# Evaluating the Coherence of Diachronic Word Embedding Similarities

Abdul Z. Abdulrahim

May 6, 2019

## Contents

# 1 Introduction

In this paper, we evaluate the coherence of diachronic word embeddings/representations from a linguistic and legal perspective. Over the years, linguists have studied the evolution of the meaning of words describing lexical semantic shifts or semantic change (Bloomfield, 1933). These lexical shifts have wider importance in society beyond linguistics, and we attempt to show how it can be applied to track semantic shifts in our laws and their interpretation — particularly with *privacy.*

The notion of a *private life* has drastically changed over the last two decades as social, economic and technological advancements continue to change we way we interact with one another. In the United States (US), this has resulted in calls for better protection of the individual and group privacy. However, the development of privacy laws in the US has predominantly been piecemeal due to its origination from tort law — and it has not been systematically reviewed or legislated as in Europe. For this reason, we investigate the judiciary's approach to interpreting privacy in corpora created from judicial opinions written by the justices of the US Supreme Court.

Using multiple approaches to generate word representations, we try to hone in on a coherent interpretation of *privacy* from the corpora with the intention of better understanding how the notion may have changed at different points in time. By conducting a genealogical account using computational linguistic methods, we introduce new insights into how to view the evolution of privacy law in the US. More importantly, we note that the slow shift in interpretation evidenced by the results shows that the US has made slower progress than Europe in its privacy case law. However, this is cautioned by the observation that the stability and coherence of the word embeddings are affected by numerous factors, e.g. the algorithm used, presence of certain opinions and much more (Antoniak & Mimno, 2018).

# 2 Literature Overview

## 2.1 A Short History of Diachronic Word Embeddings

Over time, the meanings of words continuously change reflecting complex and dynamic processes in language and society. A frequent example quoted in the diachronic word embedding literature is the change to the core meaning of words such as *gay* which has arguably changed from *carefree* to *homosexual* during the $20^{th}$ century. As such, by studying these changes we can learn more about human language, and when isolated to a smaller group generating such corpora, the way they interpret or understand these words. Given the improvement in computational techniques and availability of text data, we have seen attempts to capture these diachronic semantic shifts in a data-driven way (Kutuzov, Øvrelid, Szymanski, & Velldal, 2018). Figure 1 shows a timeline of events that have conceivably influenced the trajectory of research on this topic.

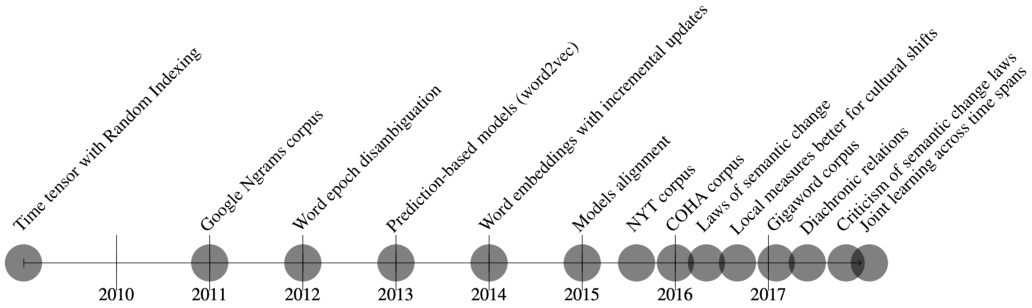
Time tensor with Random Indexing
Google Ngrams corpus
Word epoch disambiguation
Prediction-based models (word2vec)
Word embeddings with incremental updates
Models alignment
NYT corpus
COHA corpus
Laws of semantic change
Local measures better for cultural shifts
Gigaword corpus
Diachronic relations
Criticism of semantic change laws
Joint learning across time spans
2010
2011
2012
2013
2014
2015
2016
2017

Figure 1: Research timeline for distributional models in the task of tracing diachronic semantic shifts (retrieved from Kutuzov et al., 2018)

Historically, much of the theoretical work on semantic shifts had been focused on documenting and categorising types of semantic shifts — building the foundation for much of the work we see today (Stern, 1931; Bloomfield, 1933). More recently, semantic shifts are generally separated into linguistic drifts, i.e., slow and regular changes in core meaning of words; and cultural shifts, i.e. culturally determined changes in associations of a given word (Hamilton, Leskovec, & Jurafsky, 2016; Kutuzov et al., 2018). A key, and perhaps intuitive, assumption worth noting in the study of semantic shifts is that changes in a word's collocational patterns (i.e., words that co-occur together) reflect changes in word's meaning, in turn providing a usage-based account of semantics (Gries, 1999; Heine, 2009).

## 2.2 Prevalent Methods for Tracking Semantic Shifts

A common computational method used for measuring these shifts is presented in Mikolov, Sutskever, Chen, Corrado, and Dean (2013), but various works have been published on the topic with very different implementations and methods. One such approach was using the change in raw word frequencies to trace semantic shifts or linguistic change (see, inter alia, (Heine, 2009; Michel et al., 2011; Choi & Varian, 2012; Heyer, Holz, & Teresniak, 2009)).

However, semantic shifts are not always linked to the changes in word frequency, so recent publications have shown that distributional word representations perhaps provide better insights (Turney & Pantel, 2010; Baroni, Dinu, & Kruszewski, 2014). By representing the corpus with sparse or dense vectors produced from word co-occurrence counts — compressing the word counts into continuous lexical representations — we get a more efficient and convenient to algorithm (Sagi, Kaufmann, & Clark, 2011; Kulkarni, Al-Rfou, Perozzi, & Skiena, 2015).

Kim, Chiu, Hanaki, Hegde, and Petrov (2014) also introduced an innovative approach which employed prediction-based word embedding models. They trace diachronic semantic shifts using incremental updates and continuous skipgram with negative sampling (Kutuzov et al., 2018). Subsequently, Hamilton et al. (2016) evaluated the skipgram with negative sampling against the pointwise mutual information models to show the superior performance of the skipgram model — although noting that low-rank SVD approximations can perform on par on smaller datasets. The majority of publications now seen in the field use dense word representations, either in the form of SVD-factorized matrices or prediction-based embedding models (Kutuzov et al., 2018). However, it appears preference is given to prediction-based embedding algorithms (Pennington, Socher, & Manning, 2014).

## 2.3 The Alignment Problem

A common issue faced in diachronic word-embeddings is the comparison of word vectors across different time-specific models. As most modern word embedding algorithms are inherently stochastic and the resulting embedding sets are invariant under rotation, it does not make sense to compute cosine similarities of embeddings from different time-periods directly (Kutuzov et al., 2018). Given that separate learning runs may produce entirely different vectors with roughly similar pairwise similarities, even if the meaning of a word stays the same, the direct cosine similarity between its vectors from different periods can still be quite low due to the random initialisation of the two models (Kutuzov et al., 2018; Yao, Sun, Ding, Rao, & Xiong, 2018).

Many approaches to solve this issue have been proposed but to summarise they include the following. Kulkarni et al. (2015) suggests aligning the models before calculating the similarities using linear transformations preserving general vector space structure. Separately, Zhang, Jatowt, Bhowmick, and Tanaka (2015) suggests using a distance-preserving projection technique, while Eger and Mehler (2017) compared word meanings using ‘second-order embeddings’, i.e., the vectors of words’ similarities to all other words in the shared vocabulary of all models. More recently, work published by Hamilton et al. (2016) employed both ‘second-order embeddings’ and orthogonal Procrustes transformations to align diachronic models; Bamler and Mandt (2017) and Yao et al. (2018) use a different approach to learn the word embeddings across several time periods jointly, enforcing alignment across all of them simultaneously, and positioning all the models in the same vector space (Kutuzov et al., 2018).

## 2.4 Further Domain Considerations

In choosing a method to track (and align) the meaning of words in our legal corpora, we must consider the linguistic and non-linguistic factors that affect the change in the language as the driving forces of semantic shifts are varied (Blank & Koch, 1999). Therefore, in deciding on a method that represents the change we want to study, we must consider how it ties in with the type or driver of change we want to focus on — particularly in this case, whether they can tell us anything about the socio-cultural shifts in the courts.

The judicial opinions released by the courts on hearing a case serve as justification that explains the logical steps that were taken in deciding a case. While this content is significant in many ways, it is also similar to the text produced by the wider society. On the one hand, we may expect that laws derived for larger corpora by Dubossarsky, Weinshall, and Grossman (2016), on the *law of prototypicality*, Eger and Mehler (2017) on the *law of differentiation* and Hamilton et al. (2016) on the *law of conformity* to apply; but on the other, we want to understand the interpretation of our target words from a legal context. Subsequently, rather than using traditional alignment methods for diachronic representations, we opt for an option that mimics the way judiciary interprets the law.

Existing approaches to the alignment issue have been innovative in their implementations, but they do not capture an essential characteristic of our corpora, which is the doctrine of precedent. That is, in most cases, judges are bound by the preceding cases that have been decided when applying the law. On occasion, they depart from this, when there is overwhelming evidence or reason to change the law in order to reflect society and its values better. Hence, when writing their opinions on a judgement, justices predominantly stick to the meaning and interpretation set by preceding cases.

So rather than training our embeddings on time slices that are removed from one another, a better representation for our corpora is a cumulative slice that replicates the body of knowledge with which a judge is bound at the time of writing. This, in turn, should highlight what the interpretations and meanings of words have been up to a time period, and the bounds within which a judge is held in the following time period. The expectation is that this will give us a broader picture of the evolution of our laws and how judges understand the priorities of society and its laws.

Furthermore, some alignment techniques may not accurately represent the absence of change in our corpora. For instance, in some periods, we may not have any opinions that contain some of our target words because no case relating to that topic was decided. Aligning the vectors using the approach proposed by Hamilton et al. (2016) and Kulkarni et al. (2015) would drop off some of our words because it has no vector to align with. Similarly, when using the approach from Bamler and Mandt (2017) or Yao et al. (2018), in periods where our target words are absent, we are not able to detect that the absence of case law within that period means the interpretation of these words stay the same. Instead, it converges the embeddings from the previous time-period where our target word appears, and the next time-period it appears.

# 3 Corpora

## 3.1 Source

The Caselaw Access Project (CAP) is a Harvard Law School Library Innovation Lab Project that has collated all official, book-published United States case law — every volume designated as an official report of decisions by a court within the United States (Harvard Law Library, n.d.). It includes all state courts, federal courts, and territorial courts and dates back to 1658. This paper uses the opinions issued by the US Supreme Court contained in the database.

## 3.2 Preparation

In preparing the data for our algorithms, the following data wrangling and pre-processing steps were taken: (i) extracting and parsing through court documents to prepare a dataframe of opinions that includes majority, dissenting and other opinions authored by the judges of the Supreme Court; (ii) cleaning and regularising the text of opinions, e.g. lowercasing words, removing punctuation and capturing case or legislation citations; (iii) tokenising by splitting the text into sentences and the sentences into words; (iv) removing stop words; and (v) limiting our embeddings to words that occur at least 100 times at each time step (Wendlandt, Kummerfeld, & Mihalcea, 2018).

Some of the CAP opinions in our dataset, particularly the older text, are extracted using optical character recognition, hence there are inevitable minor errors that arise. We ignore these as we intend to use as minimal pre-processing as possible, so as not to lose any useful information or trigger potential algorithm sensitivities i.e. some implementations behave differently to some data pre-processing (Bojanowski, Grave, Joulin, & Mikolov, 2017). Consequently, our results demonstrate how various algorithms for word representation differ and their sensitivities given minimal data pre-processing.

### 3.3 Descriptive Characteristics

Overall, the CAP dataset contains 328,310 cases from the Supreme Court, i.e. all appeals granted and denied. A majority of these case entries are appeals to the Supreme Court which are denied, identifiable by the entry: '*court of appeal nth circuit certiorari denied*'. To enable faster computation and to focus on the opinions with material information, we filter out entries with less than 1,000 characters — bringing our case total down to 28,879. Some key characteristics of the dataset following its pre-processing are provided in table 1. Further characteristics relating to word frequencies and occurrence of Zipf's law is provided in the Appendix.

| | |
|---|---:|
| No. of opinions used: | 28,879 |
| No. of judges/opinion authors: | 112 |
| Timespan: | 1 January 1791 - 3 April 2018 |
| Average length of opinions used: | 3,573 |
| No. of unique words in opinions used: | 683,156 |
| No. of unique words with more than 100 occurrences: | 21,197 |

Table 1: Key characteristics of dataset, including the number of opinions, unique words (before and after removing words that appear fewer than 100 times), and the average number of words per opinion

# 4 Methods

In tracking the semantic shifts, we evaluate four different implementations of the word representations (described below). Conceptually, we formulate the task of discovering semantic shifts as follows. Given a time sorted corpora: [corpus 1, corpus 2, ...corpus n], we locate our target word *privacy* and its meanings in the different time periods. To evaluate the meanings, we search for the words with closest meanings to our target word by their cosine similarity. Cosine similarity is a measure of word similarity calculated by taking the cosine of the angle between two n-dimensional word vectors in an n-dimensional space, i.e. the dot product of the two word vectors divided by the product of the two word vectors' lengths (or magnitudes). The stability of the representations are evaluated by their statistical significance bounds, and the coherence is assessed by how the meaning tracks with the understanding of the laws at the relevant time period.

## 4.1 Algorithms

To establish statistical significance bounds for our representations, we train each model 50 times for each time period's corpus with a matrix dimension = 100, window = 5 and a minimum word count = 100. We calculate the cosine similarity of our target word '*privacy*' to the other words in the vocabulary, creating a similarity ranking of all the words in the vocabulary. The mean and standard deviation of the cosine similarities are calculated for the target word and vocabulary word across each set of 50 models, and we examine these metrics across different algorithms holding the corpus parameters constant.

### 4.1.1 Continuous Bag of Words (CBOW)

CBOW predicts a current word based on its context, i.e. based on the neighbouring words in a specified window. In training a CBOW model, three layers are used; an input layer to represent the context of the word, a hidden layer that represents the projection of each word from the input layer; and an output layer in which the word matrix from the hidden layer is projected. During training, the output is compared to the word itself to correct its representation based on the back propagation of the error gradient. Thus, the purpose of CBOW embedding model network is to maximise the following:

$$L_\theta = \frac{1}{T}\sum_{t=1}^{T} \log p(w_t \mid w_{t-n}, \cdots, w_{t-1}, w_{t+1}, \cdots, w_{t+n}),$$

where the model receives a window of $n$ words around the target word $w_t$ at each time step $t$. CBOW is similar to the feedforward neural net language model, where the non-linear hidden layer is removed, and the projection layer is shared for all words; thus, all words get projected into the same position (their vectors are averaged). In this approach, the order of words does not influence the projection, and it uses a continuous distributed representation of the context (Mikolov, Chen, Corrado, & Dean, 2013). We use an implementation of the CBOW algorithm included in the Python library *gensim* using the parameters specified in section 4.1.

### 4.1.2 Skip-Gram Negative Sampling (SGNS)

SGNS is a similar architecture to CBOW, but instead of predicting the current word based on the context, it tries to maximise classification of a word based on another word in the same sentence (Mikolov, Chen, et al., 2013). The input layer represents the target word, and the output layer corresponds to the context. As such, SGNS predicts the context given a word instead of predicting a word given its context like CBOW. Like CBOW, SGNS compares the output and each word in the context to correct its representation based on the back propagation of the error gradient, thus maximising:

$$L_\theta = \frac{1}{T}\sum_{t=1}^{T} \sum_{-n \leq j \leq n, \neq 0} \log p(w_{t+j} \mid w_t),$$

by summing the log probabilities of the surrounding words to the left and the right of the target word window. As with CBOW, we use an implementation from *gensim* with the parameters noted in section 4.1.

### 4.1.3 Latent Semantic Analysis (LSA)

LSA or Singular Value Decomposition (SVD), is among the more popular method for dimensionality reduction. In the LSA implementation, the word-context co-occurrence matrix is factorised into the product of three matrices $U \cdot \Sigma \times V^T$ where $U$ and $V$ are orthonormal matrices (i.e., square matrices whose rows and columns are orthogonal unit vectors) and $\Sigma$ is a diagonal matrix of eigenvalues in decreasing order (Ruder, n.d.).

The elements of the term-document matrix are weighted with TF-IDF, which measures the importance of a word to a document in a corpus by its frequency. The dense, low-rank approximation of the term-document matrix can be used to measure the relatedness of terms by calculating the cosine similarity of the relevant rows of the reduced matrix (Antoniak & Mimno, 2018). We use the *gensim* latent semantic indexing module to train our LSA model and the (dense) left singular vectors are used as the final word embeddings.

### 4.1.4 Global Vectors for Word Representation (GloVe)

GloVe is a popular method for learning word representations proposed by Pennington et al. (2014). It uses stochastic gradient updates but operates on a 'global' representation of word co-occurrence that is calculated once at the beginning of the algorithm (Pennington et al., 2014). In contrast to SGNS/CBOW, GloVe explicitly encodes meaning as vector offsets in an embedding space.

Pennington et al. (2014) illustrate that the ratio of the co-occurrence probabilities of two words is important, so encode this information as vector differences. To accomplish this, they use a weighted least squares objective $L$ that directly aims to reduce the difference between the dot product of the vectors of two words and the logarithm of their number of co-occurrences (Ruder, n.d.):

$$L = \sum_{i,j=1}^{V} f(X_{ij})\,(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2,$$

where $w_i$ and $b_i$ is the word vector and bias for word $i$; $\tilde{w}_j$, $b_j$ are the context word vector and bias respectively of word $j$; and $f(X_{ij})$ is a weighting function for the number of times word $i$ occurs in the context of word $j$. The implementation has two main steps: (i) the construction of a co-occurrence matrix $X$ from the corpus; and (ii) the factorisation of $X$ to get vectors. We apply the GloVe implementation provided in Python[1] with the default settings suggested in Pennington et al. (2014) except as noted above in section 4.1.

## 5 Results

To evaluate the results of each implementation, we consider the coherence of the top-ranked words (by similarity) to see if these correspond to landmark decisions that have been made

[1] https://github.com/maciejkula/glove-python

in relation to *privacy* over the last five decades; and we look at the stability of the representations as this gives us statistical confidence in the rankings of the nearest neighbours.

## 5.1 Embedding Coherence

The essence of privacy law derives from a right to privacy, defined broadly as ‘the right to be let alone’. It usually does not give protection to personal issues in the public interest, e.g., the actions of politicians or celebrities. By invading one’s right to privacy, it presents a basis for a lawsuit supported by the constitutional rights under the Fourth Amendment’s right to be free of unwarranted search or seizure, the First Amendment’s right to free assembly, and the Fourteenth Amendment’s due process right — which have been recognised by the Supreme Court as protecting rights to privacy within family, marriage, motherhood, procreation, and child-rearing.

Over the last five decades, the major shift we have seen has been a movement from association with bodily rights to information rights and personal data. The cases of *Griswold v. Connecticut* and *Roe v. Wade* established the connections of *privacy* to bodily rights in the 1970s, but over time cases such as *Katz v. United States* have become more relevant. Although this case was decided in 1967, it brought to head the issue of informational privacy and the legal definition of a ‘search’ of intangible property, such as electronic-based communications like telephone calls. In the era of big data, informational privacy and personal data cases are becoming more prevalent — though the legal discussion around this started in the late 1990s and early 2000s as the risks of technological surveillance and espionage became top priorities for nations (Ohm, 2009; Schneier, 2015).

### 5.1.1 CBOW Word Similarity Rankings

Given the trends noted, we assess the averaged closest words for each algorithm. These appear to pick up some, but not all, of these trends — albeit in different ways. First, table 2 shows the results from our CBOW implementation. It appears to rightly pick up the relationship between *privacy* and the *4th Amendment.* Furthermore, it seems to identify the decreasing importance of the Fourth Amendment in the adjudication of cases relating to personal data as it is still unclear whether it affords any protections for the informational privacy of individuals (Pekgözlü & Öktem, 2012). Another interesting trend is that the word *anonymity* starts to go up the ranks from 1998 onwards, perhaps confirming the increase in informational privacy and personal data cases.

An odd trend from this implementation is the increasing rank of *home-owner.* Although beyond the scope of this paper, further analysis could look at the specific occurrences in the opinions to identify if it relates to traditional privacy or informational privacy cases. Beyond that, the algorithm does what we would expect, i.e. it identifies the contexts within which *privacy* may be used. So words such as *intrusions*, *invasions* and *sanctity* are interesting for defining the concept, but do not show any useful insights into the evolution of the notion.

### 5.1.2 SGNS Word Similarity Rankings

Next, we evaluate the coherence of the SGNS results in table 2. Unsurprisingly, these are quite similar to the results from CBOW. What is interesting however is that this algorithm picks up case law relevant to *privacy.* As noted earlier, the case of *Griswold v. Connecticut* established the connection of *privacy* to bodily rights in the 1970s, which is seen in our results for 1978, but over time, the case of *Katz v. United States* appears more frequently highlighting the importance of informational privacy cases. Contrary to the CBOW algorithm, we do not see much variation in the ranking for *4th Amendment.* The SGNS algorithm also has slightly different contextual words from CBOW such as *expectation* and *search*, and it does not give the same rising ranking to *home-owner*; instead we see it decline after 1998.

### 5.1.3 LSA Word Similarity Rankings

The results from our LSA implementation in table 2 are significantly different from our other algorithms, save for the fact that it still links *privacy* and the *4th amendment.* However, given the count-based architecture of the algorithm is normalised by TF-IDF, it does not come as a surprise as it is likely to give disproportionate importance to the infrequent words across the corpora. As a result, we also observe lower cosine similarity values in comparison to the other implementations.

Table 2: The 10 closest words to the target word privacy averaged and ranked by cosine similarity for each algorithm.

**CBOW**

| 1978 | | 1988 | | 1998 | | 2008 | | 2018 | |
|---|---|---|---|---|---|---|---|---|---|
| *intrusions* | 0.76 | *intrusions* | 0.69 | *intrusions* | 0.69 | *intrusions* | 0.68 | *intrusions* | 0.67 |
| *intrusion* | 0.68 | *intrusion* | 0.67 | *sanctity* | 0.66 | *sanctity* | 0.66 | *sanctity* | 0.66 |
| *invasions* | 0.67 | *sanctity* | 0.65 | *intrusion* | 0.64 | *intrusion* | 0.64 | *intrusion* | 0.62 |
| *4th amendment* | 0.64 | *4th amendment* | 0.63 | *home-owner* | 0.63 | *home-owner* | 0.62 | *home-owner* | 0.61 |
| *associational* | 0.64 | *invasions* | 0.62 | *4th amendment* | 0.62 | *invasions* | 0.61 | *invasions* | 0.60 |
| *invasion* | 0.63 | *invasion* | 0.57 | *invasions* | 0.61 | *4th amendment* | 0.61 | *4th amendment* | 0.60 |
| *confidentiality* | 0.62 | *searches* | 0.57 | *invasion* | 0.58 | *invasion* | 0.58 | *invasion* | 0.58 |
| *sanctity* | 0.62 | *inviolability* | 0.55 | *searches* | 0.57 | *anonymity* | 0.57 | *anonymity* | 0.57 |
| *freedoms* | 0.59 | *associational* | 0.54 | *inviolability* | 0.56 | *searches* | 0.56 | *inviolability* | 0.57 |
| *searches* | 0.58 | *freedoms* | 0.54 | *anonymity* | 0.55 | *inviolability* | 0.55 | *searches* | 0.56 |

**SGNS**

| 1978 | | 1988 | | 1998 | | 2008 | | 2018 | |
|---|---|---|---|---|---|---|---|---|---|
| *intrusions* | 0.78 | *intrusion* | 0.75 | *intrusion* | 0.74 | *intrusions* | 0.74 | *intrusions* | 0.67 |
| *intrusion* | 0.74 | *intrusions* | 0.75 | *intrusions* | 0.74 | *intrusion* | 0.74 | *intrusion* | 0.66 |
| *4th amendment* | 0.70 | *4th amendment* | 0.69 | *4th amendment* | 0.68 | *4th amendment* | 0.67 | *4th amendment* | 0.62 |
| *searches* | 0.65 | *searches* | 0.64 | *searches* | 0.64 | *expectation* | 0.64 | *expectation* | 0.61 |
| *invasions* | 0.64 | *invasions* | 0.63 | *expectation* | 0.63 | *searches* | 0.63 | *searches* | 0.60 |
| *invasion* | 0.62 | *katz v. us* | 0.63 | *home-owner* | 0.63 | *invasion* | 0.61 | *invasion* | 0.60 |
| *griswold v. conn.* | 0.61 | *expectations* | 0.62 | *curtilage* | 0.62 | *katz v. us* | 0.61 | *katz v. us* | 0.58 |
| *freedoms* | 0.60 | *expectation* | 0.62 | *katz v. us* | 0.62 | *home-owner* | 0.61 | *curtilage* | 0.57 |
| *search* | 0.59 | *search* | 0.60 | *invasions* | 0.62 | *invasions* | 0.61 | *home-owner* | 0.57 |
| *eavesdropping* | 0.59 | *invasion* | 0.59 | *invasion* | 0.61 | *curtilage* | 0.60 | *expectation* | 0.56 |

**LSA**

| 1978 | | 1988 | | 1998 | | 2008 | | 2018 | |
|---|---|---|---|---|---|---|---|---|---|
| *search* | 0.29 | *search* | 0.39 | *search* | 0.41 | *search* | 0.42 | *search* | 0.41 |
| *appellant* | 0.27 | *warrant* | 0.25 | *warrant* | 0.26 | *warrant* | 0.27 | *warrant* | 0.27 |
| *materials* | 0.24 | *appellant* | 0.22 | *4th amendment* | 0.23 | *4th amendment* | 0.23 | *4th amendment* | 0.22 |
| *president* | 0.20 | *4th amendment* | 0.21 | *appellant* | 0.18 | *appellant* | 0.16 | *information* | 0.18 |
| *warrant* | 0.18 | *materials* | 0.15 | *information* | 0.15 | *information* | 0.16 | *appellant* | 0.16 |
| *papers* | 0.17 | *president* | 0.14 | *materials* | 0.13 | *materials* | 0.12 | *probable* | 0.12 |
| *4th amendment* | 0.16 | *papers* | 0.12 | *president* | 0.12 | *probable* | 0.12 | *materials* | 0.11 |
| *presidential* | 0.16 | *seizure* | 0.12 | *probable* | 0.12 | *president* | 0.11 | *seizure* | 0.11 |
| *seizure* | 0.13 | *information* | 0.11 | *seizure* | 0.11 | *searches* | 0.11 | *searches* | 0.11 |
| *privilege* | 0.12 | *presidential* | 0.11 | *searches* | 0.11 | *seizure* | 0.11 | *president* | 0.10 |

**GloVe**

| 1978 | | 1988 | | 1998 | | 2008 | | 2018 | |
|---|---|---|---|---|---|---|---|---|---|
| *intrusion* | 0.78 | *expectation* | 0.78 | *expectation* | 0.76 | *expectation* | 0.76 | *expectation* | 0.76 |
| *invasion* | 0.77 | *expectations* | 0.75 | *intrusion* | 0.72 | *intrusion* | 0.71 | *intrusion* | 0.70 |
| *invasions* | 0.71 | *intrusion* | 0.72 | *invasion* | 0.71 | *intrusions* | 0.71 | *invasion* | 0.70 |
| *instrusions* | 0.70 | *invasion* | 0.70 | *expectations* | 0.71 | *invasion* | 0.70 | *intrusions* | 0.70 |
| *invaded* | 0.65 | *instrusions* | 0.70 | *intrusions* | 0.70 | *expectations* | 0.70 | *expectations* | 0.68 |
| *protects* | 0.64 | *invasions* | 0.68 | *invasions* | 0.68 | *invasions* | 0.68 | *invasions* | 0.66 |
| *4th amendment* | 0.63 | *invaded* | 0.64 | *invaded* | 0.64 | *invaded* | 0.64 | *invaded* | 0.64 |
| *expectation* | 0.62 | *protects* | 0.60 | *protects* | 0.59 | *protects* | 0.61 | *protects* | 0.61 |
| *expectations* | 0.61 | *4th amendment* | 0.59 | *4th amendment* | 0.59 | *4th amendment* | 0.58 | *4th amendment* | 0.58 |
| *protected* | 0.59 | *invade* | 0.54 | *unwarranted* | 0.56 | *invade* | 0.55 | *informational* | 0.54 |

Similar to the CBOW and SGNS algorithms, over the five decades analysed, we do not see much variation in the rankings for the top 3 words, but we observe a hint of informational privacy becoming more important. We see this in the increasing ranking of *information* over time as *paper* drops of the top 10 rankings. The word *president* and *presidential* unexpectedly featured in our rankings, though it is suspected that this is a result of the ongoing discussion around the constitutional separation of powers and the limits of power of the executive branch when it comes to privacy and surveillance (Schneier, 2015).

### 5.1.4 GloVe Word Similarity Rankings

Finally, for our GloVe implementation, we note some similarities to both *word2vec* (CBOW and SGNS) algorithms. Though unlike these models, the results we get only provide contextual information around the definition of the notion of *privacy.* While we see the *4th amendment* feature in table 2 again, the algorithm picks up more of the variations in the context words e.g. *invade*, *invading*, *invaded* and *invasion*, making it more susceptible to the decision not to lemmatise or stem our corpus. This choice was made to evaluate the performance of each algorithm with as minimal pre-processing to the data set as possible — and to also see the unadulterated results. So, the results from the GloVe implementation are helpful to note for further iterations. That said, at the bottom of the word ranking for 2018, we still see *informational* make an appearance.

## 5.2 Embedding Stability

In figure 2, we plot the words with the highest cosine similarities (by their mean and standard deviation) to our target word *privacy* to see the variation in stability for each algorithm. We observe patterns that cause us to lower our confidence in any inferences derived from the embeddings. Figure 2 shows that the cosine similarities can vary significantly and the differences in cosine similarities between each word are not wide enough to have confidence in their rankings. We also notice that the top-ranked words can vary widely depending on the algorithm. As expected, the results from the *word2vec* algorithms are similar given their similar architecture. However, there are still minor differences in the results observed, i.e. SGNS was better at picking up case law relevant to *privacy.*

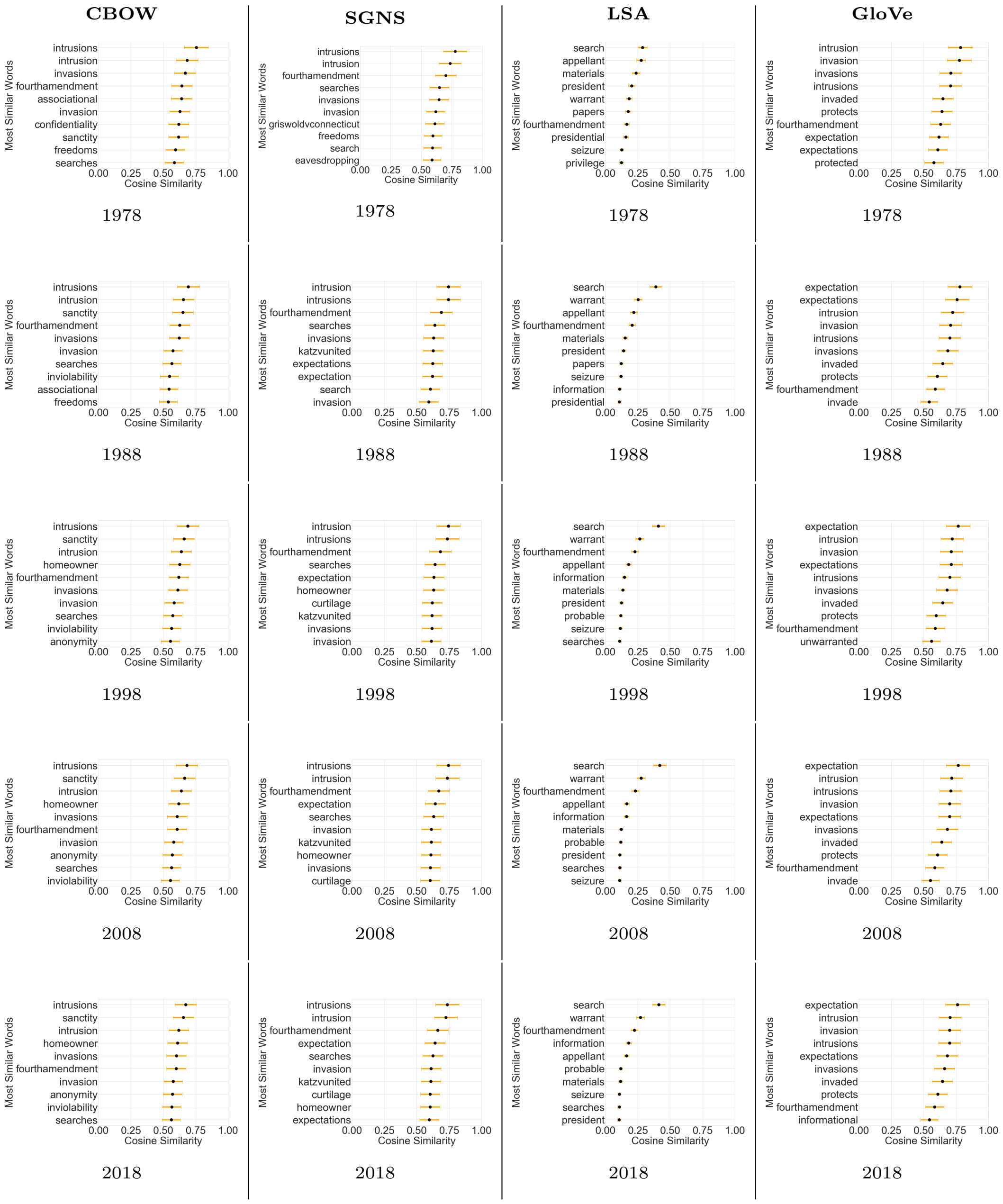

Figure 2: The most similar words with their means and standard deviations for the cosine similarities between the query word *privacy* and its 10 nearest neighbours.

As noted above, our results from the LSA model differ from the *word2vec* models, though it still picks up the relationship between *privacy* and the *4th amendment*. The model, however, exhibits significantly more stability compared to the other implementations (as shown in figure 3), which is in contrast to the results from Antoniak and Mimno (2018) where variations in a fixed setting were only slight. This may be a result of our experiment set up or the size of the corpora, but it is an interesting point to note nonetheless. While the GloVe model factorises a word-context co-occurrence matrix, bringing it closer to traditional methods such as LSA, its results appear more similar to our *word2vec*. It perhaps confirms its classification as a prediction-based model like *word2vec* in Levy, Goldberg, and Dagan (2015).

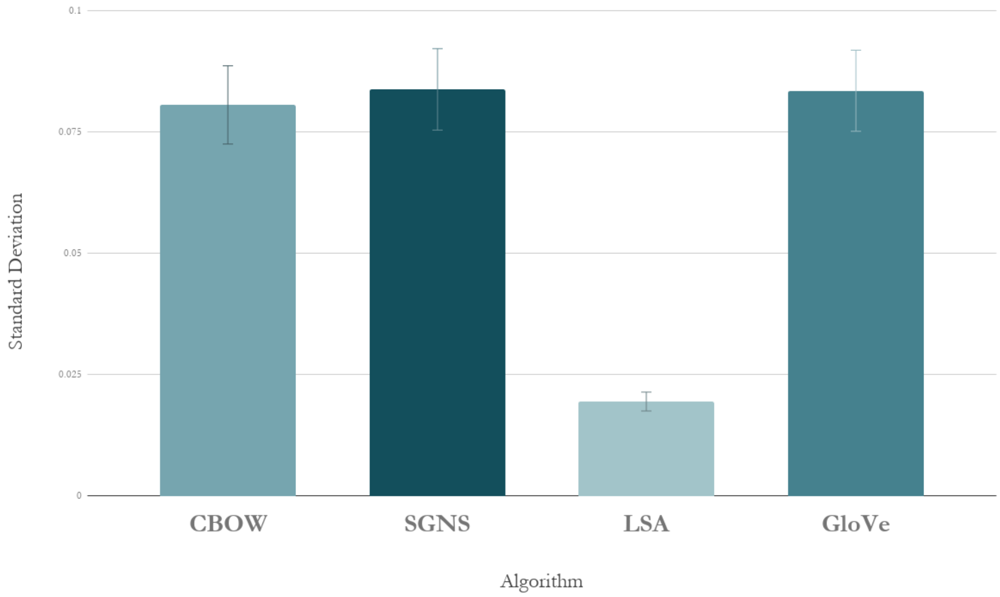

Figure 3: The mean standard deviations across time and algorithms for the 15 closest words to the target word. Larger variations indicate less stable embeddings.

These patterns of larger or smaller variations are generalised in figure 3, which shows the mean, standard deviation for the different algorithms across time, i.e. the standard deviation across the 50 runs for the target word was calculated, and then these standard deviations were averaged. The results show the average levels of variation for each algorithm, from which we observe that the *word2vec* and GloVe algorithms produce the most variability in cosine similarities and the LSA algorithm appears most stable.

# 6 Discussion

GloVe and *word2vec* algorithms have gained popularity as they have been shown to regularly and substantially outperform traditional distributional semantic models such as LSA. Many have attributed this to the 'neural' architecture, which is seen as more useful than relying solely on co-occurrence counts. Distributional semantic models are considered count-based models because they solely use the co-occurrences among words by operating on co-occurrence matrices. Neural word embedding models, in contrast, are viewed as prediction-based models as they attempt to predict surrounding words. Baroni et al. (2014) showed that in nearly all tasks prediction-based models consistently outperform count-based models, thus verifying its supposed superiority. However, the downside with prediction-based models, as demonstrated in this paper, is the variability of its word representations. A likely result of its 'neural' architecture which uses random initialisation of weights, and in turn causing the variability we observe.

Concerning GloVe, Levy et al. (2015) posit that it should be considered a prediction-based model, but the distinction is not as obvious because it factorises word-context co-occurrence matrices, which brings it close to traditional methods such as PCA and LSA. More importantly, Levy and Goldberg (2014) demonstrate that *word2vec* models also implicitly factorise word-context PMI matrices. So although on the surface distributional semantic models and neural models use different algorithms to learn word representations, both types of model fundamentally act on the same underlying statistics of the data, i.e., the co-occurrence counts between words. Given that our resulting representations from GloVe appear just as variable as the *word2vec* models, a key distinction between GloVe and LSA — and an important factor for embedding stability — may be the random initialisation used in neural models. This, of course, can be countered by fixing our random seed when initialising weights for the model or averaging over multiple runs as implemented in this paper.

Much work has been done on the hierarchies of the embedding models, but these have to be taken with a grain of salt as each algorithm might perform better than the other on specific tasks or with the right hyperparameter tuning. Levy et al. (2015) shed light on the importance of hyperparameters compared to other choices. These settings may be more important than algorithm choice as no single algorithm consistently outperforms the others on all tasks (Levy et al., 2015). Beyond hyperparameter tuning, another essential factor is potentially training our algorithms on a larger corpora (Antoniak & Mimno, 2018). Our current implementation only uses a fraction of the cases available at the federal level in the US, and further iterations should consider using cases from the High Courts and Courts of Appeal as they elaborate more on points of law that have been decided at the Supreme Court. These provide further elucidation of the interpretation of our target word and could also stabilise our embeddings.

## 7 Conclusion

From our analysis, what appears to be most interesting in the trends are the words that start to appear in a time step or move significantly up or down the ranking at each time step. Therefore, the stability of the embeddings is crucial as it affects the inferences and conclusions we can make. The results of our experiments also force us to emphasise that embeddings are not an entirely objective view of corpora (Antoniak & Mimno, 2018). There is much variability in coherence and stability as a result of the algorithm chosen, and while the *word2vec* models show the most promise in capturing the wider interpretation of our target word, much improvements can be made to all our chosen algorithms by optimising our pre-processing steps or tuning their hyperparameters (Levy et al., 2015; Antoniak & Mimno, 2018).

A further point worth noting is that while the variability we observe affects the confidence we have in making inferences, this variability can be quantified to make specific observations about an algorithm's sensitivities. Nevertheless, the use of embeddings needs to be balanced with the understanding of its limitations. As noted in Antoniak and Mimno (2018), fine-grained distinctions between cosine similarities are not reliable and smaller corpora, and longer documents are more susceptible to variation in the cosine similarities between embeddings.

We suggest for future work a comparison with other alignment methods used in diachronic word embeddings outlined in section 2.3 and a full hyperparameter sweep for the four algorithms. The alternative alignment methods and hyperparameters can substantially impact performance and coherence, so the goal would be to examine how the algorithms respond to the various changes. As with previous papers and studies on word representations, we can make no claim as to the best algorithm or model; instead we highlight the importance of making domain considerations to ensure the representations are fit-for-purpose.

# References

Antoniak, M., & Mimno, D. (2018). Evaluating the stability of embedding-based word similarities. *Transactions of the Association of Computational Linguistics*, *6*, 107–119.

Bamler, R., & Mandt, S. (2017). Dynamic word embeddings. In *Proceedings of the 34th international conference on machine learning-volume 70* (pp. 380–389).

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (Vol. 1, pp. 238–247).

Blank, A., & Koch, P. (1999). Introduction: historical semantics and cognition. *Historical semantics and cognition*, 1–16.

Bloomfield, L. (1933). *Language history: from language (1933 ed.).* Holt, Rinehart and Winston.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Choi, H., & Varian, H. (2012). Predicting the present with google trends. *Economic Record*, *88*, 2–9.

Dubossarsky, H., Weinshall, D., & Grossman, E. (2016). Verbs change more than nouns: a bottom-up computational approach to semantic change. *Lingue e linguaggio*, *15*(1), 7–28.

Eger, S., & Mehler, A. (2017). On the linearity of semantic change: Investigating meaning variation via dynamic graph models. *arXiv preprint arXiv:1704.02497*.

Gries, S. T. (1999). Particle movement: A cognitive and functional approach. *Cognitive Linguistics*, *10*, 105–146.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Harvard Law Library. (n.d.). *Caselaw access project.* Retrieved from `https://case.law/`, accessed on 2019-01-16.

Heine, B. (2009). Germanic future constructions: A usage-based approach to language change, hilpert, martin 2008. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, *33*(4), 995–1003.

Heyer, G., Holz, F., & Teresniak, S. (2009). Change of topics over time-tracking topics by their change of meaning. *KDIR*, *9*, 223–228.

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., & Petrov, S. (2014). Temporal analysis of language through neural language models. *arXiv preprint arXiv:1405.3515*.

Kulkarni, V., Al-Rfou, R., Perozzi, B., & Skiena, S. (2015). Statistically significant detection of linguistic change. In *Proceedings of the 24th international conference on world wide web* (pp. 625–635).

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: a survey. *arXiv preprint arXiv:1806.03537*.

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems* (pp. 2177–2185).

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 211–225.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., … others (2011). Quantitative analysis of culture using millions of digitized books. *science*, *331*(6014), 176–182.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Newman, M. E. (2005). Power laws, pareto distributions and zipf's law. *Contemporary physics*, *46*(5), 323–351.

Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA l. Rev.*, *57*, 1701.

Pekgözlü, İ., & Öktem, M. K. (2012). Expectation of privacy in cyberspace: The fourth amendment of the us constitution and an evaluation of the turkish case. *Sosyoekonomi*, *18*(18).

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

Ruder, S. (n.d.). *An overview of word embeddings and their connection to distributional semantic models.* Retrieved from `http://blog.aylien.com/overview-word-embeddings-history-word2vec-cbow-glove/`, accessed on 2019-04-16.

Sagi, E., Kaufmann, S., & Clark, B. (2011). Tracing semantic change with latent semantic analysis. *Current methods in historical semantics*, *73*, 161–183.

Schneier, B. (2015). *Data and goliath: The hidden battles to collect your data and control your world.* WW Norton & Company.

Stern, G. (1931). *Meaning and change of meaning; with special reference to the english language.* Wettergren & Kerbers.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, *37*, 141–188.

Wendlandt, L., Kummerfeld, J. K., & Mihalcea, R. (2018). Factors influencing the surprising instability of word embeddings. *arXiv preprint arXiv:1804.09692*.

Yao, Z., Sun, Y., Ding, W., Rao, N., & Xiong, H. (2018). Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the eleventh acm international conference on web search and data mining* (pp. 673–681).

Zhang, Y., Jatowt, A., Bhowmick, S., & Tanaka, K. (2015). Omnia mutantur, nihil interit: Connecting past with present by finding corresponding terms across time. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)* (Vol. 1, pp. 645–655).

# A Appendix: Supplementary Material

According to Zipf's Law, the most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc. This shows that the likelihood of a word occurring is inversely proportional to its rank. It is most easily observed by plotting the data on a log-log graph, with the axes being log (rank order) and log (frequency) resulting in a linear relationship.

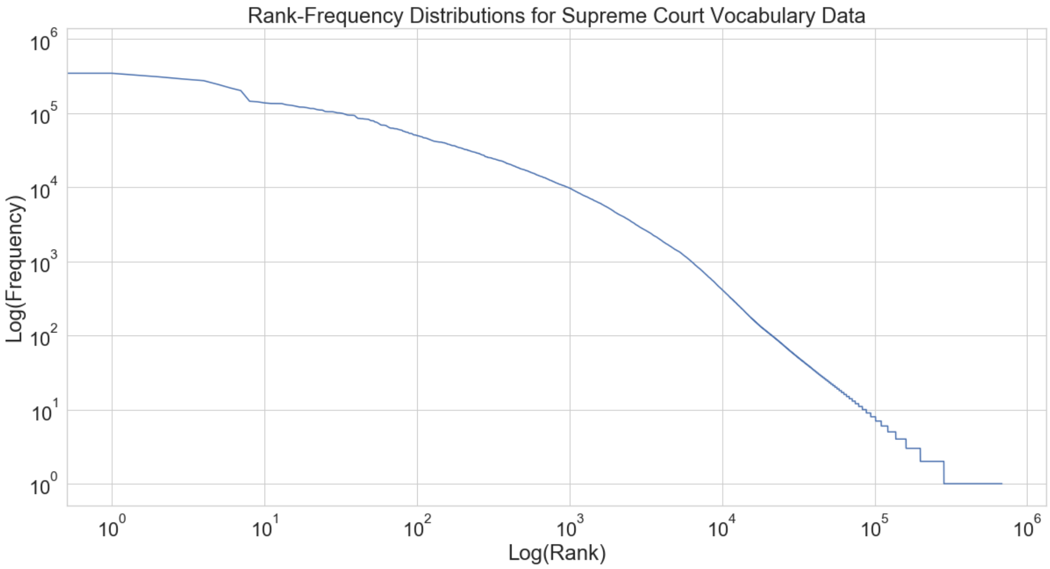

Figure 4: Plot of word frequencies on a log-log graph with log (rank order) and log (frequency) to identify a linear relationship

The distribution was imagined by Zipf to be driven by a principle of 'least effort' where speakers did not work any harder than necessary to communicate a given idea, but the basis for this relationship is still not well understood and conforms at least as well to a process of preferential attachment whereby people disproportionately attend to popular words (Newman, 2005). The relationship in figure 4 does not appear entirely linear as suggested by Zipf, particularly for the least frequent words. Table 3 shows that the relationship does not strictly hold in the top 100 words — at least with stop words removed. Instead, after the first drop between *court* and *state* we see a more gradual descent in frequency.

**Most Common Words in Corpora**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *court* | 710,418 | *new* | 104,588 | *tax* | 78,536 | *parties* | 59,246 |
| *state* | 344,987 | *justice* | 104,565 | *fact* | 76,027 | *given* | 59,049 |
| *section* | 309,751 | *company* | 104,437 | *use* | 75,614 | *constitutional* | 58,544 |
| *case* | 287,558 | *courts* | 102,953 | *supra* | 74,452 | *u.* | 57,051 |
| *states* | 274,293 | *action* | 100,953 | *circuit* | 70,987 | *person* | 57,036 |
| *act* | 243,007 | *did* | 100,656 | *jury* | 69,312 | *e.* | 56,018 |
| *law* | 218,205 | *2d.* | 99,892 | *subject* | 68,947 | *required* | 55,366 |
| *united* | 201,767 | *order* | 98,243 | *amendment* | 68,893 | *present* | 55,313 |
| *federal* | 144,657 | *held* | 96,246 | *authority* | 68,816 | *court's* | 55,180 |
| *said* | 142,292 | *public* | 94,351 | *purpose* | 68,347 | *cause* | 54,185 |
| *district* | 137,480 | *s.* | 93,754 | *laws* | 67,679 | *necessary* | 53,400 |
| *right* | 135,072 | *rule* | 93,490 | *make* | 65,971 | *business* | 53,365 |
| *shall* | 134,860 | *jurisdiction* | 92,866 | *constitution* | 64,889 | *p.* | 53,311 |
| *congress* | 134,615 | *general* | 88,508 | *issue* | 63,615 | *record* | 53,122 |
| *question* | 129,340 | *trial* | 85,136 | *petitioner* | 62,874 | *process* | 51,824 |
| *judgment* | 127,568 | *id.* | 84,896 | *suit* | 62,638 | *provisions* | 50,918 |
| *statute* | 124,228 | *f.* | 84,225 | *commission* | 62,557 | *clause* | 50,884 |
| *opinion* | 120,781 | *claim* | 84,169 | *effect* | 62,461 | *board* | 50,747 |
| *property* | 120,298 | *rights* | 83,503 | *land* | 61,729 | *commerce* | 50,621 |
| *time* | 118,560 | *decision* | 83,236 | *contract* | 61,700 | *proceedings* | 50,180 |
| *cases* | 115,531 | *defendant* | 82,822 | *stat* | 61,512 | *view* | 49,850 |
| *power* | 115,408 | *c.* | 82,621 | *title* | 61,280 | *provided* | 49,546 |
| *does* | 110,314 | *appeals* | 80,827 | *mr.* | 60,627 | *provision* | 49,326 |
| *evidence* | 109,913 | *appeal* | 79,122 | *claims* | 60,009 | *error* | 49,224 |
| *government* | 105,370 | *n.* | 78,970 | *plaintiff* | 59,835 | *supreme* | 48,874 |

Table 3: Top 100 most common words and their frequencies in the corpora used for the preparing embeddings (excludes opinions with less than 1,000 characters).