Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models

Haim Dubossarsky[1], **Eitan Grossman**[2] and **Daphna Weinshall**[3]
[1] Edmond and Lily Safra Center for Brain Sciences
[2] Department of Linguistics
[3] School of Computer Science and Engineering
The Hebrew University of Jerusalem, 91904 Jerusalem, Israel
haim.dub@gmail.com, {eitan.grossman,daphna}@mail.huji.ac.il

Abstract

This article evaluates three proposed laws of semantic change. Our claim is that in order to validate a putative law of semantic change, the effect should be observed in the genuine condition but absent or reduced in a suitably matched control condition, in which no change can possibly have taken place. Our analysis shows that the effects reported in recent literature must be substantially revised: (i) the proposed negative correlation between meaning change and word frequency is shown to be largely an artefact of the models of word representation used; (ii) the proposed negative correlation between meaning change and prototypicality is shown to be much weaker than what has been claimed in prior art; and (iii) the proposed positive correlation between meaning change and polysemy is largely an artefact of word frequency. These empirical observations are corroborated by analytical proofs that show that count representations introduce an inherent dependence on word frequency, and thus word frequency cannot be evaluated as an independent factor with these representations.

1 Introduction

The increasing availability of digitized historical corpora, together with newly developed tools of computational analysis, make the quantitative study of language change possible on a larger scale than ever before. Thus, many important questions may now be addressed using a variety of NLP tools that were originally developed to study synchronic similarities between words. This has catalyzed the evolution of an exciting new field of *historical distributional semantics*, which has yielded findings that inform our understanding of the dynamic structure of language (Sagi et al., 2009; Wijaya and Yeniterzi, 2011; Mitra et al., 2014; Hilpert and Perek, 2015; Frermann and Lapata, 2016; Dubossarsky et al., 2016). Recent research has even proposed *laws of change* that predict the conditions under which the meaning of words is likely to change (Dubossarsky et al., 2015; Xu and Kemp, 2015; Hamilton et al., 2016). This is an important development, as traditional historical linguistics has generally been unable to provide predictive models of semantic change.

However, these preliminary results should be addressed with caution. To date, analyses of changes in words' meanings have relied on the comparison of word representations at different points in time. Thus any proposed change in meaning is contingent on a particular model of word representation and the method used to measure change. Distributional semantic models typically count words and their co-occurrence statistics (*explicit* models) or predict the embedding contexts of words (*implicit* models). In this paper, we show that the choice of model may introduce biases into the analysis. We therefore suggest that empirical findings may be used to support laws of semantic change only after a proper control can be shown to eliminate artefactual factors as the underlying cause of the empirical observations.

Regardless of the specific representation used, a frequent method of measuring the semantic change a word has undergone (Gulordava and Baroni, 2011; Jatowt and Duh, 2014; Kim et al., 2014; Dubossarsky et al., 2015; Kulkarni et al., 2015; Hamilton et al., 2016) is to compare the word's vector representations between two points in time using the cosine distance:

$$cosDist(x, y) = 1 - \frac{x \cdot y}{\|x\|_2 \|y\|_2} \quad (1)$$

This choice naturally assumes that greater distances correspond to greater semantic changes. However, this measure introduces biases that may affect our interpretation of meaning change.

We examine various representations of word meaning, in order to identify inherent confounds when meaning change is evaluated using the cosine distance. In addition to the empirical evaluation, in Section 5 we provide an analytical account of the influence of word frequency on cosine distance scores when using these representations.

In our empirical investigation, we highlight the critical role of control conditions in the validation of experimental findings. Specifically, we argue that every observation about a change of meaning over time should be subjected to a control test. The control condition described in Section 2.1 is based on the construction of an artificially generated corpus, which resembles the historical corpus in most respects but where no change of meaning over time exists. In order to establish the validity of an observation about meaning change - and even more importantly, the validity of a law-like generalization about meaning change - the result obtained in a genuine experimental condition should be demonstrated to be lacking (or at least significantly diminished) in the control condition.

As we show in Section 4, some recently reported laws of historical meaning change do not survive this proposed test. In other words, similar results are obtained in the genuine and control conditions. These include the correlation of meaning change with word frequency, polysemy (the number of different meanings a word has), and prototypicality (how representative a word is of its category). These factors lie at the basis of the following proposed laws of semantic change:

- The Law of Conformity, according to which frequency is negatively correlated with semantic change (Hamilton et al., 2016).
- The Law of Innovation, according to which polysemy is positively correlated with semantic change (Hamilton et al., 2016).
- The Law of Prototypicality, according to which prototypicality is negatively correlated with semantic change (Dubossarsky et al., 2015).

Our analysis shows that these laws have only residual effects, suggesting that frequency and prototypicality may play a smaller role in semantic change than previously claimed. The main artefact underlying the emergence of the first two laws in both the genuine and control conditions may be due to the SVD step used for the embedding of the PPMI word representation (see Section 2.5).

2 Methods

The historical corpus used here is Google Books 5-grams of English fiction. Equally sized samples of 10 million 5-grams per year were randomly sampled for the period of 1900-1999 (Kim et al., 2014) to prevent the more prolific publication years from biasing the results, and were grouped into ten-year bins. Uncommon words were removed, keeping the 100,000 most frequent words as the vocabulary for subsequent model learning. All words were lowercased and stripped of punctuation.

This corpus served as the genuine condition, and was used to replicate and evaluate findings from previous studies. In this corpus, words are expected to change their meaning between decadal bins, as they do in a truly random sample of texts. According to the distributional hypothesis (Firth, 1957), one can extract a word's meaning from the contexts in which it appears. Therefore, if words' meanings change over time, as has been argued at least since Reisig (1839), it follows that the words' contexts should change accordingly, and this change should be detected by our model.

2.1 Control condition setup

Complementary to the genuine condition, a control condition was created where no change of meaning is expected. Therefore, any observed change in a word's meaning in the control condition can only stem from random "noise", while changes in meaning in the genuine condition are attributed to "real" semantic change in addition to "noise". Two methods were used to construct the corpus in the control condition:

Chronologically shuffled corpus (shuffle): 5-grams were randomly shuffled between decadal bins, so that each bin contained 5-grams from all the decades evenly. This was chosen as a control condition for two reasons. First, this condition resembles the genuine condition in size of the vocabulary, size of the corpus, overall variance in words' usage, and size of the decadal bins. Second and

crucially, words are not expected to show any apparent change in their meaning between decades in the control condition, because their various usage contexts are shuffled across decades.

One synchronous corpus (subsample): All 5-grams of the year 1999, which amount to 250 million 5-grams, were selected from Google Books English fiction. 10 million 5-grams were randomly subsampled from this selection, and this process was repeated 30 times. This is suggested as an additional control condition since the underlying assumption is always that words in the same year do not change their meaning. Again, unlike in the genuine condition, any changes that are observed based on these 30 subsamples can be attributed *only* to "noise" that stems from random sampling, rather than real change in meaning.

2.2 Measures of interest

Meaning change: Meaning change was evaluated as the cosine distance between vector representations of the same word in consecutive decades. This was done separately for each processing stage (see Section 2.5). For the subsample condition, this was defined as the average cosine distance between the vectors in all 30 samples.

Frequency: Words' frequencies were computed separately for each decadal bin as the number of times a word appeared divided by the total number of words in that decade. For the subsample control condition, it was computed as the number of times a word appeared among the 250 million 5-grams, divided by the total number of words.

2.3 Construct validity

To establish the adequacy of our control condition, we compared the meaning change scores (before log-transformation and standardization) between the genuine and the shuffled control conditions. Change scores were obtained by taking the average meaning change over all words in each decade using the representation of the final processing stage (SVD). An adequate control condition will exhibit a lower degree of change compared to the genuine condition, and is expected to show a fixed rate of change across decades (see 3a).

2.4 Statistical analysis

Following common practice (Hamilton et al., 2016), the 10k most frequent words, as measured by their average decadal bin frequencies, were used for the analysis of semantic change. Change scores and frequencies were log-transformed, and all variables were subsequently standardized.

A linear mixed effects model was used to evaluate meaning change in both the genuine and shuffled control conditions. Frequency was set as a fixed effect while random intercepts were set per word. The model attempts to account for semantic change scores using frequency, while controlling for the variability between words by assuming that each word's behavior is strongly correlated across decades and independent across words as follows:

$$\Delta w_i^{(t)} = \beta_0 + \beta_f freq_{w_i}^{(t)} + z_{w_i} + \varepsilon_{w_i}^{(t)} \quad (2)$$

Here $\Delta w_i^{(t)}$ is the semantic change score of the i'th word measured between two specific consecutive decades, β_0 is the model's intercept, β_f is the fixed-effect predictor coefficient for frequency, $z_{w_i} \sim N(0, \sigma)$ is a random intercept for the i'th word, and $\varepsilon_{w_i}^{(t)}$ is an error term associated with the i'th word. We report the predictor coefficient as well as the proportion of variance explained[1] by each model. Only statistically significant results ($p < .01$) are reported. All statistical tests are performed in R (lme4 and MuMIn packages).

2.5 Word meaning representation

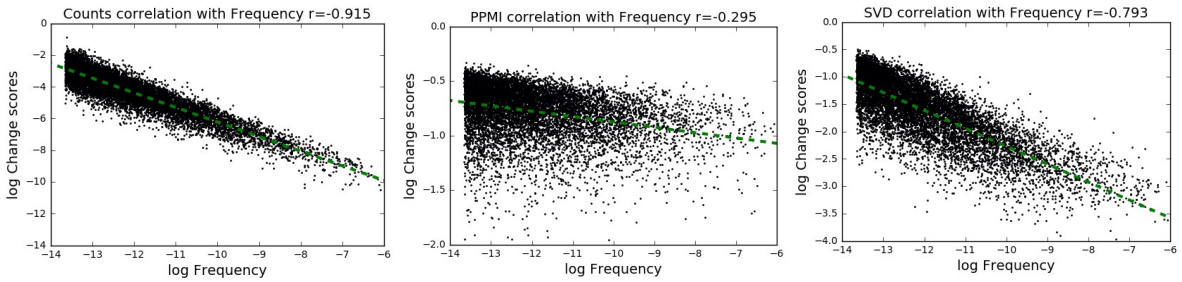
We used a cascade of processing stages based on the *explicit meaning* representation of words (i.e., word counts, PPMI, SVD, as explained below) as commonly practiced (Baroni et al., 2014; Levy et al., 2015). For each of these stages, we sought to evaluate the relationship between word frequency and meaning change, by computing the corresponding correlations between these two factors in the subsample control condition.

Counts: Co-occurrence counts were collected for all the words in the vocabulary per decade.

PPMI: Sparse square matrices of vocabulary size containing positive pointwise mutual information (PPMI) scores were constructed for each decade based on the co-occurrence counts. We used the context distribution smoothing parameter $\alpha = 0.75$, as recommended by (Levy et al., 2015), using the following procedure:

$$PPMI_\alpha(w, c) = \max\left(log\left(\frac{\hat{P}(w, c)}{\hat{P}(w)\hat{P}_\alpha(c)}\right), 0\right)$$

[1] R^2 for mixed linear models (Nakagawa and Schielzeth, 2013)



(a) (b) (c)

Figure 1: Correlations in the control condition between change scores in the year 1999 and word frequency for three word representation types, based on: (a) Counts, (b) PPMI, (c) SVD. Correlation coefficients are reported above each subplot. LS regression lines are shown in dashed green.

where $\hat{P}(w, c)$ denotes the probability that word c appears as a context word of w, while $\hat{P}(w)$ and $\hat{P}_\alpha(c) = \frac{\#(c)^\alpha}{\sum_C \#(c)^\alpha}$ denote the marginal probabilities of the word and its context, respectively.

SVD: Each PPMI matrix was approximated by a truncated singular value decomposition as described in (Levy et al., 2015). This embedding was shown to improve results on downstream tasks (Baroni et al., 2014; Bullinaria and Levy, 2012; Turney and Pantel, 2010). Specifically, the top 300 elements of the diagonal matrix of singular values Σ, denoted Σ_d, were retained to represent a new, dense embedding of the word vectors, using the truncated left hand orthonormal matrix U_d:

$$W_i^{SVD} = (U_d \cdot \Sigma_d)_i \quad (3)$$

These representations were subsequently aligned with the orthogonal Procrustes method following (Hamilton et al., 2016).

Relation to other models: (Levy and Goldberg) have shown that the Skip-Gram with Negative Sampling (SGNS) embedding model, e.g. word2vec (Mikolov et al., 2013) - perhaps the most popular model of word meaning representation, implicitly factorizes the values of the word-context PMI matrix. Hence, the optimization goal and the sources of information available to SGNS and our model are in fact very similar. We therefore hypothesize that conclusions similar to those reported below can be drawn for SGNS models.

3 Results

3.1 Confound of frequency

There are many factors that may confound the measurement of meaning change. Here we focus on frequency, and investigate the existence of an artefactual relation between frequency and meaning change. This is done by evaluating this relation in the subsample control condition. Any changes observed in this condition must be the consequence of inherent noise, since this control condition contains random samples from the same year (and the baseline assumption is that no change can be observed within the same year).

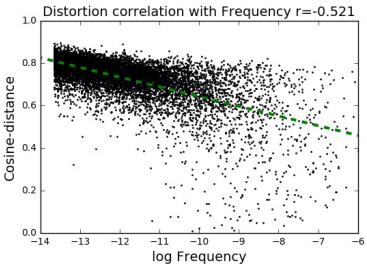
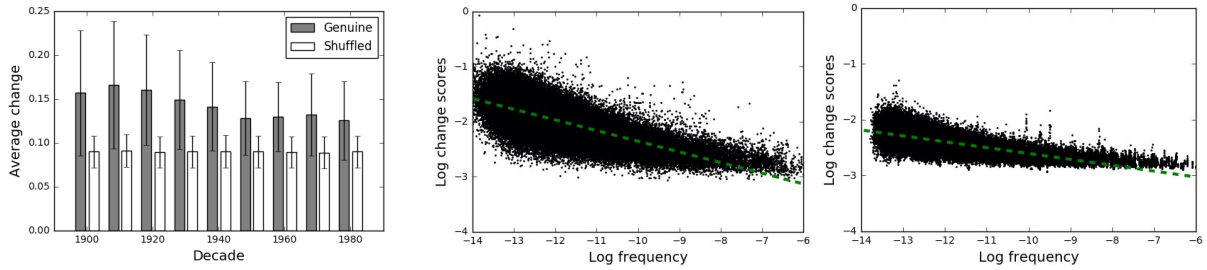


Figure 2: Cosine distances between PPMI and approximated PPMI representations (y-axis), plotted against frequency (x-axis). Correlation coefficient is reported above the plot.

We first plotted the change scores that use the representation based on word count vs. word frequency. This resulted in a robust correlation ($r = -0.915$) between the two variables, as shown in Fig. 1a (see the analytical account in Section 5). We repeated the same procedure using the PPMI representation, which showed a much weaker correlation with frequency ($r = -0.295$), see Fig. 1b.

Finally, we repeated the same procedure using the final *explicit representation* after SVD embedding[2], see Fig. 1c. Surprisingly, the negative correlation with frequency was reinstated ($r = -0.793$). To investigate how this came about,

[2]Similar results were obtained for the implicit embedding (word2vec-SGNS) described in Section 2.5.



(a) (b) (c)

Figure 3: (a) Average change score per decade for the genuine and control conditions. Bars represent standard deviations. (b-c) Change scores (y-axis), relative to their frequency (x-axis): (b) genuine historical corpus, (c) chronologically shuffled historical corpus. LS regression lines are shown in dashed green.

we computed the change in the PPMI vectors before and after the low-rank SVD embedding using the cosine-distance. As apparent from Fig. 2, it turns out that the SVD procedure distorts data in an uneven manner - frequent words are distorted less than infrequent words. Thus we demonstrate that this reinstatement of correlation between frequency and change scores is merely an artefactual consequence of the truncated SVD factorization.

3.2 Construct validity

Potential confounding factors can be addressed by comparing any experimental finding to a validated control condition. Here we validate the use of the shuffled condition as a proper control. To this end, the average change scores of words per decade in both the genuine and shuffled conditions are compared within each processing stage. In the genuine condition, words appear in different usage contexts between decades, while in the shuffled condition they do not, because the random shuffling creates a homogeneous corpus. Therefore, the validity of the control condition is established if: (a) the change scores are diminished as compared to the genuine condition; (b) change scores are uniform across decades (since decades are shuffled); (c) the variance of change scores is smaller that in the genuine condition. As seen in Fig. 3a, all these requirements are met by the control condition. Note that the change scores in the shuffled condition are all significantly positive, namely, meaning change allegedly exists in this control condition. This supports the claim that any measurement is significantly affected by unrelated noise.

Thus, we have established that the shuffled condition is a suitable control for meaning change. While validity was established for each of the processing stages, the most robust effect was seen for the PPMI representation, following by SVD and word counts.

3.3 Accounting for the frequency confound

In Section 3.1 we used the subsample control condition to establish the confounding effect of frequency on meaning change. We now examine the extent to which this frequency confound exists in a historical corpus. We do so by comparing the frequency confound between the genuine historical corpus and the shuffled historical corpus.

To visualize the frequency confound in a manner comparable to the analysis presented in Section 3.1, we again plot change scores vs. frequency, ignoring the time dimension of the data. Fig. 3b presents this plot for the genuine condition. The same analysis is repeated in the shuffled condition, see Fig. 3c.

Both plots reveal a highly significant correlation between change scores and frequency. Furthermore, the fact that the correlation coefficients are virtually identical in the genuine and shuffled conditions, with $r = -0.748$ and $r = -0.747$ respectively, suggests that they are due to artefactual factors in both conditions and not to true change of meaning over time. In fact, this pattern of results is reminiscent of the spurious pattern we see in Fig. 1c.

The relation between frequency and meaning change can also be represented by a linear mixed effect model, with the benefit that this model enables the addition of more explanatory variables to the data. The regression model found frequency to have a negative influence on change scores,

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

Table 1: Results of one-predictor and two-predictor regression analysis in all conditions.

with β_f=-0.91 and β_f=-0.75, for the genuine and shuffled conditions respectively. Importantly, frequency accounted for 67% of the variance in the change scores in the genuine condition, and was only slightly diminished in the shuffled condition, accounting for 56% of the variance. Similar results were obtained for the PPMI representation (see Table 1).

4 Revisiting previous studies

We replicated three recent results that were affected by this frequency effect, since they all define change as the word's cosine distance relative to itself at two time points. These studies report laws of semantic change that measure the role of frequency in semantic change either directly (Law of Conformity), or indirectly through another linguistic variable that is dependent on frequency (Laws of Innovation and Prototypicality).

4.1 Laws of conformity and innovation

Continuing the work described in Section 3.1, we replicated the model and analysis procedure described in (Hamilton et al., 2016), where **two predictors** were used together to explain the change scores: frequency and polysemy. Polysemy, which describes the number of different senses a word has, naturally differs among words, where some words are more polysemous than others (compare *bank* and *date* to *wine*). Following (Hamilton et al., 2016), we defined polysemy as the words' secondary connections patterns - the connections between each word's co-occurring words (using the entries in the PPMI representation for that word). The more interconnected these secondary connections are, the less polysemic a word is, and vice versa. Polysemy scores were computed using the authors' provided code[3]. We then log-transformed and standardized the polysemy scores. Next, frequency and polysemy were set as two fixed effect predictors in a linear mixed effect model, like the one described in Section 2.4.

Thus we were able to replicate the results in the genuine condition as reported in (Hamilton et al., 2016). Interestingly, the same pattern of results emerged, again, in the shuffled condition (see Table 1). Importantly, the difference in effect size between conditions, as evaluated by the explained variance of frequency and polysemy together, showed a modest effect of 8% over the shuffled condition, pointing to the conclusion that the putative effects may indeed be real, but to a far lesser extent than had been claimed. We conclude that adding polysemy to the analysis contributed very little to the model's predictive power.

Since the PPMI representation (the *explicit representation* without dimensionality reduction with SVD) seems much less affected by spurious effects correlated with frequency (see Fig. 1b), we repeated the analysis of frequency described here and in Section 3.1 while using this representation. The results are listed in Table 1, showing a similar pattern of rather small frequency effect.

4.2 Prototypicality

Prototypicality is the degree to which a word is representative of the category of which it is a member (a *robin* is a more prototypical bird than a *parrot*). According to the proposed Law of Prototypicality, words with more prototypical meanings will show less semantic change, and vice versa. Following (Dubossarsky et al., 2015), we computed words' prototypicality scores for each decade as the cos-distance between a word's vec-

[3]https://github.com/williamleif/histwords

tor and its k-means cluster's centroid, and extended the analysis to encompass the entire 20th century. The previous regression model assumed independence between words, and therefore assigned words to a random effect variable. However, when modeling prototypicality, this assumption is invalid as relations between words are what inherently define prototypicality. We therefore designed a model in which decades, rather than words, are the random effect variable.

With this analysis the prototypicality effect seems to be substantiated in two ways. First, the addition of prototypicality explains an additional 5% of the variance. Second, the effect of prototypicality meets the more stringent requirement of being diminished in the shuffle condition (see Table 1). Nevertheless, here too the effect originally reported was found to be drastically reduced after being compared with the proper control.

5 Theoretical analysis

We show in Section 5.1 that the average cosine distance between two vectors representing the same word is equivalent to the variance of the population of vectors representing the same word in independent samples, and is therefore always positive. This is true for any word vector representation.

In Sections 5.2-5.3 we prove that the average cosines distance between two *count* vectors representing the same word is negatively correlated with the frequency of the word, and positively correlated with the polysemy score of the word.

5.1 Sampling variability and the cos distance

Lemma 1. *Assume two random variables x, y of length $\|x\|_2 = \|y\|_2 = 1$, distributed iid with expected value μ and covariance matrix Σ. The expected value of the cosine distance between them is equal to the sum of the diagonal elements of Σ.*

Proof.

$$\begin{aligned} E(x-y)^2 =& E(x-\mu)^2 + E(y-\mu)^2 + \\ & 2E(x-\mu)(y-\mu) \\ =& 2\sum E(x_i-\mu_i)^2 = 2\sum Var(x_i) \\ E(x-y)^2 =& E(x^2) + E(y^2) - 2E(x\cdot y) \\ =& 2 - 2E\left(\frac{x\cdot y}{\|x\|_2\|y\|_2}\right) \\ =& 2E(cosDist(x,y)) \end{aligned}$$

It follows that

$$E(cosDist(x,y)) = \sum Var(x_i) \qquad (4)$$

□

Implication: The average cosine distance between two samples of the same random variables is directly related to the variance of the variable, or the sampling noise. This variance should be measured empirically whenever cosine distance is used, since only distances that are larger than the empirical variance can be relied upon to support significant observations.

5.2 Cos distance of count vectors: frequency

Next, we analyze the cosine distance between 2 iid samples from a normalized multinomial random variable. This distribution models the distribution of the count vector representation. Let $k_i,\ 1 \le i \le m$ denote the number of times word i appeared in the context of word w, and let m denote the size of the dictionary not including w. Let $n = \sum k_i$ denote the number of words in the count vector of w; n determines the word's frequency score. Assume that the counts are sampled from the distribution Multinomial$(n, \vec{p})$, namely

$$Prob(k_1, \cdots, k_m) = \binom{n}{k_1 \cdots, k_m} p_1^{k_1} \cdots p_m^{k_m}$$

Lemma 2. *The expected value of the cosine distance between two count vectors x, y sampled iid from this distribution is monotonically decreasing with n.*

Proof. By definition, $1 - E[cosDist(x,y)]$ equals

$$E\left[\frac{x\cdot y}{\|x\|_2\|y\|_2}\right] = \sum_i \left[E\frac{x_i}{\|x\|_2}\right]^2 = \sum_i E_i^2 \qquad (5)$$

We compute the expected value of E_i directly:

$$E_i = \sum_{(k_1,\cdots,k_m)} \frac{k_i}{\sqrt{\sum_j k_j^2}} \binom{n}{k_1 \cdots, k_m} p_1^{k_1} \cdots p_m^{k_m}$$

Using Taylor expansion:

$$\begin{aligned} \frac{k_i}{\sqrt{\sum_j k_j^2}} =& \frac{\frac{k_i}{n}}{\sqrt{(\sum_j \frac{k_j}{n})^2 - \sum_{l\neq j}\frac{k_j k_l}{n^2}}} \\ =& \frac{k_i}{n}\frac{1}{\sqrt{1 - \sum_{l\neq j}\frac{k_j k_l}{n^2}}} \\ =& \frac{k_i}{n}\left(1 + \frac{\varepsilon}{2} + O(\varepsilon^2)\right) \end{aligned} \qquad (6)$$

where $\varepsilon = \sum_{l \neq j} \frac{k_j k_l}{n^2}$.

The expected value of the 0-order term with respect to ε in (6) equals p_i, which is independent of n. We conclude the proof by focusing on the first order term with respect to ε in (6), to be denoted f_1, showing that its expected value is monotonically decreasing with n. Specifically:

$$f_1 = \sum_{\vec{k}} \sum_{l \neq j} \frac{k_i}{n} \frac{k_j}{n} \frac{k_l}{n} \binom{n}{k_1 \cdots, k_m} p_1^{k_1} \cdots p_m^{k_m}$$

We switch the summation order and compute each expression in the external sum, considering two cases separately: when $l \neq j \neq i$

$$\sum_{(k_1, \cdots, k_m)} \frac{k_i}{n} \frac{k_j}{n} \frac{k_l}{n} \binom{n}{k_1 \cdots, k_m} p_1^{k_1} \cdots p_m^{k_m} = \frac{n(n-1)(n-2)}{n^3} p_i p_j p_l$$

When $l \neq j = i$ w.l.g, we rewrite $k_i k_j = k_i(k_i - 1) + k_i$, and the sum above becomes $\frac{n(n-1)(n-2)}{n^3} p_i^2 p_l + \frac{n(n-1)}{n^2} p_i p_l$. Thus

$$f_1 = \frac{n-1}{n} p_i \left[\frac{n-2}{n} \sum_{l,j: l \neq j} p_j p_l + (1 - p_i) \right]$$

and it readily follows that f_1 is monotonically increasing with n.

Since n measures the frequency score of word w, it follows from (5) that the expected value of the cosine distance between two iid samples from the distribution of the count vector of w is monotonically decreasing with the word's frequency. □

5.3 Cos distance of count vectors: polysemy

We start our investigation of polysemy by modeling the distribution of the parameters of the multinomial distribution from which count vectors are sampled. A common prior distribution on the vector $\vec{p}^w$ in m-simplex, which defines the multinomial distribution generating the context of word w, is the Dirichlet distribution $f(\vec{p}^w; \vec{\alpha}^w) = f(p_1, \cdots, p_m; \alpha_1, \cdots, \alpha_m)$.

$\vec{\alpha}^w$ is a sparse vector of prior counts on all the words in the dictionary, by which the co-occurrence context of word w is modeled. We divide the set of none-zero indices of $\vec{\alpha}^w$ into two subsets: $i_1, \cdots, i_{m_0}$ correspond to the words which always appear in the context of w, while $j_1, \cdots, i_{m_1}$ correspond to the words which appear in the context of w in one given meaning. If w is polysemous and has two meanings, then there is a third set of indices $k_1, \cdots, k_{m_2}$ which correspond to the words appearing in the context of w in its second meaning. If w has more then two meanings, they can be modeled with additional sets of disjoint indices.

Lemma 3. *Under certain conditions specified in the proof, given two count vectors x, y sampled iid from the above distribution of w, the expected value of the cosine distance between them increases with the number of sets of disjoint indices which represent different meanings of w.*

Proof. We will prove that when w has two meanings, the expected value of the cosine distance is larger than in the case of a single meaning. The proof for the general case immediately follows.

Starting from (6) while keeping only the 0-order term in ε, it follows from the derivations in the proof of Lemma 2 that the expected cosine distance between two count vector samples of w, to be denoted M, is $1 - \sum p_i^2$. In our current model $\vec{p}$ is a random variable, and we shall compute the expected value of this random variable under the two conditions, when w has either one or two meanings.

We start by observing that, given the definition of the Dirichlet distribution, it follows that

$$\begin{aligned} E(p_i^2) &= Var(p_i) + E(p_i)^2 = \frac{\alpha_i(1+\alpha_i)}{\alpha_0(1+\alpha_0)} \\ \alpha_o &= \sum \alpha_i \\ \Longrightarrow M &= \sum E(p_i^2) = \frac{\alpha_0 + \sum \alpha_i^2}{\alpha_0(1+\alpha_0)} \end{aligned} \quad (7)$$

Considering the different sets of indices in isolation, let $\varphi_o = \sum_{i=i_1}^{i_{m_0}} \alpha_i$, $\varphi_1 = \sum_{i=j_1}^{j_{m_1}} \alpha_i$, and $\varphi_2 = \sum_{i=k_1}^{k_{m_2}} \alpha_i$. Let $\psi_o = \sum_{i=i_1}^{i_{m_0}} \alpha_i^2$, $\psi_1 = \sum_{i=j_1}^{j_{m_1}} \alpha_i^2$, and $\psi_2 = \sum_{i=k_1}^{k_{m_2}} \alpha_i^2$.

We rewrite (7) for the two conditions:

1. w has one meaning:

$$M^{(1)} = \frac{\varphi_0 + \varphi_1 + \psi_0 + \psi_1}{(\varphi_0 + \varphi_1)(1 + \varphi_0 + \varphi_1)}$$

2. w has two meanings:

$$M^{(2)} = \frac{\varphi_0 + \varphi_1 + \varphi_2 + \psi_0 + \psi_1 + \psi_2}{(\varphi_0 + \varphi_1 + \varphi_2)(1 + \varphi_0 + \varphi_1 + \varphi_2)}$$

With some algebraic manipulations, it can be shown that $M^{(1)} > M^{(2)}$ if the following holds:

$$\begin{aligned}&(\varphi_0+\varphi_1)^2\varphi_2+(\psi_0+\psi_1)\varphi_2^2 \\ &+2(\psi_0+\psi_1)(\varphi_0+\varphi_1)\varphi_2+(\psi_0+\psi_1)\varphi_2 \\ &+(\varphi_0+\varphi_1)(\varphi_2^2-\psi_2)>\psi_2(\varphi_0+\varphi_1)^2\end{aligned} \tag{8}$$

Thus when (8) holds, the average cosine distance between two samples of a certain word w gets larger as w acquires more meanings. □

(8) readily holds under reasonable conditions, e.g., when the prior counts for each meaning are similar (as a set) and much bigger than the prior counts of the joint context words (i.e., $\varphi_0 = \psi_0 = \varepsilon,\ \varphi_1 = \varphi_2,\ \psi_1 = \psi_2$).

6 Conclusions and discussion

In this article we have shown that some reported laws of semantic change are largely spurious results of the word representation models on which they are based. While identifying such laws is probably within the reach of NLP analyses of massive digital corpora, we argued that a more stringent standard of proof is necessary in order to put them on a firm footing. Specifically, it is necessary to demonstrate that any proposed law of change has to be observable in the genuine condition, but to be diminished or absent in a control condition. We replicated previous studies claiming to establish such laws, which propose that semantic change is negatively correlated with frequency and prototypicality, and positively correlated with polysemy. None of these laws - at least in their strong versions - survived the more stringent standard of proof, since the observed correlations were found in the control conditions.

In our analysis, the Law of Conformity, which claims a negative correlation between word frequency and meaning change, was shown to have a much smaller effect size than previously claimed. This indicates that word frequency probably does play a role - but a small one - in semantic change. According to the Law of Innovation, polysemy was claimed to correlate positively with meaning change. However, our analysis showed that polysemy is highly collinear with frequency, and as such, did not demonstrate independent contribution to semantic change. For similar reasons, the alleged role of prototypicality was diminished.

These results may be more consonant than previous ones with the findings of historical linguistics, as it is commonly assumed that the factors leading to semantic change are more diverse than purely distributional factors. For example, sociocultural, political, and technological changes are known to impact semantic change (Bochkarev et al., 2014; Newman, 2015). Furthermore, some regularities of semantic change have been imputed to 'channel bias', inherent biases of utterance production and interpretation on the part of speakers and listeners, e.g., (Moreton, 2008). As such, it would be surprising if word frequency, polysemy, and prototypicality were to capture *too high* a degree of variance. In other words, since semantic change may result from the interaction of many factors, small effects may be a priori more credible than large ones.

The results of our empirical analysis showed that the spurious effects of frequency were much weaker for the explicit PPMI representation unaugmented by SVD dimensionality reduction. We therefore conclude that the artefactual frequency effects reported are inherent to the type of word representations upon which these analyses are based. As the analytical proof in Section 5 demonstrates, it is count vectors that introduce an artefactual dependence on word frequency.

Intuitively, one might expect that the average value for the cosine distance between a given word's vector in any two samples would be 0. However, Lemma 1 above shows that this is not the case, and the average distance is the variance of the population of vectors representing the same word. This result is independent of the specific method used to represent words as vectors. Lemma 2 proves that the average cosine distance between two samples of the same word, when using count vector representations, is negatively correlated with the word's frequency. Thus, the role of frequency cannot be evaluated as an independent predictor in any model based on count vector representations. It remains for future research to establish whether other approaches to word representation, e.g. (Blei et al., 2003; Mikolov et al., 2013), have inherent biases.

While our findings may seem to be mainly negative, since they invalidate proposed laws of semantic change, we would like to point to the positive contribution made by articulating more stringent standards of proof and devising replicable control conditions for future research on language change based on distributional semantics representations.

References

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL*, pages 238–247.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022.

Vladimir Bochkarev, Valery Solovyev, and Sören Wichmann. 2014. Universals versus historical contingencies in lexical evolution. *Journal of The Royal Society Interface*, 11:1–23.

John A Bullinaria and Joseph P Levy. 2012. Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and svd. *Behavior research methods*, 44(3):890–907.

Haim Dubossarsky, Yulia Tsvetkov, Chris Dyer, and Eitan Grossman. 2015. A bottom up approach to category mapping and meaning change. In *NetWordS 2015 Word Knowledge and Word Usage*, pages 66–70.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2016. Verbs change more than nouns: A bottom up computational approach to semantic change. *Lingue e Linguaggio*, 1:5–25.

John Rupert Firth. 1957. *Papers in Linguistics 1934–1951*. Oxford University Press, London.

Lea Frermann and Mirella Lapata. 2016. A Bayesian model of diachronic meaning change. *TACL*, 4:31–45.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71. Association for Computational Linguistics.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of ACL*.

Martin Hilpert and Florent Perek. 2015. Meaning change in a petri dish: constructions, semantic vector spaces, and motion charts. *Linguistics Vanguard*, 1(1):339–350.

Adam Jatowt and Kevin Duh. 2014. A framework for analyzing semantic change of words across time. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 229–238.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of ACL*, pages 61–65.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2177–2185.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *TACL*, 3:211–225.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3111–3119.

Sunny Mitra, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal. 2014. That's sick dude!: Automatic identification of word sense change across different timescales. In *Proceedings of ACL*, pages 1020–1029.

Elliott Moreton. 2008. Analytic bias and phonological typology. *Phonology*, 25(1):83–127.

Shinichi Nakagawa and Holger Schielzeth. 2013. A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–142.

John Newman. 2015. Semantic shift. In Nick Rimer, editor, *The Routledge Handbook of Semantics*, pages 266–280. Routledge, New York.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 104–111. Association for Computational Linguistics.

Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

Derry Tanti Wijaya and Reyyan Yeniterzi. 2011. Understanding semantic change of words over centuries. In *Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web*, pages 35–40. ACM.

Yang Xu and Charles Kemp. 2015. A computational evaluation of two laws of semantic change. In *CogSci*.