What causes words to change meaning over time?

Ritvik Rao

One of the most interesting topics in linguistics and natural language processing pertains to the evolution of word definitions over time. How quickly do words change meaning, and what factors accelerate or decelerate such changes? One team of researchers at Stanford University attempts to answer these questions in the paper "Diachronic Word Embeddings Reveal Laws about Statistical Laws of Semantic Change." In this post, I want to talk about this paper and the questions posed, the methods used, and the conclusions reached by the researchers.

What is semantic change, and why is it important?

Semantic change is the change of the meaning of a word, and "diachronic" is something that refers to something that is historic. Putting these two concepts together, they refer to the change of word meanings over time. This phenomenon is crucial for constructing models of language and cultural evolution. There seem to be some patterns in the English language regarding semantic change: the paper offers the word "cat" as a word that has not changed meaning very much over the centuries, but the word "cast" has had a variety of meanings, including "to mold," "a collection of actors," and "a hardened bandage," despite being the same word with a constant spelling and word origin over time (Hamilton et al., 2016).

Unfortunately, a lack of historical data about word meanings has prevented academics from formulating and testing any concrete theory about semantic change. Various researchers over the years have tried to make hypotheses based on small datasets of words. For example, one paper cited by the Stanford researchers tested the hypothesis that synonyms tend to change meaning in similar ways (Xu and Kemp, 2015). However, these hypotheses generally only cover specific cases and are only tested over datasets in the English language.

The Stanford researchers wanted to test a dataset covering all types of words in multiple languages. One method that has not been explored in-depth is the use of word embeddings. A word embedding model maps words or phrases to vectors of real numbers, where the vectors provide some information about how the word or phrase is related to other words or phrases. Using an objective and robust methodology, the researchers decide to compare three different systems of word embeddings and choose the one best-suited to carry out the task of testing two unanswered questions about semantic change (Hamilton et al., 2016). They then proceed to answer those questions using the word embedding system of their choice.

The first hypothesis relates to word frequency: does word frequency have an impact on how quickly words change? Frequency is an important influence on other factors of language evolution. Changes in how words are spoken (like lenition, which is when consonants are spoken more loudly) seem to have a greater impact on words spoken more frequently, while morphological regularization (the evolution of a word to a new spelling) seems to have a lower impact on more frequently spoken words (Hamilton et al., 2016). Given these influences, it is worth asking if frequency also affects semantic change.

The second hypothesis relates to polysemy: If a word has more than one definition, is that word more likely to change one or more of its meanings? We know that polysemic words are used in

multiple different contexts because the word has many definitions, which gives that word more exposure to more people. However, it is still an open question if this extra exposure translates to a greater tendency to change. Furthermore, since polysemic words tend to be more frequent, any analysis of such words must account for frequency to determine if the diverse usage of the word (as opposed to simply the frequent use of the word) causes semantic change (Hamilton et al., 2016).

Previous analyses of semantic change

The most recent research in semantic change before the analysis of the Stanford researchers revolved around an idea called distributional semantics. The general idea behind distributional semantics is to embed each word in a vector based on which words occur near the word in a dataset reflecting a specific decade. Then, the vectors from each decade are stacked on top of each other or arranged in a similar way to provide a diagram of how the word changes over time. Our class textbook does not make much of a mention on distributional semantics or semantic change; the only reference to the topic is actually a reference to the "Diachronic Word Embeddings" paper (Jurafsky et al., 2016).

One paper that utilizes distributional semantics was titled "Tracing semantic change with Latent Semantic Analysis," published in the book "Current Methods in Historical Semantics." The study uses the Helsinki Corpus as its input, which contains 1.15 million words spread over three time periods: Early Middle English, Late Middle English, and Early Modern English. Within this corpus, the semantic change of four words (dog, do, deer, and hound) are analyzed (Sagi et al., 2011).

As the title of the paper suggests, latent semantic analysis (LSA) is used to detect semantic change. LSA, according to the paper, is the collective name for a group of methods that involve numerical representations of words based on occurrence patterns within a text or set of texts (Sagi et al., 2011).. Specifically, the paper uses a representation called a term-term matrix, which has the different word vectors as the rows and occurrences with a list of words that is considered "content-bearing" as the columns. The words that are considered "content-bearing" are holistically selected by the authors based on the subject matter of the target text. The final steps in making the matrix involve weighting the matrix values: first by using the tf-idf score, which changes the value based on the frequency of the column's context word (Sagi et al., 2011)., and then by using SVD, a way to reduce the size of the matrix and remove the number of zero values that I will talk about later on as part of the Stanford paper discussion.

Once the term-term matrix is created, two different patterns of semantic change are examined. The first pattern is broadening, a phenomenon where a word with a very specific meaning becomes more generalized over time. The second pattern is narrowing, which is the exact opposite of broadening: a word with a general meaning becomes more specific over time (Sagi et al., 2011). To examine these patterns, the authors compare the vectors of the words they are examining (dog, do, deer, and hound) to the vectors of the context words they have chosen, and determine the context density of each word over different eras. A high context density means the word appears in only one or a few contexts, while a low density means the word appears in many contexts and therefore is likelier to have more meanings (Sagi et al., 2011). This comparison happens with the cosine similarity, which is a measure of the angle between two vectors. There is an inverse relationship between the

Table 1. Mean angle between context vectors for target words in different periods in the *Helsinki Corpus* (standard deviations are given in parentheses, sample size given below the mean)

	n	Unknown composition date (<1250)	Early Middle English (1150–1350)	Late Middle English (1350–1500)	Early Modern English (1500–1710)
dog	130			12.8 (13.5) *n* = 12	24.7 (10.4) *n* = 118
do	4298		10.3 (13.5) *n* = 1000	13 (9.5) *n* = 1000	24.5 (11.2) *n* = 1000
deer	61	38.7 (17.6) *n* = 16	20.6 (18.2) *n* = 22		20.5 (9.8) *n* = 23
hound	36			22.8 (14.2) *n* = 21	16.4 (11.6) *n* = 15
science	79			13.5 (13. 3) *n* = 22	28.3 (12.2) *n* = 57

angle measurement and context density, and therefore a direct relationship between angle measurement and word versatility.

The result of the experiment on the Helsinki corpus produced the following table of values to the right (Sagi et al., 2011). As seen in the table, the words dog and do have higher angles, which means they have more meanings and are therefore more generalized than before. However, the words deer and hound have become more specific over time as their angles have decreased.

The goal of the LSA paper was to show that, to some extent, semantic change over time could be captured with a word embedding method, and the results for each of the target words display this. However, even the LSA paper admits that "while LSA captures some of the variability in meaning exhibited by words in context, it does not capture all of it." The LSA method is still very simplistic since it exclusively relies on basic occurrence patterns with little weighting, which can lead to some misleading results. The hypothesis of simply demonstrating that two common patterns of semantic change could be detected with LSA was also simplistic. Additionally, the paper only explores the semantic change of a small number of words (4) and only uses the English language. This is much different from the "Diachronic Word Embeddings" paper, which uses text from four languages, rigorously selects a more complex word embedding system, and tests hypotheses about what types of words change rather than simply demonstrating that change happens.

Approach

The Stanford research team tests their hypotheses with a two-step process (Hamilton et al., 2016).

The first step is to choose a word embedding system that is the most suitable for the task at hand. To be specific, the best word embedding system should be able to detect whether pairs of words moved closer or further apart in meaning over time, to correctly choose which words in a dataset have changed meaning the most over time, and to detect the similarity of words within a time period. The researchers consider three embedding systems (Hamilton et al., 2016).

The first system is called positive point-wise mutual information (PPMI). The formula for calculating each cell in a PPMI matrix is as follows:

$$\mathbf{M}_{i,j}^{\text{PPMI}} = \max\left\{\log\left(\frac{\hat{p}(w_i, c_j)}{\hat{p}(w)\hat{p}(c_j)}\right) - \alpha, 0\right\}, \tag{1}$$

The p-hat is the probability that the word to which the row vector corresponds appears near the context word represented by the column, and the alpha is a smoothing value. The minimum value is 0, which prevents values from becoming too extreme and allows for an emphasis of positive word-word relationships instead of negative ones.

The second system is called SVD. SVD is the same method used in the LSA paper mentioned above. To make an SVD embedding, the authors start with a PPMI matrix. They then transform it by solving the following equation for U and sigma (this is called obtaining the truncated singular value decomposition):

$$\mathbf{M}^{\text{PPMI}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}$$

The final step is to come up with the word embedding with this equation:

$$\mathbf{w}_i^{\text{SVD}} = \left(\mathbf{U}\mathbf{\Sigma}^{\gamma}\right)_i$$

Where gamma is an eigenvalue weighting parameter in the range [0,1]. In short, the method takes a PPMI matrix and creates a denser matrix with fewer zeros, which improves efficiency during the testing step of the experiment.

The third system is called skip-gram with negative sampling (SGNS). SGNS represents each word in not one but two vectors: a word vector w and a context vector c. SGNS does not attempt to directly calculate co-occurrence relationship patterns; rather, the method uses a machine learning concept called stochastic gradient descent (SGD) to predict co-occurrence patterns. The SGD optimizes the embeddings so that:

$$\hat{p}(c_i|w_i) \propto \exp(\mathbf{w}_i^{\text{SGNS}} \cdot \mathbf{c}_j^{\text{SGNS}}),$$

Where exp is the exponent of the Euler number. To avoid calculating extremely large dot products and extremely large exponents, the authors randomly select some context words such that exp(w*c) is small and then running the optimization.

The authors then test these systems based on two characteristics: synchronic accuracy and diachronic validity (Hamilton et al., 2016). Synchronic accuracy is the ability to capture the similarity of words within an individual time period. To test this, the authors use a similarity benchmark defined by another paper and use text from the 1990's in one of the datasets. The SVD embedding performed the best, followed by PPMI and SGNS; this finding seemed to concur with another study.

Diachronic validity is the ability for a system to quantify semantic changes over time. The authors chose to test this in two different ways. The first way is to check if each method can detect known historical shifts after being trained on the dataset. A set of 28 different known shifts was used; each shift was defined with a target word, a relation word, a shift direction, and the start of the shift. For example, one of the tests was to see if the system could detect the word "fatal" moving towards the word "lethal" in meaning starting before 1800. Each of the three systems was trained on two datasets. Once these shifts were tested, the results shown on the left were obtained. It is clear that each of the tests were highly accurate in detecting the shifts. However, at 95% confidence, only the SVD model tested on the more refined COHA dataset produced a significant result (Hamilton et al., 2016).

Method	Corpus	% Correct	%Sig.
PPMI	ENGALL	96.9	84.4
	COHA	100.0	88.0
SVD	ENGALL	100.0	90.6
	COHA	100.0	96.0
SGNS	ENGALL	100.0	93.8
	COHA	100.0	72.0

The second diachronic validity test is to use each of the three methods to find the 10 words with the most semantic change from 1900s to 1990s by measuring the cosine distance for each word between each decade. The authors limited the test to words with a frequency of greater than 1 in 100,000 to allow the test to run faster. The end result for each method is below (Hamilton et al., 2016):

Method	Top-10 words that changed from 1900s to 1990s
PPMI	know, got, would, decided, think, stop, remember, **started**, must, wanted
SVD	harry, **headed**, **calls**, **gay**, wherever, male, **actually**, special, cover, naturally
SGNS	**wanting**, **gay**, **check**, **starting**, **major**, **actually**, touching, harry, **headed**, romance

In this table, bolded words represent known significant changes in meaning, underlined words have not changed in meaning but are used in more diverse contexts ("false discoveries"), and all other words are what are known as "corpus artifacts," which are words that show up in the text because they are frequently used in places such as cover pages and advertisements as opposed to regular text. The PPMI model does not seem to be great at picking out the biggest semantic changes, and in fact produces many false discoveries. This caused the authors to eliminate that model as a possibility for their analysis. The SVD model was second best, but the best was clearly SGNS. The conclusion from this testing was that SVD is better at synchronic accuracy and SGNS is better at diachronic validity, and the authors acknowledged this tradeoff in their paper. They ultimately chose

to use SVD embeddings to test their hypotheses but mentioned that similar results can be obtained with SGNS (Hamilton et al., 2016).

The final step for the authors was to test the hypotheses they had formulated. In a similar fashion to the LSA study, a word's rate of semantic change is measured using the cosine similarity (Hamilton et al., 2016). The cosine similarity between two vectors is the dot product of the vectors divided by the product of the vector lengths or norms. The larger the value of the similarity, the more the word has changed over time. Unlike the LSA study, the Stanford researchers wanted to demonstrate factors that could affect semantic change, so they needed to derive a formula more complex than a simple cosine similarity; ideally, a function that would have weights that could be trained with a regression analysis. The new devised function is below:

$$\tilde{\Delta}^{(t)}(w_i) = \beta_f \log\left(f^{(t)}(w_i)\right) + \beta_d \log\left(d^{(t)}(w_i)\right) + \beta_t + z_{w_i} + \epsilon_{w_i}^{(t)} \;\; \forall w_i \in \mathcal{V}, t \in \{t_0, ..., t_n\}, \quad (7)$$

This function is log-scaled on the x-axis to ensure a zero mean and unit variance. The function f(w) is the frequency of a word. The function d(w) measures the polysemy of the word. Each of the beta terms represents the weights being tested: β_f for the frequency weight, β_d for the polysemy weight, and β_t for the time/decade weight. The constant z is a random intercept value, and the epsilon constant is the error term. By design, the function accounts for both frequency and polysemy, which accounts for the positive relation between polysemy and frequency, allowing the researchers to separate the effect of polysemy and frequency on semantic change. Finally, the researchers chose to only run the regression test for the 10,000 most frequent words to ensure decade-to-decade data for each word and did not include proper nouns whose meanings can change more unpredictably (Hamilton et al., 2016).

The result of the regression analysis showed two results that supported the researchers' hypotheses. The first result was the "law of conformity:" words that appear more frequently are less likely to undergo semantic change and vice versa. This result was achieved because the researchers observed the following relationship:

$$\Delta(w_i) \propto f(w_i)^{\beta_f}$$

The decade-to-decade gradient has an exponential relation to the frequency of the word, but β_f was observed to be in the range of [-1.26, -0.27], making the relationship inverse (Hamilton et al., 2016).

The second result was the "law of innovation:" Highly polysemic words are more likely to exhibit semantic change. This is shown by the following relation, similar to the one for frequency:

$$\Delta(w_i) \propto d(w_i)^{\beta_d}$$

In this case, β_d was observed to be in the range of [0.37, 0.77], making the relationship direct (Hamilton et al., 2016).

From these findings, the authors were able to conclude that the law of conformity and the law of innovation accounted for between 48 and 88% of all semantic change. They theorized that people are more likely to use less frequent words in weird ways while facing more social pressure to use common words correctly. They also postulated that the law of innovation holds because polysemous words are more likely to have rare definitions that make the word's meaning unstable according to the law of conformity (Hamilton et al., 2016).

Dataset

A corpus is a collection of words and texts from many different sources, and they are generally used as inputs for NLP models. The "Diachronic Word Embeddings" paper used the following corpuses:

Name	Language	Description	Tokens	Years	POS Source
ENGALL	English	Google books (all genres)	8.5×10^{11}	1800-1999	(Davies, 2010)
ENGFIC	English	Fiction from Google books	7.5×10^{10}	1800-1999	(Davies, 2010)
COHA	English	Genre-balanced sample	4.1×10^{8}	1810-2009	(Davies, 2010)
FREALL	French	Google books (all genres)	1.9×10^{11}	1800-1999	(Sagot et al., 2006)
GERALL	German	Google books (all genres)	4.3×10^{10}	1800-1999	(Schneider and Volk, 1998)
CHIALL	Chinese	Google books (all genres)	6.0×10^{10}	1950-1999	(Xue et al., 2005)

As stated above, the main source of the corpuses is Google Books, but the COHA corpus is a genre-balanced sample that was curated by the authors to ensure a fair balance of all types of text. There are four languages represented, something the authors wanted to do so that they could generalize their theories across languages (Hamilton et al., 2016). For English, along with COHA, there is also and EngFic corpus that is consists of the fictional works from the EngAll corpus. The drawback to using Google Books as a source is that it increases the frequency of "corpus artifacts," which are not works written by authors such as prose or poetry, but instead are ancillary data (such as text from advertisements) that can be considered as noise and is not particularly useful to the researchers. The Stanford team attempts to rectify this by curating the COHA text set, which includes more normal text. The tradeoff is that COHA is orders of magnitude smaller than the other corpuses, even though it is higher quality.

Along with the English corpuses, the authors used corpuses from the French, German, and Chinese languages, each of which is analogous to EngAll. The authors use these to test their theories on multiple languages, as opposed to only English in older research.

Models

I described the models used in the Approach section. Each of the three models relies on supervised creation. PPMI and SVD models are not created with any machine learning techniques, while skip-gram uses gradient descent, a type of supervised ML. PPMI and SVD both use direct probability calculations as the basis for their features: the construction of both types of embeddings involve directly counting the occurrences of context words within a fixed window (Hamilton et al., 2016). SGNS uses predictions of occurrence patterns calculated by SGD, which may possibly vary based on the step size during optimization. PPMI and SVD are deterministic since they should always produce the same matrix for the same corpus, while SGNS is not deterministic.

Why did the authors choose this strategy?

One advantage to the authors' approach is that it included a rigorous test of choosing a word embedding that best suited the problem of exploring the effects of frequency and polysemy on semantic change. As opposed to choosing just one model, as was done in the LSA paper, the Stanford researchers chose three different state-of-the-art systems and created objective benchmarks tailored for their task to make an informed selection of the optimal model. By creating those benchmarks, the researchers provide a robust system that can be used by others in the future.

Another strength of the paper was the choice of multiple data sets from multiple languages and genres. The variables being changed with the other corpuses are language (English, Spanish, French, and German), genre (EngFic), and text quality (COHA). This choice allowed the researchers to generalize the laws they discovered to cover all human language, which broadens the future scope of research on semantic change.

One disadvantage is the author's method of measuring polysemy. The authors chose to measure polysemy by quantifying the number of diverse contexts in which a word appeared (Hamilton et al., 2016). The problem with this measure is that is has a bias towards words that have the same meaning but are common function words that appear in many contexts (like the word "also"). The authors accept this bias due to the simplicity of the measure, but they offer little assurance that it will not be a confounding variable in the data.

Possibly the most glaring issue in the paper is the apparent lack of a control corpus. It was a good choice to choose a variety of datasets, but it is crucial to implement a control experiment (for example, analysis over a single decade) to prove that the chosen models can detect meaningful change. Without a control, the authors risk the effect of confounding variables creating what seems like meaningful data.

The reason that this task is meaningful in the NLP world is that it postulates two major sources of semantic change that need to be accounted for, or at least more rigorously tested, in future research. One such paper is "Tracking word semantic change in biomedical literature." As the title implies, the paper analyzes semantic change in a specific genre (biomedical literature) and attempts to see if the application of the law of conformity created by the Stanford researchers can be used in this context. Interestingly, the paper cannot find much evidence to support the law of conformity in biomedical literature (Yan & Zhu, 2018). Another use of this research is in the analysis of societal attitudes over time. One of the authors of this paper published another paper that used word embeddings to quantify 100 years of gender and ethnic stereotypes. The laws created in the "Diachronic Word Embeddings" paper were taken into account when creating a representation to demonstrate that the change in word embeddings of occupation and demographic data in the U.S. Census reflect a change in societal attitudes toward gender and ethnicity (Garg et al., 2018). In short, whenever someone does research that uses semantic change, the effect of word frequency and polysemy must be accounted for.

Class topics that appear in this paper

The major class topic that appears in this paper is the one regarding word embeddings. Word embeddings are useful to this research because very large corpuses of text are involved, and word embeddings can be created without any preprocessing or marking of the text, which means that building the models takes less time. Specifically, we covered the skip-gram algorithm in-depth, and this paper uses a variation of skip-gram as one of their possible representations of data. Just like in class, skip-gram uses logistic regression to calculate probabilities that estimate co-occurrence patterns within the corpus used (Jurafsky et al., 2016). Instead of the word embeddings representing the direct probability that one word will appear near a context word, the embeddings represent weights used in the logistic regression, and each step of the gradient descent uses the previous step's result as inputs. The main difference between the skip-gram used in the paper versus the one used in class is the negative sampling (Hamilton et al., 2016). This means that rather than building embeddings with all words and all context words, the algorithm randomly samples non-context words (negative examples) and uses them during training to increase the speed of the algorithm.

One other use of regression analysis (a concept we learned in class) is during the training of the gradient function (the normalized semantic displacement score) during the main experiment of analyzing the effect of frequency and polysemy on semantic change. The weights included in the function were trained over many epochs on each of the six corpuses to provide a range of the values of each of the weights. In general, a negative weight signaled a negative correlation while a positive weight signaled a positive correlation.

Follow-up work to this paper

As mentioned in the previous section, researchers in NLP are making use of the large corpuses now available from places such as Google Groups and are trying to analyze semantic change in genres such as scientific research and in fiction, while also trying to see if semantic change reflects changes in culture and society. There is still ongoing debate about what exactly causes semantic change, and not everyone agrees with the conclusions of the Stanford researchers. Many other researchers are either coming up with alternate laws of semantic change or are questioning the usefulness of these laws altogether.

About a year after this paper was released, a team of researchers from The Hebrew University of Jerusalem wrote a paper that analyzed three laws of semantic change that were previously proposed: the laws of conformity and information proposed in "Diachronic Word Embeddings" and the law of prototypicality proposed in another paper. The Hebrew University researchers wanted to re-test the laws with what they considered a more suitable experiment, one with a control group and other testing groups to properly remove confounding variables (Dubossarsky et al., 2017).

The genuine condition group consisted of 10 million 5-grams from the Google Books corpus, randomly sampled and evenly distributed from 1900 to 1990. "Genuine condition" in this context means a group where semantic change is expected. Since these 5-grams are supposed to represent randomly sampled text, words are expected to change meaning from decade to decade. Word context changes should also change, since word contexts change with word meanings (Dubossarsky et al., 2017).

The control group was created such that no change of meaning is expected. The control group was constructed in two ways. First, 5-grams from the Google Books corpus were randomly shuffled and placed into bins so that each bin had words from the entire range of decades. This method was chosen so that a word's many contexts from one decade would be shuffled across many bins, which would make the word appear as if it did not change very much within one bin. Next, ten million 5-grams from Google Books from the year 1999 were randomly sampled thirty times. Since words in the same year are not supposed to change their meaning, this means that in the control group, any detection of semantic change will just be noise (Dubossarsky et al., 2017).

Once the datasets were curated, the experiment was carried out. The researchers used the same three models from the "Diachronic Word Embeddings" paper and also used regression analysis to train weights to measure the impact of frequency and frequency plus polysemy together. They then calculated the explained variance of each category. Some of their results are shown below (Dubossarsky et al., 2017).

		PPMI + SVD		PPMI	
		Genuine	Shuffled	Genuine	Shuffled
Frequency	β	-0.91	-0.75	-0.29	0.06
(one-predictor)	explained variance (σ^2)	67%	56%	8%	0%
Frequency +	β frequency	-1.22	-1.12	-0.69	0.53
Polysemy	β polysemy	0.43	0.40	0.49	-0.52
(two-predictor)	explained variance (σ^2)	68%	60%	9%	4%
Frequency +	β frequency	-0.71	-0.70	-0.02	0.07
Prototypicality	β polysemy	0.22	0.21	0.12	0.02
(two-predictor)	explained variance (σ^2)	65%	60%	2%	0%

As seen in this table, the effect of frequency and polysemy seems high when considering both the PPMI and SVD models, but low when considering just the PPMI model. This seems to suggest that frequency and polysemy do have an impact, but just one that is smaller than previously indicated. To ensure that the data is consistent, the Hebrew University researchers calculated polysemy the same way as in the Stanford paper. They found that in the genuine condition, they were able to recreate the results from the "Diachronic Word Embeddings" paper. However, they also found that for some reason, the shuffled control group also returned similar results, and that

there was not a large difference in explained variance between the control and the genuine group (Dubossarsky et al., 2017). Furthermore, there does seem to be a big difference in results based on the type of model chosen. From this, the authors concluded that while frequency and polysemy do seem to have an impact on semantic change, this impact is not nearly as large as the 48 to 88% impact proposed by the Stanford researchers. Additionally, the measured impact of frequency and polysemy is greatly affected by the type of model chosen, where some model types seem to show less of an impact. In my opinion, this newer research is correct that the laws are not that strong since there is little difference between the control and genuine group. I disagree with their conclusion that the type of model makes much of an impact because they use the PPMI model as one of their testing models, which was rejected by the Stanford team as being a bad representation of this task.

Works Cited

Dubossarsky, H., Weinshall, D., & Grossman, E. (2017, September). Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 1136-1145).

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2011. Tracing semantic change with latent semantic analysis. In Kathryn Allan and Justyna A. Robinson, edi- tors, *Current Methods in Historical Semantics*, page 161. De Gruyter Mouton, Berlin, Germany.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635-E3644.

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

Jurafsky, D., & Matrin, J. (2019). *Speech & language processing*. Stanford University.

Yan, E., & Zhu, Y. (2018). Tracking word semantic change in biomedical literature. *International journal of medical informatics*, *109*, 76-86.

Yang Xu and Charles Kemp. 2015. A computational evaluation of two laws of semantic change. In *Proc. 37th Annu. Conf. Cogn. Sci. Soc.*