Where New Words Are Born: Distributional Semantic Analysis of Neologisms and Their Semantic Neighborhoods

Maria Ryskina[1] **Ella Rabinovich**[2] **Taylor Berg-Kirkpatrick**[3]
David R. Mortensen[1] **Yulia Tsvetkov**[1]

[1]Language Technologies Institute, Carnegie Mellon University,
{mryskina,dmortens,ytsvetko}@cs.cmu.edu
[2]Department of Computer Science, University of Toronto, ella@cs.toronto.edu
[3]Computer Science and Engineering, University of California, San Diego, tberg@eng.ucsd.edu

Abstract

We perform statistical analysis of the phenomenon of *neology*, the process by which new words emerge in a language, using large diachronic corpora of English. We investigate the importance of two factors, semantic sparsity and frequency growth rates of semantic neighbors, formalized in the distributional semantics paradigm. We show that both factors are predictive of word emergence although we find more support for the latter hypothesis. Besides presenting a new linguistic application of distributional semantics, this study tackles the linguistic question of the role of language-internal factors (in our case, sparsity) in language change motivated by language-external factors (reflected in frequency growth).[1]

1 Introduction

Natural languages are constantly changing as the context of their users changes (Aitchison, 2001). Perhaps the most obvious type of change is the introduction of new lexical items, or *neologisms* (a process called "neology"). Neologisms have various sources. They are occassionally coined out of whole cloth (*grok*). More frequently, they are loanwords from another language (*tahini*), derived words (*unfriend*), or existing words that have acquired new senses (as when *web* came to mean 'World Wide Web' and then 'the Internet'). While neology has long been of interest to linguists (§2), there have been relatively few attempts to study it as a global, systemic phenomenon. Computational modeling and analysis of neology is the focus of our work.

What are the factors that predict neology? Certainly, social context plays a role. Close interaction between two cultures, for example, may result in increased borrowing (Appel and Muysken, 2006). We hypothesize, though, that there are other factors involved—factors that can be modeled more directly. These factors can be understood in terms of **supply** and **demand**.

Bréal (1904) introduced the idea that the distribution of words in semantic space tends towards uniformity. This framework predicts that new words would emerge where they would repair uniformity—where there was a space not occupied by a word. This could be viewed as supply-driven neology. Next, demand plays a role as well as supply (Campbell, 2013): new words emerge in "stylish" neighborhoods, corresponding to domains of discourse that are increasing in importance (reflected by the increasing frequency of the words in those neighborhoods).

We operationalize these ideas using distributional semantics (Lenci, 2018). To formalize the hypothesis of supply-driven neology for computational analysis, we measure **sparsity of areas in the word embedding space** where neologisms would later emerge. The demand-driven view of neology motivates our second hypothesis: **neighborhoods in the embedding space containing words rapidly growing in frequency** are more likely to produce neologisms. Both hypotheses are defined more formally in §3.

Having formalized our hypotheses in terms of word embeddings, we test them by comparing the distributions of the corresponding metrics for a set of automatically identified neologisms and a control set. Methodology of the word selection and hypothesis testing is detailed in §4. We discuss the results in §5, demonstrating evidence for both hypotheses, although the demand-driven hypothesis has more significant support.

2 Background

Neology Specific sources of neologisms have been studied: lexical borrowing (Taylor and Grant,

[1]The code and word lists are available at https://github.com/ryskina/neology

Proceedings of the Society for Computation in Linguistics (SCiL) 2020, pages 43-52.
New Orleans, Louisiana, January 2-5, 2020

2014; Daulton, 2012), morphological derivation (Lieber, 2017), blends or portmanteaus (Cook, 2012; Renner et al., 2012), clippings, acronyms, analogical coinages, and arbitrary coinages, but these studies have tended to look at neologisms atomistically, or to explicate the social conditions under which a new word entered a language rather than looking at neologisms in systemic context.

To address this deficit, we look back to the seminal work of Michel Bréal, who introduced the idea that words exist in a semantic space. His work implies that, other things being equal, the semantic distribution of words tends towards uniformity (Bréal, 1904). This is most explicit in his law of differentiation, which states that near synonyms move apart in semantic space, but has other implications as well. For example, this principle predicts that new words are more likely to emerge where they would increase uniformity. This could be viewed as supply-driven neology—new words appear to fill gaps in semantic space (to express concepts that are not currently lexicalized).

In linguistic literature neology is often associated with new concepts or domains of increasing importance (Campbell, 2013). Just as there are factors that predict where houses are built other than the availability of land, there are factors that predict where new words emerge other than the availability of semantic space. Demand, we hypothesize, plays a role as well as supply.

Most existing computational research on the mechanisms of neology focuses on discovering sociolinguistic factors that predict acceptance of emerging words into the mainstream language and growth of their usage, typically in online social communities (Del Tredici and Fernández, 2018). The sociolinguistic factors can include geography (Eisenstein, 2017), user demographics (Eisenstein et al., 2012, 2014), diversity of linguistic contexts (Stewart and Eisenstein, 2018) or word form (Kershaw et al., 2016). To the best of our knowledge, there is no prior work focused on discovering factors predictive of the emergence of new words rather than modeling their lifecycle. We model language-external processes indirectly through their reflection in language, thereby capturing phenomena evident of our hypotheses through linguistic analysis.

Distributional semantics and language change Word embeddings have been successfully used for different applications of the diachronic analysis of language (Tahmasebi et al., 2018). The closest task to ours is analyzing meaning shift (tracking changes in word sense or emergence of new senses) by comparing word embedding spaces across time periods (Kulkarni et al., 2015; Xu and Kemp, 2015; Hamilton et al., 2016; Kutuzov et al., 2018). Typically, embeddings are learned for discrete time periods and then aligned (but see Bamler and Mandt, 2017). There has also been work on revising the existing methodology, specifically accounting for frequency effects in embeddings when modeling semantic shift (Dubossarsky et al., 2017).

Other related questions where distributional semantics proved useful were exploring the evolution of bias (Garg et al., 2018) and the degradation of age- and gender-predictive language models (Jaidka et al., 2018).

3 Hypotheses

This section outlines the two hypotheses we introduced earlier from the linguistic perspective, formalized in terms of distributional semantics.

Hypothesis 1 *Neologisms are more likely to emerge in sparser areas of the semantic space.* This corresponds to the supply-driven neology hypothesis: we assume that areas of the space that contain fewer semantically related words are likely to give birth to new ones so as to fill in the 'semantic gaps'. Word embeddings give us a natural way of formalizing this: since semantically related words have been shown to populate the same regions in embeddings spaces, we can approximate semantic sparsity (or density) of a word's neighborhood as the number of word vectors within a certain distance of its embedding.

Hypothesis 2 *Neologisms are more likely to emerge in semantic neighborhoods of growing popularity.* Here we formalize our demand-driven view of neology, which assumes that growing frequency of words in a semantic area is a reflection of its growing importance in discourse, and that the latter is in turn correlated with emergence of neologisms in that area. In terms of word embeddings, we again consider nearest word vectors as the word's semantic neighbors and quantify the rate at which their frequencies grow over decades (formally defined in §4.4).

4 Methodology

Our analysis is based on comparing embedding space neighborhoods of neologism word vectors and neighborhoods of embeddings of words from an alternative set. Automatic selection of neologisms is described in §4.2, and in §4.4 we detail the factors we control for when selecting the alternative set. In §4.1 we describe the datasets used in our experiments. Our data is split into two large corpora, HISTORICAL and MODERN; we additionally require the HISTORICAL corpus to be split into smaller time periods so that we can estimate word frequency change rate. Embedding models are trained on each of the two corpora, as described in §4.3. We compare the neighborhoods in the HISTORICAL embedding space, but due to the nature of our neologism selection process, many neologisms might not exist in the HISTORICAL vocabulary. To locate their neighborhoods, we adapt an approach from prior work in diachronic analysis with word embeddings: we learn an orthogonal projection between HISTORICAL and MODERN embeddings to align the two spaces in order to make them comparable (see Hamilton et al., 2016), and use projected vectors to represent neologisms in the HISTORICAL space. Finally, §4.5 describes the details of hypothesis testing: statistics we choose to quantify our two hypotheses and how their distributions are compared.

4.1 Datasets

We use the Corpus of Historical American English (COHA, Davies, 2002) and the Corpus of Contemporary American English (COCA, Davies, 2008), large diachronic corpora balanced by genre to reflect the variety in word usage. COHA data is split into decades; we group COHA documents from 18 decades (1800-1989) to represent the HISTORICAL English collection and use full COCA 1990-2012 corpus as MODERN.

The obtained HISTORICAL split contains 405M tokens of 2M types, and MODERN contains 547M tokens of 3M types.[2]

4.2 Neologism selection

We rely on a usage-based approach to extract the set of neologisms for our analysis, choosing the words based on their patterns of occurrence in our datasets. It can be seen as an approximation to selecting words based on their earliest recorded use dates, as these dates are also determined based on the words' usage in historical corpora. This analogy is supported by the qualitative analysis of the obtained set of neologisms, as discussed in §6.

We limit our analysis to nouns, an open-class lexical category. We identify nouns in our corpora using a part-of-speech dictionary, collected from a POS-tagged corpus of English Wikipedia data (Wikicorpus, Reese et al., 2010), and select words that are most frequently tagged as 'NN'.

We additionally filter candidate neologisms to exclude words that occur more frequently in capitalized than lowercased form; this heuristic helps us remove proper nouns missed by the POS tagger.

We select a set of neologisms by picking words that are substantially more frequent in the MODERN corpus than in the HISTORICAL one. It is important to note that while we use the term "neologism," implying a word at the early stages of emergence, with this method we select words that have entered mainstream vocabulary in MODERN time but might have been coined prior to that. We consider a word w to be a neologism if its ratio $f_m(w)/f_h(w)$ is greater than a certain threshold; here $f_m(\cdot)$ and $f_h(\cdot)$ denote word frequencies (normalized counts) in MODERN and HISTORICAL data respectively. Empirically we set the frequency ratio threshold equal to 20.

We rank words satisfying these criteria by their frequency in the MODERN corpus and select the first 1000 words to be our neologism set; this is to ensure that we only analyze words that subsequently become mainstream and not misspellings or other artifacts of the data.

4.3 Embeddings

Our hypothesis testing process involves inspecting semantic neighborhoods of neologisms in the HISTORICAL embedding space. However, many neologisms are very infrequent or nonexistent in the HISTORICAL data, so we approximate their vectors in the HISTORICAL space by projecting their MODERN embeddings into the same coordinate axes.

We learn Word2Vec Skip-Gram embeddings[3] (Mikolov et al., 2013) of the two corpora

[2] Statistics accompanying the corpora state that entire COHA dataset contains 385M words, and COCA contains 440M words; we assume the discrepancy is explained by tokenization differences.

[3] Hyperparameters: vector dimension 300, window size 5, minimum count 5.

and use orthogonal Procrustes to learn the aligning transformation:

$$\mathbf{R} = \arg\min_{\mathbf{\Omega}} \|\mathbf{\Omega}\mathbf{W}^{(m)} - \mathbf{W}^{(h)}\|,$$

where $\mathbf{W}^{(h)}, \mathbf{W}^{(m)} \in \mathbb{R}^{|V|\times d}$ are the word embedding matrices learned on the HISTORICAL and MODERN corpora respectively, restricted to the intersection of the vocabularies of the two corpora (i.e. every word embedding present in both spaces is used as an anchor). To project MODERN word embeddings into the HISTORICAL space, we multiply them by the obtained rotation matrix $\mathbf{R}$.

4.4 Control set selection

To test our hypotheses, we collect an alternative set of words and analyze how certain statistical properties of their neighbors differ from those of neighbors of neologisms. At this stage it is important to control for non-semantic confounding factors that might affect the word distribution in the semantic space. One such factor is word frequency: it has been shown that embeddings of words of similar frequency tend to be closer in the embedding space (Schnabel et al., 2015; Faruqui et al., 2016), which results in very dense clusters, or hubs, of words with high cosine similarity (Radovanović et al., 2010; Dinu et al., 2014). We choose to also restrict our control set to only include words that did not substantially grow or decline in frequency over the HISTORICAL period in order to prevent selecting counterparts that only share similar frequency in the MODERN subcorpus (e.g., due to recent topical relevance), but exhibit significant fluctuation prior to that period. In particular, we refrain from selecting words that emerged in language right before our HISTORICAL-MODERN split.

We create the alternative set by pairing each neologism with a non-neologism counterpart that exhibits a stable frequency pattern, while controlling for word frequency and word length in characters. Length is chosen as an easily accessible correlate to other factors for which one should control, such as morphological complexity, concreteness, and nativeness. We perform the pairing only to ensure that the distribution of those properties across the two sets is comparable, but once the selection process is complete we treat control words as a set rather than considering them in pairs with neologisms.

Following Stewart and Eisenstein (2018), we formalize frequency growth rate as the Spearman correlation coefficient between timesteps $\{1, \ldots, T\}$ and frequency series $f_{(1:T)}(w)$ of word w. In our setup, timesteps $\{1, \ldots, 18\}$ enumerate decades from 1810s to 1980s, and $f_t(\cdot)$ denote word frequencies in the corresponding t-th decade of the HISTORICAL data.

Formally, for each neologism w_n we select a counterpart w_c satisfying the following constraints:

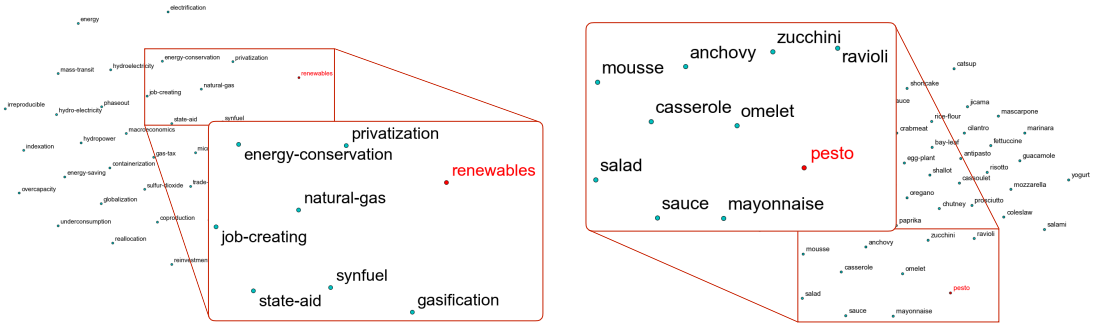
- Frequencies of the two words in the corresponding corpora are comparable: $f_m(w_n)/f_h(w_c) \in (1-\delta, 1+\delta)$, where δ was set to 0.25;
- The length of the two words is identical up to 2 characters;
- The Spearman correlation coefficient r_s between decades $\{1, \ldots, 18\}$ and the control word frequency series $f_{(1:18)}(w_c)$ is small: $|r_s\left(\{1:18\}, f_{(1:18)}(w_c)\right)| \leq 0.1$

These words, which we will refer to as *stable*, make up our default and most restricted control set. We will also compare neologisms to a *relaxed* control set, omitting the stability constraint on the frequency change rate but still controlling for length and overall frequency, to see how neologisms differ from non-neologisms in a broader perspective.

4.5 Experimental setup

We evaluate our hypotheses by inspecting neighborhoods of neologisms and their stable control counterparts in the HISTORICAL embedding space, viewing them as proxy for neighborhoods in the underlying semantic space. Since many neologisms are very infrequent or nonexistent in the HISTORICAL data, we approximate their vectors in the HISTORICAL space with their MODERN embeddings projected using the transformation described in §4.3. The neighborhood of a word w is defined as the set of HISTORICAL words for which cosine similarity between their HISTORICAL embeddings and v_w exceeds the given threshold τ; v_w denotes a projected MODERN embedding if w is a neologism or a HISTORICAL embedding if it is a control word.[4]

[4]Cosine similarity is chosen as our distance metric since it is traditionally used for word similarity tasks in distributional



(a) Semantic neighborhood of the word *renewables*.

(b) Semantic neighborhood of the word *pesto*.

Figure 1: Neighborhoods of projected MODERN embeddings of two neologisms (shown in red), *renewables* and *pesto*, in the HISTORICAL embedding space, visualized using t-SNE (Maaten and Hinton, 2008). Figure 1a shows an example of a neighborhood exhibiting frequency growth: words like *synfuel* or *privatization* have been used more towards the end of the HISTORICAL period. The neighborhood also includes *natural-gas* that can be seen as representing a concept to be replaced by *renewables*. The word *pesto* (Figure 1b) is projected into a neighborhood of other food-related words, most of which are also loanwords, several from the same language; it also has its hypernym *sauce* as one of its neighbors.

The two factors we need to formalize are semantic sparsity of the neighborhoods and increase of popularity of the topic that the neighborhood represents. We use sparsity in the embedding space as a proxy for semantic sparsity and approximate growth of interest in a topic with frequency growth of words belonging to it (i.e. embedded into the corresponding neighborhood). For the neighborhood of each word w, we compute the following statistics, corresponding to our two hypotheses:

1. *Density of a neighborhood* $d(w, \tau)$: number of words that fall into this neighborhood $d(w, \tau) = |\{u : \text{cosine}(v_w, v_u) \geq \tau\}|$

2. *Average frequency growth rate of a neighborhood* $r(w, \tau)$: as defined in the previous subsection, we compute the Spearman correlation coefficient between timesteps and frequency series for each word in the neighborhood and take their mean:

$$r(w,\tau) = \frac{1}{d(w,\tau)} \times \\ \times \sum_{u:\text{cosine}(v_w,v_u)\geq\tau} r_s\left(\{1:18\}, f_{(1:18)}(u)\right)$$

In our tests, we compare the values of those metrics for neighborhoods of neologisms and neighborhoods of control words and estimate the significance of each of the two factors for a range of neighborhood sizes defined by the threshold τ. We test whether means of the distributions of those statistics for the neologism and the control set differ and whether each of the two is significant for classifying words into neologisms and controls.

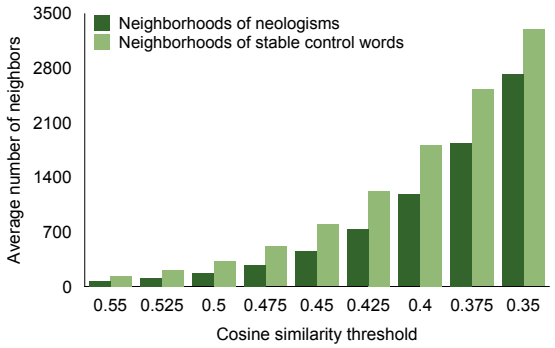
As mentioned in §4.2, our vocabulary is restricted to nouns, and we only consider vocabulary noun neighbors when evaluating the statistics.[5] Since we project all neologism word vectors from MODERN to HISTORICAL embedding space, for neologisms occurring in the HISTORICAL corpus we might find a HISTORICAL vector of the neologism itself among the neighbors of its projection; we exclude such neighbors from our analysis. We cap the number of nearest neighbors to consider at 5,000, to avoid estimating statistics on overly large sets of possibly less relevant neighbors.

5 Results

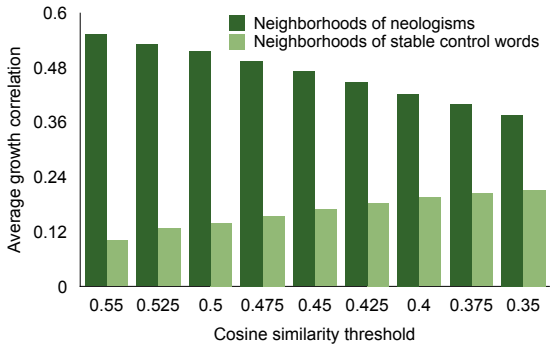
Following the experimental setup described in §4.5, we estimate the contribution of each of the hypothesized factors employing strictly constrained and relaxed control sets. We start by analyzing how the distributions of those statistics differ for neologisms and stable controls, both by

semantics (Lenci, 2018). We have also observed the same results when repeating the experiments with the Euclidean distance metric.

[5] Here we refer to the vocabulary of words participating in our analysis, not the embedding model vocabulary; embeddings are trained on the entire corpora.



(a) Average HISTORICAL word vector density in the neighborhoods of neologisms and stable control set words.



(b) Average frequency growth rate of HISTORICAL word vectors in the neighborhoods of neologisms and stable control set words.

Figure 2: Number of HISTORICAL word vectors within a certain cosine distance of a word and average growth rate of frequency (represented by Spearman correlation coefficient) of those HISTORICAL words, averaged across neologism (darker) and stable control word (lighter) sets. Projected neologism vectors appear in lower-density neighborhoods compared to control words, and neighbors of neologisms exhibit a stronger growth trend than those of the control words, especially in smaller neighborhoods.

comparing their sample means and by more rigorous statistical testing. We also evaluate the significance of the factors using generalized linear models for both stable and relaxed control sets.

5.1 Comparison to stable control set

First, we test our hypotheses on 720 neologism-stable control word pairs (not all words are paired in the stable control setting due to its restrictiveness).

Figure 2 demonstrates the values of density and frequency growth rate for a range of neighborhood sizes, averaged over neologism and control sets. Both results conform with our hypotheses: Figure 2a shows that on average the projected neologism has fewer neighbors than its stable counterpart, especially for larger neighborhoods, and Figure 2b shows that, on average, frequencies of neighbors of a projected neologism grow at a faster rate than those of a counterpart. Interestingly, we find that neighbors of stable controls still tend to exhibit small positive growth rate. We attribute it to the general pattern that we observed: about 70% of words in our vocabulary have positive frequency growth rate. We believe this might be explained by the imbalance in the amount of data between decades (e.g. 1980s sub-corpus has 20 times more tokens than 1810s): some words might not occur until later in the corpus because of the relative sparsity of data in the early decades.

As we can see from Figure 2a, neighborhoods of larger sizes (corresponding to lower values of the threshold) may contain thousands of words, so the statistics obtained from those neighborhoods might be less relevant; we might only want to consider the immediate neighborhoods, as those words are more likely to be semantically related to the central word. It is notable that the difference in the growth trends of the neighbors is substantially more prominent for smaller neighborhoods (Figure 2b): average correlation coefficient of immediate neighbors of stable words also falls into stable range as we defined it, while immediate neighbors of neologisms exhibit rapid growth.

5.2 Statistical significance

To estimate the significance and relative contribution of the two factors, we fit a generalized linear model (GLM) with logistic link function to the corresponding features of neologism and control word neighborhoods:[6]

$$y(w) \sim (1 + \exp(-\beta_0^{(\tau)} - \\ - \beta_d^{(\tau)} \cdot d(w,\tau) - \beta_r^{(\tau)} \cdot r(w,\tau)))^{-1}$$

where y is a Bernoulli variable indicating whether the word w belongs to the neologism set (1) or the control set (0), and τ is the cosine similarity threshold defining the neighborhood size.

Table 1 shows how the coefficients and p-values for the two statistics change with the neighborhood size. We found that when comparing with

[6]We use the implementation provided in the MATLAB Statistics and Machine Learning Toolbox.

Neighborhood size	Stable control set				Relaxed control set			
	Density		Growth		Density		Growth	
	$\beta_d^{(\tau)} \times 10^4$	p-value	$\beta_r^{(\tau)} \times 10$	p-value	$\beta_d^{(\tau)} \times 10^4$	p-value	$\beta_r^{(\tau)}$	p-value
Large ($\tau = 0.35$)	1.98	8.25×10^{-5}	1.84	2.35×10^{-80}	-1.07	5.63×10^{-4}	0.61	2.83×10^{-34}
Medium ($\tau = 0.45$)	0.20	8.29×10^{-1}	1.16	2.92×10^{-80}	-3.67	4.00×10^{-10}	0.46	6.19×10^{-46}
Small ($\tau = 0.55$)	6.90	2.90×10^{-2}	0.70	1.61×10^{-68}	-8.92	4.01×10^{-5}	0.28	1.19×10^{-36}

Table 1: Values of the GLM coefficients and their p-values for different neighborhood cosine similarity thresholds τ. $\beta_d^{(\tau)}$ and $\beta_r^{(\tau)}$ denote the coefficients for density and average frequency growth respectively for neighborhoods defined by τ. Comparing the results for the stable and relaxed control sets, we find that for the stable controls density is only significant in larger neighborhoods, but without the stability constraint both factors are significant for all neighborhood sizes.

the stable control set, average frequency growth rate of the neighborhood was significant for all sizes, but neighborhood density was significant at level $p < 0.01$ only for the largest ones.[7] We attribute this to the effect discussed in the previous section: difference in average frequency growth rate between neighbors of neologisms and stable words shrinks as we include more remote neighbors (Figure 2b), so for large neighborhoods frequency growth rate by itself is no longer predictive enough.

We also evaluate the significance of features for the relaxed control set without the stability constraint on 1000 neologism-control pairs. We have repeated the experiment with 5 different randomly sampled relaxed control sets (results for one showed in Table 1). For medium-sized neighborhoods ($0.4 \leq \tau \leq 0.5$) density variable is always significant at $p < 0.01$, but densities of largest and smallest neighborhoods were rejected in several runs. With more variance in the control set, differences in neighborhood frequency growth rate between neologisms and controls are less prominent than in the stable setting, so density plays a more important role in prediction.[8]

Growth feature weights $\beta_r^{(\tau)}$ are always positive and density feature weights $\beta_d^{(\tau)}$ are negative in the relaxed setting (where density is significant). This matches our intuition that neighborhood frequency growth and sparsity are predictive of neology.

Comparing sample means of density and growth rates between neologisms and each of the 5 randomly selected relaxed control sets (as we did for stable controls in Figure 2) demonstrated that neologisms still appear in sparser neighborhoods than the controlled counterparts. The difference in frequency growth rate between the neologism and control word neighborhoods is also observed for all control sets (although it varies noticeably between sets), but it no longer exhibits an inverse correlation with neighborhood size.

6 Discussion

We have demonstrated that our two hypotheses hold for the set of words we automatically selected to represent neologisms. To establish validity of our results, we qualitatively examine the obtained word list to see if the words are in fact recent additions to the language. We randomly sample 100 words out of the 1000 selected neologisms and look up their earliest recorded use in the Oxford English Dictionary Online (OED, 2018). Of those 100 words, eight are not defined in the dictionary: they only appear in quotations in other entries (*bycatch* (quotation from 1995), *twentysomething* (1997), *cross-sex* (1958), etc.) or do not occur at all (*all-mountain*, *interobserver*, *off-task*). Of the remaining 92 words, 78 have been first recorded after the year 1810 (i.e. since the beginning of the HISTORICAL timeframe), 44 have been first recorded in the twentieth century, and 21 words since 1950. However, some of the words dating back to before 19th century have only been recorded in their earlier, possibly obsolete sense: for example, while there is evidence of the word *software* being used in 18th century, this usage corresponds to its obsolete meaning of 'textiles, fabrics', while the first recorded use in its currently dominant sense of 'programs essential to the operation of a computer system' is dated 1958. To account for such semantic neologisms, we can count

[7] Applying Wilcoxon signed-rank test to the series of neighborhood density and frequency growth values for neologism and stable control sets showed the same results.

[8] Detailed results of the regression analysis and collinearity tests can be found in the repository. No evidence of collinearity was found in any of the experiments.

the first recorded use of the newest sense of the word; that gives us 82, 58 and 31 words appearing since 1810, 1900 and 1950 respectively.[9] This leads us to assume that most words selected for our analysis have indeed been neologisms sometime over the course of the HISTORICAL time.

We would also like to note that the results of this examination may be skewed due to factors for which lexicography may not account: for example, many words identified as neologisms are compound nouns like *countertop* or *soundtrack* that have been written as two separate words or joined with a hyphen in earlier use. There is also considerable spelling variation in loanwords, e.g. *cuscusu*, *cooscoosoos*, *kesksoo* were used interchangeably before the form *couscous* was accepted as the standard spelling. Specific word forms might also have different life cycles: while the word *music* existed in Middle English, the plural form *musics* in a particular sense of 'genres, styles of music' is much more recent.

Qualitative examination of the neologism set reveals that new words tend to appear in the same topics; for example, many words in our set were related to food, technology, or medicine. This indirectly supports our second hypothesis: rapid change in these spheres makes it likely for related terms to substantially grow in frequency over a short period of time. One example of such a neighborhood is shown in Figure 1a: the neologism *renewables* appeared in a cluster of words related to energy sources — a topic that has been more discussed recently. There is also some correlation between the topic and how new words are formed in it: most food neologisms are so-called cultural borrowings (Weinreich, 2010), when the name gets loaned from another culture together with the concept itself (e.g. *pesto*, *salsa*, *masala*), while many technology neologisms are compounds of existing English morphemes (e.g. *cyber+space*, *cell+phone*, *data+base*).

We also consider nearest neighbors (HISTORICAL words with highest cosine similarity) of the neologisms to ensure that they are projected into the appropriate parts of the embedding space. Examples of nearest neighbors are shown in Table 2. We saw different patterns of how the concept represented by the neologism relates to concepts represented by its neighbors. For example, some terms for new concepts appear next to related concepts they succeeded and possibly made obsolete: e.g. *email:letter*, *e-book:paperback*, *database:card-index*. Other neologisms emerge in clusters of related concepts they still equally coexist with: *hip-hop:jazz*, *hoodie:turtleneck*; most cultural borrowings fall under this type (see the neighborhood of *pesto* in Figure 1b). Both those patterns can be viewed as examples of a more general trend: one concept takes place of another related one, whether in terms of fully replacing it or just taking its place as the dominant form.

Neologism	Nearest HISTORICAL neighbors	
email	telegram	letter
pager	beeper	phone
blogger	journalist	columnist
sitcom	comedy	movie
spokeswoman	spokesman	director
sushi	caviar	risotto
rehab	detoxification	aftercare

Table 2: Nearest HISTORICAL neighbors of projected MODERN embeddings for a sample of emerging words. We can see that words get projected into semantically relevant neighborhoods, and nearest neighbors can even be useful for observing the evolution of a concept (e.g. *pager:beeper*).

Other interesting effects we observed include lexical replacement (a new word form replacing an old one without a change in meaning, e.g. *vibe:ambience*), tendency to abbreviate terms as they become mainstream (*biotech:biotechnology*, *chemo:chemotherapy*), and the previously mentioned changes in spellings of compounds (*lifestyle:life-style*, *daycare:day-care*).

7 Conclusion

We have shown that our two hypothesized factors, semantic neighborhood sparsity and its average frequency growth rate, play a role in determining in what semantic neighborhoods new words are likely to emerge. Our analyses provide more support for the latter, conforming with prior linguistic intuition of how language-external factors (which this factor implicitly represents) affect language change. We also found evidence for the former, although it was found less significant.

Our contributions are manifold. From a computational perspective, we extend prior research

[9] For all words that have one or more senses marked as a noun, we only consider those senses. Out of the 92 listed words, only three do not have nominal senses, and for two more usage as a noun is marked to be rare.

on meaning change to a new task of analyzing word emergence, proposing another way to obtain linguistic insights from distributional semantics. From the point of view of linguistics, we approach an important question of whether language change is affected by not only language-external factors but language-internal factors as well. We show that internal factors—semantic sparsity, specifically—contribute to where in semantic space neologisms emerge. To the best of our knowledge, our work is the first to use word embeddings as a way of quantifying semantic sparsity. We have also been able to operationalize one kind of external factor, technological and cultural change, as something that can been measured in corpora and word embeddings, paving the way to similar work with other kinds of language-external factors in language change.

An admittable limitation of our analysis lies in its restricted ability to account for polysemy, which is a pervasive issue in distributional semantics studies (Faruqui et al., 2016). As such, semantic neologisms (existing words taking on a novel sense) were not a subject of this study, but they introduce a potential future direction. Additional properties of word's neighbors can also be correlated with word emergence, both language-internal (word abstractness or specificity) and external; these can also be promising directions for future work. Finally, our future plans include exploration of how features of semantic neighborhoods are correlated with word obsolescence (gradual decline in usage), using similar semantic observations.

Acknowledgments

We thank the BergLab members for helpful discussion, and the anonymous reviewers for their valuable feedback. This work was supported in part by NSF grant IIS-1812327.

References

Jean Aitchison. 2001. *Language Change: Progress Or Decay?* Cambridge University Press.

René Appel and Pieter Muysken. 2006. *Language contact and bilingualism*. Amsterdam University Press.

Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings. In *International Conference on Machine Learning*, pages 380–389.

Michel Bréal. 1904. *Essai de sémantique:(science des significations)*. Hachette.

Lyle Campbell. 2013. *Historical Linguistics: an Introduction*. MIT Press, Cambridge, MA.

Paul Cook. 2012. Using social media to find English lexical blends. In *Proceedings of the 15th EURALEX International Congress (EURALEX 2012)*, pages 846–854, Oslo, Norway.

Frank E. Daulton. 2012. Lexical borrowing. In *The Encyclopedia of Applied Linguistics*. American Cancer Society.

Mark Davies. 2002. *The Corpus of Historical American English (COHA): 400 million words, 1810-2009*. Brigham Young University.

Mark Davies. 2008. *The corpus of contemporary American English*. BYE, Brigham Young University.

Marco Del Tredici and Raquel Fernández. 2018. The road to success: Assessing the fate of linguistic innovations in online communities. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1591–1603.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2014. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1136–1145.

Jacob Eisenstein. 2017. Identifying regional dialects in on-line social media. *The Handbook of Dialectology*, pages 368–383.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2012. Mapping the geographical diffusion of new words. In *NIPS Workshop on Social Network and Social Media Analysis*.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2014. Diffusion of lexical change in social media. *PloS one*, 9(11):e113114.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 30–35.

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

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1489–1501.

Kokil Jaidka, Niyati Chhaya, and Lyle Ungar. 2018. Diachronic degradation of language models: Insights from social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 195–200.

Daniel Kershaw, Matthew Rowe, and Patrick Stacey. 2016. Towards modelling language innovation acceptance in online social networks. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 553–562. ACM.

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. International World Wide Web Conferences Steering Committee.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397.

Alessandro Lenci. 2018. Distributional models of word meaning. *Annual review of Linguistics*, 4:151–171.

Rochelle Lieber. 2017. Derivational morphology.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Michael Proffitt, editor. 2018. *OED Online*. Oxford University Press. http://www.oed.com/.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(Sep):2487–2531.

Samuel Reese, Gemma Boleda, Montse Cuadros, Lluís Padró, and German Rigau. 2010. Wikicorpus: A word-sense disambiguated multilingual Wikipedia corpus. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Vincent Renner, Franois Maniez, and Pierre Arnaud, editors. 2012. *Cross-disciplinary perspectives on lexical blending*. De Gruyter Mouton, Berlin.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307.

Ian Stewart and Jacob Eisenstein. 2018. Making" fetch" happen: The influence of social and linguistic context on nonstandard word growth and decline. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4360–4370.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of computational approaches to diachronic conceptual change. *arXiv preprint arXiv:1811.06278*.

John R Taylor and Anthony P. Grant. 2014. *Lexical Borrowing*. Oxford University Press, Oxford.

Uriel Weinreich. 2010. *Languages in contact: Findings and problems*. Walter de Gruyter, The Hague.

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