HEBREW UNIVERSITY OF JERUSALEM

Semantic change at large: A computational approach for semantic change research

Thesis for the degree
"Doctor of Philosophy"
in
Brain Sciences: Computation and Information Processing

by:
Haim DUBOSSARSKY

Submitted to the Senate of the Hebrew University of Jerusalem

April 2018

This work was carried out under
the supervision of
Doctor Eitan GROSSMAN
and
Professor Daphna WEINSHALL

HEBREW UNIVERSITY OF JERUSALEM

Abstract

Edmond and Lily Safra Center for Brain Sciences

Doctor of Philosophy

Semantic change at large:
A computational approach for semantic change research

by Haim DUBOSSARSKY

The search for hitherto undiscovered regularities and laws of semantic change is today made possible by two recent and independent developments: first, the upsurge in the availability of digitized historical corpora of texts; and second, novel computational methods that allow the automatic semantic processing of these corpora. Compared to traditional research methods in semantic change, this dynamic duo of massive corpora and innovative computational methods has the potential to lead to novel insights about semantic change, both in discovering regular patterns of change and in identifying their explanatory factors, based on large-scale data-driven analysis.

In a series of studies, I demonstrate that this approach is not only feasible, but also fruitful for semantic change research. My research papers show that this approach extends the scope of research beyond known phenomena of semantic change, uncovering several regularities of semantic change, and complements existing theories with more objective and reliable analyses.

However, these studies have also demonstrated the importance of methodological issues. As this field of research has just emerged, fundamental methodological concerns were raised which must be addressed in order to make an objective, reliable, and genuine contribution to the research field.

Two papers were dedicated to tackle these methodological issues, and their results promise to put this computational approach on a solid methodological footing right from the beginning. The studies not only call into question earlier results, but also provide clear, feasible, and replicable validation routines to ensure objective and reliable testing. Importantly, these methodological accentuations may benefit the NLP community at large, as they have applications that go beyond semantic change research.

Letter of Contribution

Chapter 2 - Verbs change more than nouns: A bottom-up computational approach to semantic change.

Haim Dubossarsky, Daphna Weinshall and Eitan Grossman.

HD proposed the idea for the main analysis, processed the corpora, trained the model, and performed the critical analysis. DW supervised the analyses. HD and EG interpreted the results and proposed a linguistic theory for them. HD and EG wrote the manuscript: HD the technical parts to communicate them to the linguistic community, and EG the linguistic-theoretical parts.

Chapter 3 - A bottom up approach to category mapping and meaning change.

Haim Dubossarsky, Yulia Tsvetkov, Chris Dyer and Eitan Grossman

HD proposed the theoretical research question. YT and CD positioned it in the right computation-theoretic background, and suggested the model that would allow to investigate this question. HD processed the corpora, trained the model, made the diachronic comparisons, and devised the critical analysis. HD, YT and CD interpreted the results on the computational side, and EG contributed the pivotal linguistic interpretation. HD and EG wrote the manuscript, and YT wrote part of the methodology.

Chapter 4 - Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models.

Haim Dubossarsky, Eitan Grossman and Daphna Weinshall.

HD speculated about the magnitude of the problem in the current research, and together with DW outlined the required experiments to test it as a research question. HD planned and carried out the experiments, analyzed their results, replicated the results of previous studies, and conducted the critical comparisons. HD interpreted the results together with DW. DW contributed to the theoretical part. HD, DW and EG wrote the manuscript, and emphasize its critical importance to the research field.

Chapter 5 - Coming to Your Senses: on Controls and Evaluation Sets in Polysemy Research.

Haim Dubossarsky, Matan Ben-Yosef, Eitan Grossman and Daphna Weinshall.

HD identified the misconception in the current research, and suggested to phrase it as a research hypothesis. HD and DW devised the set of experiments to investigate the hypothesis. HD collected and preprocessed the corpora, trained the models, and conducted the critical comparisons. HD and MB replicated the results of previous models, and manipulating their code for the critical simulations. HD analyzed the results from all the experiments and simulations. DW contributed to the theoretical part. HD and DW interpreted the results, and drew conclusions. HD, DW and EG wrote the manuscript, and positioned the conclusions in the right research context.

"In the beginning was the Word, and the Word was with God, and the Word was God."

John, 1:1.

Contents

Chapter 1

Introduction

At the heart of this PhD dissertation is a new research paradigm for the study of semantic change, i.e., changes in the meaning of lexical items. This paradigm is characterized by a focus on large-scale, data-driven analyses based on modern computational tools, in order to address research questions stemming from linguistic research. This approach complements existing paradigms, and in so doing, enriches the toolbox of semantic change research. At the same time, this dissertation addresses fundamental methodological issues that are crucial not only for the emerging field of NLP research on semantic change, but for research in NLP more broadly. In particular, this dissertation constitutes a contribution to methodological issues inherent to research in the framework of DISTRIBUTIONAL SEMANTICS.

The structure of this chapter is as follows. Section 1.1 provides a definition of semantic change. Section 1.2 presents essential background about semantic change research. Section 1.3 states our research objectives in light of the needs that are derived from current research state. Section 1.4 briefly describes our research papers, and how they contribute to the research field in accordance with the objectives outlined before.

1.1 Semantic change

The meanings of words – or, more precisely, lexical items – are liable to change over time. For example, *girl* originally used to denote a child of either sex, but since the 15th century only refers to a young female (Bybee 2015, p. 202). *Broadcast* used to describe a method of scattering seeds less than a hundred years ago, but nowadays refers to the transmission of information by radio or television. The phenomenon of changes in the meaning of words over time is part of what is called *semantic change*, and is known at least since the work of Reisig (1839). Semantic change, to use a classic definition, may

be described as "innovations which change the lexical meaning rather than the grammatical function of a form" (Bloomfield 1933, p. 425)[1].

A formal definition of semantic change would be as follows: given a historical corpus that contains texts from n different time periods in a chronological order $[\mathcal{C}_{t1}, \mathcal{C}_{t2}, ..., \mathcal{C}_{tn}]$, semantic change is any difference between the meanings associated with a given word at *t1*, *t2* ... *tn*. A model of semantic change aims first of all to discover which words changed their meaning between any two time-periods. Given this information, additional tasks can be undertaken in order to answer further questions, such as the size or extent of the changes, their characteristics, and their dependence on other factors.

1.2 Research paradigms in semantic change

1.2.1 Overview

Until recently, research on semantic change was largely based on a traditional paradigm that used philological tools informed by linguistic analysis. For the most part, linguists interested in semantic change have studied changes in the denotation or connotation of individual words or constructions (or small groups thereof) as they occur in historical corpora. The introduction of modern quantitative and computational tools to the field did not change the basically TOP-DOWN nature of this research paradigm.

While a top-down research approach allows a fine-grained examination of semantic change, it also imposes limitations on what one can learn about semantic change. The fundamental limitation is the size and the representativeness of the dataset to be generalized over. Since previous studies were done on a SMALL SCALE and were based on ultimately anecdotal instances of semantic change, it is not at all certain to what extent the semantic changes collected in surveys are representative of semantic change in general. Furthermore, analyses based on small and anecdotal datasets cannot turn up large-scale regularities of semantic change.

As a consequence of their top-down and small-scale nature, these analyses tend to draw conclusion on a subjective and qualitative evaluation of a limited number of examples, and rarely deal with counterexamples. As such, statistical methodologies are rarely brought to bear in support of theoretical claims, and the importance of statistical significance in hypotheses

[1] For several overviews of semantic change, see, e.g., Bybee (2015, pp. 188-208), Hock and Joseph (2009, pp. 205-240), Newman (2015, pp. 266-280), or Traugott and Dasher (2002).

testing is somewhat foreign to the field. On the other hand, large-scale, bottom-up analyses demand STATISTICALLY-RIGOROUS research methods in order to reach sound and reliable findings. However, although it is clear that statistically-rigorous research methods are crucial for large-scale, bottom-up research, the methods themselves – and their application to NLP analyses – are not self-evident, and they require dedicated research.

Over the years, several research paradigms in semantic change research have developed. These paradigms differ in their research methods with respect to the three distinctions that were made above, namely (i) whether they are top-down or bottom-up, (ii) small or large scale, and (iii) whether they employ statistically-rigorous methods. While it may seem these paradigms show a clear developmental trajectory, as some have preceded the others and the later ones may improve on the older ones, the boundaries between them may be obscure. Importantly, they are all still used in contemporary research.

In the following paragraphs, I briefly survey the main research paradigms for semantic change, comparing them according to the distinctions made above (see Table 1.1 for a summary).

Paradigm	**Research type**	**Research design**	**Scale**	**Statistical methods**
TRADITIONAL	Qualit.	Top-down	Small	None
QUANTITATIVE	Quantit.	Top-down	Small, but full corpora	Simple statistical tests
COMPUTATIONAL	Quantit.	Bottom-up	Initially small, then large	Initially none, then rigorous

TABLE 1.1: Comparative summary of the research paradigms

1.2.2 The traditional paradigm

Many if not most studies of semantic change have focused primarily on describing and classifying the major types of change in meaning into taxonomies. This could be called the TRADITIONAL paradigm. According to most taxonomies, the major types of change involve (a) changes in the extension of meaning, i.e., either *widening* or *narrowing*, and (b) changes in the connotation of meaning, i.e., which may become either more positive (*amelioration*), or more negative (*pejoration*). See the examples in Table 1.2, taken from English:

Widening	*bird*	'young bird' > 'any kind of bird'
Narrowing	*meat*	'food' > 'edible flesh'
Amelioration	*knight*	'servant boy' > 'nobleman'
Pejoration	*awful*	'inspiring wonder' > 'terribly bad'

TABLE 1.2: Major taxonomies of semantic change

This paradigm is characterized by a top-down approach, where examples of individual words that underwent a certain type of semantic change are searched for in historical corpora and are either classified according to an existing taxonomy, or used to define a new one. The traditional paradigm provides us with invaluable source for understanding potential regularities and tendencies of semantic change, upon which principled theoretical work could be based, and it is an active and valued research paradigm until today (see, for example, the large database of semantic changes documented by Zalizniak et al. (2012)). Theoretical work in this vein is mostly limited to identifying the mechanisms that may explain *how* semantic change came into being, i.e., the claim that a metaphorical process has led *broadcast* to change its meaning from "casting seeds" into "casting audio and video signals".

However, knowing *how* the meaning of words changes diachronically is not equivalent to explaining *why* such changes occur. In this respect, this paradigm does not usually generate explanatory theories of semantic change. In fact, what it considers as *causes* for semantic change are actually *motivations* thereof which only explain the need to adjust the meaning of words to support the ever changing objects and ideas around us in the technical, scientific, political or sociocultural domains of the environment (Blank 1999). As such, the true *causes* of semantic change that comprise, for example, the linguistics pre-conditions that may promote or repress words from undergoing semantic change, was mostly neglected in actual research.

Cognitive semantics, and in particular theories of prototype semantics, did pay some attention to this question. These theoretical approaches emphasize the role that the relations between words may have on semantic change. Significantly, it was argued that the differences in prototypicality of words in their category (e.g., a robin or a dove is more prototypical in the category of birds, than a peacock or an ostrich which are more peripheral) can explain many cases of semantic change (Koch 2016). For example, Geeraerts (1985, 1992) maps related words into semantic categories, analyzes these categories

over time, and concludes that prototypical semantic areas are more stable over time than peripheral ones.

However, as discussed above, these generalizations use a top-down approach, and are based on fine-grained studies of small and anecdotal datasets. Moreover, their claims about regularities of semantic change are basically untestable. Specifically it is not clear whether these are claims about inviolable rules or tendencies. If they are inviolable rules, what would constitute counter-evidence? And if they are tendencies, how strong are they, and how might one investigate this question? It is important to point out that there is still no comprehensive database of documented semantic changes from a wide range of languages, and the databases that do exist, such as Zalizniak et al. (2012), are problematic from several points of view.

1.2.3 The quantitative paradigm

The paradigm described above generally bases its identification of semantic changes on the qualitative evaluation of words' usage-context in historical corpora. Such approach is informative, as it provides examples that are easily and naturally appreciated as supporting claimed taxonomies or mechanisms of change. However, this type of analysis is, in the end, subjective, due to its dependence on the particular set of examples selected, as well as on those that were left out. This makes it difficult to assess the magnitude or importance of the reported semantic changes, but moreover make them questionable in terms of reliability (i.e., representativeness).

A second approach, which we call the QUANTITATIVE PARADIGM, puts numbers to these changes by quantifying changes in the words' distributional statistics over time. This approach has been operationalized in a number of ways. A relatively straightforward operationalization involves measuring change in words' token frequencies (Bybee 2006). A more sophisticated method was introduced by Hilpert (2006), who adapted Stefanowitsch and Gries (2003) collostructional analysis for diachronic analysis. In this approach, the strength of association between two words is modeled by their statistical dependency (i.e., how often they appear together relative to what is expected from their general frequency). These association strengths are computed for each period and compared diachronically. For example, the association between *apple* and *phone* was not evident in the 1990s, and only emerged in the 2000s. The diachronic differences in association strength are

evaluated statistically, which puts the question of whether or not a semantic change took place on a statistically objective footing.

The quantitative paradigm has been used in a number of studies, albeit generally with a top-down approach[2]. One use has been to identify cases where semantic change took place for subsequent philological analysis (Fonteyn and Hartmann 2016; Smirnova 2012; Tantucci et al. 2017). Another has been to identify empirically well-founded stages in diachronic corpora (Gries and Hilpert 2008). A third use has been to test hypotheses that derive from a theory (Geeraerts et al. 2011). Moreover, word frequency was even proposed as a possible explanation for semantic change, as it has been argued that high frequency has a *bleaching* effect on a word's meaning (Bybee 2006), an effect similar to broadening (see Table 1.2).

While these works extract distributional statistics about words from corpora, their analyses tend to be on a small scale, diachronically examining only a few words or grammatical constructions at a time.

1.2.4 The computational paradigm

According to the distributional hypothesis (Harris 1954) similar words appear in similar contexts. Consequently, it has been suggested that the meaning of a word can be extracted from its usage contexts, or in the words of Firth "you shall know a word by the company it keeps." (1957, p. 11). This is the basis of most, if not all, current corpus-based computational models that rely on contextual information to represent the meanings of words. We call this approach the COMPUTATIONAL PARADIGM.

It should be noted that all the paradigms discussed above comply, in one way or another, with the distributional hypothesis. Crucially though, none of the above paradigms directly address the issue of word *representation*. Instead, the meaning of words is implicitly assumed either subjectively (e.g., in the traditional paradigm), or according to the association strengths (e.g., in the quantitative paradigm). In contrast, the computational paradigm necessarily represents word meanings explicitly, and only then defines semantic change as measurable differences between these representations. For that reason, the following focuses on the computational paradigm approach for meaning representation and semantic change.

[2]Bochkarev et al. (2014), used a bottom-up approach to study global changes in languages, but completely ignored the level of individual words.

Vector space models (VSM)[3] are usage-based models specifically based on NLP tools that were developed under the distributional hypothesis to extract word meanings from their contextual information. These models use the local context window around a given word to capture information about the words that co-occur with it. They then represent each word as a continuous vector in a high-dimensional space. These vectors are used extensively – in fact, almost exclusively – in a variety of NLP domains, such as semantic word similarity, analogy solving, synonym detection, and information retrieval (Baroni et al. 2014; Mikolov et al. 2013; Pennington et al. 2014), as well as in chunking, named entity recognition and sentiment analysis (Guo et al. 2014; Socher et al. 2013; Turian et al. 2010), among others. Importantly, these vector representations are the primary tool with which computational approaches – including the one adopted in this dissertation – investigate, in a data-driven fashion, how words' meanings change over time.

In line with the distributional hypothesis discussed above, these vector representations map semantically similar words, if these appear in similar contexts, to proximate points in the vector space (Turney and Pantel 2010). It is customary to use the cosine-distance between the vectors of two words as an estimate of the degree of their semantic similarity. For example, in the toy example shown in Figure 1.1 below, *automobile* and *car* are close to each other in the vector space, relative to *horse*. This will be reflected in a small cosine-distance between the former pair, and a larger one between either of them and horse.

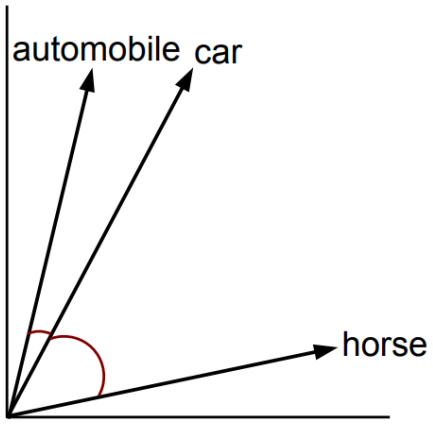


FIGURE 1.1: Three word vectors in the semantic space created by the VSM

[3]We refer to both explicit count-based models (e.g., point-wise mutual information, PPMI) as well as implicit predictive models (e.g., word2vec) by this term.

Following the same logic, the cosine-distance is often viewed – and used – as an estimate of the degree or amount of semantic change. Specifically, the semantic change that a particular word has undergone is measured by the cosine-distance between its vectors at two time points. Therefore, the larger the cosine-distance the greater the semantic change, and vice versa. Or, more formally:

$$\Delta w^{t_0 \rightarrow t_1} = 1 - \frac{v_w^{t_0} \cdot v_w^{t_1}}{\|v_w^{t_0}\| \cdot \|v_w^{t_1}\|} \tag{1.1}$$

Where $v_w^{t_0}$ and $v_w^{t_1}$ are the vector representations of the word w at two time points, t_0 and t_1, respectively.

This idea might seem intuitively compelling, as it looks like a natural extension of the synchronic comparison between two different words. However, its validity as an accurate metric for semantic change has never been tested. This poses an acute problem for claims about semantic change that are based on this measure (which comprise the vast majority of studies).

Computational studies of semantic change

Among the first to take a computational approach to linguistically motivated questions of semantic change were Sagi et al. (2009) and Wijaya and Yeniterzi (2011). Despite their pioneering nature, these studies, and others that soon followed (Heylen et al. 2015; Hilpert and Perek 2015), have analyzed only a small number of examples or a single construction, thus not exploiting the potential of combining massive corpora with modern computational tools in a large-scale bottom-up analysis of an entire lexicon.

Since these first pioneering works, the computational paradigm has become increasingly popular in the NLP research community, as additional studies brought this research field into the big-data era (Cook and Stevenson 2010; Frermann and Lapata 2016; Gulordava and Baroni 2011; Jatowt and Duh 2014; Kim et al. 2014; Kulkarni et al. 2015; Schlechtweg et al. 2017). These studies focused on developing techniques to detect words that underwent semantic change, or a specific type thereof, but neglected almost completely the potential insights for the linguistic analysis of meaning, primarily the *how's* and *why's* of semantic change.

Recently though, several studies have addressed this research lacuna and reported phenomena that can only be observed in a large-scale analysis. These

studies have proposed law-like generalizations or regularities of semantic change:

- The Diachronic Prototypicality Effect (Dubossarsky et al. (2015), and see Chapter 3)
- The Diachronic Word-Class Effect (Dubossarsky et al. (2016), and see Chapter 2)
- The Law of Conformity (Hamilton et al. 2016)
- The Law of Innovation (Hamilton et al. 2016)

However, all of these studies (and virtually all studies in the field[4]), use a small number of hand-picked examples as a basis for arguing for the soundness of their models. In the absence of a gold standard evaluation set for words that have undergone semantic change, it is difficult to objectively assess the quality of any proposed model of semantic change, or to compare different models.

Crucially, these above-mentioned studies have relied on the cosine-distance between a word's respective vector-representations at two time points as their sole metric to measure semantic change. As noted above, despite its wide and apparently uncontroversial use, the validity of this metric as an accurate estimate for semantic change has never been tested. As a result, previously reported findings might not be accurate. As studies that are based on this unvalidated metric accumulate, without being evaluated against a gold standard test for semantic change, the problem becomes increasingly acute. I address these two methodological problems in my work (see Section 1.4).

A multi-sense semantic representation

A fundamental property of VSMs is that each word is represented by a single vector. This may seem surprising, given the prevalence of polysemy in natural language. Lexical items typically – and perhaps usually – have multiple senses or meanings. The word *cell*, for example, has distinct senses or meanings related to biology, incarceration, and telecommunications. Current models collapse these different senses into a single *global* vector and are thus unable to explicitly represent distinct senses of a word.

[4]Except, perhaps, Frermann and Lapata (2016) who used a task of novel-sense detection as a proxy.

This global approach to word representation is especially disadvantageous for semantic change research, as polysemy has a fundamental role in semantic change (Traugott and Dasher 2002). Comparing a word's global vectors, and determining that it underwent semantic change, presumably tells us little about the nature of such a change. It does not tell us what aspects of its meaning have changed, whether new senses have emerged, whether some senses have dropped from usage and so on. It seems that researchers have presumed that this information resides in the original usage contexts of the word, but is impaired by the global representation approach.

In light of this drawback, several studies advocate the use of sense-specific vector representations for general NLP purposes. These vectors arguably capture the various senses of a word by creating a distinct representation for each sense. Intuitively, sense-specific vectors should be more accurate than global vectors, and this improvement in accuracy should be reflected in performance gains in downstream tasks.

Several studies reported such performance gains when using sense-specific vectors, which they conclude is due to the utility of polysemic information that these vectors arguably capture (Huang et al. 2012; Li and Jurafsky 2015; Neelakantan et al. 2014). However, such a conclusion can only be drawn if two conditions are met: (1) if sense-specific vectors truly represent polysemic information; and (2) if sense-specific vectors improve performance. While the first condition was not addressed in these studies, the second condition was consistently met in all of them. As a result, it remains to be demonstrated that these performance gains can be attributed to the representation of polysemy. Only when the first condition is validated can these vectors be reliably used in a more ecological model of semantic change that takes polysemy into account. I address this validation step in my work (see Section 1.4).

1.3 Research objectives

This PhD dissertation aims to advance semantic change research with state-of-the-art computational techniques in two ways. First, it aims to demonstrate that it is both possible and fruitful to employ a large-scale bottom-up computational approach to semantic change research. Second, it aims to make sure that this emerging research field stands on a methodologically solid footing from its beginning.

As the survey in Section 1.2 points out, research on semantic change, whether it was conducted in traditional or more contemporary paradigms

could greatly benefit from a bottom-up, large-scale analyses that are carried out over an entire lexicon. A bottom-up approach basically seeks to articulate generalizations not on the basis of prior theoretical assumptions but on the basis of naturally-occurring language production, which is also not liable to the subjective judgments of individual researchers. Large-scale analyses address the fundamental limitation of the representativeness of the data-set to be generalized over. Since previous studies are based on small and anecdotal cases of semantic change, it is not at all certain to what extent the semantic changes collected in surveys are representative of semantic change in general. The investigation of entire lexicons not only provides more examples of semantic change as grist for the theoretical mill, but also allows one to identify possible regularities of change that may be evident only on a large scale.

These desiderata are today made possible by two recent independent developments. The first is an upsurge in the availability of digitized historical corpora of texts. The second is novel computational methods that allow the automatic semantic processing of these corpora. This dynamic duo of massive corpora and innovative computational methods has the potential to lead to novel insights about semantic change, both in discovering regular patterns of change and in identifying their linguistics causes, based on large-scale data-driven analyses.

It should be noted that in comparison to the first two paradigms that use small-scale analysis, the computational paradigm has also some disadvantages. While the former carefully choose their findings to fit their conclusions (which is also their main methodological weakness), the latter includes many unexpected findings due to its large-scale bottom-up approach. Naturally, these findings comprise many intuitive examples of words known to have undergone semantic change, in addition to counter-intuitive and surprising examples that may help challenge existing theories and advance research. However, it also expected to find "noisy" and incorrect examples of words that were mistakenly found to have undergone semantic change. For example, a naïve computational semantic change model that is based only on the distributions of words in context may find that *president* undergoes a recurrent semantic change once in every 4 or 8 years. Therefore, while the carefully chosen examples that are used in small-scale analyses are usually sufficient to support an argument, more rigorous research methods must be employed to handle the type of data involved in large-scale bottom-up analysis. Moreover, since each paradigm has its advantages and disadvantages,

they should be seen as complementary rather than competing accounts.

The initial attempts to employ the computational paradigm on a large scale have demonstrated the importance of methodological issues. The place of quantitative hypothesis testing is more central in this emerging field than in its parent fields of research: first, the quantitative turn in modern linguistics has been relatively recent, and second, NLP research has no tradition of empirical hypothesis testing, which is instead based on modeling work. Thus, fundamental methodological issues must be addressed in order to make an objective, reliable, and genuine contribution to the research field.

Primarily, the absence of a gold standard evaluation set for semantic change makes it problematic to assess the quality of any proposed model of semantic change objectively. Furthermore, the reliance on an unvalidated semantic change metric poses an additional hurdle, especially as this metric becomes increasingly popular. Lastly, validating that words polysemy is reliably represented by sense-specific vectors would facilitate their diachronic analysis in finer, more ecological models of semantic change. By contributing to the resolution of these methodological issues, this dissertation aims to provide general guidelines and frameworks for hypothesis testing in NLP research that go beyond the specific problem of semantic change.

1.4 Current work

The first paper (see Chapter 2) in this dissertation examines whether words corresponding to different parts of speech (POS), i.e., Nouns, Verbs and Adjectives, differ in their rates of semantic change. This is the first time such a question could be experimentally tested, as it required large-scale bottom-up analysis over an entire lexicon which was impossible until recently. In this article, we analyze the semantic change rates of a very large sample of nouns, verbs and adjectives, and compare their rates of semantic change throughout the decades of the 20th century. This study set out to demonstrate the usefulness of our approach for semantic change research as it focuses on regularities of semantic change that may only be uncovered on a large scale.

The second paper (see Chapter 3) tackles a theoretical question: are there linguistic factors that make certain words more prone to semantic change? We address this question by mapping semantic change patterns over an entire lexicon in order to examine why certain words change more than others. Similarly to our first paper, we take a whole-lexicon approach and analyze the semantic change rates of the 7000 most frequent words in the lexicon

throughout the decades of the 20th century. We compare the results to previous theoretical accounts that proposed causal factors in semantic change (which were based on a very small-set of hand-picked examples). Thus, this study further demonstrates the potential of our large-scale bottom-up approach to contribute to the investigation of the linguistic causes that are involved in semantic change by rigorously testing an existing theoretical hypothesis on a much larger scale.

The third paper (see Chapter 4) addresses the problem of evaluating models of semantic change in the absence of gold standard evaluation sets. As a result of this lacuna, it is necessary to find a way to validate the results of these models. Specifically, we examine the standard metric in semantic change research, i.e., the cosine-distance, which despite being widely used has never been directly validated. We developed a control condition that is crafted specifically for this task under which we test this metric, in addition to an analytical investigation, and revisited previous studies that relied on this metric in their analyses. Overall, this paper provides a critical validation analysis for a pivotal methodological procedure in semantic change research. Importantly, it proposes a general framework to circumvent the problem of an absent gold standard evaluation set for semantic change, by using control conditions. Such a framework may also be used to verify the reliability of results reported in prior art, in addition to make future contributions stand on a more solid methodological footing. Consequently, the findings of our analyses are not limited to semantic change research, and may be useful for the NLP research community at large.

Our fourth paper (see Chapter 5) aims to test the feasibility of using sense-specific vectors as a "next generation" model for semantic change, due to their presumably more ecological representation of meaning. Specifically, we set to investigate the validity of the claim that sense-specific vectors truly represent polysemic information, a claim that was based on performance gains obtained using these vectors. This is a necessary step before such vectors can be used as a viable tool to study changes in sense representations over time. We rigorously tested this claim, first by dismantling it to its constituent elements, and then testing each element using a combination of an appropriate control condition, an analytical analysis and computational simulations. Overall, this paper provides a critical reevaluation of prominent results in polysemy analysis and representation, a reevaluation that is not limited to semantic change research and further promotes the use of carefully designed control conditions and validation routines in NLP research at large.

References

Baroni, Marco, 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*, pp. 238–247.

Blank, Andreas (1999). "Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change". In: *Historical Semantics and Cognition*. Ed. by Andreas Blank and Peter Koch. Mouton de Gruyter, pp. 61–89. DOI: 10.1515/9783110895100.335.

Bloomfield, L (1933). *Language*. University of Chicago Press.

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

Bybee, Joan (2006). *Frequency of Use and the Organization of Language*. Oxford: Oxford University Press, p. 375.

— (2015). *Language change*. Cambridge, UK: Cambridge University Press, p. 292.

Cook, Paul and Suzanne Stevenson (2010). "Automatically Identifying Changes in the Semantic Orientation of Words." In: *Proceedings of LREC*.

Dubossarsky, Haim, 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*, pp. 66–70.

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

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

Fonteyn, Lauren and Stefan Hartmann (2016). "Usage-based perspectives on diachronic morphology: a mixed-methods approach towards English ing-nominals". In: *Linguistics vanguard* 2.1, pp. 1–12. DOI: 10.1515/lingvan-2016-0057.

Frermann, Lea and Mirella Lapata (2016). "A bayesian model of diachronic meaning change". In: *Transactions of the Association for Computational Linguistics* 4, pp. 31–45.

Geeraerts, Dirk (1985). "Cognitive restrictions on the structure of semantic change". In: *Historical Semantics*, pp. 127–153.

— (1992). "Prototypicality effects in diachronic semantics: A round-up". In: *Diachrony within Synchrony: language, history and cognition*. Ed. by G. Kellermann and M. D. Morissey. Frankfurt am Main: Peter Lang, pp. 183–203.

Geeraerts, Dirk, Caroline Gevaerts, and Dirk Speelman (2011). "How anger rose: hypothesis testing in diachronic semantics". In: *Current Methods in Historical Semantics*. Ed. by Justyna Robynson and Kathryn Allan. Topics in English Linguistics [TiEL]. Berlin/New York: Mouton de Gruyter, pp. 109–32.

Gries, Stefan Th and Martin Hilpert (2008). "The identification of stages in diachronic data: variability-based neighbour clustering". In: *Corpora* 3.1, pp. 59–81. ISSN: 1749-5032. DOI: 10.3366/E1749503208000075.

Gulordava, Kristina 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*. Association for Computational Linguistics, pp. 67–71.

Guo, Jiang, Wanxiang Che, Haifeng Wang, and Ting Liu (2014). "Revisiting embedding features for simple semi-supervised learning". In: *Proceedings of EMNLP*, pp. 110–120.

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

Harris, Zellig S (1954). "Distributional structure". In: *Word* 10.2-3, pp. 146–162.

Heylen, Kris, Thomas Wielfaert, Dirk Speelman, and Dirk Geeraerts (2015). "Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis". In: *Lingua* 157, pp. 153–172.

Hilpert, Martin (2006). "Distinctive collexeme analysis and diachrony". In: *Corpus Linguistics and Linguistic Theory* 2.2, pp. 243–256.

Hilpert, Martin and Florent Perek (2015). "Meaning change in a petri dish: constructions, semantic vector spaces, and motion charts". In: *Linguistics Vanguard* 1.1, pp. 339–350.

Hock, H H and B D Joseph (2009). *Language History, Language Change, and Language Relationship: An Introduction to Historical and Comparative Linguistics*. Trends in Linguistics. Studies and Monographs [TiLSM]. De Gruyter.

Huang, Eric H, Richard Socher, Christopher D Manning, and Andrew Y Ng (2012). "Improving word representations via global context and multiple word prototypes". In: *Proceedings of ACL*. Association for Computational Linguistics, pp. 873–882.

Jatowt, Adam 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*, pp. 229–238.

Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov (2014). "Temporal Analysis of Language through Neural Language Models". In: *Proceedings of ACL*, pp. 61–65.

Koch, Peter (2016). "Meaning change and semantic shifts". In: *The Lexical Typology of Semantic Shifts*. Ed. by Päivi Juvonen and Maria Koptjevskaja-Tamm. Chap. 2, pp. 21–66.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena (2015). "Statistically significant detection of linguistic change". In: *Proceedings of WWW*, pp. 625–635.

Li, Jiwei and Dan Jurafsky (2015). "Do multi-sense embeddings improve natural language understanding?" In: *arXiv preprint arXiv:1506.01070*.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig (2013). "Linguistic regularities in continuous space word representations". In: *Proceedings of NAACL*, pp. 746–751.

Neelakantan, Arvind, Jeevan Shankar, Re Passos, and Andrew Mccallum (2014). "Efficient nonparametric estimation of multiple embeddings per word in vector space". In: *Proceedings of EMNLP*.

Newman, John (2015). "Semantic shift". In: *The Routledge Handbook of Semantics*. Ed. by Nick Rimer. New York: Routledge, pp. 266–280.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). "Glove: Global vectors for word representation". In: *Proceedings of EMNLP*, pp. 1532–1543.

Reisig, Christian Karl (1839). *Vorlesungen über lateinische Sprachwissenschaft*. Leipzig: Lehnhold.

Sagi, Eyal, 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*. Association for Computational Linguistics, pp. 104–111.

Schlechtweg, Dominik, Stefanie Eckmann, Enrico Santus, Sabine Schulte im Walde, and Daniel Hole (2017). "German in Flux: Detecting Metaphoric Change via Word Entropy". In: *Proceedings of CoNLL*, pp. 354–367.

Smirnova, Elena (2012). "On some Problematic Aspects of Subjectification". In: *Language Dynamics and Change* 2.1, pp. 34–58. ISSN: 2210-5832.

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts (2013). "Recursive deep models for semantic compositionality over a sentiment treebank". In: *Proceedings of EMNLP*, pp. 1631–1642.

Stefanowitsch, Anatol and Stefan Th Gries (2003). "Collostructions: Investigating the interaction of words and constructions". In: *International journal of corpus linguistics* 8.2, pp. 209–243.

Tantucci, Vittorio, Jonathan Culpeper, and Matteo Di Cristofaro (2017). "Dynamic resonance and social reciprocity in language change: the case of Good morrow". In: *Language Sciences*. ISSN: 0388-0001.

Traugott, Elizabeth Closs and Richard B Dasher (2002). *Regularity in semantic change*. Vol. 97. Cambridge University Press.

Turian, Joseph, Lev Ratinov, and Yoshua Bengio (2010). "Word representations: a simple and general method for semi-supervised learning". In: *Proceedings of ACL*. Association for Computational Linguistics, pp. 384–394.

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

Wijaya, Derry Tanti 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*. ACM, pp. 35–40.

Zalizniak, Anna S., Maria Bulakh, Dmitrij Ganenkov, Ilya Gruntov, Timur Maisak, and Maxim Russo (2012). "The catalogue of semantic shifts as a database for lexical semantic typology". In: *Linguistics* 50.3, pp. 633–669. ISSN: 00243949. DOI: 10.1515/ling-2012-0020.

Chapter 2

The Diachronic Word-Class Effect

Verbs change more than nouns: A bottom up computational approach to semantic change

Published

Dubossarsky, H., Grossman, E., & Weinshall, D. (2016). Verbs change more than nouns: A bottom up computational approach to semantic change. *Lingue e Linguaggio*, 5-25.

* Affiliations appear in the last page of the published paper

VERBS CHANGE MORE THAN NOUNS: A BOTTOM-UP COMPUTATIONAL APPROACH TO SEMANTIC CHANGE

HAIM DUBOSSARSKY DAPHNA WEINSHALL EITAN GROSSMAN

ABSTRACT: Linguists have identified a number of types of recurrent semantic change, and have proposed a number of explanations, usually based on specific lexical items. This paper takes a different approach, by using a distributional semantic model to identify and quantify semantic change across an entire lexicon in a completely bottom-up fashion, and by examining which distributional properties of words are causal factors in semantic change. Several independent contributing factors are identified. First, the degree of prototypicality of a word within its semantic cluster correlated inversely with its likelihood of change (the "Diachronic Prototypicality Effect"). Second, the word class assignment of a word correlates with its rate of change: verbs change more than nouns, and nouns change more than adjectives (the "Diachronic Word Class Effect"), which we propose may be the diachronic result of an independently established synchronic psycholinguistic effect (the "Verb Mutability Effect"). Third, we found that mere token frequency does not play a significant role in the likelihood of a word's meaning to change. A regression analysis shows that these effects complement each other, and together, cover a significant amount of the variance in the data.

KEYWORDS: semantic change, distributional semantics.

1. THE PROBLEM OF SEMANTIC CHANGE

Lexical semantic change - change in the meanings of words - is a basic fact of language change that can be observed over long periods of time. For example, the English word girl originally indicated a child of either sex, but in contemporary English, it refers only to a female child. Bybee shows the turning point was the fifteenth century, after the conventionalization of the word *boy* to refer to a male child, which "cut into the range of reference for girl" (Bybee 2015: 202). But semantic change is also "an undeniable and ubiquitous facet of our experience of language" (Newman 2015: 267), with words acquiring new

senses, developing new polysemies, and entirely new meanings, in time-frames that can be observed even by casual observation by speakers. For example, recent changes in technology have led to novel meanings of words like *navigate*, *surf*, and *desktop* (Newman 2015: 266). Speakers and listeners may even be aware of "mini" semantic change in real time, when they experience an innovative use of an existing word.

Linguists have identified some recurring types of semantic change. Some of the major types include the textbook examples of change in scope, e.g., *widening* (Latin *caballus* 'nag, workhorse' > Spanish *caballo* 'horse') or narrowing (*hound* 'canine' > 'hunting dog'), or in connotation (*amelioration* or *pejoration*). However, the systematic search for an explanatory theory of semantic change was largely neglected until Geeraerts (1985, 1992) and Traugott & Dasher (2002), who both claimed that semantic change is overwhelmingly regular. Moreover, both Geeraerts and Traugott have claimed that semantic change – like language change in general – is rooted in and constrained by properties of human cognition and of language usage.

Contemporary research identifies different kinds of regularity in semantic change as *tendencies of change*, which are asymmetries with respect to the directions in which change is more likely to occur. For example, Traugott & Dasher (2002) propose that semantic change regularly follows the pathway: objective meaning > subjective meaning > intersubjective meaning. It has also been suggested that concrete meanings tend to develop into more abstract ones (Bloomfield 1933; Haspelmath 2004; Sweetser 1990). See the following examples:

(1) *see* 'visual perception' > 'understanding'

(2) *touch* 'tactile perception' > 'feel'

(3) *head* 'body part' > 'chief'

Another often-observed regularity is that semantic change overwhelmingly tends to entail polysemy, in which a word or expression acquire new senses that co-exist with the older conventionalized senses (e.g., a new sense for *surf* has emerged since the 1990s). These new senses can continue to co-exist stably with the older ones or to supplant earlier senses, thereby "taking over" the meaning of the word.

The existence of such regularities and asymmetries, or "unidirectional pathways of change", has been taken as evidence that language change is not random. Moreover, these asymmetries call for explanations that are plausible in terms of what we know about human cognition and communication. Numerous such explanations have been offered, from Traugott & Dasher's (2002) influential Neo-Gricean account to other pragmatically-based accounts (for an

overview, see Grossman & Noveck 2015). However, while such accounts may offer potentially convincing explanations for observed changes, there is to date no empirically-grounded theory that can explain – or predict – which words are likely to undergo semantic change, and why this is so, across an entire lexicon.

This last point is the focus of the present article. While historical linguists have painstakingly accumulated much data about – and proposed explanations for – cross-linguistically recurrent pathways of semantic change (e.g., body-part term > spatial term), the data and explanations are usually specific to a particular group of words. For example, the explanations proposed for the development of body-part terms into spatial terms cannot necessarily be generalized to words of other semantic classes. In fact, the question posed in this article – what are the specific properties of words that make them more or less prone to semantic change? – has been almost entirely neglected in historical linguistic research. Furthermore, most studies of attested pathways of change tend to focus on their descriptive semantics, and have tended to ignore their distributional properties.

Nonetheless, some work in this direction can be found in earlier structuralist and cognitivist theories of semantic change, which emphasized the role of the structure of the lexicon in explaining semantic change. For example, it has often been assumed that changes in words' meanings are due to a tendency for languages to avoid ambiguous form-meaning pairings, such as homonymy, synonymy, and polysemy (Anttila 1989; Menner 1945). On the other hand, when related words are examined together, it has been observed that one word's change of meaning often "drags along" other words in the same semantic field, leading to parallel change (Lehrer 1985). These seemingly contradictory patterns of change lead to the conclusion that if ambiguity avoidance is indeed a reason of semantic change, its role is more complex than initially assumed.

However, what is common to both ideas – the putative tendency to avoid ambiguous form-meaning pairings and the equally putative tendency for words in the same semantic domain to change in similar ways – is the observation that changes in a word's meaning may result from – or cause – changes in the meaning of a semantically related word. The idea that words should be examined relative to each other, and that these relations play a causal role in semantic change is elaborated by Geeraerts (1985, 1992), who maps related words into clusters, and based on Rosch's prototype theory (1973), establishes which words are the prototypical or peripheral exemplars within each cluster. Geeraerts analyzes these clusters diachronically, finds characteristic patterns of change due to meaning overlap, and concludes that prototypical semantic areas are more stable diachronically than peripheral ones. While Geeraert's

ideas are promising for studies of semantic change, they are based on case-studies hand-picked by the linguist, and are not based on large-scale corpora (Geeraerts 2010). This is a lacuna in the research field of semantic change, which we have addressed in a previous article (Dubossarsky et al. 2015) by articulating a method for identifying and quantifying semantic change across an entire lexicon, represented by a massive historical corpus.

Our aim in the present article is to evaluate whether other distributional properties of words are indeed implicated in semantic change. Specifically, we examine whether words of different parts-of-speech or word classes change at different rates. We assume that the null hypothesis is that there is no difference between word class assignment and rate of change. However, we predict that there will indeed be differences, based on the fact that different word classes prototypically encode cognitively different things: nouns prototypically encode entities, verbs prototypically encode events, and adjectives prototypically encode properties. Moreover, different word classes can have significantly different collocational properties, i.e., they occur in different types and ranges of contexts. Finally, Sagi et al. (2009), one of the only studies to tackle this question, found that in 19th century English, a small selection of verbs showed a higher rate of change than nouns.

It is important to stress that at no time do we, or any of the above works cited as far as we know, claim that semantic change is governed by a single factor. In fact, it is clear that previous work on semantic change is likely to be correct in supposing that social, historical, technological, cognitive, communicative, and other factors are implicated in semantic change. The question is how to tease them apart and understand their respective contributions. This paper demonstrates that an observable property of words, i.e., their part-of-speech or word class assignment, is indeed implicated in semantic change. Moreover, we demonstrate that this effect is in addition to another effect which we have argued for earlier, namely, that the position of a word within its semantic cluster – interpreted as its degree of prototypicality.

The structure of the paper is as follows: in Section 2, we sketch the methodology used, and in Section 3, we describe the experiment conducted. In Section 4 we discuss the results, and in Section 5 we analyze possible interactions with other factors. Section 6 is devoted to discussion on the results and their implications. Section 7 provides concluding remarks, focusing on directions for future research.

2. METHODOLOGY

2.1 The role of input frequency

There are numerous ways of representing lexical meaning. Computational models developed for representing meaning excel in what computational approaches do best and classical historical linguistics does poorly, namely, the large-scale analysis of language usage and the precise quantitative representation of meaning. At the heart of these models lies the "distributional hypothesis" (Firth 1957; Harris 1954), according to which the meaning of words can be deduced from the contexts in which they appear.

We employ a distributional semantic modeling (DSM) approach to represent word meanings. DSM collects distributional information on the co-occurrence profiles of words, essentially showing their collocates (Hilpert 2006; Stefanowitsch & Gries 2003), i.e., the other words with which they co-occur in specific contexts. Traditionally, this is done by representing each word in terms of its collocates across an entire lexicon. This type of model has the advantage of providing an explicit (or direct) quantitative measure of a word's meaning, and is informative in that it tells us which words do or do not occur with a given word of interest. However, since most words occur with a limited range of collocates, most of the words in a lexicon will co-occur with most other words in the lexicon zero times. As such, these kinds of representations are sparse. This can be seen in the following illustrative example bellow, where only ten words collocate with the word *pan*, while the rest of the vocabulary (i.e., *surf, sky, dress, hat, call*, etc.) does not.

collocations	*pot*	*fry*	*cook*	*egg*	*bacon*	*cake*	*butter*	*oil*	*stir*	*stove*	*surf*	*sky*	*dress*	*hat*
#	87	69	61	55	51	49	23	19	17	9	0	0	0	0

TABLE 1. WORDS COLLOCATIONS STATISTICS FOR THE WORD *PAN* (ILLUSTRATIVE EXAMPLE)

This type of representation is usually further analyzed, e.g., by normalizing the word counts to frequencies, or with more sophisticated statistical methods, e.g., tf-idf or point mutual information. However, for our purposes, such models are inadequate, because in the end they tell us only whether a word co-occurs with another word or not. In order to understand the relationship of a word with the rest of the words in an entire lexicon, other types of models are necessary.

These models are the more recent ones that exploit machine-learning and neural network tools to learn the distributional properties of words automatically. Unlike traditional models, they do so by representing words in terms of the interaction of multiple properties. However, the specific contribution of

each property, when taken on its own, is opaque; as such, the quantitative representation of a word's meaning is implicit. Of the available recent models of this type, we chose a recently developed skip-gram word2vec model (Mikolov et al. 2013c, 2013d). This word2vec model has been fruitfully applied to distributional semantic corpora research, and scores high in semantic evaluation tasks (Mikolov et al. 2013a). As we will show, proof-of-concept can also be found in our results.

The word2vec model captures the meaning of words through dense vectors in an n-dimensional space. Every time a word appears in the corpus, its corresponding vector is updated according to the collocational environment in which it is embedded, up to a fixed distance from that word. The update is carried out such that the probability in which these words predict their context is maximized (Figure 1a.). As a result, words that predict similar contexts would be represented with similar vectors. In fact, this is much like linguistic items in a classical structuralist paradigm, whose interchangeability at a given point or "slot" in the syntagmatic chain implies that they share certain aspects of function or meaning, i.e., the Saussurian notion of "value" (Figure 1b.). It is worth noticing that if taken individually, the vectors' dimensions are opaque; only when the full range of dimensions is taken together do they capture the meaning of a word in the semantic hyper-space they occupy.

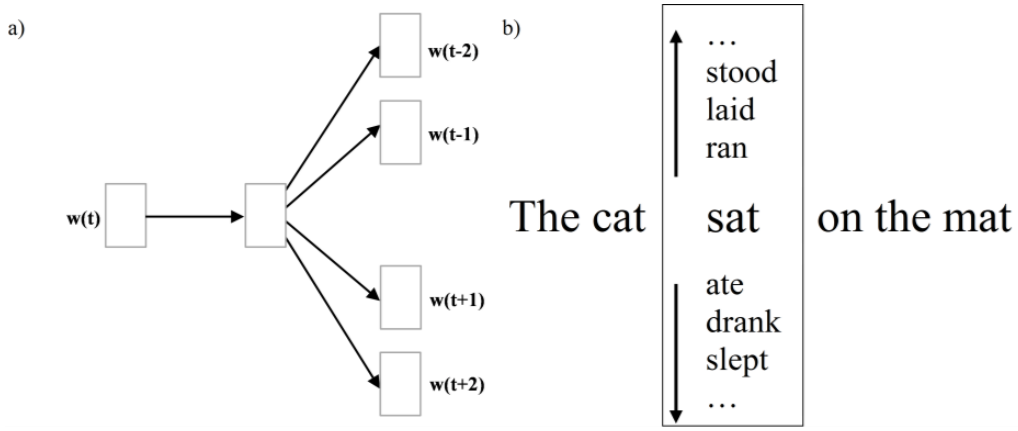


FIGURE 1. (A) WORD2VEC SKIP-GRAM ARCHITECTURE.
GIVEN A WORD, W(T), THE MODEL PREDICTS THE WORDS THAT PRECEDE AND PROCEED IT IN A WINDOW OF 4 WORDS, W(T-2),W(T-1),W(T+1),W(T+2) (MIKOLOV ET AL. 2013B).
(B) AN EXAMPLE OF THE CLASSICAL STRUCTURALIST PARADIGM.

While it may be surprising for linguists that one would choose to rely on a model whose individual dimensions are opaque, this is not a major concern, since it is well-established that words assigned similar vectors by the model are in fact semantically related in an intuitive way; for a recent demonstration,

see Hilpert & Perek (2015), which looks at the collocates of a single construction in English. The similarity between vectors is evaluated quantitatively, and defined as the cosine distance between the vectors in the semantic hyperspace. Short distances are considered to reflect similarity in meaning: related words are closer to each other in the semantic space (Turney 2006; Mikolov et al. 2013d; Levy & Goldberg 2014). In fact, this is reflected in the words' nearest neighbors in the semantic space that often capture synonymic, antonymic or level-of-category relations.

Although the model uses the entire lexicon for training, the accuracy of the meaning representations that is captured in the corresponding vectors is expected to diminish for less frequent words. This is simply because these words do not appear frequently enough for the model to learn their corresponding contexts. Therefore, only the most frequent words in the corpus, excluding stop-words, are defined as words-of-interest and are further analyzed. These words represent the entire lexicon.

2.2 Corpus

A massive historical corpus is required to train distributional semantic models. This is because the words whose distributional properties we are interested in must appear frequently enough in each time period in order to collect enough statistical information about their properties. Clearly, the time resolution of any analysis on such models is limited by the nature of the historical corpus: the finer the tagging for time, the finer the analysis can be.

Google Ngrams is the best available historical corpus for our purposes, as it provides an unprecedented time resolution – year by year – on a massive scale; the second largest historical corpus is about 1000 times smaller. Tens of millions of books were scanned as part of the Google Books project, and aggregated counts of Ngrams on a yearly resolution from those books are provided.

We used a recently published syntactic-Ngram dataset (Goldberg & Orwant 2013), where the words[1] are analyzed syntactically using a dependency

[1] The present study deals with word forms rather than lexemes. While this is possibly a shortcoming, it is shared by most NLP studies of massive corpora. Furthermore, the issue is less likely to affect English, with its relatively poor morphology, than other languages. Nevertheless, one might speculate about the effects of this. For example, it might be that the meaning of a specific verb forms in the corpus will be narrower than that of specific noun forms, overall, in an analysis based on word forms than in one based on lexemes. While it would be of considerable interest to conduct an experiment to determine the effect of using word forms versus lexemes, the issue has never been dealt with explicitly in computational linguistics, as far as we know, and it is beyond the scope of the present paper. We thank an anonymous reviewer for bringing this issue to our attention.

parser in their original sentences. The dataset provides aggregated counts of syntactic Ngrams on a yearly resolution that includes their part-of-speech (POS)[2] assignments as well. The dataset distinguishes content words, which are meaning-bearing elements, from functional markers[3] that modify the content words. Therefore, a syntactic Ngram of order N includes exactly N content words and few optional function markers. We used syntactic Ngrams of 4 content words from the English fiction books,[4] and aggregated them over their dependency labels to provide POS Ngrams. The following is an example POS Ngram from the corpus.

(4) and_CC with_IN sanction_NN my_PR tears_NN gushed_VB out_RB

Verbs, nouns, and adjectives below a certain frequency threshold, and all the rest of the POS assignment, lose their tags. In this Ngram, only *tears* retains it.

The historical corpus is sorted diachronically, with 10 million POS Ngrams (about 50 million words) per year for the years 1850-2000. When the number of POS Ngrams in the corpus for a given year was bigger than that size, due to the increasing number of published and scanned books over time, a random subsampling process was conducted to keep a fixed corpus size per year. This resulted in a corpus size of about 7.5 billion words. Only the words-of-interest, the most frequent words in the corpus, retain their POS assignment, while the rest of the words reverted to their original word forms. All words were lowered case.

2.3 Diachronic Analysis

After initialization, the model is trained incrementally, one year after the other, for the entire historical corpus (POS-tagged and untagged words alike). In this way, the model's vectors at the end of one year's training are the starting point of the following year's training, which make them comparable diachronically. The model is saved after each year's training, so that the words' vectors could be later restored for synchronic and diachronic analyses.

The words vectors are compared diachronically in order to detect semantic change. Based on the affinity between similarity in meaning and similarity in vectors described in §2.1, semantic change is defined here as the difference between a word's two vectors at two time points. This allows us to quantify

[2] We use the term "part-of-speech" abbreviated POS, in the context of Natural Language Processing tagging, and the term "word class" otherwise.
[3] These include the following dependency labels: *det, poss, beg, aux, auxpass, ps, mark, complm* and *prt.*
[4] From the 2nd version of Google books.

semantic change in a straightforward fashion: the bigger the distance between the two vectors of a given word, the bigger the semantic change that this word underwent over that period of time. Specifically, the comparison is defined as the cosine distance between the word's two vectors according to equation 1, with 0 being identical vectors and 2 being maximally different. This is carried out for the entire lexicon.

$$(1) \qquad \Delta w^{t^0 \to t^1} = 1 - \frac{{v_w}^{t_0} \cdot {v_w}^{t_1}}{\|{v_w}^{t_0}\| \cdot \|{v_w}^{t_1}\|}$$

where ${v_w}^{t_0}$ and ${v_w}^{t_1}$ are the word's w vectors at two time points, t_0 and t_1, respectively.

In the following section, we present an experiment that investigates the relationship between word class assignments and likelihood of change.

3. EXPERIMENT

In this experiment, we evaluate the hypothesis that different parts of speech change at different rates. As noted above, we assume that the null hypothesis is that there is no difference between part of speech assignment and rate of change. However, we predict that there will indeed be differences, based on the fact that different parts of speech prototypically encode cognitively different things: nouns prototypically encode entities, verbs prototypically encode events, and adjectives prototypically encode properties. Moreover, different parts of speech can have significantly different collocational properties, i.e., they occur in different types and ranges of contexts. Finally, pilot studies of this question (Sagi et al. 2009) have indicated that some verbs show a higher rate of change than some nouns.

The word2vec model[5] was initialized with the length of vector set to 52, which means that the words' contexts are captured in a 52-dimension semantic hyper-space. The model was trained over the POS-tagged English fiction corpus (see §2.2), using the method described above (see §2.3). Words that appeared less than 10 times in the entire corpus were discarded from the lexicon and were ignored by the model.

The vectors of the 2000 most frequent verbs, nouns and adjectives (6000 in total) as they appear in the corpus were defined as the words-of-interest, and restored from the model at every decade from 1900 till 2000. For each word, the cosine distances between its vectors at every two consecutive decades were computed using equation (1). This resulted in 6000x10 semantic

[5] We used genism python library for its word2vec implementation (Řehůřek & Sojka 2010).

change scores that represent the degree of semantic change that each word underwent in every decade throughout the twentieth century (e.g., 1900-1910, 1910-1920, until 1990-2000). The average semantic change scores of each POS assignment were compared between groups.

4. RESULTS

Figure 2 shows the average semantic change for the different POS assignment groups at ten decades throughout the twentieth century. The results were submitted to a two-way ANOVA with POS *assignment* and *decade* as the independent variables. The first main effect, also clearly visible, is that the POS assignment groups differ in their rates of semantic change over all the decades ($F_{(2,59970)} = 6464$, $\eta = .177$, p-value <.001). The second main effect is that the semantic change rate appears to differ throughout different decades across all POS assignment groups ($F_{(9,59970)} = 576$, $\eta = .08$, p-value <.001). The interaction between the variables was found to be significant as well ($F_{(18,59970)} = 14.34$, $\eta = .004$, p-value <.001). This means that the rate of semantic change along the decades is not uniform across the POS assignment groups. However, the effect size of the first two variables reported above is robust, accounting for 17.7% and 8% of the overall variance in the words semantic change, respectively, which render these variables highly meaningful. In contrast, the effect size of the aforementioned interaction accounts for only 0.4% of the variance, which makes it unimportant, albeit statistically significant.

In order to evaluate the source of the first main effect – the difference in the rate of semantic change between the POS assignment, we conducted permutation tests as a post-hoc analysis on the pairs *verbs-nouns* and *nouns-adjectives*. The permutation tests created null hypotheses for each pair by assigning words to one of the two POS group randomly, then computing the differences between the averages of the two groups, and repeating the process 10,000 times for each decade. These distributions were later compared to the real differences in the average semantic change in each decade, so that their statistical significance could be evaluated. The permutation tests corroborate what is visibly clear from the descriptive pattern of the results (all p-values <.001), that *verbs* change more than *nouns*, and *nouns* change more than *adjectives*.

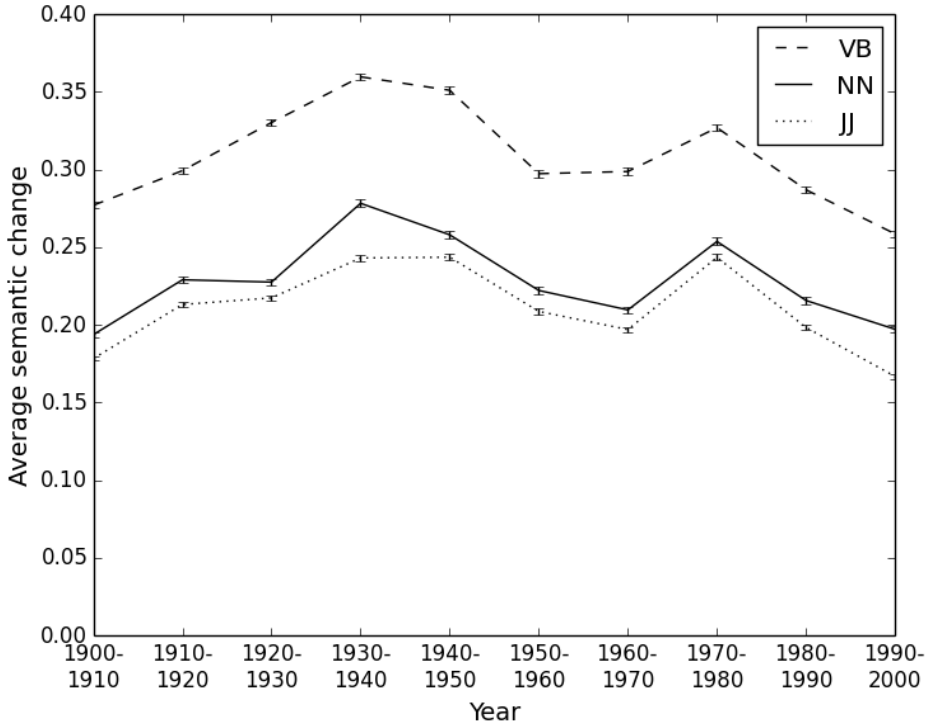


FIGURE 2. AVERAGE SEMANTIC CHANGE RATES THROUGHOUT THE DECADES IN THE TWENTIETH CENTURY FOR DIFFERENT POS ASSIGNMENT GROUPS. BARS REPRESENT STANDARD ERRORS.

5. INTERACTION WITH OTHER FACTORS: FREQUENCY AND PROTOTYPICALITY

In previous work, at least two observable properties of words have been argued to be causally implicated in semantic change, word frequency and prototypicality. We wanted to test their joint involvement in semantic change in light of the aforementioned findings.

5.1 Frequency

Frequency is often linked to language change, but its exact effects still remain to be worked out (Bybee 2006, 2010). While frequency clearly facilitates reductive formal change in grammaticalization and in sound change, it also protects morphological structures and syntactic constructions from analogy (e.g., irregular verbs forms are more frequent). Since no explicit hypothesis has been made regarding the role of frequency in semantic change per se, we set out to test the hypothesis that frequency plays some role in semantic

change. The null hypothesis was that there is no correlation between words' frequencies and their degree of semantic change.

Token frequencies were extracted from the entire corpus (about 7.5 billion words) and served as the words frequencies. The degrees of semantic change were taken from the results reported in §4 above.

In general, frequency was not found to correlate with the degree of words' semantic change over the ten decades in the twentieth century. Only four decades (1900-1910; 1910-1920; 1950-1960; 1960-1970) showed significant (p-value <.01) correlations. However, such correlations are so small, with maximum correlation coefficient <.07, that in terms of their effect size they account for less than 0.5% of the variance in the semantic change scores. Similar results were obtained when the analysis was repeated for each POS assignment group separately. Most correlations were statistically insignificant, and the ones that were significant were very small. Overall these results suggest that frequency plays little or no role in semantic change. We think that this result is surprising, since frequency is often thought to correlate with the degree of entrenchment of linguistic items in the mental lexicon (Bybee 2010). As such, one might hypothesize that words with high token frequency might be "protected" from semantic change. However, this hypothesis is counter-indicated by the results of our experiment. It may be that token frequency is, in the end, mainly responsible for coding asymmetries (Haspelmath 2008) and does not contribute much to semantic change per se.

5.2 Prototypicality

One of the model's inherent properties is that similar words have similar vectors (see §2.1). This makes the vectors ideal for clustering, where each cluster captures the words' "semantic landscape," as Hilpert & Perek (2015) call it. Importantly, it turned out that these clusters exhibit an internal structure, with some words closer to the center and others further away. In Dubossarsky et al. (2015) we analyzed this structure, and interpreted the distance of a word from its cluster center to reflect its degree of prototypicality, which is the degree by which a word resembles its category prototype. Crucially, this prototypicality was found to play an important role in semantic change, as the further a word is from its category's prototype, the more likely it is to undergo change.

We employ the methodology described in Dubossarsky et al. (2015) to the current dataset. Specifically, for each decade we cluster the 6000 word vectors using 1500 clusters, and compute the words' distances from their cluster centroids. This resulted in ten "prototypicality scores" for each word.

In Table 2, we present two clusters as examples. In each cluster, the words are sorted in prototypicality order (distance from their cluster's center). As a

result, said and chamber/room, appear at the tops of their lists, and constitute the most prototypical exemplars in their clusters, *verbs of utterance and enclosed habitats for humans* (see Dubossarsky et al. 2015 for further examples).

said_VB, 0.06	chamber_NN, 0.04
exclaimed_VB, 0.08	room_NN, 0.04
answered_VB, 0.08	drawing_NN, 0.05
added_VB, 0.11	bedroom_NN, 0.06
whispered_VB, 0.13	kitchen_NN, 0.07
cried_VB, 0.14	apartment_NN, 0.1
murmured_VB, 0.15	
growled_VB, 0.16	
repeated_VB, 0.2	
muttered_VB, 0.25	

TABLE 2. TWO WORD CLUSTERS, WITH POS TAGS AND DISTANCES FROM THEIR CENTROID, SORTED IN ASCENDING ORDER OF THE LATTER.

We used this approach to extend our previous finding that focused on semantic change in only one decade (1950-1960) to the entire twentieth century. Indeed, prototypicality at the beginning of each of the ten decades was related to the semantic change the words underwent by the end of that decade. Correlation coefficients ranged between r=.27 and r=.35, with average coefficient of r=.32 (all p-values <.001). This means that the farther a word is from the prototypical center of its category, the more likely it is to undergo semantic change, and attests to the meaning-conserving nature of prototypicality in semantic change. This could be called the "Diachronic Prototypicality Effect".

5.3 Regression analysis

It is intuitively clear that semantic change is not induced solely by a single factor, and that different factors may also be involved. Therefore, we wanted to evaluate the interaction between the two factors that were proven to be involved in semantic change, word class assignment and prototypicality.

In order to discern the contribution of these two factors, whether they complement each other or are to a large extent redundant, they were submitted to a multiple linear regression analysis. Prototypicality, as distance from centroid, and POS assignment were the independent variables, and the semantic change scores was the dependent variable. Regression analyses were conducted for these variables at each of the ten decades, and also pulled over all the decades.

Table 3 shows the contribution of each of the two variables in accounting for the semantic change in each of the ten decades examined as well as overall

the decades (all the results reported were statistically significant p-value <.01). The results show that the two variables account for a fair amount of the variance in semantic change, between 21%-29%. Although both variables account for a large part of semantic change when taken individually, POS plays a larger role. Prototypicality, despite playing a lesser role, accounts for a substantial amount of the variance in semantic change as well, which exactly reflects its correlation coefficients' values reported above.

Crucially, prototypicality's unique contribution to the variance in semantic change, over and above what is being explained by POS, is smaller than its individual contribution. This indicates that the two variables overlap to a certain degree, and are not fully independent. However, the fact that prototypicality adds a substantial and unique explanatory power to the regression model suggests that different independent causal elements are involved in semantic change. Our variables are unable to capture these elements in a fully independent form, but different choice of variables, at a different linguistic level, perhaps could. Nevertheless, the results support the hypothesis that the different factors involved in semantic change can be ultimately teased apart.

Variables \ Decades	1	2	3	4	5	6	7	8	9	10	pulled
POS + Prototypicality	29	24	28	25	23	22	23	19	21	22	22
POS	24	17	23	19	19	16	20	12	14	17	17
Prototypicality	10	12	10	11	7	11	8	12	12	9	10
Δ Prototypicality	5	7	5	6	4	6	3	7	7	5	5

TABLE 3. PERCENTAGES OF THE EXPLAINED VARIANCE IN SEMANTIC CHANGE WITH DIFFERENT COMBINATIONS OF VARIABLES THROUGHOUT THE 10 DECADES, AND PULLED OVER THE DECADES.

6. DISCUSSION

In the above section, we have argued that the word class assignment of a word is a distinct and significant contributing factor to the likelihood for its meaning to change over time. While, as we have noted above, the null hypothesis is that part of speech assignment does not play a role in semantic change, it is nonetheless reasonable that verbs change at a faster rate than nouns, and that both change at a faster rate than adjectives.

For an explanation, we turn to psycholinguistic research that indicates that in particular contexts, verb meanings are more likely to be reinterpreted than noun meanings. In this section, we restrict ourselves to the noun-verb asymmetry, leaving adjectives for future research. Early work on this topic (Gentner 1981) identified a processing effect known as "verb mutability"

which basically says that "the semantic structures conveyed by verbs and other predicate terms are more likely to be altered to fit the context than are the semantic structures conveyed by object-reference terms" (Gentner & France 1988: 343). Broadly speaking, this effect states that when language users are confronted with semantically implausible utterances, e.g., *the lizard worshipped*, they are more likely to reinterpret the verb's meaning than that of the collocate noun. While it would have been possible for lizard to be reinterpreted as meaning slimy man, in fact, experimental subjects preferentially reinterpreted the verb as meaning, e.g., look at the sun or some other action that lizards actually do.[6] Similarly, given the utterance *the flower kissed the rock*, English speakers did not reinterpret the meaning of the nouns, e.g., a flower-like and rock-like person kissing, but rather of the verb, interpreting *kissed* as describing an act of gentle contact (Gentner & France 1988: 345).

The verb mutability effect requires explanation. Several types of explanations have been proffered which mostly have to do with the inherent semantic and formal properties of nouns as opposed to verbs:

1. Nouns outnumber verbs in utterances (Gentner & France 1988).
2. Verbs are typically more polysemous than nouns (Gentner & France 1988).
3. Verbs are typically predicates, while nouns establish reference to objects (Gentner & France 1988).
4. Nouns concepts are more internally cohesive than verb representations (Gentner & France 1988).
5. Nouns are learned earlier than verbs, and presumably for this reason are more stable (Gentner & Boroditsky 2001).

However, all of these explanations have problems (Gentner & France 1988; Fausey et al. 2006; Ahrens 1999).

Our results do not allow us to take a position on the ultimate causal factors underlying the verb mutability effect, nor do we assume that it is universal.[7]

[6] Another line of research that may contribute to an explanation of this phenomenon is generally known as coercion, in which the meaning of a construction is "type-shifted" in appropriate contexts. For example, while the verb *know* in English has a stative default interpretation, when combined with an adverb like *suddenly*, e.g., *Suddenly, she knew it*, it takes on an inchoative meaning. Michaelis (2004) has provided a detailed theory of coercion in the framework of Construction Grammar, focusing on aspectual coercion. What we observe from the literature on coercion, although the point is not made explicitly therein, is that it is the event whose semantics is adjusted to fit the context, rather than the referring expressions.

[7] For example, Ahrens (1999) shows that the verb mutability effect observed in Mandarin is different from that observed in English, and Fausey et al. (2006) found that Japanese does not show a robust noun-verb asymmetry.

Rather, we opportunistically embrace the observation that in English, the language investigated here, this effect has been shown to be robust. Under the assumption that diachronic biases are ultimately rooted in synchronic "online" performance or usage, we expect that the tendency of verbs' meanings to be more frequently adapted to contexts of semantic strain than the meanings of their noun collocates should show up as a diachronic bias.

In fact, this is the leading hypothesis in most theories of semantic change: the interpretive strategies of language users, specifically listeners, are what lead to semantic reanalysis. For example, Bybee et al. (1994) propose that listeners' inferences cause some types of semantic change observed in grammaticalization. Traugott & Dasher (2002) make a similar argument, couching their theory in Neo-Gricean pragmatics. Detges & Waltereit (2002) propose a "Principle of Reference", according to which listeners interpret contextual meanings as coded meanings, and Heine (2002) talks about "context-induced reinterpretation". However, closest to the type of effect discussed here is Regina Eckardt (2009) principle of "Avoid Pragmatic Overload", which says that when listeners are confronted with utterances with implausible presuppositions, they may be coerced into a form-meaning remapping.[8]

Essentially, all of these theories argue that the ways in which listeners interpret semantically implausible utterances lead to biases in semantic change, and, ultimately, the appearance of "pathways" of semantic change. The verb mutability effect identified by Gentner (1981) may be one kind of synchronic interpretative bias implicated in the diachronic asymmetry observed in the present article: in terms of synchronic processing, verbs are more semantically mutable than nouns; correspondingly, in terms of diachronic change over time, verbs undergo more semantic change than nouns. However, the bridge between synchronic processing and diachronic change is not an obvious one. What does seem to be clear is that one would need an appropriate model of memory that would allow individual tokens of utterances, with their contextual meanings, to be stored as part of the representation of a word; for an example, see the exemplar-based model proposed in detail by Bybee (2010).

We would like to point out that we do not think that it is necessarily the word class as a structural label that is implicated in semantic change. Rather, we suspect, along with previous researchers, that this is but a proxy for another asymmetry: verbs, nouns, and adjectives prototypically encode different concepts, with verbs prototypically denoting events, nouns denoting entities, and adjectives denoting properties (Croft 1991, 2000, 2001). It is highly plausible

[8] Grossman et al. (2014) and Grossman & Polis (2014) have applied the latter to long-term diachronic changes in Ancient Egyptian, which provides some necessary comparative data from a language other than the well-studied western European languages.

that the diachronic asymmetry observed in this article is the result of the semantics of the concepts prototypically encoded by a word class rather than the formal appurtenance to a word class per se.

7. CONCLUSIONS

In this paper, we have proposed that a computational approach to the problem of semantic change can complement the toolbox of traditional historical linguistics, by detecting and quantifying semantic change over an entire lexicon using a completely bottom-up method. Using a word2vec model on a massive corpus of English, we characterized word meanings distributionally, and represented it as vectors. Defining the degree of semantic change as the cosine distance between two vectors of a single word at two points in time allowed us to characterize semantic change. While in earlier work (Dubossarsky et al. 2015), we argued that the degree of semantic change undergone by a word was found to correlate inversely with its degree of prototypicality, defined as its distance from its category's center, in the present article we argued that the degree of semantic change correlates with its word class assignment: robustly, verbs change more than nouns, and nouns change more than adjectives. A regression analysis showed that although these effects are not entirely independent from each other, they nevertheless complement each other to a large extent, and together account for about 25% of the variance found in the data. Interestingly, token frequency on its own did not play a role in semantic change.

These results are both reasonable and surprising. They are reasonable because part-of-speech assignment is probably a proxy for the prototypical meanings denoted by the different parts of speech. While verbs, nouns, and adjectives are formal categories of English ("descriptive categories," Haspelmath 2010), and as such, may encode non-prototypical meanings (e.g., the English word *flight* denotes an event rather than an entity), the majority of frequently encountered nouns are likely to denote entities, verbs to denote events, and adjectives to denote properties. Our results indicate that the inherent prototypical semantics of parts-of-speech does indeed influence the likelihood of word meanings to change, individually and aggregately across a lexicon.

We have addressed one part of the diachronic data observed, by relating the diachronic noun-verb asymmetry to the findings of experimental psychology: verbs not only change more than nouns over time, their meanings are also more likely to be changed in online synchronic usage, especially under conditions of "semantic strain," i.e., when language users are confronted with semantically implausible collocations. Under the assumption that semantic

change over time is the result of "micro-changes" in synchronic usage, we think it is plausible that the "verb mutability effect" may be part of a real causal explanation for the diachronic noun-verb asymmetry. To the extent that this assumption is correct, it provides further evidence for the need for rich models of memory, possibly along the lines of Bybee's exemplar-based model.

Obviously, much remains for future research. The findings presented here are for a particular language over a particular time period. The most urgent desideratum, therefore, is cross-linguistic investigation. Since the computational tools used here require massive corpora, such cross-linguistic research would demand either larger corpora for more languages, or the development of computational tools that could deal adequately with smaller corpora. Another direction for future research is to continue to identify and tease apart the causal factors implicated in semantic change: while our findings account for a considerable amount of the variance found in the data, they hardly account for all of it. It is likely that further causal factors will be found both in purely distributional factors, the semantics of individual lexical items (given a finer-grained semantic tagging), and extra-linguistic factors. For example, our results show a lack of uniformity in the total amount of change across decades in the twentieth century, a finding that may be related to that of (Bochkarev et al. 2014), which showed that the total amount of change in the lexicons of European languages over the same time period correlated with actual historical events.

Despite the preliminary and language-specific nature of our results, we believe that this study makes a real contribution to the question of semantic change, by showing that a bottom-up analysis of an entire lexicon can identify and quantify semantic change, and that the interaction of the causal factors identified can be evaluated.

REFERENCES

Ahrens, K. (1999). The mutability of noun and verb meaning. *Chinese Language and Linguistics* 5. 335–371.

Anttila, R. (1989). *Historical and comparative linguistics*. Amsterdam: John Benjamins.

Bloomfield, L. (1933). *Language*. New York: Henry Holt.

Bochkarev, V., V. Solovyev & S. Wichmann (2014). Universals versus historical contingencies in lexical evolution. *Journal of The Royal Society Interface*. 1–23.

Bybee, J. (2006). *Frequency of Use and the Organization of Language*. Oxford: Oxford University Press.

Bybee, J. (2010). *Language, Usage and Cognition*. Cambridge, UK: Cambridge University Press.

Bybee, J. (2015). *Language change*. Cambridge, UK: Cambridge University Press.

Bybee, J., R. Perkins & W. Pagliuca (1994). *The evolution of grammar: Tense, aspect, and modality in the languages of the world*. Chicago: University of Chicago Press.

Croft, W. (1991). *Syntactic categories and grammatical relations: The cognitive organization of information*. Chicago: University of Chicago Press.

Croft, W. (2000). Parts of speech as language universals and as language-particular categories. In P. Vogel & B. Comrie (eds.), *Approaches to the typology of word classes*, 65–102. Berlin: Mouton de Gruyter.

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press.

Detges, U. & R. Waltereit (2002). Grammaticalization vs. reanalysis: A semantic-pragmatic account of functional change in grammar. *Zeitschrift für Sprachwissenschaft* 21(2). 151–195.

Dubossarsky, H., Y. Tsvetkov, C. Dyer & E. Grossman (2015). A bottom up approach to category mapping and meaning change. In V. Pirrelli, C. Marzi & M. Ferro (eds.), *Word Structure and Word Usage. Proceedings of the NetWordS Final Conference*, 66–70. Pisa.

Eckardt, R. (2009). APO—avoid pragmatic overload. In M.-B. Mosegaard Hansen & J. Visconti (eds.), *Current trends in diachronic semantics and pragmatics*, 21–42. Bingley: Emerald.

Fausey, C. M., H. Yoshida, J. Asmuth & D. Gentner (2006). The verb mutability effect: Noun and verb semantics in English and Japanese. In *Proceedings of the 28th annual meeting of the Cognitive Science Society*, 214–219.

Firth, J. R. (1957). *Papers in Linguistics 1934–1951*. London: Oxford University Press.

Geeraerts, D. (1985). Cognitive restrictions on the structure of semantic change. In J. Fisiak (ed.), *Historical Semantic*, 127–153. Berlin: Mouton de Gruyter.

Geeraerts, D. (1992). Prototypicality effects in diachronic semantics: A round-up. In G. Kellermann & M. D. Morissey (eds.), *Diachrony within Synchrony: language, history and cognition*, 183–203. Frankfurt am Main: Peter Lang.

Geeraerts, D. (2010). *Theories of Lexical Semantics*. Oxford: Oxford University Press.

Gentner, D. (1981). Some interesting differences between verbs and nouns. *Cognition and brain theory* 4(2). 161–178.

Gentner, D. & L. Boroditsky (2001). Individuation, relativity, and early word learning. In M. Bowerman & S. C. Levinson (eds.), *Language acquisition and conceptual development*, 215–256. Cambridge, UK: Cambridge University Press.

Gentner, D. & I. M. France (1988). The verb mutability effect: Studies of the combinatorial semantics of nouns and verbs. In S. L. Small, G. W. Cottrell & M. K. Tanenhaus (eds.), *Lexical ambiguity resolution: Perspectives from psycholinguistics, neuropsychology, and artificial intelligence*, 343–382. San Mateo, CA: Kaufmann.

Goldberg, Y. & J. Orwant (2013). A Dataset of Syntactic-Ngrams over Time from a Very Large Corpus of English Books. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, 241–247. Atlanta, Georgia, USA.

Grossman, E., G. Lescuyer & S. Polis (2014). Contexts and Inferences. The grammaticalization of the Later Egyptian Allative Future. In E. Grossman, S. Polis, A. Stauder & J. Winand (eds.), *On Forms and Functions: Studies in Ancient Egyptian Grammar*. Hamburg: Kai Widmaier Verlag.

Grossman, E. & I. Noveck (2015). What can historical linguistics and experimental pragmatics offer each other? *Linguistics Vanguard* 1(1). 145–153.

Grossman, E. & S. Polis (2014). On the pragmatics of subjectification: the emergence and modalization of an Allative Future in Ancient Egyptian. *Acta Linguistica Hafniensia* 46(1). 25–63.

Harris, Z. S. (1954). Transfer Grammar. *International Journal of American Linguistics* 20(4). 259–270.

Haspelmath, M. (2004). On directionality in language change with particular reference to grammaticalization. In O. Fischer, M. Norde & H. Perridon (eds.), *Up and down the cline: The nature of grammaticalization*, 17–44. Amsterdam: Benjamins.

Haspelmath, M. (2008). Creating economical morphosyntactic patterns in language change. In J. Good (ed.), *Language universals and language change*, 185–214. Oxford: Oxford University Press.

Haspelmath, M. (2010). Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86(3). 663–687.

Heine, B. (2002). On the role of context in grammaticalization. In I. Wischer & G. Diewald (eds.), *New reflections on grammaticalization*, 83–101. Amsterdam & Philadelphia, PA: John Benjamins.

Hilpert, M. (2006). Distinctive collexeme analysis and diachrony. *Corpus Linguistics and Linguistic Theory* 2(2). 243–256.

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

Lehrer, A. (1985). The influence of semantic fields on semantic change. In J. Fisiak (ed.), *Historical Semantic, Historical Word formation*, 283–296. Berlin: Mouton.

Levy, O. & Y. Goldberg (2014). Dependencybased word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, vol. 2, 302–308. Baltimore, Maryland.

Menner, R. J. (1945). Multiple meaning and change of meaning in English. *Language* 21. 59–76.

Michaelis, L. A. (2004). Type shifting in construction grammar: An integrated approach to aspectual coercion. *Cognitive linguistics* 15(1). 1–68.

Mikolov, T., K. Chen, G. Corrado & J. Dean (2013a). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations (ICLR)*. Scottsdale, Arizona, USA.

Mikolov, T., Q. V. Le & I. Sutskever (2013b). Exploiting Similarities among Languages for Machine Translation. *ArXiv preprint arXiv:1309.4168*.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado & J. Dean (2013c). Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, 3111–3119.

Mikolov, T., W. Yih & G. Zweig (2013d). Linguistic Regularities in Continuous

Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751.

Newman, J. (2015). Semantic shift. In N. Rimer (ed.), *The Routledge Handbook of Semantics*, 266–280. New York: Routledge.

Řehůřek, R. & P. Sojka. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA.

Rosch, E. H. (1973). Natural categories. *Cognitive psychology* 4(3). 328–350.

Sagi, E., S. Kaufmann & B. Clark (2009). Semantic Density Analysis : Comparing word meaning across time and phonetic space. In *Proceedings of the EACL 2009 Workshop on GEMS: Geometrical Models of Natural Language Semantics*, 104–111.

Stefanowitsch, A. & S. T. Gries (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243.

Sweetser, E. (1990). *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure*. Cambridge, UK: Cambridge University Press.

Traugott, E. C. & R. B. Dasher (2002). *Regularity in Semantic Change*. Cambridge, UK: Cambridge University Press.

Turney, P. D. (2006). Similarity of semantic relations. *Computational Linguistics* 32(3). 379–416.

Haim Dubossarsky
The Edmond and Lily Safra Center for Brain Sciences (ELSC)
Hebrew University of Jerusalem
Jerusalem 91905
Israel
email: haim.dub@gmail.com

Daphna Weinshall
School of Computer Science and Engineering
Hebrew University of Jerusalem
Jerusalem 91905
Israel
email: daphna@cs.huji.ac.il

Eitan Grossman
Department of Linguistics
Hebrew University of Jerusalem
Jerusalem 91905
Israel
email: eitan.grossman@mail.huji.ac.il

Chapter 3

The Diachronic Prototypicality Effect

A bottom up approach to category mapping and meaning change.

Published

Dubossarsky, H., Tsvetkov, Y., Dyer, C. & Grossman, E. (2015). A bottom up approach to category mapping and meaning change. In *Word Structure and Word Usage. Proceedings of the NetWordS Final Conference* (eds. Pirrelli, V., Marzi, C. & Ferro, M.) 66–70.

A bottom up approach to category mapping and meaning change

Haim Dubossarsky
The Edmond and Lily Safra Center for Brain Sciences
The Hebrew University of Jerusalem
Jerusalem 91904, Israel
haim.dub@gmail.com

Yulia Tsvetkov
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213 USA
ytsvetko@cs.cmu.edu

Chris Dyer
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213 USA
cdyer@cs.cmu.edu

Eitan Grossman
Linguistics Department and the Language, Logic and Cognition Center
The Hebrew University of Jerusalem
Jerusalem 91904, Israel
eitan.grossman@mail.huji.ac.il

Abstract

In this article, we use an automated bottom-up approach to identify semantic categories in an entire corpus. We conduct an experiment using a word vector model to represent the meaning of words. The word vectors are then clustered, giving a bottom-up representation of semantic categories. Our main finding is that the likelihood of changes in a word's meaning correlates with its position within its cluster.

1 Introduction

Modern theories of semantic categories, especially those influenced by Cognitive Linguistics (Geeraerts and Cuyckens, 2007), generally consider semantic categories to have an internal structure that is organized around prototypical exemplars (Geeraerts, 1997; Rosch, 1973).

Historical linguistics uses this conception of semantic categories extensively, both to describe changes in word meanings over the years and to explain them. Such approaches tend to describe changes in the meaning of lexical items as changes in the internal structure of semantic categories. For example, (Geeraerts, 1999) hypothesizes that changes in the meaning of a lexical item are likely to be changes with respect to the prototypical 'center' of the category. Furthermore, he proposes that more salient (i.e., more prototypical) meanings will probably be more resistant to change over time than less salient (i.e., less prototypical) meanings.

Despite the wealth of data and theories about changes in the meaning of words, the conclusions of most historical linguistic studies have been based on isolated case studies, ranging from few single words to few dozen words. Only recently though, have usage-based approaches (Bybee, 2010) become prominent, in part due to their compatibility with quantitative research on large-scale corpora (Geeraerts et al., 2011; Hilpert, 2006; Sagi et al., 2011). Such approaches argue that meaning change, like other linguistic changes, are to a large extent governed by and reflected in the statistical properties of lexical items and grammatical constructions in corpora.

In this paper, we follow such usage-based approaches in adopting Firth's famous maxim "You shall know a word by the company it keeps," an axiom that is built into nearly all diachronic corpus linguistics (see Hilpert and Gries, 2014 for a state-of-the-art survey). However, it is unclear how such 'semantic fields' are to be identified. Usually, linguists' intuitions are the primary evidence. In contrast to an intuition-based approach, we set out from the idea that categories can be extracted from a corpus, using a 'bottom up' methodology. We demonstrate this by automatically categorizing the entire lexicon of a corpus, using clustering on the output of a word embedding model.

We analyze the resulting categories in light of the predictions proposed in historical linguistics regarding changes in word meanings, thus providing a full-scale quantitative analysis of changes in the meaning of words over an entire corpus. This approach is distinguished from previous research by two main characteristics: first, it provides an exhaustive analysis of an entire corpus; second, it is fully bottom-up, i.e., the categories obtained emerge from the data, and are not in any way based on linguists' intuitions. As such, it provides an independent way of evaluating linguists' intuitions, and has the potential to turn up new, unintuitive or even counterintuitive

In Vito Pirrelli, Claudia Marzi, Marcello Ferro (eds.): *Word Structure and Word Usage*. Proceedings of the NetWordS Final Conference, Pisa, March 30-April 1, 2015, published at http://ceur-ws.org

facts about language usage, and hence, by hypothesis, about knowledge of language.

2 Literature review

Some recent work has examined meaning change in large corpora using a similar bottom-up approach and word embedding method (Kim et al., 2014). These works analyzed trajectories of meaning change for an entire lexicon, which enabled them to detect if and when each word changed, and to measure the degree of such changes. Although these works are highly useful for our purposes, they do not attempt to explain why words differ in their trajectories of change by relating observed changes to linguistic parameters.

Wijaya and Yeniterzi (2011) used clustering to characterize the nature of meaning change. They were able to measure changes in meaning over time, and to identify which aspect of meaning had changed and how (e.g., the classical semantic changes known as 'broadening,' 'narrowing,' and 'bleaching'). Although innovative, only 20 clusters were used. Moreover, clustering was only used to describe patterns of change, rather than as a possible explanatory factor.

3 Method

A distributed word vector model was used to learn the context in which the words-of-interest are embedded. Each of these words is represented by a vector of fixed length. The model changes the vectors' values to maximize the probability in which, on average, these words could predict their context. As a result, words that predict similar contexts would be represented with similar vectors. This is much like linguistic items in a classical structuralist paradigm, whose interchangeability at a given point or 'slot' in the syntagmatic chain implies they share certain aspects of function or meaning.

The vectors' dimensions are opaque from a linguistic point of view, as it is still not clear how to interpret them individually. Only when the full range of the vectors' dimensions is taken together does meaning emerges in the semantic hyperspace they occupy. The similarity of words is computed using the cosine distance between two word vectors, with 0 being identical vectors, and 2 being maximally different:

$$(1) \qquad 1 - \frac{\sum_{i=1}^{d} W_i \times W'_i}{\sqrt{\sum_{i=1}^{d} (W_i)^2} \times \sqrt{\sum_{i=1}^{d} (W'_i)^2}}$$

Where d is the vector's dimension length, and W_i and W_i' represent two specific values at the same vector point for the first and second words, respectively.

Since words with similar meaning have similar vectors, related words are closer to each other in the semantic space. This makes them ideal for clustering, as word clusters represent semantic 'areas,' and the position of a word relative to a cluster centroid represents its saliency with respect to the semantic concept captured by the cluster. This saliency is higher for words that are closer to their cluster centroid. In other words, a word's closeness to its cluster centroid is a measure of its prototypicality. To test for the optimal size of the 'semantic areas,' different numbers of clusters were tested. For each the clustering procedure was done independently.

To quantify diachronic word change, we train a word vector model on a historical corpus in an orderly incremental manner. The corpus was sorted by year, and set to create word vectors for each year such that the words' representations at the end of training of one year are used to initialize the model of the following year. This allows a yearly resolution of the word vector representations, which are in turn the basis for later analyses. To detect and quantify meaning change for each word-of-interest, the distance between a word's vector in two consecutive decades was computed, serving as the degree of meaning change a word underwent in that time period (with 2 being maximal change and 0 no change).

Having two representational perspectives – synchronic and diachronic – we test the hypothesis that words that exhibit stronger cluster saliency in the synchronic model – i.e., are closer to the cluster centroid – are less likely to change over time in the diachronic model. We thus measure the correlation between the distance of a word to its cluster centroid at a specific point in time and the degree of change the word underwent over the next decade.

4 Experiment

We used the 2nd version of Google Ngram of fiction English, from which 10 millions 5-grams were sampled for each year from 1850-2009 to serve as our corpus. All words were lower cased.

Word2vec (Mikolov et al., 2013) was used as the distributed word vector model. The model was initiated to 50 dimensions for the word vectors' representations, and the window size for context set to 4, which is the maximum size giv-

en the constraints of the corpus. Words that appeared less than 10 times in the entire corpus were discarded from the model vocabulary. Training the model was done year by year, and versions of the model were saved in 10 year intervals from 1900 to 2000.

The 7000 most frequent words in the corpus were chosen as words-of-interest, representing the entire lexicon. For each of these words, the cosine distance between its two vectors, at a specific year and 10 years later, was computed using (1) above to represent the degree of meaning change. A standard K-means clustering procedure was conducted on the vector representations of the words for the beginning of each decade from 1900 to 2000 and for different number of clusters from 500 until 5000 in increments of 500. The distances of words from their cluster centroids were computed for each cluster, using (1) above. These distances were correlated with the degree of change the words underwent in the following ten-year period. The correlation between the distance of words from random centroids of different clusters, on the one hand, and the degree of change, on the other hand, served as a control condition.

4.1 Results

Table 1 shows six examples of clusters of words. The clusters contain words that are semantically similar, as well as their distances from their cluster centroids. It is important to stress that a centroid is a mathematical entity, and is not necessarily identical to any particular exemplar. We suggest interpreting a word's distance from its cluster's centroid as the degree of its proximity to a category's prototype, or, more generally, as a measure of prototypicality. Defined in this way, *sword* is a more prototypical exemplar than *spear* or *dagger*, and *windows*, *shutters* or *doors* may be more prototypical exemplars of a *cover of an entrance* than *blinds* or *gates*. In addition, the clusters capture near-synonyms, like *gallop* and *trot*, and level-of-category relations, e.g., the modal predicates *allowed, permitted, able*. The very fact that the model captures clusters and distances of words which are intuitively felt to be semantically closer to or farther away from a category prototype is already an indication that the model is on the right track.

sword, 0.06 *spear, 0.07* *dagger, 0.09*	*allowed, 0.02* *permitted, 0.04* *able, 0.06*
shutters, 0.04 *windows, 0.05* *doors, 0.08* *curtains, 0.1* *blinds, 0.11* *gates, 0.13*	*hat, 0.03* *cap, 0.04* *napkin, 0.09* *spectacles, 0.09* *helmet, 0.13* *cloak, 0.14*
gallop, 0.02 *trot, 0.02*	*handkerchief, 0.14* *cane, 0.15*

Table 1: Example for clusters of words using 2000 clusters and their distance from their centroids.

Figure 1 shows the analysis of changes in word meanings for the years 1950-1960. We chose this decade at random, but the general trend observed here obtains over the entire period (1900-2000). There is a correlation between the words' distances from their centroids and the degree of meaning change they underwent in the following decade, and this correlation is observable for different number of clusters (e.g., for 500 clusters, 1000 clusters, and so on). The positive correlations ($r>.3$) mean that the more distal a word is from its cluster's centroid, the greater the change its word vectors exhibit the following decade, and vice versa.

Crucially, the correlations of the distances from the centroid outperform the correlations of the distances from the prototypical exemplar, which was defined as the exemplar that is the closest to the centroid. Both the correlations of the distance from the cluster centroid and of the distance from the prototypical exemplar were significantly better than the correlations of the control condition (all *p's < .001 under permutations tests*).

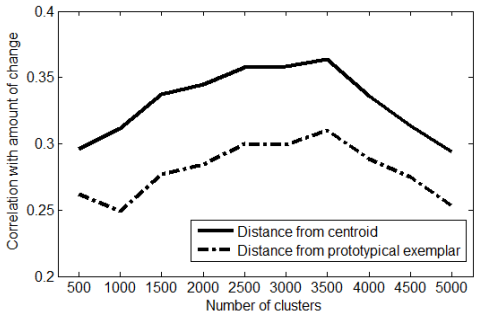


Figure 1. Change in the meanings of words correlated with distance from centroid for different numbers of clusters, for the years 1950-1960.

In other words, the likelihood of a word changing its meaning is better correlated with the distance from an abstract measure than with the distance from an actual word. For example, the likelihood of change in the *sword-spear-dagger* cluster is better predicted by a word's closeness

to the centroid, which perhaps could be conceptualized as a non-lexicalized 'elongated weapon with a sharp point,' than its closeness to an actual word, e.g., *sword*. This is a curious finding, which seems counter-intuitive for nearly all theories of lexical meaning and meaning change.

The magnitude of correlations is not fixed or randomly fluctuating, but rather depends on the number of clusters used. It peaks for about 3500 clusters, after which it drops sharply. Since a larger number of clusters necessarily means smaller 'semantic areas' that are shared by fewer words, this suggests that there is an optimal range for the size of clusters, which should not be too small or too large.

4.2 Theoretical implications

One of our findings matches what might be expected, based on Geeraert's hypothesis, mentioned in Section 1: a word's distance from its cluster's most prototypical exemplar is quite informative with respect to how well it fits the cluster (Fig. 1). This could be taken to corroborate Roschian prototype-based views. However, another finding is more surprising, namely, that a word's distance from its real centroid, an abstract average of the members of a category by definition, is even better than the word's distance from the cluster's most prototypical exemplar.

In fact, our findings are consonant with recent work in usage-based linguistics on attractors, 'the state(s) or patterns toward which a system is drawn' (Bybee and Beckner, 2015). Importantly, attractors are 'mathematical abstractions (potentially involving many variables in a multidimensional state space)'. We do not claim that the centroids of the categories identified in our work are attractors – although this may be the case – but rather make the more general point that an abstract mathematical entity might be relevant for knowledge of language and for language change.

In the domain of meaning change, the fact that words farther from their cluster's centroid are more prone to change is in itself an innovative result, for at least two reasons. First, it shows on unbiased quantitative grounds that the internal structure of semantic categories or clusters is a factor in the relative stability over time of a word's meaning. Second, it demonstrates this on the basis of an entire corpus, rather than an individual word. Ideas in this vein have been proposed in the linguistics literature (Geeraerts, 1997), but on the basis of isolated case studies which were then generalized.

5 Conclusion

We have shown an automated bottom-up approach for category formation, which was done on an entire corpus using the entire lexicon.

We have used this approach to supply historical linguistics with a new quantitative tool to test hypotheses about change in word meanings. Our main findings are that the likelihood of a word's meaning changing over time correlates with its closeness to its semantic cluster's most prototypical exemplar, defined as the word closest to the cluster's centroid. Crucially, even better than the correlation between distance from the prototypical exemplar and the likelihood of change is the correlation between the likelihood of change and the closeness of a word to its cluster's actual centroid, which is a mathematical abstraction. This finding is surprising, but is comparable to the idea that attractors, which are also mathematical abstractions, may be relevant for language change.

Acknowledgements

We thank Daphna Weinshall (Hebrew University of Jerusalem) and Stéphane Polis (University of Liège) for their helpful and insightful comments. All errors are, of course, our own.

Reference

Joan Bybee. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.

Joan Bybee and Clay Beckner. 2015. Emergence at the cross linguistic level. In B. MacWhinney and W. O'Grady (eds.), *The handbook of language emergence*, 181-200. Wiley Blackwell.

Dirk Geeraerts. 1997. *Diachronic prototype semantics. A contribution to historical lexicology*. Oxford: Clarendon Press.

Dirk Geeraerts. 1999. Diachronic Prototype Semantics. A Digest. In: A. Blank and P. Koch (eds.), *Historical semantics and cognition*. Berlin & New York: Mouton de Gruyter.

Dirk Geeraerts, and Hubert Cuyckens (eds.). 2007. *The Oxford handbook of cognitive linguistics*. Oxford: Oxford University Press.

Dirk Geeraerts, Caroline Gevaerts, and Dirk Speelman. 2011. How Anger Rose: Hypothesis

Testing in Diachronic Semantics. In J. Robynson and K. Allan (eds.), *Current methods in historical semantics*, 109-132. Berlin & New York: Mouton de Gruyter.

Martin Hilpert. 2006. Distinctive Collexeme Analysis and Diachrony. *Corpus Linguistics and Linguistic Theory*, 2 (2): 243–256.

Martin Hilpert and Stefan Th. Gries. 2014. Quantitative Approaches to Diachronic Corpus Linguistics. In M. Kytö and P. Pahta (eds.), *The Cambridge Handbook of English Historical Linguistics*. Cambridge: Cambridge University Press, 2014.

Yoon Kim, Yi-I Chiu, Kentaro Haraki, Darshan Hegde, and Slav Petrov. 2014. Temporal Analysis of Language through Neural Language Models. *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, 61-65. Baltimore, USA.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic Regularities in Continuous Space Word Representations. *Proceedings of NAACL-HLT 2013:* 746–751. Atlanta, Georgia.

Eleanor H. Rosch. 1973. Natural Categories. *Cognitive Psychology* 4 (3): 328–350.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2011. Tracing semantic change with latent semantic analysis. In K. Allan and J.A. Robinson (eds.), *Current methods in historical semantics*, 161-183. Berlin & New York: Mouton de Gruyter.

Derry T. 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 (DETECT '11*) 35-40. Glasgow, United Kingdom.

Chapter 4

Avoiding methodological pitfalls in semantic change research

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

Published

Dubossarsky, H., Grossman, E. & Weinshall, D. (2017). Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models. In *Empirical Methods in Natural Language Processing*, 1147–1156.

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:

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

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1147–1156
Copenhagen, Denmark, September 7–11, 2017.

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

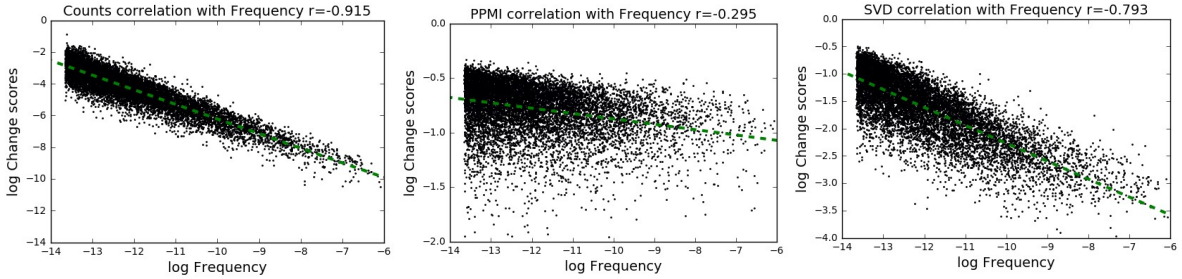
We used a cascade of processing stages based on the *explicit meaning* representation of words (i.e., word counts, PPMI, SVD, as explained below) as commonly practiced (Baroni et al., 2014; Levy et al., 2015). For each of these stages, we sought to evaluate the relationship between word frequency and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

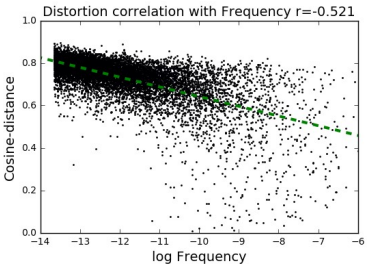
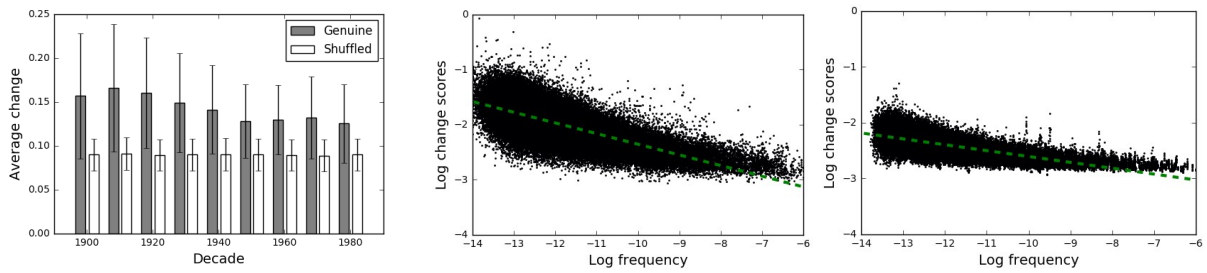


Figure 2: Cosine distances between PPMI and approximated PPMI representations (y-axis), plotted against frequency (x-axis). Correlation coefficient is reported above the plot.

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.

$$\begin{aligned} E(x-y)^2 =& E(x-\mu)^2 + E(y-\mu)^2 + \\ & 2E(x-\mu)(y-\mu) \\ =& 2\sum E(x_i-\mu_i)^2 = 2\sum Var(x_i) \end{aligned}$$

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

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

$$E(p_i^2) = Var(p_i) + E(p_i)^2 = \frac{\alpha_i(1+\alpha_i)}{\alpha_0(1+\alpha_0)}$$

$$\alpha_o = \sum \alpha_i$$

$$\Longrightarrow M = \sum E(p_i^2) = \frac{\alpha_0 + \sum \alpha_i^2}{\alpha_0(1+\alpha_0)} \quad (7)$$

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:

$$(\varphi_0 + \varphi_1)^2\varphi_2 + (\psi_0 + \psi_1)\varphi_2^2 \quad (8)$$
$$+2(\psi_0 + \psi_1)(\varphi_0 + \varphi_1)\varphi_2 + (\psi_0 + \psi_1)\varphi_2$$
$$+(\varphi_0 + \varphi_1)(\varphi_2^2 - \psi_2) > \psi_2(\varphi_0 + \varphi_1)^2$$

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*.

Chapter 5

Sense-specific word representation, a misconception?

Coming to Your Senses: on Controls and Evaluation Sets in Polysemy Research.

Under review

Dubossarsky, H., Ben-Yosef, M., Grossman, E. & Weinshall, D. Coming to Your Senses: on Controls and Evaluation Sets in Polysemy Research.

Coming to Your Senses: on Controls and Evaluation Sets in Polysemy Research

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

Abstract

The point of departure of this article is the claim that sense-specific vectors capture polysemy, which is based on performance gains in gold standard evaluation tests such as word similarity tasks. We demonstrate that this claim, at least as it is instantiated in prior art, is unfounded in two ways. Furthermore, we provide empirical data and an analytic discussion that may account for the previously reported improved performance. First, we show that ground-truth polysemy degrades performance in word similarity tasks. Therefore word similarity tasks are not suitable as an evaluation test for polysemy representation. Second, random assignment of words to senses is shown to improve performance in the same task. This and additional results point to the conclusion that performance gains as reported in previous work may be an artifact of random sense annotation, which is equivalent to sub-sampling and multiple estimation of word vector representations. Theoretical analysis shows that this may on its own be beneficial for the estimation of word similarity, by reducing the bias in the estimation of the cosine distance.

1 Introduction

Polysemy is a fundamental feature of natural languages, which typically have many polysemic words. *Chair*, for example, can refer to either a piece of furniture or to a person in charge of a meeting. Therefore both theoretical linguistics and computational linguistics seek to establish principled methods of identifying the senses that together constitute the meaning of words.

It is commonly assumed or claimed that standard word embeddings are unable to capture polysemy (Iacobacci et al., 2015), which results in suboptimal performance in gold standard evaluation tests such as word similarity tasks, and potentially hamper performance in downstream tasks. The corollary assumption is that sense-specific representations will show improved performance on these evaluation tests. This assumption is conceptually attractive, since it makes sense that sense-specific representations are more accurate than global representations. For example, in translating 'chair,' it is reasonable that performance should improve if the two senses are represented separately. This view is supported by several studies (Huang et al., 2012; Neelakantan et al., 2014; Chen et al., 2014; Li and Jurafsky, 2015), which consistently argue that sense-specific representations lead to improved performance in word similarity tasks.

In contrast, recent studies (Arora et al., 2016; Sun et al., 2017) have argued against this claim, and show that global representations are able to capture polysemic information to a great extent. Additionally, they demonstrate that this information is in fact easily accessible for evaluation tests.

Ideally, claims about polysemy should be evaluated using a gold standard evaluation set that is tailored specifically for polysemic words. As such a set does not exist, tasks involving word similarity tests have been used as a proxy (see Section 2). The underlying hypothesis is that enriching word vector representations with polysemic information should express itself in performance gains in these tasks. Unfortunately, this hypothesis has never been tested directly, and the ability of word similarity tasks to directly benefit from polysemic information must first be validated if they are to serve as genuine evaluation sets in research on polysemy. Until this is done, the validity of any

reported positive effects of sense-specific representations on evaluation tests is to be treated with caution.

In this paper our first aim is to assess the validity of *word similarity tasks* as proper evaluation tests for polysemic word representations. We use two independent corpora in order to obtain polysemic vectors: (i) a sense-annotated corpus, and (ii) an artificially-induced annotated corpus, constructed by using an established method that we modify for our purposes. Surprisingly, our analyses show that even the most accurate sense-specific word vectors do not improve performance. In fact, peak performance is achieved when polysemic information is *ignored*, i.e., when all different annotated senses are collapsed to a single word, as they naturally appear in a normal text. These counter-intuitive results indicate that the word similarity tasks are not a suitable test to evaluate polysemic representations.

Although these negative results point to the inadequacy of the evaluation test, they may also point to a suboptimal representation of polysemy by the sense-specific vectors. We therefore ask whether the representation of polysemy in existing models indeed adequately captures the phenomenon it is meant to measure. We provide evidence that currently-used sense-specific representations might not adequately capture polysemy.

We thus identify two independent pitfalls in NLP research on polysemy representation: first, the inadequacy of currently-used evaluation tests, and second, current polysemic representations. Given these conclusions, a serious question arises: why might inaccurate polysemic representations would show superior performances in inadequate evaluation tests?

An alternative explanation for the reported effects may lie in an inherent property of sense-specific representations. The procedure of assigning a word occurrence to a particular sense amounts to a sampling procedure. This sampling procedure itself, regardless of its validity, may be the true source of the reported performance gains. To test this hypothesis, we created a control condition in which word occurrences are randomly assigned to different senses. Determining that an effect is attributable to *genuine polysemy* can only be established if a similar effect is lacking or significantly reduced in this control condition.

We demonstrate that performance gains are indeed obtained for a corpus with randomly assigned senses. In addition, we modify the polysemy representation model proposed by (Li and Jurafsky, 2015) to randomly assign words to senses, and observe that the effect size remains unchanged between the original and random conditions.

In support of our empirical findings, we discuss the difficulty of obtaining an unbiased estimator for the cosine distance between two normalized random variable vectors. This may provide a partial explanation for the empirical findings, under the assumption that words are better represented as a population of vectors. Specifically, the true source of the reported performance gains may be an artifact of a purely statistical benefit that derives from the assignment of words to particular senses, or separate sub-samples, which subsequently reduces the bias of the similarity estimator.

2 Background

Previous attempts to use polysemic information for enriching word representation report marked performance gains (Huang et al., 2012; Neelakantan et al., 2014; Chen et al., 2014; Li and Jurafsky, 2015). These studies use normal unannotated corpora, and therefore have to disambiguate the different senses of words before exploiting any sense-specific information. This approach produces (i) global vectors that represent a word's meaning as a single vector (with no subdivision into distinct senses), as well as (ii) sense-specific vectors representing individual senses of words, determined in the disambiguation step, as separate vectors. For example, such approaches would represent the meaning of *chair* as a single vector, as well as distinct vectors for each of its multiple senses, e.g., "chair (person)" and "chair (furniture)."

In order to evaluate performance, the vectors created by the models are evaluated using two standard word similarity tasks, WordSim-353 (Finkelstein et al., 2001) and Stanford's Contextual Word Similarities (SCWS) (Huang et al., 2012). These tasks comprise pairs of words and the similarity scores assigned to them by human annotators. For example, the similarity between *table* and *chair* might be rated as 0.8 (i.e., human annotators found these words to be very similar, but not perfectly so), while the similarity between *table* and *tree* might be rated as 0.3 (not very similar). The models used by Huang (2012) and others produce similarity scores for each word pair by

computing the cosine-distance between the word vectors for each pair. The model's performance is then evaluated as the rank-order similarity in the order of pairs (Spearman correlation) between the human annotators' scores and the scores produced by the model. In line with the assumption discussed above, one would predict that the rank-order similarity produced by the sense-specific vectors should outperform the one produced by the global vectors. In particular, more accurate sense-specific vectors should produce better results in these tasks; conversely, better performance on these tasks is interpreted as indicating that polysemy has been captured more accurately.

Computing word similarity is straightforward when each word is represented as a single vector, but it is less so when the meaning of a polysemic word is represented by multiple sense-specific vectors. This problem of *matching* the senses relevant for a specific word pair, i.e., matching the "person" sense of *chair* with the correct sense of the word *meeting*, poses a major hurdle for meaningful comparison, and has been tackled in three different ways: (i) *average* over all similarity scores between all the different possible pairs; (ii) *weighted average* over these scores according to the probability assigned to each of them by the disambiguation model; or (iii) *selection* of the most suitable sense according to the disambiguation model, and using only the corresponding similarity score.

Common sense suggests that the third approach should outperform the others, as it is based on the clearest distinction between the relevant and non-relevant senses. However, previous studies all report results that do not conform to the naïve prediction. Rather, across studies, the best results are obtained for *average* and *weighted average*, followed by *global* (ignoring polysemy), while *selection* falls far behind the others. This counter-intuitive observation suggests that the observed benefit may be less related to sense disambiguation than previously supposed.

3 Task validation

Generally, before any task can be used as an evaluation testbed for polysemy discovery algorithms or polysemous representations, we argue that the *task* itself should be validated as suitable (or not) for the intended purpose. We propose the following **task validation** methodology: (i) Start by identifying a corpus where polysemic information is known for a significant number of words. (ii) Compute two sets of word representations: $\mathcal{A}_1$ - which computes a single representation for all words in the corpus, and $\mathcal{A}_2$ - which computes multiple representations for each polysemic word in the corpus based on the different *known* senses of the word. (iii) Evaluate the task using the two representation sets $\mathcal{A}_1$ and $\mathcal{A}_2$. Only if significant performance gains can be shown when using $\mathcal{A}_2$ as compared to $\mathcal{A}_1$, the task can be used to evaluate polysemy representation.

3.1 Polysemy induction

A major drawback of the proposed methodology is that such annotated corpora are scarce, and the largest among them is still small. We therefore articulate a methodology to generate a **task validation test** from **any corpus**, even without prior annotation of polysemy. To this end, we use a variant of the *pseudo-words* approach (Gale et al., 1992), which we call *polysemy induction*. This allows us to subject any task, including the *word similarity tasks*, to a *task validation test* which is based on a much larger corpus.

More specifically, we induce polysemy in a natural corpus by randomly selecting pairs of words, then collapsing every pair of words into a single word-form while keeping their "sense tags" as *polysemy annotation*. For example *ring* and *table* may be collapsed to a single word with two senses, $table^1$ and $table^2$ respectively. The new corpus is polysemic with respect to the collapsed words, while all other words keep a single sense. This corpus has most of the features of a natural corpus, but unlike most natural corpora (and all large corpora), it contains polysemy annotation.

With this corpus, we follow the methodology for *task validation* described above: Let $\mathcal{A}_1$ denote a set of word representations whereby every word (collapsed or natural) has a single representation, constructed without the use of polysemy annotation. Let $\mathcal{A}_2$ denote a set of word representations whereby each collapsed word has two representations, each corresponding to one natural word from the original pair of words that have been merged. For task evaluation, the task is performed while using one of the two sets of representations. Only if $\mathcal{A}_2$ leads to a significant performance gain in the task as compared to $\mathcal{A}_1$, the task under examination should be considered adequate to eval-

uate polysemy computation, i.e., the accurate disambiguation and representation of word senses.

3.2 Methods

Word embedding model (*word2vec)* skip-gram model (Mikolov et al., 2013) is used to obtain vector representations for words. The model was separately trained over the corpus, the first time producing sense-specific vectors according to the annotated senses, and a second time producing global vectors by ignoring the annotated senses and collapsing all their occurrences to a single word. Throughout the analyses we used an embedding size of 300d, a window size of 5 words from each size of the target word, negative-sampling of 5 words, and an initial learning rate of 0.025.

Sense-annotated corpus We use OntoNotes (Weischedel et al., 2013), the largest available corpus annotated for word senses. This allows us to circumvent the problem of first disambiguating the words' senses, and thus to directly test the utility of using polysemic information in word vector representations. The corpus contains 1.5 million English tokens, comprising about 50k English word types, of which 8675 word types are sense-annotated. Because the annotation is not uniform throughout the corpus (words are not annotated every time they appear), which can bias the analysis described below, we extract a subset of the corpus by removing sentences where polysemous words are not annotated, thus removing 40% of the corpus. Stopwords as well as words occurring less than 10 times are ignored by the word representation models. All words are lowercased.

Sense-induced corpus Wikipedia dump (04/2017) is the original corpus from which a polysemic version is induced by randomly pairing words into a single word-form (see Section 3.1). Stopwords and infrequent words (<300 tokens) are ignored by the word representation models.

Evaluation The utility of polysemic word representations is evaluated on the two word similarity tasks described. Crucially, the problem of *matching* the relevant sense in these tests (described in Section 2) is tackled by taking the *average* of the sense-specific representations and comparing it to the *global* word representations[1].

[1]Recall that *average* was reported to be one of the best performing matching methods is previous work.

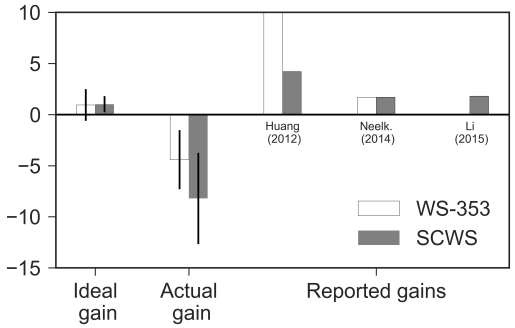


Figure 1: Summary of the results reported in Table 2, showing the *difference* between the performance of the vanilla *global* representation and the performance of various polysemy representation methods. 5 polysemy methods are shown. *Ideal gain*: when using ground-truth polysemy and known matching. *Actual gain*: when using ground-truth polysemy but no given matching, the condition which best describes a natural polysemous corpus. *Reported gain*: 3 reported results from the literature on the representation of polysemy. Color (dark and white) marks the tasks.

3.3 Results

Results clearly demonstrate that global representations are significantly superior to sense-specific representations in both evaluation tests and across corpora, as shown in Tables 1,2.

	GLOBAL	AVERAGE
WS-353	44.7 (0.7)	41.3 (0.3)
SCWS	64.0 (0.4)	62.6 (0.6)

Table 1: OntoNotes scores on two word similarity tasks. GLOBAL ignores sense information, while AVERAGE uses sense information by taking the average of the sense-specific vectors. Standard deviation for 10 independent runs of the vector model shown in parentheses.

Fig. 1 summarizes the results. We compare the performance of each polysemy representation method to the results of the vanilla global representation, which serves as the baseline that they aim to improve. Thus we show the difference between the performance of each method and this baseline. This difference is expected to be positive, if the method indeed improves performance over baseline.

As clearly (and surprisingly) seen in Fig. 1, this

	IDEAL GLOBAL	COLLAPSED GLOBAL	AVERAGE
WS-353			
INDUCED P1	70	68.3 (1.5)	61.0 (0.7)
INDUCED P2	70	68.1 (0.5)	63.9 (2.8)
INDUCED P3	70	70.8 (0.7)	69.1 (1.1)
HUANG		22.8	71.3
NEELAKANTAN		69.2	70.9
RAND. SENSES	70		69.8
SCWS			
INDUCED P1	65.7	64.2 (0.5)	50.1 (1.1)
INDUCED P2	65.7	64.8 (0.3)	56.7 (1.6)
INDUCED P3	65.7	66 (0.1)	63 (0.3)
HUANG		58.6	62.8
NEELAKANTAN		65.5	67.2
LI		64.6	66.4
RAND. SENSES	65.7		67.3

Table 2: Polysemy induced word similarity scores. IDEAL GLOBAL uses the original Wikipedia scores with complete disambiguation of senses in the training set and the evaluation test, COLLAPSED GLOBAL uses the collapsed Wikipedia scores when sense information is ignored, while AVERAGE uses the average over sense-specific representations. Several random pairing parameters were used to induce polysemy: INDUCED P1: 10k words, INDUCED P2: 6k words, INDUCED P3: 2k words. Standard deviation for 10 independent runs in parentheses. Previous results are reported for comparison.

is not the case for the condition that simulates the clearest polysemy representations, the condition termed *Actual gain* in the figure. Thus in the ideal control condition, based on *induced polysemy*, the word similarity tasks fail to demonstrate the value of polysemy representation in improving performance over the baseline. This stands in marked contrast to the results reported in the relevant literature, the 3 sets of bars which are marked in Fig. 1 by *Reported gains*. This seems to indicate that the reported gains do not reflect effective polysemy representation by these methods of true, but rather some other unknown factor which benefits performance.

Interestingly (and reassuringly), when we measure performance gain with the ideal method, which has access to the ground-truth polysemy *and* the correct sense label for each word in a word-pair in the word similarity tasks, we see performance gains (albeit small) over the vanilla method (Fig. 1, *Ideal gain*). This information is only available in the *induced polysemy* condition, and is generally not available in a natural corpus. It suggests that the failure to obtain performance gains in the word similarity tasks when using real ground-truth polysemy may be due to the need for additional information when computing similarity between words with multiple representations.

3.4 Discussion

The main result of the analysis described above is negative, demonstrating that *word similarity tasks* are not suitable to serve as gold standard tests for polysemy representation. However, the methodology we developed for *polysemy induction* constitutes a positive contribution, as it can be used to effectively test any task for its utility in the evaluation of polysemy representation while using state-of-the-art corpora. This may lead to the discovery of suitable tasks which can serve as gold standard evaluation tests for polysemy. Moreover, the use of polysemy induction adds yet another type of control to the NLP toolbox; such controls are still rarely implemented in NLP studies (but see (Dubossarsky et al., 2017)).

4 The statistical signature of polysemous representations

The reported lack of a positive effect could also stem from the inaccurate representation of polysemy by the sense-specific vectors. To evaluate this alternative, we look at the statistics of the pairwise similarity between the different sense representations. We start from the observation that polysemy is inherently defined by word senses that are distinguishable from each other. We therefore expect sense-specific representations of the same word to show a smaller resemblance to each other when genuine polysemy is captured, as compared to arbitrary assignment to senses.

4.1 Sense-specific vectors sets

Using the Wikipedia corpus, we investigated and compared 4 ways to obtain sense-specific word vector representations:

1. WIKI-INDUCED: sense-specific vectors obtained by way of polysemy induction (see Section 3.1).

2. CRP-ORIGINAL: the representations described in (Li and Jurafsky, 2015), reproduced using the provided code[2].

[2]https://github.com/jiweil/mutli-sense-embedding

3. CRP-RANDOM: representation based on the random assignment of senses, obtained by modifying the code provided by (Li and Jurafsky, 2015). We changed only the words' sense assignment process, in order to achieve random assignment of senses which is based on the same sense distribution as in the original model (see details in Section 6.1).

4. WIKI-RANDOM: random sense annotation (see details in section 6.1).

4.2 Comparing the different sets

For each word in each of the four polysemic vectors sets, the average cosine-distance between its different sense-specific vectors is computed. The distribution of these average distances within a specific set is defined as its "polysemic signature", which is then compared across sets.

The results are shown in Fig. 2, revealing marked differences in the polysemic signature of the four sense-specific vector sets. High polysemic signature is seen for the induced polysemy set, which is the only set with guaranteed semantically different senses. The two random sets have a rather small polysemic signature, while the model originally designed to recover true polysemy (Li and Jurafsky, 2015) has an intermediate polysemic signature, which is much closer to the signature of the two random sense sets.

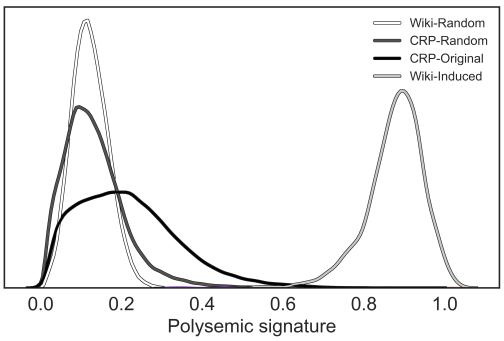


Figure 2: Density distribution of poylsemic signatures for the four sets, see text for details.

4.3 Discussion

The broader implications of these results on our research hypothesis can be understood in the context of the findings reported in Section 3.3. The results described in Fig. 1 demonstrate a marked difference between previously reported effects of improved performance, and the *actual gain* condition which shows a *worsening* of performance in the same task when polysemic information is included. The results demonstrated in Fig. 2 can be described as a negative image of those presented in Fig. 1. Specifically, the *actual gain* condition of the *induced-polysemy* has the largest polysemic signature as compared to the other conditions.

Together, these results indicate that the condition that demonstrates polysemy most clearly shows the poorest performance in the evaluation tests. The converse is also true: the conditions that demonstrate polysemy rather poorly show heightened performance in the evaluation tests. A gold standard for polysemy representation should entail that given optimal vector representations, performance on the evaluation tests would be optimal, and vice versa. Since our results demonstrate that the directions of optimal vector representation and optimal test performance are opposite, we are lead to the following conclusions: First, methods which provide improvement in the word similarity tasks may not necessarily be suitable for the recovery of polysemous vector representations. Second, the word similarity task is not a suitable evaluation test for studying the recovery of polysemy.

5 Theoretical discussion

In this section we recall and analyze some properties of the cosine distance, and describe how they may partially account for the empirical observations discussed in this article. The crucial point is to model the contextual representation of words as a distribution over some vector space.

Let X_i denote the random variable which captures the contextual representation of word i. Let $\{X_i^l, X_j^l \in \mathbb{R}^d\}_{l=1}^L$ denote a sample of such representations for words i, j respectively, where L denotes the sample size. d corresponds to the dimension of the vector space when using word2vec representations, or the number of words in the dictionary when using explicit representations (e.g., PPMI) (see Section 3.2). To simplify the analysis, we further assume that $\|X^l\| = 1 \; \forall l$.

The similarity between two words i, j can be plausibly measured (as customarily done) by the cosine distance between their contextual representations, namely, $\mathbb{E}[X_i X_j]$:

$$\mathbb{E}[X_i X_j] = \mathbb{E}[X_i]\,\mathbb{E}[X_j] + \mathrm{cov}(X_i, X_j) \quad (1)$$

Thus the average distance is not equivalent to the

distance between the average representations, with an additional bias term - $\mathrm{cov}(X_i, X_j)$ - which reflects the statistical dependence between the two vector representations X_i, X_j. This term is significant, because the contextual representation of two words is likely to exhibit strong dependence, especially when the words are more similar.

This is where the problem lies. In the process of generating words' representations, we start from a sample of sentences and generate a single representation. This representation is essentially our estimate of $\mathbb{E}[X_i]$ for word i. When multiplying two such representations in order to compute the cosine distance between them, we obtain an estimate for $\mathbb{E}[X_i]\mathbb{E}[X_j]$, which is *not* a good estimate for $\mathbb{E}[X_i X_j]$ because of the bias term in (1).

Ideally, in order to provide an unbiased estimate of $\mathbb{E}[X_i X_j]$, we should divide the sample of sentences into mini-batches, compute the appropriate contextual representation for both words i, j from each mini-batch, and then directly estimate $\mathbb{E}[X_i X_j]$ by taking the average multiplication of the corresponding representations in each mini-batch. Interestingly, in the process of generating polysemous representations, whether relying on true polysemy or arbitrary polysemy, we essentially accomplish the same goal: for each word, a mini-batch is replaced by the subset of sentences in which only one of the word's meanings is present[3].

If sense matching (see Section 2) is achieved by way of *average* or *weighted average*, it implies that our estimate of word similarity should improve with the number of senses used in the analysis, especially when the assignment is arbitrary. Of course, any improvement is hampered by the deterioration in the quality of the contextual representation computed from the smaller mini-batch sample, and therefore improvement is only expected for a small number of real or artificial "senses".

6 Performance gain revisited

The empirical findings presented so far converge on the conclusion that the performance gains reported in prior art may not stem from the utility of polysemic information, as previously claimed, but are the result of an alternative source. In the theoretical discussion we argue that random sense annotation is equivalent to sub-sampling and multiple estimation of contextual vector representations, and that this alone may be beneficial for the estimation of word similarity. A reasonable conclusion may be promoted, that sub-sampling and multiple vector estimation may have produced the reported performance gains. In this section we test this hypothesis directly.

In order to do so, we propose a simple control condition, in which senses are randomly assigned to words in a corpus, and sense-specific vectors are produced in the same way as before. Determining that an effect is reliably attributed to *genuine polysemy* can only be established if a similar effect is lacking or significantly reduced in this control condition.

6.1 Random sense assignment

We investigated two ways to achieve random sense assignment:

Sim1: Sampling from a known distribution. For the entire corpus and vocabulary (100k words), we assigned senses at random from a known probability distribution (note that (Neelakantan et al., 2014; Li and Jurafsky, 2015) also took this entire vocabulary approach). After trying both the uniform and multinomial distributions, and with different numbers of possible senses for each distribution, we empirically found that the results differed only slightly between the conditions.

Sim2: Sampling from an unknown distribution. To test the hypothesis more directly against the sense distribution used in prior work, we first reproduced sense-specific vectors using the model and code described in (Li and Jurafsky, 2015). We kept their Chinese-Restaurant-Process probabilistic mechanism, where senses are assigned to words based on the similarity of their contexts. We only shuffled the elements of the final vector of sense assignments produced by the model. In addition and for further comparisons, we used the original code unchanged to reproduce another set of global and sense-specific vectors.

6.2 Performance boost due to word sampling

The results of Sim1 show a marked performance gain for the sense-specific vectors that were produced at random as compared to the global vectors (see Table 2 under dashed-line). In fact, the effect reported in (Neelakantan et al., 2014) is replicated almost exactly, perhaps due to the fact that they also used a fix number of senses for each word as

[3] For the purpose of this discussion we ignore sentences in which a word appears more than once.

we did in this simulation. Furthermore, the results of Sim2 in Table 3 are almost identical in the original and random condition. This means that randomly assigning words to senses does not weaken the effect, even though it did change the "polysemic signature" (see Fig. 2).

	ORIGINAL		RANDOM	
	GLOBAL	AVE.	GLOBAL	AVE.
WS-353	61.0	67.8	60.5	68.2
SCWS	58.9	66.2	57.4	66.2

Table 3: Word similarity test scores based on the original (Li and Jurafsky, 2015) method, and on its modification with shuffled assignment.

Taken together, these two independent control conditions clearly show that an effect of the same magnitude as previously reported in several studies emerges under random sense assignment. Therefore, our findings strongly undermine the assumption that the reported effects are in any way related to actual polysemy.

7 Summary and Discussion

Here we investigate the validity of polysemy representation methods. First, we question the validity of using word similarity tasks to evaluate the performance of such methods. To test the claim that resolving the polysemy of words improves performance in these tasks, we used real-world polysemy in two independent conditions: (i) a human-annotated corpus that tags word senses and (ii) a corpus in which polysemy was induced in a controlled artificial fashion. In both conditions, the performance in the word similarity tasks *deteriorated*. This implies that previously observed improvements in performance in these tasks stem from one or both of the following: (i) the tasks are not suitable for evaluating sense-specific vectors, or (ii) the commonly produced sense-specific vectors do not accurately capture polysemy.

With regards to the first hypothesis, if the word similarity task is inadequate to evaluate polysemy, why would it show high gains for polysemic representations? One possibility is that polysemic informations per se is not required to drive such effects, and instead these effects are artificially caused by the procedures by which polysemous vectors are created. To prove this, we set out to demonstrate that even representations that bear no polysemic information could nonetheless yield improved performance due to a methodological artifact. We thus created a control condition, in which we randomly assign word occurrences to senses, and found that randomly-produced sense-specific vectors indeed showed a marked improvement in performance. Since this effect cannot stem from polysemy (which is lacking in this condition), it may only be the result of a methodological artifact - the sampling procedure entailed by the assignment of words to senses. Note that this experiment provides evidence against the validity of the evaluation task, but does not address the validity of representations, as both accurate or inaccurate representations would result in similar performance due to the same sampling artifact.

The existence of a sampling artifact is supported by our theoretical discussion, showing that multiple vector sampling can lower the bias of the estimator of the cosine distance between two vectorial random variables. The underlying assumption is that a better model for contextual word representation should employ a population of vectors. Thus, we demonstrate both experimentally and theoretically that the evaluation test does not provide a marker for the utility of polysemy, but is artifactually influenced by the procedure by which polysemic representations are created.

To address the hypothesis that polysemic representations may be inaccurate, we measured the similarity between word senses, as reflected by their sense-specific vectors. We found that sense specific vectors in previous studies were more similar to vectors obtained by assigning random polysemy, than to vectors obtained by polysemy induction. Nevertheless, we note that our proposed polysemic signature should be taken as a coarse "rule of thumb", and more systematic manners for evaluating the accuracy of sense-specific representations should be employed in order to fully address this hypothesis.

Essentially, the findings reported here mean that there is no solid empirical foundation to the claim that sense-specific vectors improve performance in evaluation tests. In fact, they corroborate the general impression that polysemic representations do not improve performance on most downstream tasks (Li and Jurafsky, 2015). It may be the case that sense-specific vectors can or will show heightened performance on evaluation tests, or improve downstream tasks, but this will have to be demonstrated on the basis of tasks whose suitability has been properly validated, and any effects reported

will have to be supported by demonstrating that these effects are absent or strongly reduced in a properly articulated control condition.

References

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. Linear algebraic structure of word senses, with applications to polysemy. *arXiv preprint arXiv:1601.03764*.

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *Proceedings of EMNLP*, pages 1025–1035.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of EMNLP*, pages 1136–1145.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM.

William A Gale, Kenneth W Church, and David Yarowsky. 1992. Work on statistical methods for word sense disambiguation. In *AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, volume 54, page 60.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. Sensembed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 95–105.

Jiwei Li and Dan Jurafsky. 2015. Do multi-sense embeddings improve natural language understanding? *arXiv preprint arXiv:1506.01070*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Arvind Neelakantan, Jeevan Shankar, Re Passos, and Andrew Mccallum. 2014. Efficient nonparametric estimation of multiple embeddings per word in vector space. In *Proceedings of EMNLP*.

Yifan Sun, Nikhil Rao, and Weicong Ding. 2017. A simple approach to learn polysemous word embeddings. *arXiv preprint arXiv:1707.01793*.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*.

Chapter 6

Discussion and Conclusions

This dissertation has presented a new research paradigm for the study of semantic change. It presents the first large-scale bottom-up approach that uses Vector Space Models (VSM) and focuses on research questions originating in traditional linguistics. While computational approaches have some drawbacks (see Section 1.3), they also have important advantages. First, a bottom-up approach allows more objective research, as it prevents bias in the selection of examples. Second, large-scale studies make research more reliable, as findings are more credible when they are not based on a small set of examples (even if objectively chosen). These two aspects of computational approaches may contribute to a higher standard of research in semantic change.

In addition, we address fundamental methodological issues which are critical to both this emerging field and to the NLP research community at large. The fact that these type of research questions were never tested on a large scale, combined with the lack of an empirical research tradition in NLP, have made the development and adaptation of rigorous research methodologies of paramount importance.

The benefit of our approach for semantic change research

In two experiments (Chapters 2, 3), we were able to show the usefulness of a straightforward diachronic analysis of vector representations over an entire lexicon. In Chapter 2, we examined whether words with different parts-of-speech (POS) tags, i.e., Nouns, Verbs and Adjectives, differ in their rates of semantic change. This is the first time such a question could be experimentally tested, as it required a large-scale, bottom-up analysis over an entire lexicon, which was impossible until recently. We report robust and systematic differences in the rates of semantic change that are associated with different POS types. Specifically, throughout the 20th century we find that verbs show greater changes than nouns and adjectives. This regularity in semantic change, which we call the DIACHRONIC WORD-CLASS EFFECT, is a completely novel finding, which uncovers a covert regularity of semantic change that may only be observed on a large scale.

The novelty of such a finding is further attested by the fact that there was no linguistic theory that had predicted it (or was able to account for it). In the end, it was a

theory from psycholinguistics that suggested a synchronic cognitive mechanism that we propose is the source of the word-class effect. According to the verb mutability effect (Gentner and France 1988), language speakers prefer to modify the meaning of a verb, rather than that of a noun, when they encounter an utterance with semantic mismatch (e.g., *the lizard worshiped*). We hypothesize that this synchronic preference has an accumulated effect, which may explain why overtime verbs change more than nouns.

Ultimately, the fact that an "external" theoretical explanation was required further demonstrates the potential of this bottom-up approach in contributing to semantic change research. This novel finding extends beyond existing or expected results and provides surprising ones, and thus is able to nourish theoretical breakthroughs.

The second paper (see Chapter 3) tackles a theoretical question: are there linguistic factors that make certain words more prone to semantic change? The results obtained emphasize the importance of the semantic relations between words to their likelihood of change. According to the DIACHRONIC PROTOTYPICALITY EFFECT, the degree of proximity of words to their category's prototype (e.g., the proximity of a robin or a peacock to a prototypical 'bird'), plays a decisive role in semantic change. Specifically, the more prototypical a word is (i.e., the more similar it is to its category prototype), the more "protected" it is from semantic change, and vice versa. This finding corroborates Geeraerts (1985, 1992) results, which were based on the descriptive analysis of only two verbs. Thus, this study further demonstrates the utility of our large-scale, bottom-up approach in contributing to the research for the linguistic causes of semantic change by rigorously testing and corroborating an existing theoretical hypothesis on a much larger scale.

The notion that words do not change in isolation but rather in an intricate manner that involves their relations with other words is not a novel one. For example, it has often been assumed that changes in words' meanings are due to a tendency for languages to avoid ambiguous form-meaning pairings, such as homonymy, synonymy, and polysemy (Anttila 1989; Menner 1945). On the other hand, when related words are examined together, one word's change of meaning often "drags along" other words in that semantic field, leading to parallel change (Lehrer 1985). The proposed diachronic prototypicality effect joins a recent study of Xu and Kemp (2015) in the renewal of interest in the idea that the inter-relations between words are central to predicting their trajectories of semantic change.

We hope that our research papers will contribute to the study of semantic change by tightly coupling the latest developments in NLP with linguistic questions. Our first publications were credited in recent studies that extended our approach to other aspects of semantic change or to other languages (Ponti et al. 2017; Rodda et al. 2017).

Our first two studies were unique in the landscape of this field of research. At the time when the research was carried out we gave little attention to methodological issues. Only after the studies were completed, and several other works appeared,

did we notice important methodological drawbacks that went unnoticed at first. These are, primarily, the lack of gold standard evaluation sets for semantic change and unvalidated metrics for semantic change. Traditionally, research on semantic change was not based on rigorous statistical analysis to test its hypotheses, simply because its small scale and subjective nature does not lend itself to such analyses. Likewise, statistical hypothesis testing has not been relevant for the vast majority of NLP works, most of which did not test theories. Only when these two streams of research are combined in our empirical framework do certain methodological issues surface. This is the background for last two studies in this dissertation, which take a more methodological approach, as the necessity and importance of dealing with these problems became clearer.

Important methodological guidelines to this research field and a reminder to others

The studies in Chapters 4 and 5 critically examine fundamental methodological issues in current research. Our third paper addresses the problem of assessing models of semantic change in the absence of gold standard evaluation sets, as well as validating their results. Specifically, we examine the standard metric that is used to evaluate semantic change – the cosine-distance between a word's vectors at two periods – which despite being widely used has never been directly validated. We report that this method adds a bias to the estimate of semantic change, and that the size of this bias is inversely proportional to the words' frequency (i.e., more frequent words produce a smaller bias, and vice versa). Ultimately, this bias comprises the lion's share of the semantic change scores that are commonly reported across many studies.

With respect to the former laws of semantic change that were proposed, including our own (see bullet points in Section 1.2.4), we report that the Diachronic Prototypicality Effect was significantly diminished, and its true size is half what we previously reported in Chapter 3, explaining 5 % of the total semantic changes of the whole lexicon as opposed to the 10 % that was previously reported. Significantly, the laws proposed by Hamilton et al. (2016) were far more impaired by this bias, i.e., the Law of Conformity plunged to about 15 % of its original reported size, and the Law of Innovation vanished completely. Importantly, the Diachronic Word-Class Effect that we reported in Chapter 2 is exempt from this scrutiny. This is because the words' frequency, which is the driving source of this bias, is equated between the three groups of word classes, and therefore cannot serve as an alternative account for the results.

Of course there is nothing innately biased in the cosine-distance as a metric per se. Only its interaction with certain types of word vector representation is what turns it into a biased estimator for semantic change. This is demonstrated by the lack of bias for word representations that are based on Positive Pointwise Mutual

Information (PPMI), as opposed to the large bias observed for the more common representation types (PPMI + SVD and SGNS-word2vec). Clearly, for the purpose of evaluating semantic change using the cosine-distance, PPMI representations should be used. We conclude, however, that if other representations are chosen, their results should be critically compared to a proper control condition as our paper demonstrates. Importantly, we propose a general framework to circumvent the problem of assessing models' quality in the absent of gold standard evaluation sets, as well as to verify the reliably of results obtained using such models. This framework is based on a critical comparison to a control condition and is not limited to semantic change research, but may be used in various research domains in NLP.

Interestingly, the very basis of our approach – the assumption that the distance between word vectors represent the semantic similarity between the words – has been called into question lately. Mimno and Thompson (2017), for example, have shown that vectors representations are sensitive to the hyper-parameters used in the training phase of the predictive word-embedding models (e.g., word2vec, Glove), which adds a substitutional stochastic artefact to the distances between these vectors. Other studies have analyzed an alternative approach to word meaning representations that is based on the words' closest neighbors in the embedding space, and represents semantic changes in these words according to changes to each word's closest neighbors. These studies conclude that this approach, perhaps the second most popular after the approach we use, leads to unstable word representations (Antoniak and Mimno 2017) and has low reliability (i.e., repeated initializations of the model led to markedly different results), and is specifically problematic for diachronic use (Hellrich and Hahn 2016).

Our results supplement these findings in two ways: first, by reporting a critical analysis on the other, more popular metric of semantic change that has similar reservations; and second, by advocating the importance of a principled methodological work that needs to be done in order to test the validity and reliability of any proposed variable before putting it to use in actual research. As this field of large-scale computational approach to semantic change research is still nascent, such methodological critics are well expected, but also paramount for the advancement of the research field at large.

Our last research paper (see Chapter 5) aims to improve word vectors such that they include polysemic information in a more accessible way for semantic change research. In this paper, we aim to test the feasibility of using sense-specific vectors as a more informed model for semantic change, due to their presumably more ecological representation of meaning. Specifically, we set to investigate the validity of the claim that sense-specific vectors truly represent polysemic information, a claim that was based on performance gains obtained in word-similarity tasks using these vectors. This is a necessary step before such vectors could be used as a viable tool to study changes in sense representations over time.

The crux of our paper is the validity of word-similarity tasks in evaluating polysemic word representation, because these are the main, almost solely tasks in which performance gains for sense-specific vectors were observed. In a series of critical evaluations, we found that performance gains similar in size to gains previously reported for sense-specific vectors (Huang et al. 2012; Li and Jurafsky 2015; Neelakantan et al. 2014) were obtained for vectors that were produced using random sense annotations. In contrast, we found that sense-specific vectors that were based on the most accurate sense distinctions in fact worsen performance compared to global word representations. This dissociation between sense-specific vectors and performance gains in word-similarity tasks undermines the claim that the latter are due to true polysemy distinctions that are captured by the sense-specific vectors, although it cannot completely rule out that sense-specific vectors do capture polysemy to some extent.

Importantly, our analysis does show that multiple estimations of word vectors through sub-sampling lead to performance gains, presumably as sub-sampling averages out random bias that is captured by the word vectors. This positive finding may be the true source of the performance gains obtained for sense-specific vectors, as their production is equivalent in some sense to a sub-sampling procedure.

Based on these two findings we conclude that the ability of current models of sense-specific vectors to truly capture polysemy is questionable at best, and probably false. As a result, these vectors cannot be used to in sematic change research to study changes in sense representations over time. Interestingly, recent studies have argued against the notion that polysemy must be represented using a distinct vector per sense, and showed that global representations are able to capture polysemic information to a great extent (Arora et al. 2016; Sun et al. 2017).

Overall, this paper provides a critical reevaluation of prominent results and common assumptions in polysemy research that are not limited to semantic change research. Our analyses and findings further promote the use of carefully designed control conditions and validation routines in NLP research at large. This joins our previous paper in emphasizing the importance of a meticulous evaluation, and sometimes re-evaluation, of working hypotheses that are considered fundamental, almost axiomatic, in the research field.

References

Antoniak, Maria and David Mimno (2017). "Evaluating the Stability of Embedding-based Word Similarities". In: *Tacl* 6, pp. 107–119.

Anttila, Raimo (1989). *Historical and comparative linguistics*. Vol. 6. John Benjamins Publishing.

Arora, Sanjeev, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski (2016). "Linear algebraic structure of word senses, with applications to polysemy". In: *arXiv preprint arXiv:1601.03764*.

Geeraerts, Dirk (1985). "Cognitive restrictions on the structure of semantic change". In: *Historical Semantics*, pp. 127–153.

— (1992). "Prototypicality effects in diachronic semantics: A round-up". In: *Diachrony within Synchrony: language, history and cognition*. Ed. by G. Kellermann and M. D. Morissey. Frankfurt am Main: Peter Lang, pp. 183–203.

Gentner, Dedre and Ilene M France (1988). "The verb mutability effect: Studies of the combinatorial semantics of nouns and verbs". In: *Lexical ambiguity resolution*. Elsevier, pp. 343–382.

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

Hellrich, Johannes and Udo Hahn (2016). "Bad company—neighborhoods in neural embedding spaces considered harmful". In: *Proceedings of COLING: Technical Papers*, pp. 2785–2796.

Huang, Eric H, Richard Socher, Christopher D Manning, and Andrew Y Ng (2012). "Improving word representations via global context and multiple word prototypes". In: *Proceedings of ACL*. Association for Computational Linguistics, pp. 873–882.

Lehrer, Adrienne (1985). "The influence of semantic fields on semantic change". In: *Historical Semantics, Historical Word Formation* 29, p. 283.

Li, Jiwei and Dan Jurafsky (2015). "Do multi-sense embeddings improve natural language understanding?" In: *arXiv preprint arXiv:1506.01070*.

Menner, Robert J (1945). "Multiple meaning and change of meaning in English". In: *Language*, pp. 59–76.

Mimno, David and Laure Thompson (2017). "The strange geometry of skip-gram with negative sampling". In: *Proceedings of EMNLP*, pp. 2873–2878.

Neelakantan, Arvind, Jeevan Shankar, Re Passos, and Andrew Mccallum (2014). "Efficient nonparametric estimation of multiple embeddings per word in vector space". In: *Proceedings of EMNLP*.

Ponti, Edoardo Maria, Elisabetta Jezek, Bernardo Magnini Fondazione, and Bruno Kessler (2017). "Distributed Representations of Lexical Sets and Prototypes in Causal Alternation Verbs". In: *Italian Journal of Computational Linguistics*.

Rodda, Martina A., Marco S.G. Senaldi, and Alessandro Lenci (2017). "Panta rei: Tracking semantic change with distributional semantics in ancient Greek". In: *Italian Journal of Computational Linguistics* 3.1, pp. 11–24. ISSN: 16130073.

Sun, Yifan, Nikhil Rao, and Weicong Ding (2017). "A simple approach to learn polysemous word embeddings". In: *arXiv preprint arXiv:1707.01793*.

Xu, Yang and Charles Kemp (2015). "A Computational Evaluation of Two Laws of Semantic Change." In: *CogSci*.

שינוי משמעות הולך בגדולות:
גישה חישובית למחקר של שינוי סמנטי

Semantic change at large: A computational approach for semantic change research

מאת: חיים דובוסרסקי מנחים: ד"ר איתן גרוסמן ופרופ' דפנה ויינשל

תקציר

החיפוש אחר דפוסים וחוקים המתארים מדוע וכיצד מילים משנות משמעותן לאורך זמן, טומן בחובו פוטנציאל רב לאור שתי התפתחויות חדשות בתחום: זמינותם ההולכת וגוברת של קורפוסים היסטוריים ממוחשבים, וחידושים בשיטות החישוביות המאפשרות עיבוד סמנטי אוטומטי. בהשוואה לשיטות מחקר מסורתיות בחקר השינוי הסמנטי, השילוב הזה מאפשר לאתר תבניות שינוי חדשות, ולזהות את גורמיהן – גילויים המבוססים על יתרונם של ניתוחים הנעשים בקנה מידה רחב על כלל מופעיהם של המילים בקורפוסים הטקסטואליים.

בסדרה של עבודות הדגמתי כי גישה חדשה זאת למחקר של שינוי סמנטי אינה רק אפשרית, אלא גם נושאת פרי. עבודות המחקר שלי הראו שגישה זאת יכולה להרחיב את היריעה מעבר לתופעות המוכרות, ולחשוף דפוסים לא מוכרים של שינוי סמנטי, כמו גם לבסס תיאוריות קיימות של שינוי סמנטי באמצעות ניתוחים מדוייקים ואמינים יותר.

בד בבד, מחקר זה גם חשף את חשיבותם של נושאים מתודולוגיים. מאחר ומדובר בתחום מחקרי חדש, ביקורות מתודולוגיות מסוג זה הינן צפויות, ואף חיוניות על מנת לבסס שדה מחקר זה על אדניים אמינות.

שני מאמרי האחרונים שהוקדשו לבעיות המתודולוגיות הללו, מבטיחים שמחקר השינוי הסמנטי הנעשה באמצעות גישות חישוביות יהיה מבוסס על מתודולוגיות אמינות כבר מראשיתו. במחקרים אלה לא רק ביקרתי את תקופתן של תוצאות קודמות, אלא גם הצעתי דרכים סדורות שנועדו להבטיח ניתוח אובייקטיבי ואמין של תוצאות המחקר. הדגשים המתודולוגיים הללו יכולים להועיל לקהילת ה-NLP בכללותה, משום שתובנותיהם והיישומים הקשורים בהם תקפים עבור תחומים החורגים מעבר לשינוי סמנטי.

עבודה זאת נעשה בהדרכתם של:

ד"ר איתן גרוסמן

ופרופ' דפנה ויינשל

האוניברסיטה העברית בירושלים

שינוי משמעות הולך בגדולות

גישה חישובית למחקר של שינוי סמנטי

חיבור לשם קבלת תואר

"דוקטור בפילוסופיה" ב-

מדעי המח: חישוב ועיבוד מידע

מאת

חיים דובוסרסקי

הוגש לסנט האוניברסיטה העברית בירושלים

אפריל 2018 אייר תשע"ח