbf{A}$), tf-idf weighting and BERT contextual embeddings (Devlin et al., 2018).[5] We further explore performance variation as a function of sample size. To quantify uncertainty in our metric, we use block bootstrapping—resampling individual instances of the focal word.[6] Figure 3 summarizes our results.

While tf-idf does improve with larger sample sizes, it never catches up to our approach.

[5]BERT is a substantially more complicated embedding method which produces contextually-specific em-

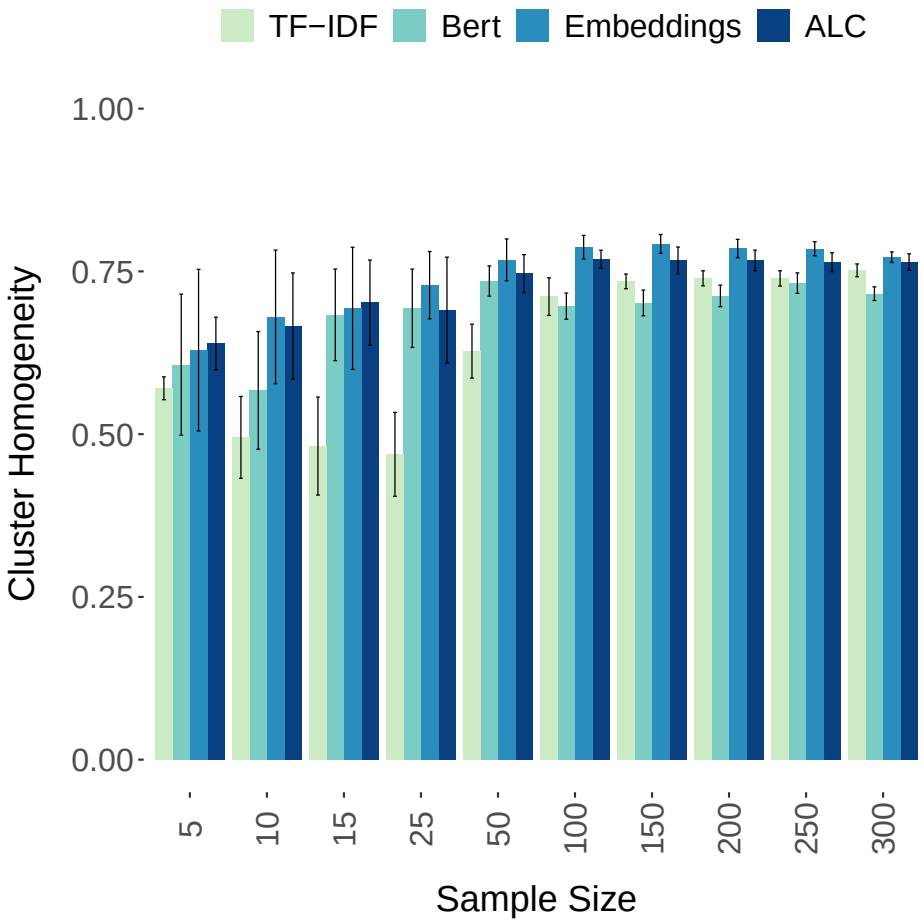


Figure 3: Cluster homogeneity as a function of sample size and word vector method.

Somewhat surprisingly, our approach also outperforms the transformer-based BERT embeddings.[7] On the other hand, simple averaging of embeddings does seem to perform equally well. Does this mean the linear transformation that distinguishes ALC from simple averaging is redundant? To evaluate this, we look at nearest neighbors using both methods. Table 3 displays these results. We observe that simple averaging of embeddings produces mainly

beddings and uses word order information.

[6]Note here that we are treating the $\mathbf{A}$ matrix as fixed and thus we are not incorporating uncertainty in those estimates. The hope is that this is a relatively small d^2 number of parameters, relative to the number of tokens in the corpus on which it is trained, and thus the uncertainty contribution will be small. This is an important concern for cases where users train their own $\mathbf{A}$ if they are using a small corpus.

[7]This may be a result of BERT being optimized for sentence embeddings more than embeddings for an individual word. Nonetheless, it is surprising given that BERT-based models currently lead almost every natural language process benchmark task. Even at comparable performance though there would be reason not to use BERT models simply based on computational cost and comparative complexity.

stopwords as nearest neighbors. ALC, on the other hand, outputs nearest neighbors aligned with the meaning of each term, `Trump` is associated with president Trump while `trump` is largely associated with whist (the card game) terms. This serves to highlight the importance of the linear transformation **A** in the ALC method.

Trump		**trump**	
Embeddings	**ALC**	**Embeddings**	**ALC**
but	president	but	declarer
that	assailed	only	spades
even	clinton	even	colloquies
because	bush	one	suitors
the	presidents	because	counterclaims
would	assailing	that	reprove
not	impeaching	they	emboldens
what	upbraided	same	rationales
when	alluded	well	overbid
also	barack	the	frontmen

Table 3: Top 10 nearest neighbors using simple averaging of embeddings and ALC.

While this example is a relatively straightforward case of polysemy, we also know that the meaning of `Trump`, the surname, underwent a significant transformation once Donald J. Trump was elected president of the United States in November 2016. This is a substantially harder case since the person being referred to is still the same, even though the contexts it is employed in—and thus in the sense of the distributional hypothesis, the meaning—has shifted. But as we show in Supporting Information B, ALC has no problem with this case either, returning excellent cluster homogeneity and nearest neighbors.

The good news for the `Trump` examples is that ALC can produce reasonable embeddings even from single instances. Next we demonstrate that each of these individual instances can be treated as an observation in a hypothesis-testing framework. Before doing so, while readers may be satisfied about the performance of ALC in small samples, they may wonder about its performance in *large* samples. That is, whether it converges to the inferences one would make from a 'full' corpus model as the number of instances increases; the answer is 'yes' and we provide more details in Supporting Information C.

4.2 à la Carte on Text embedding regression model: `conText`

Recall the original statement of the relationship between the embedding of a focal word and the embeddings of the words within its context: $\mathbf{v}_w = \mathbf{A}\mathbb{E}[\mathbf{u}_w]$. Here we note that because the matrix $\mathbf{A}$ is constant we can easily swap it into the expectation and then calculate the resulting expectation conditional on some covariate X: $\mathbb{E}[\mathbf{A}\mathbf{u}_w|X]$. In particular, this can be done implicitly through a multivariate regression procedure. In the case of word meanings in discrete subgroups, this is exactly the same as the use of ALC applied above.

To illustrate our set up, suppose that each $\mathbf{v}_{w_i}$ is the embedding of a particular instance of a given word in some particular context, like `Trump`. Each is of some dimension, D and thus each 'observation' in this setting is a $1 \times D$ embedding vector. We can stack these to produce an outcome variable $\mathbf{Y}$ which is of dimensions n (the number of instances of a given word) by D. The usual multivariate matrix equation is then:

$$\underbrace{\mathbf{Y}}_{n\times D} = \underbrace{\mathbf{X}}_{n\times p+1}\underbrace{\boldsymbol{\beta}}_{p+1\times D} + \underbrace{\mathbf{E}}_{n\times D}$$

where $\mathbf{X}$ is a matrix of p covariates and includes a constant term, while $\boldsymbol{\beta}$ is a set of p coefficients and an intercept (all of dimension D). Then $\mathbf{E}$ is an error term.

To keep matters simple, suppose that there is a constant and then one binary covariate indicating group membership (in the group, or not). Then, the coefficient $\boldsymbol{\beta_0}$ (the first row of the matrix $\boldsymbol{\beta}$) is equivalent to averaging over all instances of the target word belonging to those not in the group. Meanwhile, $\boldsymbol{\beta_0} + \boldsymbol{\beta_1}$ (the second row of $\boldsymbol{\beta}$) is equivalent to averaging over all instances of the target word that belong to the group (i.e. for which the covariate takes the value 1, as opposed to zero). In the more general case of continuous covariates, this provides a model-based estimate of the embedding among all instances at a given level of the covariate space.

The key outputs from this à la Carte on Text (`conText`) embedding 'regression' model are:

- the coefficients themselves, $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$. These provide the estimated embeddings for the word in question, across the groups. We can then take the cosine distance between these implied embeddings and the (pre-trained) embeddings of other words to obtain the nearest neighbors for the two groups.

- the (Euclidean) norms of the coefficients. In the categorical covariate case, these tell us how different one group is to another in a *relative* sense. While the magnitude of this difference is not directly interpretable, we can nonetheless comment on whether it is statistically significantly different from zero. To do this, we use a variant of covariate assignment shuffling suggested by Gentzkow, Shapiro and Taddy (2019). In particular, we randomly shuffle the entries of the $\mathbf{Y}$ column and run the regression many (here 100) times. Each time, we record the norms of the coefficients. We then compute the proportion of those values that are larger than the *observed* norms (i.e. with the true group assignments). This is the empirical p-value.

Note that, if desired, one can obtain the sampling distribution (and thus standard errors) of the (normed) coefficients via non-parametric bootstrap. This allows for comments on the *relative* size of differences in embeddings across and within groups as defined by their covariates. We now show how the `conText` model may be used in a real estimation problem.

4.3 Our Framework in Action: Pre-Post Election Hypothesis Testing

We can compare the change in the usage of the word `Trump` to the change in the usage of the word `Clinton` after the 2016 election. Given Trump won the election and subsequently became President—a major break with respect to his real-estate/celebrity past—we expect a statistically significant change for `Trump` relative to any changes in the usage of `Clinton`.

We proceed as follows: for each target word-period combination—`Clinton` and `Trump`, pre-election (2011–2014) and post-election (2017-2020)— we embed each individual instance

of the focal word from our *New York Times* corpus of leading article paragraphs, and estimate the following model:

$$Y = \beta_0 + \beta_1 \text{Trump} + \beta_2 \text{Post\_Election} + \beta_3 \text{Trump} \times \text{Post\_Election} + \epsilon \tag{2}$$

where Trump is an indicator variable equal 1 for `Trump` instances, 0 otherwise. Likewise Post_Election is a dummy variable equal 1 for 2017-2020 instances of `Trump` or `Clinton`. As before, this is simply a regression-based estimator for the individual sub-groups. We will use permutation for hypothesis testing.

Figure 4 plots the norm of the $\hat{\beta}$s. The significant positive value on the `Trump x Post_Election` coefficient indicates the expected additional shift in the usage of `Trump` post-election over and above the shift in the usage of `Clinton`.

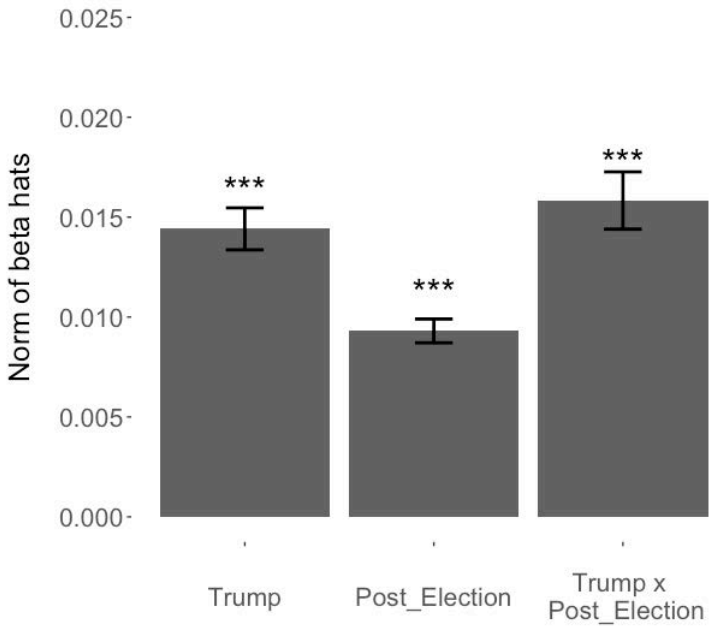


Figure 4: Relative semantic shift from `Trump`. Values are the norm of $\hat{\beta}$ and bootstrap confidence intervals.

While this news is encouraging, readers may wonder how the `conText` regression model performs relative to a 'natural' alternative—specifically, a full embeddings model fit to each

use of the term by covariate value(s). This would require the entire corpus (rather than just the instances of `Trump` and `Clinton`) and would be computationally slow, but we might reasonably believe it would yield more accurate (and different inferences). As we demonstrate in Supporting Information D, inferences are similar and our approach is more stable by virtue of holding constant embeddings for all words but the focal word.

5 Results

We now turn from proof-of-concept to substantive use cases, beginning with notions of partisan differences in the United States.

5.1 Partisan and Gender Differences

We want to evaluate partisan and gender differences in the usage of a given term in Congress. Our focus is a set of target words known to be politically charged: `abortion`, `immigration` and `marriage`. We also include three stopwords—`and`, `the` and `but`—in our target set as comparison, for which we do not anticipate partisan differences.

We estimate the following multivariate model for each of our words:

$$Y = \beta_0 + \beta_1 \texttt{Party} + \beta_2 \texttt{Gender} + \epsilon. \tag{3}$$

To form the dependent variable, we build a corpus of instances and their respective contexts and embed each realization individually using ALC. To form the righthand side, we use indicator variables (Republican or otherwise; Male or otherwise). We use permutation to approximate the null and bootstrapping to quantify the sampling variance.

Note again that magnitudes have no natural absolute interpretation, but can be compared relatively: that is, a larger coefficient on X_i relative to X_j implies the difference in embeddings for the groups defined by i is larger than the difference in the groups as defined by j. Our actual results are displayed in Figure 5. The 'gender' coefficient is the average difference

across the gender classes, controlling for party. The 'party' coefficient is the average difference across the parties, controlling for gender.

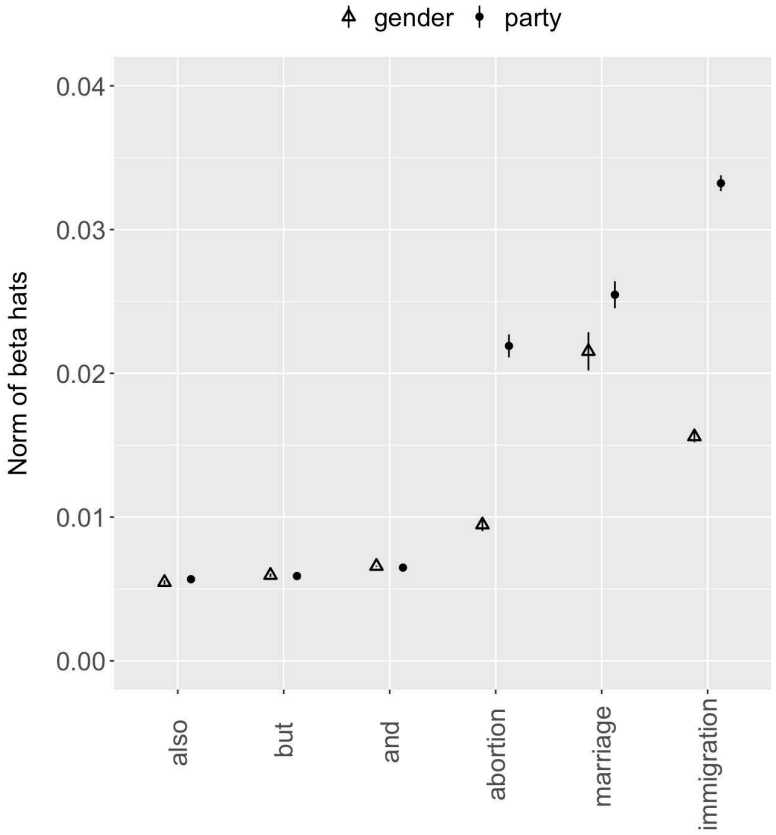


Figure 5: Differences in word meaning by gender and party: generally, different genders in the same party have more similar understanding of a term, than the same gender across parties.

As expected, the differences across parties and across genders, is much larger for the more political terms—relative to function words. But, in addition, embeddings differ more by party than they do by gender. That is, on average, males and non-males *within* a party have more similar understandings of the terms than males and non-males *across* parties.

The "most partisan" target in our set is `immigration`. Table 4 shows the top 10 nearest neighbors for each party. One reading of these nearest neighbors is that Democrats were pushing for reform of existing laws while Republicans were mainly arguing for enforcement. This fits with contemporary understandings of US politics, and suggests some validity of our

general approach.

Democrats	reform, overhauling, legislation, overhaul, enact, enacting, reforming, immigration, entitlement, revamp
Republicans	immigration, laws, enacting, enforcing, enacted, legislation, illegals, legislations, enforcement, enact

Table 4: Top 10 nearest neighbors for the target term `immigration`.

5.2 The Meaning of 'Empire'

We return to the second substantive case study we laid out in Section 2. Recall that our plan was to compare the embedding of `Empire` in the UK and US context for the period 1935–2010. The setup for this is similar to our partisan differences over time example except the group indicator captures corpus membership—*Congressional Record* or *Hansard*. Table 5 summarizes the document collections in aggregate. In the estimation that follows we use the top (most frequent) 5000 tokens of the combined corpus (i.e. combining the corpora).

	UK	US
source	*Hansard*	*Congressional Record*
period	1935–2010	1935-2010
# speeches	4.4 million	10.4 million
# tokens	717 million	1.3 billion
# unique tokens	5000	5000

Table 5: Description of *Hansard* and *Congressional Record* corpora for comparing embedding of `empire`

The multivariate regression analogy here is

$$Y = \beta_0 + \beta_1 \texttt{Congressional Record} + \epsilon \tag{4}$$

estimated for every year of the period. Interest focuses on the (normed) value of β_1: when this rises, the use of `Empire` is becoming less similar across the corpora (Congress is becoming more distinctive). The time series of the β_1s is given in Figure 6. The basic summary is that, sometime around 1947-48, there was a once-and-for-all increase in the distance between US and UK understandings of `Empire`. We confirmed this with a structural break test (in the sense of Bai and Perron, 1998).

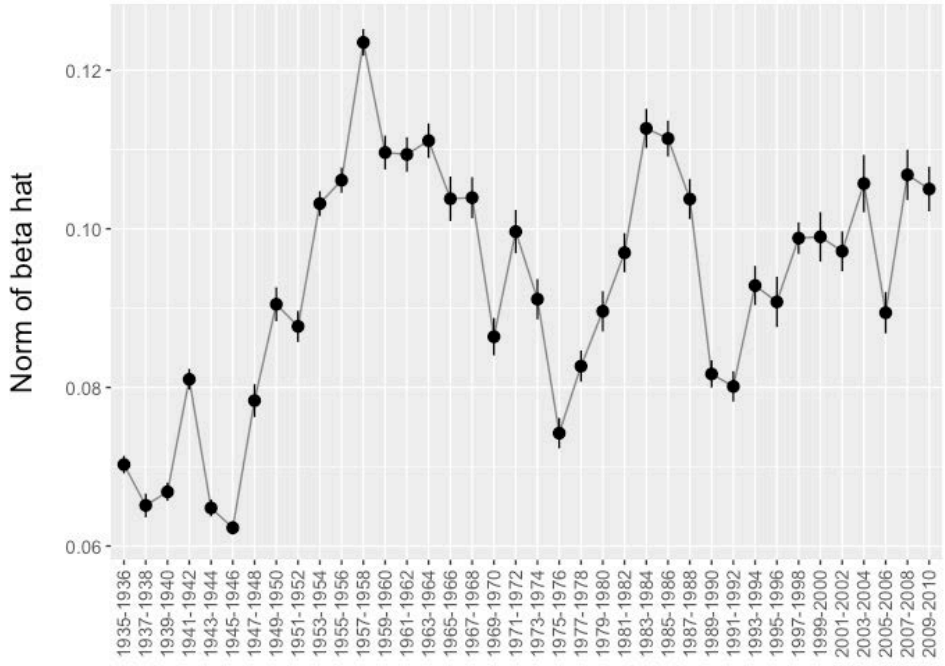
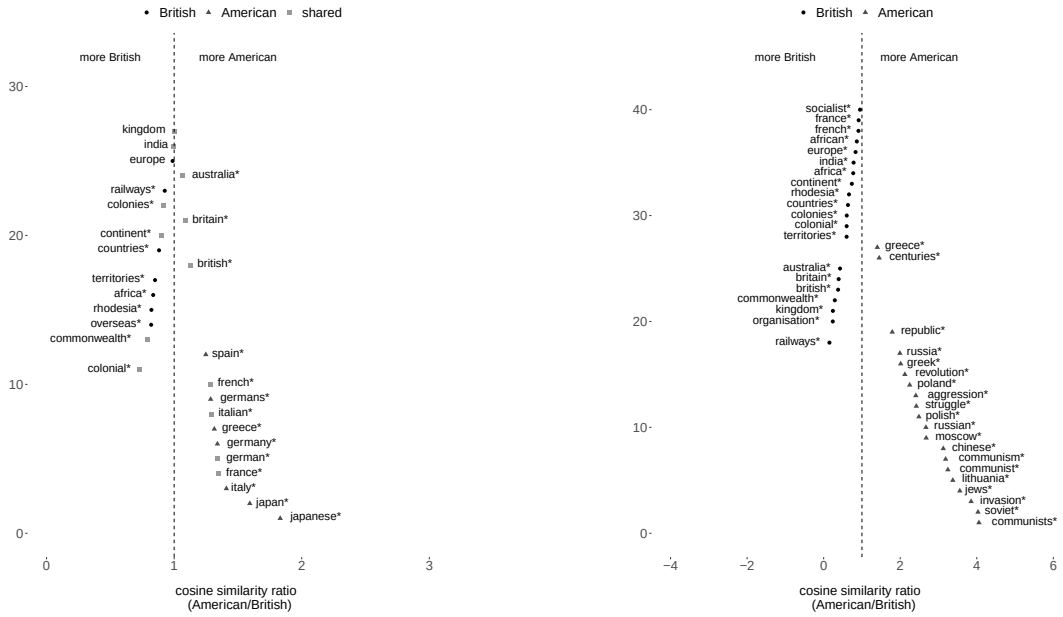


Figure 6: Distance between British and American understanding of `Empire`, 1935–2010: larger values imply the uses are more different.

To understand more about the substance of the change, consider Figure 7. There we report the 'most American' and 'most British' (with reference to the parliaments) terms from the period either side of the split in the time series. Specifically, we calculate the cosine similarity between the ALC embedding for `Empire` and each nearest neighbor in the UK and US corpus. The x-axis is the *ratio* of these similarities across the corpora: when it is large, the word is relatively closer to the US understanding of `Empire` than to the UK one. An asterisk by the term implies that ratio's deviation from 1 is statistically significantly larger than its permuted value, $p < 0.01$. The y-axis reports the rank of the word in terms of

distance from 0: so, words near the bottom of the plot are more distinct than those near the top.

The main observation is that in the pre-period, while British and American MPs have distinct terms, they talk about `Empire` primarily in connection with the old European powers: Britain, Spain, Italy, Germany, and France. By contrast, the vocabularies are radically different in the post-break period. In particular, the UK parliament continues to talk of the "British" empire (and its travails in "India" and "Rhodesia"), but the US focus has switched. For the American legislators, the most distinct nearest neighbors are now "invasion", "Soviet" and "communists", with explicit references to eastern European nations like "Lithuania". Clearly then, US understandings of empire are specifically with respect Soviet imperial ambitions.



(a) Pre-1948 (b) Post-1948

Figure 7: UK and US discussions of `Empire` diverged after 1948: most US and UK nearest neighbors pre and post estimated breakpoint.

5.3 Open-Ended Responses

Recall that in the case of the *ANES*, we are interested in responses to the query "what do you think is the most important political problem facing the United States today?" To examine the answers to this question, we embed the full response (so, not simply the context around a target word) using ALC. We then regress that response embedding on whether the respondent self-identified as 'Conservative' or 'Liberal' (we dropped all other identities for this exercise). In a way analogous to Figure 7, we report the "most Conservative" and "most Liberal" nearest neighbors. We do this in Figure 8. Some issues are nearest neighbors for both: for example, `homelessness` and `unemployment`. But there are stark differences for others: issues of `immigration` are named by conservatives, while liberals focus on equal rights, noting `racism`, `sexism` and `homophobia`. This roughly accords with our priors, but is nonetheless informative.

6 Limitations and Challenges

All new methods have limitations: we now candidly discuss those for our case, and how future efforts might mitigate those.

First, when we talk of "meaning", this is purely in the sense of the distributional hypothesis: it is about co-occurrences, not some deeper psychological understanding. For the cases we studied, this was enough. But whether this is sufficiently subtle for a given purpose is up to the researcher. We think the interpretability of our nearest neighbor output will help scholars decide whether this operationalization works in their case.[8]

Second, understanding how a focal word's meaning differs across covariate values requires that the meaning of other words is relatively fixed. In practice, we found excellent performance even when the assumption does not hold exactly. But obviously there must be limits.

[8]The word embeddings are simply a lower dimensional representation of the co-occurrences of that word with all other words in the vocabulary. Thus the empirical claim that an embedding is different at some different time is simply a smoothed estimate of whether the distribution of words surrounding the focal word has changed.

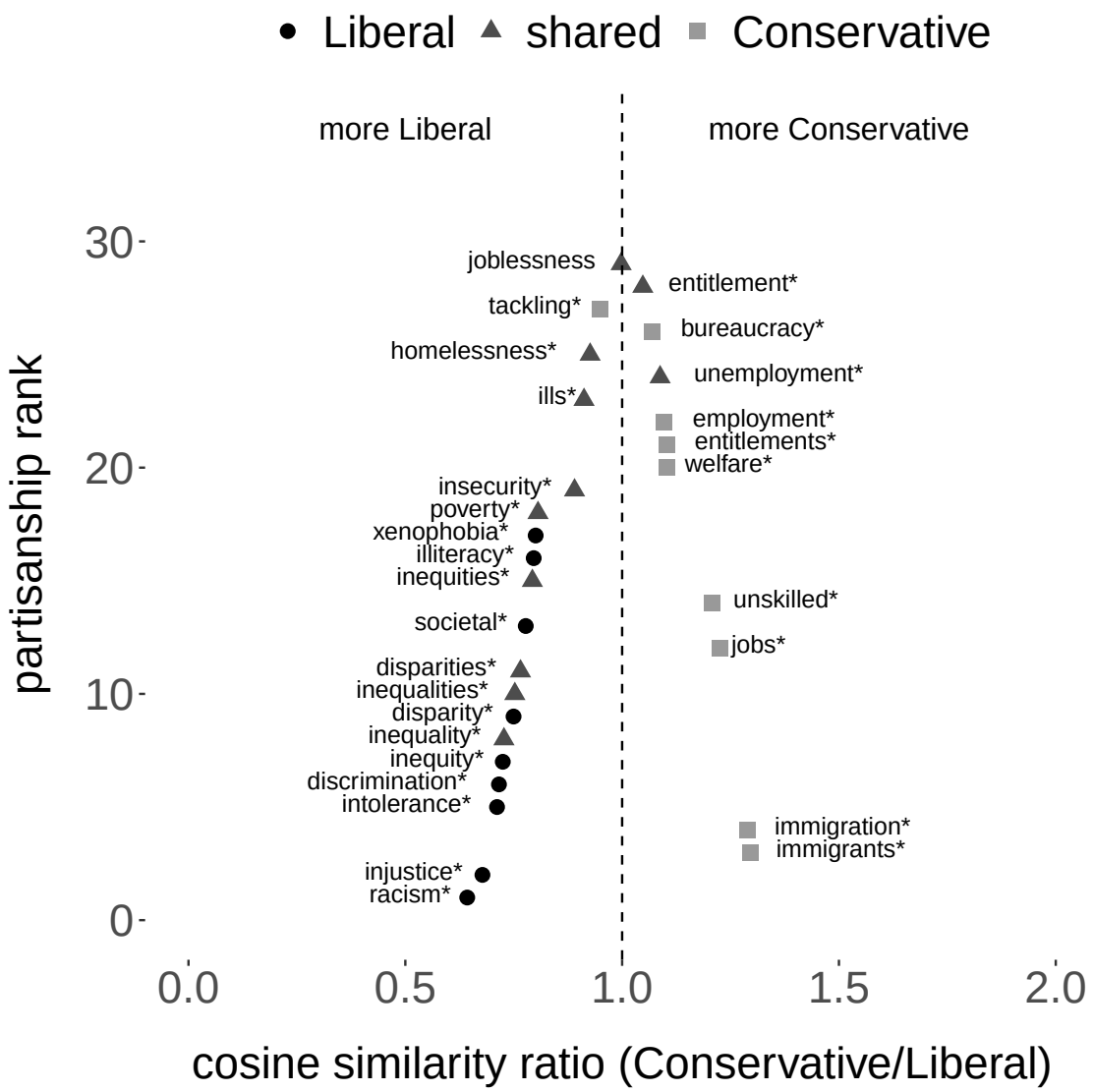


Figure 8: Distance between (self identified) liberal and conservatives understandings of "the most important political problem facing the United States today" differ markedly. The 'most liberal' terms are to the bottom left; most conservative are to the bottom right.

Future work could consider second-order information—which words co-occur with words that co-occur with the focal words—but it is unclear whether this will provide sufficient performance gains to justify the complexity.

Third, our approach is simple with essentially no parameters to tune, but this means it uses few data points in high dimensions to draw conclusions. This is always difficult (see Gentzkow, Shapiro and Taddy, 2019), and means, for example, our estimates of the norms

are biased for rare words (a problem we address with the covariate shuffling). As with any quantitative method, our advice would be that users do not over-interpret very rare instances. Related to this, future work might explore the behavior of both the intermediate (e.g. the weighting matrix) and end (e.g. the norms) products of the technique under different data quantity constraints.[9]

7 Conclusion

"Contextomy"—the art of quoting of context to ensure that a speaker is misrepresented—has a long and troubling history in politics (McGlone, 2005). As a strategy, it works, because judicious removal of surrounding text can so quickly and easily alter how audiences perceive a central message. Understanding how context affects meaning is thus of profound interest in our polarized times. But it is difficult—to measure and model. This is especially true in politics, where our corpora may be small and our term counts low. Yet we simultaneously want statistical machinery which allows us to speak of statistically significant effects of covariates. This paper begins to address these problems.

Specifically, we proposed a flexible approach to study differences in semantics between groups and over time using high-quality pre-trained embeddings: the `conText` embedding regression model. We showed that it has advantages over previous efforts, and that it can reveal new things about politics. We explained how controversial terms divide parties not simply in their use rates or the way they are attached to topics of debate, but in their very meaning. Similarly, we showed that understandings of terms like "empire" are not fixed, even in the relatively short-term, and instead develop in-line with superpower positions in international relations. We then demonstrated that open-ended responses allow us to see the dividing lines between voters of different ideological stripes. All of our analyses can be implemented using the `conText` software package in `R` (see Supporting Information E).

[9] A related issue for finite populations is dependence among observations and the impact on the bootstrap (Booth, Butler and Hall, 1994; Mashreghi et al., 2016). We address this by resampling at the instance-level, but for some cases we may want to resample at the document or author level.

We built our framework on the ALC embedding strategy. But our general approach is not inextricably connected to this particular method for estimating contextually specific meanings. We used it because it is transparent, efficient, and computationally simple. We introduced a regression framework for understanding word meanings using individual instance embeddings as observations. This may be easily extended to more complex functional forms.

As social scientists develop further methods to study these problems, we expect that this will drive sharper questions which will in turn spur better methods. We hope that the `conText` model that we have laid out here can provide a useful foundation for future work.

References

Antoniak, Maria and David Mimno. 2018. "Evaluating the stability of embedding-based word similarities." *Transactions of the Association for Computational Linguistics* 6:107–119.

Arora, Sanjeev, Yuanzhi Li, Yingyu Liang, Tengyu Ma and Andrej Risteski. 2016. "A latent variable model approach to pmi-based word embeddings." *Transactions of the Association for Computational Linguistics* 4:385–399.

Arora, Sanjeev, Yuanzhi Li, Yingyu Liang, Tengyu Ma and Andrej Risteski. 2018. "Linear algebraic structure of word senses, with applications to polysemy." *Transactions of the Association for Computational Linguistics* 6:483–495.

Aslett, Kevin, Nora Webb Williams, Andreu Casas, Wesley Zuidema and John Wilkerson. 2020. "What Was the Problem in Parkland? Using Social Media to Measure the Effectiveness of Issue Frames." *Policy Studies Journal* .

Austin, John Langshaw. 1962. *How to do things with words*. Oxford: Oxford University Press.

Bai, Jushan and Pierre Perron. 1998. "Estimating and testing linear models with multiple structural changes." *Econometrica* pp. 47–78.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent and Christian Janvin. 2003. "A neural probabilistic language model." *The journal of machine learning research* 3:1137–1155.

Booth, James G, Ronald W Butler and Peter Hall. 1994. "Bootstrap methods for finite populations." *Journal of the American Statistical Association* 89(428):1282–1289.

Caliskan, Aylin, Joanna J Bryson and Arvind Narayanan. 2017. "Semantics derived automatically from language corpora contain human-like biases." *Science* 356(6334):183–186.

Chong, Dennis and James N Druckman. 2007. "Framing theory." *Annual Review of Political Science* 10:103–126.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2018. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* .

Di Carlo, Valerio, Federico Bianchi and Matteo Palmonari. 2019. Training temporal word embeddings with a compass. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33 pp. 6326–6334.

Faruqui, Manaal, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy and Noah A Smith. 2014. "Retrofitting word vectors to semantic lexicons." *arXiv preprint arXiv:1411.4166* .

Firth, John Rupert. 1957. *Studies in linguistic analysis*. Wiley-Blackwell.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky and James Zou. 2018. "Word embeddings quantify 100 years of gender and ethnic stereotypes." *Proceedings of the National Academy of Sciences* 115(16):E3635–E3644.

Geertz, Clifford. 1973. "Thick description: Toward an interpretive theory of culture." *Turning points in qualitative research: Tying knots in a handkerchief* 3:143–168.

Gentzkow, Matthew, Jesse M Shapiro and Matt Taddy. 2019. "Measuring group differences in high-dimensional choices: method and application to congressional speech." *Econometrica* 87(4):1307–1340.

Gentzkow, Matthew, J.M. Shapiro and Matt Taddy. 2018. "Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts.".
URL: *https://data.stanford.edu/congress_text*

Gonen, Hila, Ganesh Jawahar, Djamé Seddah and Yoav Goldberg. 2020. Simple, Interpretable and Stable Method for Detecting Words with Usage Change across Corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics pp. 538–555.
URL: *https://www.aclweb.org/anthology/2020.acl-main.51*

Grimmer, Justin. 2010. "A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases." *Political Analysis* 18(1):1–35.

Grimmer, Justin and Brandon M Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political analysis* 21(3):267–297.

Hamilton, William L, Jure Leskovec and Dan Jurafsky. 2016. "Diachronic word embeddings reveal statistical laws of semantic change." *arXiv preprint arXiv:1605.09096* .

Han, Rujun, Michael Gill, Arthur Spirling and Kyunghyun Cho. 2018. Conditional word embedding and hypothesis testing via bayes-by-backprop. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. pp. 4890–4895.

Hennessy, Peter. 1992. *Never Again: Britain 1945-1951*. Penguin UK.

Hinton, Geoffrey E et al. 1986. Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society*. Vol. 1 Amherst, MA p. 12.

Hobbs, William R. 2019. "Text Scaling for Open-Ended Survey Responses and Social Media Posts." *Available at SSRN 3044864* .

Hopkins, Daniel J. 2018. "The exaggerated life of death panels? The limited but real influence of elite rhetoric in the 2009–2010 health care debate." *Political Behavior* 40(3):681–709.

Khodak, Mikhail, Nikunj Saunshi, Yingyu Liang, Tengyu Ma, Brandon Stewart and Sanjeev Arora. 2018. "A la carte embedding: Cheap but effective induction of semantic feature vectors." *arXiv preprint arXiv:1805.05388* .

Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde and Slav Petrov. 2014*a*. Temporal Analysis of Language through Neural Language Models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. Baltimore, MD, USA: Association for Computational Linguistics pp. 61–65.
URL: *https://www.aclweb.org/anthology/W14-2517*

Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde and Slav Petrov. 2014*b*. "Temporal analysis of language through neural language models." *arXiv preprint arXiv:1405.3515* .

Kozlowski, Austin C, Matt Taddy and James A Evans. 2019. "The geometry of culture: Analyzing the meanings of class through word embeddings." *American Sociological Review* 84(5):905–949.

Krosnick, Jon A. 1999. "Survey research." *Annual review of psychology* 50(1):537–567.

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*. pp. 625–635.

Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski and Erik Velldal. 2018. "Diachronic word embeddings and semantic shifts: a survey." *arXiv preprint arXiv:1806.03537* .

Lauretig, Adam. 2019. Identification, Interpretability, and Bayesian Word Embeddings. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*. Minneapolis, Minnesota: Association for Computational Linguistics pp. 7–17.

Levy, Omer and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*. pp. 2177–2185.

Lucas, Christopher, Richard A Nielsen, Margaret E Roberts, Brandon M Stewart, Alex Storer and Dustin Tingley. 2015. "Computer-assisted text analysis for comparative politics." *Political Analysis* 23(2):254–277.

Mashreghi, Zeinab, David Haziza, Christian Léger et al. 2016. "A survey of bootstrap methods in finite population sampling." *Statistics Surveys* 10:1–52.

McGlone, Matthew S. 2005. "Contextomy: the art of quoting out of context." *Media, Culture & Society* 27(4):511–522.

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*. pp. 3111–3119.

Monroe, Burt L, Michael P Colaresi and Kevin M Quinn. 2008. "Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict." *Political Analysis* 16(4):372–403.

Mutz, Diana C. 2011. *Population-based survey experiments*. Princeton University Press.

Park, Baekkwan, Kevin Greene and Michael Colaresi. 2020. "Human rights are (increasingly) plural: Learning the changing taxonomy of human rights from large-scale text reveals information effects." *American Political Science Review* 114(3):888–910.

Pennington, Jeffrey, Richard Socher and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. pp. 1532–1543.

Quinn, Kevin M, Burt L Monroe, Michael Colaresi, Michael H Crespin and Dragomir R Radev. 2010. "How to analyze political attention with minimal assumptions and costs." *American Journal of Political Science* 54(1):209–228.

Rheault, Ludovic and Christopher Cochrane. 2019. "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora." *Political Analysis* pp. 1–22.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G Rand. 2014. "Structural topic models for open-ended survey responses." *American Journal of Political Science* 58(4):1064–1082.

Roberts, Margaret E, Brandon M Stewart and Edoardo M Airoldi. 2016. "A model of text for experimentation in the social sciences." *Journal of the American Statistical Association* 111(515):988–1003.

Rodman, Emma. 2019. "A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors." *Political Analysis* pp. 1–25.

Rodriguez, Pedro L and Arthur Spirling. 2021. Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research. Technical report Journal of Politics.

Rudolph, Maja, Francisco Ruiz, Susan Athey and David Blei. 2017. Structured embedding models for grouped data. In *Advances in Neural Information Processing Systems*. pp. 251–261.

Sanders, David and David Patrick Houghton. 2016. *Losing an empire, finding a role: British foreign policy since 1945*. Macmillan International Higher Education.

Skinner, Quentin. 1969. "Meaning and Understanding in the History of Ideas." *History and theory* 8(1):3–53.

Tversky, Amos and Daniel Kahneman. 1981. "The framing of decisions and the psychology of choice." *science* 211(4481):453–458.

Verba, Sidney and Gabriel Almond. 1963. *The civic culture: Political attitudes and democracy in five nations*. Princeton, NJ: Princeton University Press.

Wu, Patrick Y, Walter R Mebane Jr, Logan Woods, Joseph Klaver and Preston Due. 2019. "Partisan Associations of Twitter Users Based on Their Self-descriptions and Word Embeddings." http://www-personal.umich.edu/~wmebane/partisanassociations_wumebanewoodsklaverdue_apsa2019.pdf.

Yin, Zi, Vin Sachidananda and Balaji Prabhakar. 2018. "The global anchor method for quantifying linguistic shifts and domain adaptation." *arXiv preprint arXiv:1812.10382* .

Online Supporting Information:
Embedding Regression: Models for Context-Specific Description and Inference in Social Science

Contents (Appendix)

A Rodman: Details on Sample Sizes

A key challenge in Rodman's (2019) approach is that there is relatively little data (per time slice) to estimate embeddings from. Table 6 presents the number of instances of each theme word for each period. Note that in almost 30% of the word-era combinations, there are fewer than 10 observations. Producing meaningful embeddings given these sample sizes is generally difficult.

	1855–	1880–	1905–	1930–	1955–	1980–	2005–
`african_american`	63	27	79	171	274	45	22
`gender`	4	41	374	560	460	258	284
`german`	1	2	62	512	13	2	2
`race`	5	15	76	188	190	34	38
`treaty`	3	1	143	216	30	3	1
Total Documents	80	102	496	1137	660	259	371

Table 6: Number of instances of each category word in the Rodman corpus by 25 year time slice. All documents have the word `equality`. Many of the counts are quite low leading to a serious challenge for word embeddings.

B The Presidential Transition in Meaning

The meaning of `Trump`, the surname, underwent a significant transformation once Donald J. Trump was elected president of the United States in November 2016. This is a substantially harder case since the person being referred to is still the same, even though the meaning has shifted.

Using ALC, we embed a random sample of 100 mentions of `Trump` from 2001–2014 and 2017–2020, which we label celebrity `Trump` and president `Trump`, respectively. We do the same two cluster routine as above and inspect the 10 nearest neighbors—these are given in Table 7. As we would expect, `Trump` in 2001–2014 is mentioned in the context of casinos and

real-estate terms while `Trump` in 2017–2020 is mentioned in the context of terms associated with his presidency.

celebrity Trump	trump, ivanka, ivana, wynn, donald, casino, casinos, resorts, taj, caesars
President Trump	president, assailing, clinton, bush, impeach, impeachment, presidential, impeached, appointee

Table 7: Top 10 nearest neighbors of the transformed cluster centroids.

In Figure 9 label the mentions of celebrity `Trump` and president `Trump`, respectively (results projected down to two-dimensions for visualization purposes). While the two groups overlap, as would be expected given mentions are all of the same person, it is clear mentions of `Trump` tend to cluster by period.

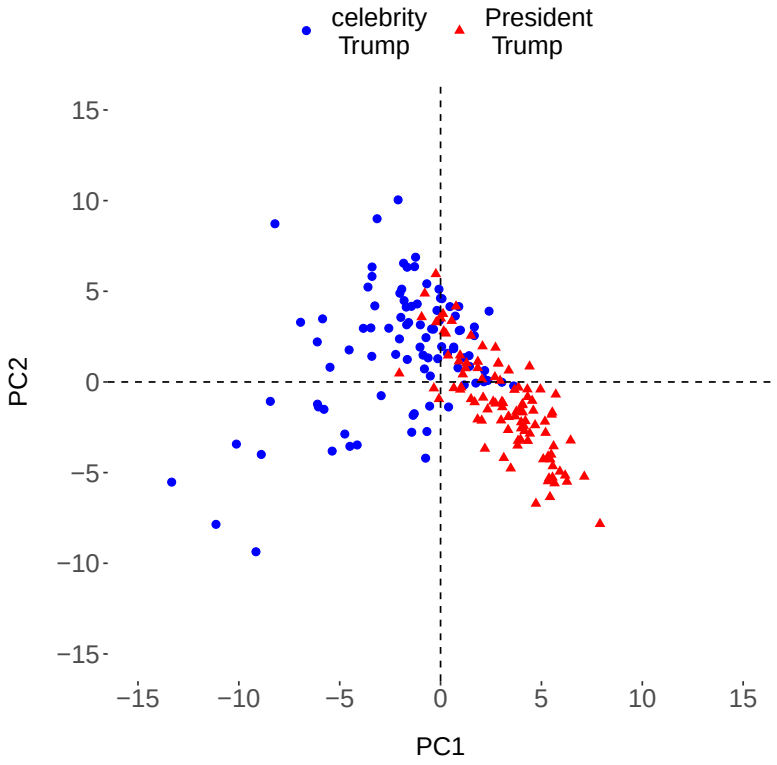
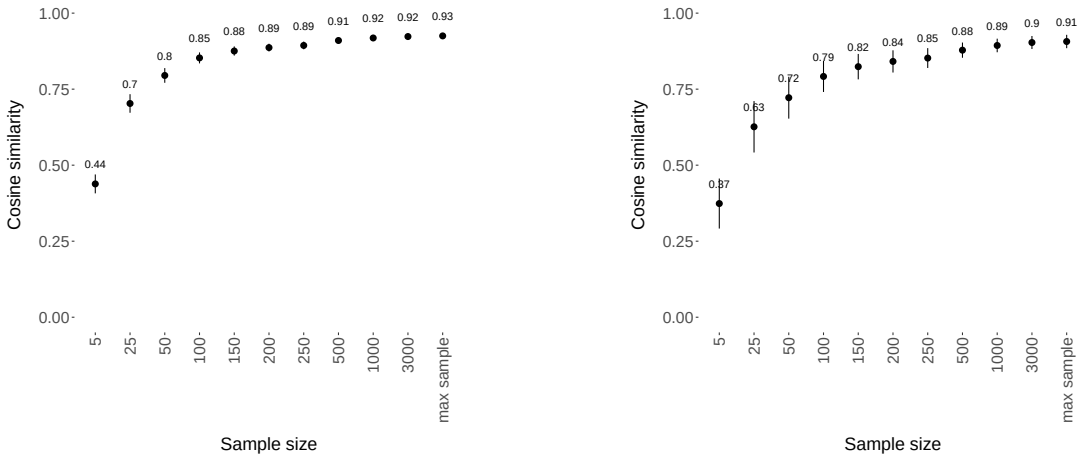


Figure 9: Each observation represents a single realization of a context. Contexts for celebrity `Trump` include mentions of `Trump` in the New York Times during the period 2001-2014, while contexts for President `Trump` include mentions of `Trump` in the New York Times during the period 2017-2020.

C Asymptotic Behavior

In this exercise we evaluate the asymptotic performance of our approach. That is, we want to know whether—and how quickly—ALC embeddings converge to embeddings from a fully trained, full corpus `GloVe` model, as we increase the number of instances ALC has access to. Obviously, we would hope that as the sample approaches the whole corpus, ALC 'looks like' a full corpus model.

For our corpus we use the *Congressional Record*. We begin by estimating a full `GloVe` embeddings model and a corresponding transformation matrix A. Next we select a set of 20 target words from the corpus vocabulary, including 10 politics terms and 10 randomly sampled terms, and estimate their corresponding ALC embeddings. We vary the number of instances, from 5 to the total number of instances of each term.[10] Finally, we compute the cosine similarity between each ALC embedding and its corresponding embedding in the full `GloVe` model. Figure 10 plots the results separately for the politics and random set of terms. We see that for both sets the ALC embeddings quickly converge to within a margin of error of the `GloVe` embeddings as the number of instances used to estimate the ALC embedding increases. This is expected and welcome behavior. In the case of the politics terms, with as few as fifty instances we see an average cosine similarity value of 0.8.[11]



(a) Mean over 10 politics terms

(b) Mean over 10 random terms

Figure 10: Cosine similarity between a full `GloVe` (full corpus) embeddings model and ALC as a function of sample size.

[10]The set of politics terms are: `democracy`, `freedom`, `equality`, `justice`, `immigration`, `abortion`, `welfare`, `taxes`, `republican` and `democrat`. The set of random terms are: `adopt`, `appreciate`, `deserve`, `governments`, `however`, `insert`, `proposals`, `reduceds`, `temporary` and `thus`.

[11]Note, we do not expect this value to converge fully to 1 as the transformation matrix A is itself a regression estimate.

D Benchmarking Embedding Regression against group-specific 'full' embeddings

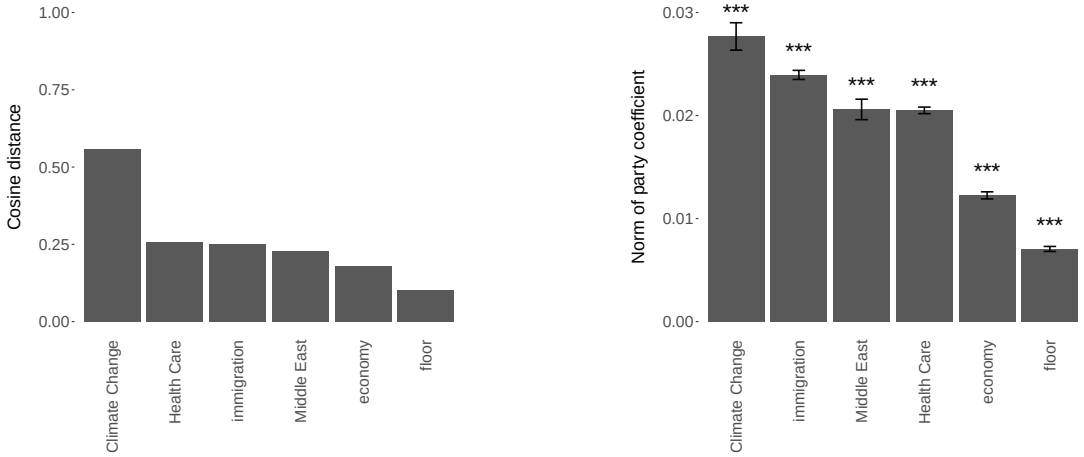
An alternative to our *regression* approach to quantifying group differences is to estimate a full `GloVe` embeddings model for each group's use of a term. For any given word this can be done by tagging (literally, slightly altering) the word in the corpus such that it appears differently for each different group. Estimating a full `GloVe` model on this tagged corpus yields group-specific embeddings for the tagged words. We can then use these embeddings to quantify group differences. This is computationally costly but provides us with a straightforward benchmark for our approach. Specifically we are interested in comparing inferences when applying both approaches to the following task: ranking a set of terms according to partisanship (in use).

For this exercise we use the Congressional Record corpus, sessions 111^{th} - 114^{th} (the Obama years). As target words we use: `immigration`, `economy`, `climatechange`, `healthcare`, `middleeast` and, as a non-political control word we use `floor`.[12]

We tag every instance of a target word in the corpus with the party of its corresponding speaker, so for example, given a particular instance of `immigration` in a speech, we replace it with `immigrationd` if the author of the speech is a Democrat and with `immigrationr` if the author is a Republican. Given party specific embeddings for each target word we quantify partisanship using cosine distance, the higher the cosine distance, the more partisan the term. To quantify partisanship using our preferred approach we simply run a regression with party as a covariate and compute the norm of the resulting coefficient, the higher the norm of the party coefficient, the more partisan the term.

Figure 11 plots both sets of results. Broadly speaking, the inferences one would draw from each are similar. On the one hand, `Climate Change` is clearly the most partisan issue while, as expected, our control term `floor` is the least partisan according to both models. `economy` stands out as the second least partisan according to both models. The remaining terms are similarly ranked except our approach suggests `immigration` is somewhat more partisan than `Health Care` and `Middle East`. All in all, the inferences from both approaches are not wildly different. In contrast to estimating a full `GloVe` embeddings model however, our approach is much faster, more stable—the solution does not vary across runs—and allows us to speak to the significance and sampling variance of our estimates.

[12] In the corpus we replace any mentions of `middle east` with `middleeast`, `health care` with `healthcare`, `immigrants` and `immigrant` with `immigration` and `climate change` with `climate change`.

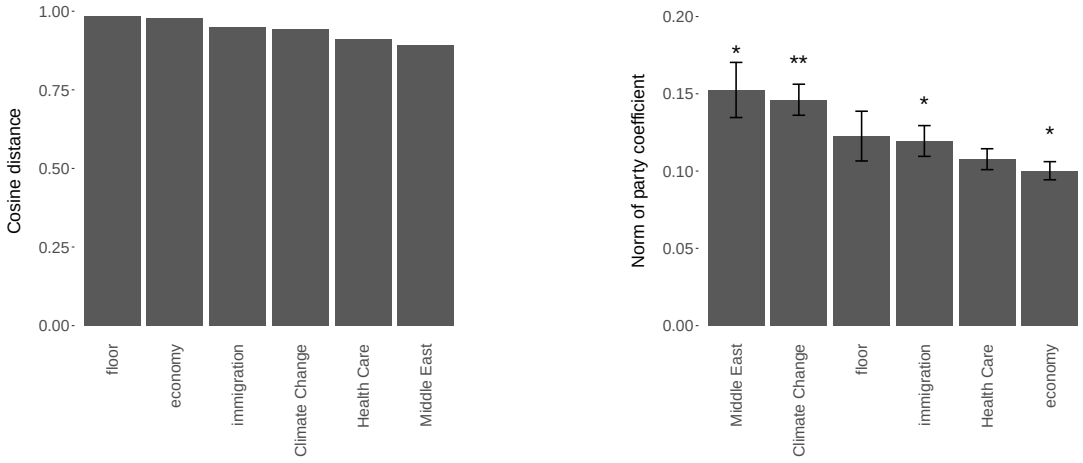


(a) `GloVe` 'regression' (b) `conText` regression

Figure 11: Partisan differences using the Congressional Record corpus (Sessions 111$^{\text{th}}$ - 114$^{\text{th}}$).

Next we compare each model's performance with a significantly reduced sample, specifically one in which each target word appears in no more than five documents.[13] Our goal with this exercise is to compare how both methods fare in a small-sample world, relative to inferences using the full corpus. Figure 12 plots both sets of results. In the case of the full GloVe model we see results are now flipped, with `floor` and `economy` showing the largest partisan differences. In contrast, the ALC results are comparable to the full-sample case. While `floor` shows a larger norm, it is not significant, and `Climate Change` remains the most significantly partisan of the target words. Combined, these results serve to highlight the added value of our approach, yielding similar inferences as the full embeddings model at a fraction of the cost and more robust in small-sample scenarios.

[13]To build this corpus we identify for each target word all documents containing the word and randomly sample five of these. We exclude from this sample any document containing multiple target words. Documents that do not contain any of the target words remain part of the corpus.



(a) GloVe

(b) ALC

Figure 12: Partisan differences using the Congressional Record corpus (Sessions 111$^{\text{th}}$ - 114$^{\text{th}}$), including only 5 instances of each target word.

E Software

To facilitate applying the methods presented in this paper we put together an R package – `conText`. The main function `conText` follows generic `R lm( )` and `glm( )` syntax in terms of $\sim$ operator. Please refer to the quick start guide to get started using the package. As with any package, we had to make a couple of design decisions that are worth noting here. First, ALC embeddings are computed using the available pre-trained context word embeddings. If a given context word is not available in the provide pre-trained embeddings, then that context word is simply ignored and the average is taken over the set of available context embeddings. Second, we've found that in practice limiting the candidate nearest neighbors to the set of words in the provided contexts, significantly reduces noise (non-sensical nearest neighbors such as misspelled words etc.). Whenever exploring nearest neighbors you can use the parameter `candidates` to delimit the set of nearest neighbors. Finally, we have made available—or simply more accessible—the GloVe pre-trained embeddings used in most of the examples in this paper along with their corresponding transformation matrix.[14] We are often asked when is it appropriate to use these pre-trained embeddings and their corresponding transformation matrix rather than estimate ones own. Unfortunately, there is no hard-and-fast rule for this, it comes down to how distinct you think your corpus is relative to the corpus used to train these embeedings (Wikipedia 2014 and Gigaword 5).

[14]The original GloVe embeddings computed by the Stanford NLP Group can be found here while the original transformation matrix computed by Khodak et al. (2018) can be found here.