

Word Embeddings and Semantic Shifts in Historical Spanish: Methodological Considerations

Hai Hu, Patrícia Amaral, Sandra Kübler
Indiana University
{huhai, pamaral, skuebler}@indiana.edu

Abstract

Word embeddings have recently been applied to detect and explore changes in word meaning on large historical corpora. While word embeddings are useful in many Natural Language Processing tasks, there are a number of questions that need to be addressed concerning accuracy and applicability of these methods for historical data. There is scarce literature on the stability and replicability of these embeddings, especially on small corpora, which are common in historical work. It also remains unclear whether methods used to evaluate embeddings in contemporary data can also be used for historical data sets.

Our overarching goal is to use word embeddings for investigating semantic shifts in the history of Spanish. In the work presented here, we focus on methodological questions that arise: We first examine the stability and applicability of three commonly used word embedding models on a small corpus of medieval and classical Spanish. Comparing our results with a study on the word *algo* as a test case, we show that a rank-averaging method can produce more stable results from the embeddings. We corroborate previous theoretical work while demonstrating the applicability of our method when training word embeddings on small corpora for the analysis of semantic change. Second, we investigate how best to evaluate different embeddings models. We show that an existing analogy test cannot be used without modification. Our new analogy test, consisting of roughly ten thousand questions for medieval and classical Spanish, will be released with the article.

1 Introduction

Word meaning changes over time, a process that linguists call *semantic change*.¹ The study of semantic change remains the least developed and most challenging field within historical linguistics (Hock and Joseph, 2009). One of the main difficulties in the field has been finding formal tools and methods that can harness the potential of digitally available data to study changes in meaning. A growing number of computational studies of semantic change rely on distributional methods to model semantic shifts (Hamilton et al., 2016a,b; Luo et al., 2019, among others; see Tang (2018); Tahmasebi et al. (2018) for an overview). Such methods make it possible for a field that has largely relied on the linguist’s intuitions to develop studies that are replicable and objectively measurable (Sagi et al., 2012).

However, distributional methods used so far in historical semantics require very large amounts of data. Previous research using word embeddings to model semantic change has generally focused on languages with large resources, such as English, and used data from the 19th and 20th century,² i.e., periods from which large sets of digitized data are available, for example, from Google Books (Michel et al., 2011), and in which the older texts generally follow spelling norms that are very similar to the modern standards. However, we

¹In Natural Language Processing, the term “lexical semantic change” is generally used.

²Exceptions are Rodda et al. (2017); Perrone et al. (2019); McGillivray et al. (2019), who investigate semantic change in Ancient Greek.

do not know whether we can apply the models from prior research (Hamilton et al., 2016a,b) in a low resource setting, for example, for cases in which only a much smaller dataset from an older period of time is available. It is also unclear whether these methods—which have been successful largely on relatively clean corpora of morphologically poor languages such as English and Mandarin Chinese (e.g. Hamilton et al., 2016b)—can still be useful in situations where morphological inflection as well as orthographical variation abounds.

Our overarching goal is to use word embeddings for investigating instances of semantic change in the history of Spanish. As a test case, here we examine the semantic shifts of the Spanish word *algo*, which was both a noun meaning ‘goods, possessions’ (e.g. *tenía muy grand algo* ‘he had many possessions’) and an indefinite pronoun of inanimate reference ‘something’ (e.g. *dar algo* ‘to give something’) in the medieval period. While the noun disappeared from the language after the Middle Ages, the pronominal meaning ‘something’ remained over time and is still found in contemporary Spanish. In addition, the word has also become a degree adverb meaning ‘a bit’ (e.g. *estaba algo cansada* ‘she was a bit tired’), with the earliest attestations being from the end of the 15th century. These three meanings have been identified and described in an existing corpus-based study that analyzes the semantic and syntactic changes undergone by the word (Amaral, 2016). This work used concordances as well as data from historical vocabularies and etymological dictionaries of Spanish. While this work necessarily has limitations, it provides us with a set of meanings of the word *algo* that have been attested and dated (approximately) in the history of Spanish. In the work presented here, we compare the results obtained from word embedding algorithms with this previous study, in order to assess the validity of the distributional methods for this type of research. While the focus of our study is restricted in terms of languages (Spanish) and to one word exhibiting semantic change (*algo*), our goal is to zoom in on methodological concerns, which are of importance not only in our specific setting, but more generally for all research using word embeddings in historical settings. Determining the validity and scalability of these methods is potentially relevant to studies of semantic change across languages and to digital humanities research using historical texts more broadly.

Our focus on one word is intentional and can be found in other studies in the field (e.g., Hengchen (2017), who adopts topic modelling in Dutch newspaper texts to trace the semantic change of one word). From a semantics perspective, it allows for an in-depth analysis of meaning. Since one of our goals is to evaluate the usability and accuracy of these methods, it is imperative that we compare the results with previous linguistic studies that provide detailed philological knowledge of the word and its possible interpretations, like Amaral (2016). In addition, the study of *algo* is representative of the type of words of interest to historical linguists. Much NLP work on semantic change focuses on semantic shifts of content words that reflect cultural or socio-historical changes (e.g., Rodda et al. (2017); McGillivray et al. (2019)). Even studies that do not just include content words analyze semantic shifts of words that retain their part-of-speech over time, e.g., Hamilton et al. (2016a); Luo et al. (2019). On the other hand, in historical linguistics, semantic change that is associated with syntactic change (e.g., involving change in syntactic category) is common and central, as it results from language-internal mechanisms of change. Such mechanisms, which are cross-linguistically attested, are crucial for linguistic theories. Our purpose with this case study is not so much to gather information on *algo* per se, but rather to show the types of information that can be obtained from word embeddings for historical linguistics and the challenges associated with these methods. If we look in depth at the neighbors obtained for one particular word, we can see the types of semantic relations tackled by the algorithms and reflect on what they reveal about the contexts in which change occurs.

In this article, we focus on three methodological questions that arise when using word embedding models on a small corpus of medieval and classical Spanish. We experiment with a historical corpus that is an order of magnitude smaller than those used in previous studies to obtain word embeddings.

More specifically, we investigate the following research questions:

1. Replicability: How replicable are the results of word embeddings models? Depending on the type

of embeddings we choose, results may vary considerably. Thus, we need to first document how well different methods for creating word embeddings can produce replicable results, and under which settings. Can we counteract the lack of replicability by averaging over the ranking of different runs of an algorithm?

2. Accuracy: How do we determine the accuracy of the different embedding models on our data? Can we use the currently available resources such as the Google analogy tests (Mikolov et al., 2013a) developed for contemporary languages? Or can we adapt the tests adequately?
3. Usability of embeddings: Given the low resource situation, can we find (enough) meaningful words in the embeddings to draw conclusions about semantic change? Do our findings correspond to prior knowledge?

To answer these questions, we use three commonly used architectures to create word embeddings models (Skip-Gram with Negative Sampling (SGNS) (Mikolov et al., 2013b,a), Global Vectors for Word Representation (GloVe) (Pennington et al., 2014), and Singular Value Decomposition (SVD) (Golub and Reinsch, 1971)) on a historical corpus of Spanish³. This allows us to address methodological questions relevant for research in diachronic semantics that uses word embeddings. To evaluate the applicability of these methods, we compare our results with those of Amaral (2016).

The remainder of the article is organized as follows. We first review relevant literature on using word embeddings to study semantic change (section 2), and then present our data sets in section 3. In section 4, we describe the experimental setup and explain the three algorithms. Section 5 reports and discusses the results with regard to their applicability in our setting. We conclude with a summary of the implications of our findings.

2 Background and Related Work

2.1 Previous Approaches to Semantic Change

At least since Bréal (1897) linguists have been interested in changes in meaning, and Ullmann (1959) investigated the causes underlying semantic change. Studying semantic change involves challenges, such as finding the tools to precisely model the interaction between semantic content, utterance context, and the mechanisms that lead to the generation of new meanings (Deo, 2015). There is no consensus regarding the best approach or the best methods to conduct such investigations. Functionalist/usage-based approaches, which focus on the different functions that language performs and on patterns of language use, tend to rely on general cognitive mechanisms (e.g. metaphor, metonymy) to explain semantic shifts⁴ (Sweetser, 1990; Bybee et al., 1994). On the other hand, formal approaches rely on model-theoretic semantics to identify the truth-conditions of the different meanings of a word over time, and explain semantic shifts by reanalysis and general pragmatic principles (Eckardt, 2006). Functionalist approaches have a long tradition in the field and have contributed evidence for the regularity of semantic shifts in grammaticalization across languages (Traugott, 1982; Hopper and Traugott, 2003). However, from a methodological point of view both functionalist/usage-based and formal approaches largely rely on the intuitions of the historical linguist and on the analysis of a small number of corpus examples (Traugott and Dasher, 2002; Eckardt, 2006). For this reason, identifying word meanings at different synchronies and explaining the transitions between them ultimately relies on an individual’s judgments.

A new approach to studying changes in word meaning comes from distributional semantics. The underlying principle behind distributional semantics is that looking at the surrounding context of a word is a way of capturing its meaning (Harris, 1954). For example, if we do a corpus search for the words that occur after

³Our code and data is available at <https://github.com/pamaral1604/SemChangeSpanish>.

⁴By “semantic shift” we mean a change from meaning *a* to meaning *b*, where *a* is attested prior to *b*.

the verb *eat* we find nouns denoting either food or meals (e.g. *food, meat, fish* and *lunch, dinner, breakfast*). If we compare *eat* with *consume* or *ingest* we see that they are similar because they share some of the objects they can take. Hence, we can say that *eat, consume, and ingest* are similar in meaning. Under this approach, “the semantic similarity of lexical items is a function of their distribution in linguistic contexts” (Lenci, 2018, p.153).

The idea that the meaning of a word can be modeled by “the company it keeps” (Firth, 1957) has important consequences for diachronic semantics. If meaning is a function of the co-occurrence patterns of a word, then changes in meaning can be traced by analyzing changes in the distribution of a word over time (Boleda, 2020, among others). Such an approach is data-driven and replicable, hence allowing the field to move away from solely intuition-based approaches. Specifically, a distributional approach to semantic change may provide objective measures of change and replicable processes to test extant hypotheses about semantic shifts. Early attempts at using distributional methods to study changes in word meaning adopted Latent Semantic Analysis (Sagi et al., 2012) and other collocation-based approaches (Gulordava and Baroni, 2011), showing the promise of distributional methods to study meaning change.

2.2 Computational Approaches to Semantic Change

The introduction of neural network representations in the field has become prevalent in more recent research. There is a sizeable number of recent studies using word embeddings to explore semantic change in English (Kim et al., 2014; Hamilton et al., 2016b,a; Del Tredici et al., 2019; Boukhaled et al., 2019). However, there is considerably less research with word embeddings in other languages; exceptions include German (Hellrich, 2019; Hamilton et al., 2016a,b; Schlechtweg et al., 2018), Ancient Greek (Rodda et al., 2017), and Mandarin Chinese (Hamilton et al., 2016b), see Tang (2018) for an overview. This work has not only shown the promise of these methods for the analysis of historical data, but has also revealed how computational measures can now be used to tackle questions regarding changes in word meaning that are of theoretical importance.

Hamilton et al. (2016b) established methodology to quantify semantic change and to evaluate these state-of-the-art approaches against known facts on semantic change. They provided objective measures of semantic similarity that are relevant for diachronic studies. In their work, the authors compared several word embeddings algorithms, SVD, PPMI, SGNS, and evaluated them against known historical changes in English, German, French, and Chinese. With respect to synchronic accuracy (i.e., the ability to capture word similarity within the same time period), SVD performed best, followed by PPMI and SGNS. As for diachronic validity (i.e., the ability to quantify semantic changes over time), the methods performed differently depending on the tasks. First, in detecting known shifts, SGNS performed the best on a large dataset, but SVD performed best on a smaller dataset (the COHA corpus). In the second task, discovering new shifts from data, SGNS performed best, followed by SVD, and PPMI performed much worse; only one out the top 10 words predicted by PPMI has actually undergone a semantic shift. These three algorithms will be used in the present study and will be explained in more detail in section 4.2. Other work on diachronic word embeddings, such as by Hamilton et al. (2016b), shows that SGNS with Orthogonal Procrustes is noisier than other approaches (c.f. Dubossarsky et al. (2017)).

Building on the work by Hamilton et al. (2016b) and addressing one of the central questions in historical linguistics – what causes semantic change? – Hamilton et al. (2016a) show how computational measures based on word embeddings can be used to disentangle changes in meaning caused by cultural shifts from those changes caused by mechanisms of change (e.g. grammaticalization and subjectification, (Traugott, 1982)). While the former are often found with nouns and reflect technological innovations and societal changes (e.g. *cell* from ‘prison cell’ to ‘mobile phone’) and are thus unpredictable, the latter show regular patterns across languages and hence are of interest to historical linguists (e.g. *actually* from ‘in actuality, in reality’ to an adverb expressing the speaker’s attitude, as in “I actually agree”). This work shows that on

the one hand, a global measure of change (i.e. “how far a word has moved in semantic space between two time-periods” (Hamilton et al., 2016a, p. 2117)) is a good indicator of semantic change caused by regular mechanisms of change. On the other hand, a local neighborhood measure (i.e. changes in the nearest semantic neighbors of a word within each decade) better accounts for culturally-induced changes in word meaning. Note that although the English examples of the two types of change compared by the authors in the paper (the words *actually* and *gay*) can be found in data from the 20th century, changes in meaning caused by regular mechanisms of change usually take place over several centuries, as is the case for the word *algo* examined here. Studying such changes that span over centuries requires having access to data from early periods of a language, which poses specific problems for the use of these computational methods.

Hellrich and Hahn (2017) and Hellrich (2019) investigate the replicability of word embeddings based on machine learning architectures such as SGNS and GloVe using data from historical English and German. Their work shows that results based on SGNS or GloVe are not replicable, i.e., they produce different embeddings and consequently different lists of closest neighbors every time they are run, even if the data and the settings are kept the same. There are several sources of instability of the embeddings, as described by Hellrich (2019, ch. 4), for example, the downsampling method used, the random initialization of vectors, and the random seed. Hellrich and Hahn (2017) recommend using SVD since this algorithm proves to be completely replicable in its results. However, they do not consider the usability of the methods in their investigation.

The work by Hellrich and Hahn is one of very few investigating the issue of replicability.⁵ However, their work focuses on a setting where a very large corpus (the Google Books corpus (Michel et al., 2011)) is available. We agree that replicability is an important issue in research, but our focus is on small data sets. While such data sets are more likely to produce unstable word embeddings, for some languages they are often the only accessible data source in historical research, especially for changes taking place over many centuries. Therefore, we examine the replicability and stability issue with regard to a small corpus of medieval and classical Spanish.

In another line of work, Schlechtweg et al. (2018) propose a general framework for annotating diachronic shifts in word meaning, where annotators are asked to rate the relatedness of sentences containing a target word from two time periods using a scale of 1 to 4. They show that their method can be used to predict different types of changes in word meaning, and that they can reach a high inter-annotator agreement. The resulting dataset of German Diachronic Usage Relatedness (DUREl) was used by Schlechtweg et al. (2019) to evaluate different methods of detecting words that underwent semantic change in DUREl. They found that the best embeddings models are SGNS, PPMI, and SVD. In 2020, Schlechtweg et al. (2020) organized a shared task on the unsupervised detection of semantic change, which has advanced the knowledge on technical approaches. While this task focuses on finding words whose meaning has changed, we start from a known instance of semantic change and focus on determining the contexts in which the change occurs. This is, by necessity, more detail oriented and requires the interpretation of a semanticist.

There exists work on creating embeddings from corpora that span multiple time periods. Since these corpora cannot be compared directly, they need special treatment, either by creating separate embeddings models per time stamp and then aligning them (Zhang et al., 2015, 2016; Hamilton et al., 2016b) or by training models for multiple time periods jointly (Bamler and Mandt, 2017; Yao et al., 2018; Dubossarsky et al., 2019). Dubossarsky et al. (2019) have shown that using Temporal Referencing (treating the target word at different time periods as different lexical items) is superior to aligning the embeddings of those time periods after training them separately. However, as for the work by Hellrich and Hahn, these methods are tested with large-scale and homogeneous corpora, and it remains to be examined whether and how they can be applied in our case where the size of the historical texts is an order of magnitude smaller.

⁵Another work that investigates the stability of word embeddings (specifically, SGNS, GloVe, PPMI, and LSA) focusing on small corpora is by Antoniak and Mimno (2018). However, their study is not on historical data.

3 Data Sets

In order to address our research questions and compare the validity of our methods for contemporary and historical data resources, we selected the following two data sets:

3.1 Contemporary Spanish: Spanish Billion Word Corpus (SBW)

For the contemporary data, we used an unannotated corpus of contemporary Spanish of almost 1,5 billion words (Cardellino, 2016). This corpus combines different Spanish corpora and other resources available on the web, including texts from Wikipedia and WikiBooks, texts in Spanish from United Nations documents, and a Spanish treebank.⁶ The corpus was originally collected to create Spanish word embeddings. We also used the word relation test set for Spanish from the same resource with adaptations, as described in section 5.2.2.

3.2 Medieval and Classical Spanish: Chronicles Corpus

Although diachronic corpora of Spanish are available (*Corpus del Español – CDE*, *Corpus Diacrónico del Español – CORDE*, *Corpus de Documentos Españoles Anteriores a 1800 – CODEA*),⁷ they can be consulted only through web interfaces and are not fully downloadable. For this reason, these large data sets cannot be used to conduct studies with word embeddings.

Our case study pertains to a change that requires data from medieval to contemporary Spanish. For the older data we used the section of Spanish Chronicle Texts (*Textos Cronísticos Españoles*), a set of 49 manuscripts and printed books with production dates ranging from 1200 to 1627, with the majority of the texts being from the 1300s-early 1500s, and containing approximately 7 million words. As noted by the editors (Gago Jover (2011)), these texts are not only an invaluable resource for historical linguists and lexicographers, but also for historians and other social scientists. The texts are freely downloadable from the *Digital Library of Old Spanish Texts* (Gago Jover (2011)), a resource on Hispanomedievalism that provides detailed and reliable information about textual sources and editions. The transcription norms are also available online.⁸ The Chronicle texts were chosen mainly because the pre-processing of narrative texts presented fewer challenges than that of legal documents or theatre plays (for details on pre-processing see section 4.1). Mostly, the narrative texts had a common structure that allowed for consistent extraction while the other texts were more variable in nature and structural properties. Using a small corpus is also more realistic overall for research problems related to historical questions.

4 Experimental Setup

4.1 Pre-processing of Corpora

The Chronicles corpus consists of transcribed text from paleographic editions of the medieval and classical works. As the text is composed of faithful transcriptions of the manuscripts and printed works, the raw data needs a considerable amount of pre-processing. First, we removed the marked indices for the folio and Column Boundary (CB), as well as all other meta information about the text that are not helpful for training word embeddings. We then removed all the brackets, including “<>” which indicate abbreviations, “<<>>”

⁶For a complete list of sources, see <https://crscardellino.github.io/SBWCE/>.

⁷These resources can be found in Davies (2001), Real Academia Española (nd), GITHE (2015). To our knowledge the only freely downloadable corpus of historical Spanish is IMPACT, see Sánchez-Martínez et al. (2013) but it contains texts first printed between 1481-1748, and for this project texts produced earlier, starting in the 1200s, were required. For a thorough overview of current corpora of Spanish and the possibilities they afford, see Benito Moreno (2019).

⁸See <http://hispanicseminary.org/manual/HSMS-manual.pdf> for their transcription manual.

[fol. 1r]
 {CB1.
 {IN5.) Estas canonicas fizo escribir el Reuerent
 en lh<es>u xp<ist>o padre don fray garcia de Eugui ob<is>po
 de Bayona delos fechos que fuero<n> fechos anti-
 gament en espan~a segunt se trueba por sc<r><<i>>pto
 en diuersos libros antigos & por que mellor se
 p<ar>ta deuedes saber que los sabios antigos p<ar>tiero<n> todos los t<iem>pos
 pasados despues que dios formo ad adam en vj hedades et
 por esto aqui digamos que cosa es hedat. Et Responden los
 sabios antigos que antigam<en>t qu<an>do porel mundo achaesc'ia
 algun grant. fecho estrayn~o que nu<n>qua oviessse achaec'ido
 fazien enel dep<ar>timj<en><<to>>. del t<iem>po hedat & clamaua<n> hedat al t<iem>po
 pasado & exo mesmo clamaua<n>. hedat al t<iem>po por venir et
 agora digamos dela p<r><<i>>mera hedat & qu<an>tos an~os turo.
 {RUB. La p<r><<i>>mera hedat}
 {IN3.) Deuedes saber. que la p<r><<i>>mera hedat enpesco
 qu<an>do n<uest>ro sen~or dios creo el mundo et formo. a adam
 et turo esta p<r><<i>>mera hedat fasta el diluuiio que noe
 / por mandamj<en><<to>>. de n<uest>ro s<en><<or>>. dios se puso. con. sus. iij. fillos et
 con sus mulleres en larq<u><<a>>. et fuero<n>. por todos. viij<<o>>. p<er>sonas et
 ouo en esta p<r><<i>>m<<e>><r>a hedat. segunt la biblia que oy es et segu<n>t
 el conto que fazen los judios mil dc.l.vj. an~os. mas.
 segunt. los. lxx. jnterpretadores. dela ley. obo. ij. mil.cc.
 l<<o>>. an~os destos. lxx. jnterpretadores dela ley dezir sea aua<n>t

Figure 1: Initial text of the *Crónica General de España* (Escorial: Monasterio X.II.22), as obtained from the Digital Library of Old Spanish Texts.

amen estas canonicas fizo escribir el reuerent en ihesu xpisto padre don fray garcia de eugui obispo de bayona delos fechos que fueron fechos antigament en españa segunt se trueba por scripto en diuersos libros antigos y por que mellor se parta deuedes saber que los sabios antigos partieron todos los tiempos pasados despues que dios formo ad adam en vj hedades et por esto aqui digamos que cosa es hedat et responden los sabios antiguos que antigament quando porel mundo achaesc'ia algun grant fecho estrayño que nunca oviessse achaec'ido fazien enel departimjento del tiempo hedat y clamauan hedat al tiempo pasado y exo mesmo clamauan hedat al tiempo por venir et agora digamos dela primera hedat y quantos años turo deuedes saber que la primera hedat enpesco quando nuestro señor dios creo el mundo et formo a adam et turo esta primera hedat fasta el diluuiio que noe por mandamjento de nuestro senor dios se puso con sus iij fillos et con sus mulleres en larqua et fueron por todos viijo personas et ouo en esta primera hedat segunt la biblia que oy es et segunt el conto que fazen los judios mil dc.l.vj años mas segunt los lxx jnterpretadores dela ley obo ij mil.cc lo años destos lxx jnterpretadores dela ley dezir sea auant

Figure 2: The text from Figure 1 after pre-processing.

which indicate superscripts, “[???” or “(???)” for illegible text, etc. Next, nasalized letters and the “&” symbol (“y” in Spanish) are normalized. Finally, all punctuation was removed and all letters lower cased. One possible way to address data sparsity resulting from the small corpus size is lemmatization. However, we decided against lemmatization since it is very difficult to perform it reliably on the historical data due to the high variability in spelling and morphology.

A sample of the unprocessed corpus is shown in Figure 1, and the same text after pre-processing is shown in Figure 2, for comparison. As can be seen from the first word in Figure 2, which is the final word of the preceding text, the processed corpus consists of continuous text that does not separate the different works. While the automatic pre-processing entails loss of information (e.g. paragraph boundaries, sentence

	...	cute	play	...	grass	milk	...
<i>dog</i>		40	50		1	5	
<i>cat</i>		38	62		0	8	
<i>cow</i>		4	5		30	50	
<i>goat</i>		2	3		35	25	

Table 1: A toy word-word co-occurrence matrix, with *dog* and *cat* appearing in similar contexts, and *cow* and *goat* respectively.

boundaries), the result is more normalized text, which makes it possible to apply the word embedding algorithms. It is possible that there are OCR errors in the text, but we did not correct them since we have no way of knowing which of the variations are spelling variations and which ones are OCR errors. However, they would not have any effect on the analysis we are performing since we use a frequency threshold when creating the word embeddings, thus eliminating very infrequent words such as OCR errors.

The contemporary corpus, SBW, is already pre-processed, and therefore we did not do any additional pre-processing.

4.2 Architectures for Creating Embeddings

4.2.1 Representing Word Meaning

Word embeddings are a representation of lexical meaning based on the distributional hypothesis, which states that words with similar meaning occur in similar contexts, as described in section 2 for the verbs *eat* and *consume*. Additionally, the hypothesis states that the similarity in word meaning is correlated with the similarity of the word contexts. Thus we can use such contexts to represent the meaning of words.

A simple way of representing such contextual meaning would be to take the co-occurrence counts of words in a text. This creates a word-word matrix, where each cell in the matrix denotes the frequency of a word (rows) and its context words (columns) in a text, see Table 1 for a small example. For instance, in the first cell, the number 40 indicates that *cute* and *dog* co-occur 40 times in a paragraph, and that *cute* and *cat* co-occur 38 times. On the other hand, *cute* is far less frequent in the vicinity of *cow* and *goat* in this text. The last two columns of the matrix show the opposite pattern with respect to frequency of co-occurrence with the words *grass* and *milk*. Based on this table, the meaning of the word *dog* would consist of its vector of numbers: $\langle \dots, 40, 50, \dots, 1, 5, \dots \rangle$.

However, such tables are very sparse matrices since there will be many dimensions (rows and columns), and most of the cells will have zero values. In order to make the representations more usable, we employ dimensionality reduction techniques. By reducing the dimensions, we obtain dense representations, which hopefully abstract away from accidental differences in counts and represent the similarities between words in a compact way. If we reduce the dimensionality further to 2, we can display the word meanings, as shown in the left graph in Figure 3.⁹ We see that not only words denoting animals (in red) cluster together, but also subclusters: *dog* and *cat* appear next to each other, and there is a cluster of ruminants and one of insects and arachnids. This closeness in the semantic space, a visual representation of the distribution of these words, reflects our intuition that domestic animals, like dogs and cats, are similar in some respects, and different from grass eaters and insects.

Thus, we have a very useful method not only to visualize but also to quantify the semantic distance between words. For example, we can use any existing similarity metric to compute the similarity of two

⁹300 dimension GloVe model downloaded from <https://github.com/stanfordnlp/GloVe>, showing a small corner of the visualization, based on t-SNE (Maaten and Hinton, 2008) dimensionality reduction, of the top 0.1% of the vocabulary plus animals in red.

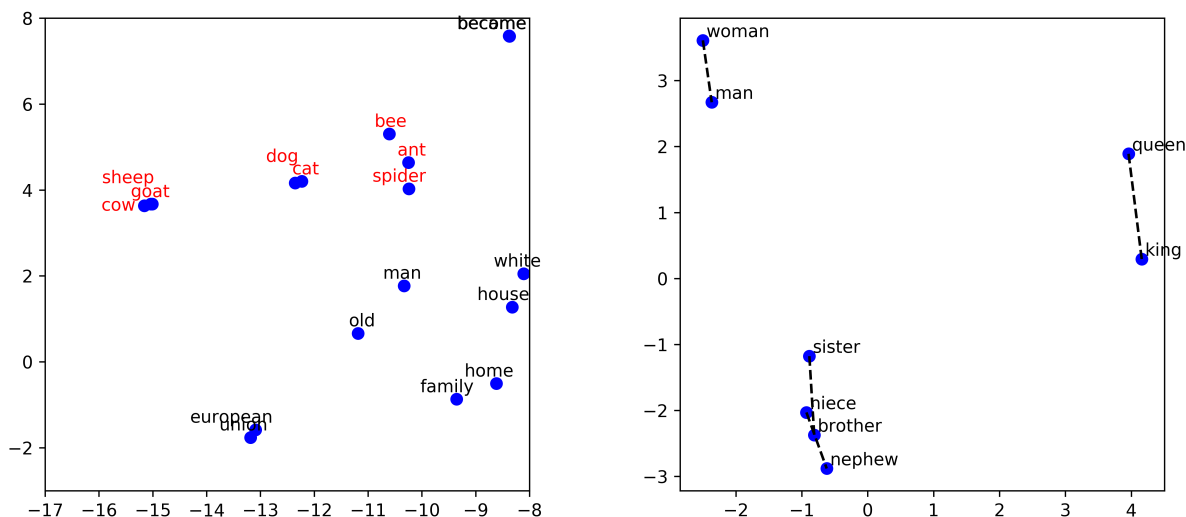


Figure 3: Example semantic space from a GloVe model trained on 6 billion tokens in English, showing clusters of different animals, using t-SNE for dimensionality reduction (left); visualization of analogy relations, using PCA for dimensionality reduction (right). Note that t-SNE favors cluster coherence whereas PCA preserves distances.

words, as represented by their vectors of numbers. We adopt cosine similarity to measure the “closeness” of two words in the semantic space. The cosine similarity is a common measure for similarity using concepts from linear algebra. If we interpret a vector as coordinates in an n -dimensional space, the cosine represents the angle between two vectors, where a small angle is equivalent to similar vectors while a large angle means dissimilar vectors.

Using the word vectors, we can also examine analogy relations between pairs of words, as shown in Figure 3 (right),¹⁰ where an almost parallel pair of lines can be drawn between *man* : *woman* and *king* : *queen*, suggesting that such analogies are captured in the vector space.

One of the first dimensionality reduction techniques that was used to create word embeddings is Singular Value Decomposition (SVD) which was used for latent semantic indexing (Landauer et al., 1998). More recent models use neural networks to create these dimensions. These are trained on tasks such as “predict the next word after word x ”, and then we use the activation of cells in the neural network as the dimensions to represent word meaning. The task is not directly relevant, but is chosen so that it forces the neural network to focus on the context of a word. However, neural networks require extremely large data sets for training. Skip-Grams with Negative Sampling (SGNS), also often called word2vec, and Global Vectors for Word Representation (GloVe) are two popular early models. For the purpose of investigating semantic shifts, embeddings are generally used to compute the similarity between words to find the neighbors of the word in question (Hamilton et al., 2016b). Note that the choice of specific neural network model, the task, the definition of context, and the choice of similarity metric all influence the results.

There are several informative resources for explaining the embeddings and their hyper-parameters, for example, Levy et al. (2015), Kutuzov et al. (2018), chapter 6 of Jurafsky and Martin (2019), and we refer interested readers to those for a more detailed description.

¹⁰Same GloVe model, but using PCA for dimensionality reduction.

4.2.2 Word embedding algorithms used in this study

Skip-Gram with Negative Sampling (SGNS) This model was introduced by Mikolov et al. (2013a,b). The model is based on a logistic regression model with the task of assigning a probability to a word and a context, for example, how likely it is that the word *cow* occurs with the context words *cute* to the left and *plays* to the right. SGNS uses positive examples, i.e., examples that occur in the text, and negative examples, i.e., a combination of word and context that does not occur; the latter is called “negative sampling”, hence the name of the model. Since the number of possible negative samples is much larger than the positive samples, their ratio can be controlled, by a hyper-parameter.

Global Vectors for Word Representation (GloVe) As indicated by its name, GloVe (Pennington et al., 2014) uses a more global view of word co-occurrences. It also uses ratios of probabilities rather than the actual probabilities, as SGNS does. GloVe has been shown to perform on a par with SGNS (Pennington et al., 2014; Levy et al., 2015), and has been used in studies of semantic change (Hellrich, 2019).

Singular Value Decomposition (SVD) This type of embeddings first computes the Positive Pointwise Mutual Information (PPMI) for each cell. PPMI is an extension of Pointwise Mutual Information (Church and Hanks, 1990), which measures the “surprise” of seeing two words together. To obtain a dense matrix, Singular Value Decomposition (SVD, Golub and Reinsch (1971)) is used.

Algorithms and Hyper-Parameters For SVD, we use Omer Levy’s hyperword implementation.¹¹ For GloVe, we use the original implementation.¹² For SGNS, we use the `gensim` implementation for SGNS¹³ as it is easy to train and provides interfaces to compute accuracy of analogy tests.

We use the following hyper-parameters for all experiments, as discussed in more detail in Sec. 5.2.1. For all three algorithms trained on the SBW corpus: minimal word frequency threshold = 20; window size = 14, vector dimensions = 300. $c.d.s = 0.75$. For all three algorithms trained on the Chronicles corpus: minimal word frequency threshold = 20; window size = 10, vector dimensions = 50. Specifically, for SVD models: $c.d.s$ (context distribution smoothing) = 1.

5 Methodological Investigations

5.1 Replicability

5.1.1 Replicating Prior Results

For a semanticist who uses lists of similar words based on embeddings to determine semantic shifts, a lack of replicability, as reported by Hellrich (2019), is a major concern: If there is variability between different runs of the same embeddings algorithm with the same settings, this potentially results in different lists of closely related words every time. This means that the interpretation of the semanticist would be based on one specific run of the algorithm, and may look different if a different run had been used. We ask: Will the randomness in the embeddings result in qualitatively different results and interpretations of the data? Do the results by Hellrich and Hahn (2017) on German texts from the 19th century translate to our small set of Spanish texts, which additionally exhibit a high degree of spelling variation? We investigate this by running each algorithm with the same settings three times but with different seeds, and then compare the lists of the 10 words determined by the resulting embeddings to be the most similar words to *algo*.

¹¹<https://bitbucket.org/omerlevy/hyperwords/>

¹²<https://github.com/stanfordnlp/GloVe>

¹³<https://radimrehurek.com/gensim/index.html>

Algorithm seed	Analogy accuracy	Most similar words
SBW corpus		
SGNS 1	45.7	nada, realmente, eso, mucho, bastante, cosa, alguien, porque, aspecto, demasiado
SGNS 2	44.6	nada, realmente, mucho, eso, bastante, cosa, alguien, porque, aspecto, tan
SGNS 3	45.6	nada, realmente, eso, mucho, bastante, cosa, alguien, porque, tan, demasiado
GloVe 1	42.8	nada, eso, parece, cosa, mucho, poco, parecido, bastante, cierto, porque
GloVe 2	43.1	nada, eco, parece, bastante, poco, mucho, porque, cosa, realmente, cierto
GloVe 3	43.5	nada, eso, parece, mucho, alguien, cosa, proque, bastante, poco, realmente
SVD	25.0	nada, eso, mucho, pensar, realmente, imaginarlo, bueno, quizá, parece, quizás
Chronicles corpus		
SGNS 1	27.3	aueres, pro, heredades, demas, mayordomos, ualiesse, dones, soldadas, prometer, heredamientos
SGNS 2	28.1	aueres, abenagit, dones, criar, heredades, demas, pro, mayordomos, soldadas, conducho
SGNS 3	25.5	aueres, demas, pro, soldadas, ualiesse, criar, dones, abondados, enbiolos, enbargo
GloVe 1	19.9	demas, auer, nada, quanto, farie, dones, dar, daua, pan, comer
GloVe 2	19.8	dones, sabor, demas, quanto, ganar, auer, aueres, comer, pan, nada
GloVe 3	20.2	auer, ganar, quanto, dones, nada, demas, dar, darie, farie, sabor
SVD	23.8	nada, mester, quanto, dar, ello, demas, gelos, gelo, recabdo, rentas

Table 2: 10 most similar words to *algo* from three embedding models trained on the SBW and Chronicles corpus, across 3 runs per model. Accuracy results are based on the MTS+ours test set, as described in section 5.2.

Table 2¹⁴ shows the results of our experiments on the contemporary corpus SBW and the historical Chronicles corpus, focusing on the 10 most similar words.¹⁵ For both corpora, the SVD results are stable: SVD is deterministic since no seed can be set. For this reason, we show the results only once in the table. For SGNS and GloVe, we see some differences in ranking on the contemporary corpus, and a minor variation in SGNS between *aspecto*, *demasiado*, and *tan*.

The historical results, however, show a significant degree of variation for both SGNS and GloVe. SGNS consistently chooses *aueres* as the most similar word, but for the next positions, there is no agreement. If we consider the first 10 words as an unordered set, 4 words appear in all 3 lists, and 4 more in two. For GloVe, the three runs share five common words (*demas*, *auer*, *nada*, *quanto*, and *dones*), and another seven words occur in 2 lists. However, note that the overlap would have been significantly lower if we had looked at only 5 words per run. We also computed the common words among the lists of the 50 most similar words from the three runs, to see if there is less variation in larger sets. For SGNS, the three sets have 28 words in common, for GloVe 35.

Our results corroborate the findings by Hellrich (2019) that the gensim implementation of SGNS will produce different results unless all of the following criteria are met: 1) Set the random seed, 2) set the

¹⁴The second column, analogy accuracy, will be explained in section 5.2. We give this information here in order to provide a complete view of our results.

¹⁵A more typical approach would be to use the cosine distance between vectors of the same word in different time bins to assess semantic change. However, we do not focus on finding words that have changed meaning, but rather we focus on an in-depth understanding of how the word has changed, thus this method is less relevant for us.

set 1 of 3 SGNS runs		set 2 of 3 SGNS runs		set 3 of 3 SGNS runs	
word	Rank	word	Rank	word	Rank
aueres	1.00	aueres	1.67	aueres	1.00
pro	4.00	pro	2.00	heredades	4.33
demas	4.00	demas	3.33	pro	5.67
dones	5.67	dones	4.33	abenagit	6.67
heredades	6.67	heredades	5.67	heredamjentos	8.00
soldadas	7.00	criar	7.33	dones	9.00
criar	7.00	soldadas	7.67	enbiolos	10.33
ualiesse	8.67	heredamjentos	10.00	soldadas	10.33
abenagit	9.33	abenagit	12.00	demas	12.33
mayordomos	10.67	prometer	17.33	ualiesse	14.67

Table 3: *Rank* of words among 3 sets of SGNS runs, each set consisting of 3 runs.

PYTHONHASHSEED, and 3) do not parallelize, use only a single thread. For GloVe, in order to obtain a completely deterministic result, we need to set the seeds for the `shuffle` and `glove` commands. If we do not take these precautions, we have significant differences between the different runs.

5.1.2 Creating more Stable Word Lists

The replicability results lead to the question of whether we can stabilize the results by averaging the words’ ranks over n runs instead of randomly choosing the results from one run.

For this purpose, we compute the average rank for every word w in the ranked lists of the neighbors of the word in question:¹⁶

$$Rank(w) = \frac{1}{n} \sum_{i=1}^n rank(w, model_i)$$

For example, assuming $n = 3$, if a word is rated to be the most similar word by three runs, it is assigned the rank of 1. Another example: If a word is ranked as the most similar (rank 1), the 7th, and the 11th most similar word by the three runs, its averaged rank would be 6.33. Note that the rank only determines the list of words we use for the manual inspection; the number per se is inconsequential.

Since it is not obvious which value we should assume for n , we determine the best n experimentally: We run each model 9 times, using different seeds, and then we randomly sample subsets of these 9 runs of different size in order to determine the variability across subsets with the same number of runs. In other words, we sample n runs from the 9 runs, with $1 \leq n \leq 6$. We repeat the sampling 3 times, thus obtaining 3 samples, each sample containing n ranked lists of neighbors. We then average over these n runs. Finally we examine the standard deviation of the mean ranks from the 3 samples to determine at which n the standard deviation flattens out. Figure 4 visualizes the results for SGNS and GloVe for both corpora.

First, there is a (sharp) decrease in the variance as n increases, especially for the small Chronicles corpus. Since we need a single n across all settings, we decided to use $n = 3$ since this is where the curves flatten out. Note that the mean ranks in Figure 4 are for the top 30 neighbors.

We then look at the word lists, showing the mean ranks for each sample (for $n = 3$) using SGNS on the Chronicles corpus, as shown in Table 3. We see from these rankings that the results are more stable: Seven words are shared by all three lists while only three words are not shared by the three sets.

This approach of averaging over three runs has several advantages: First, it can help us delete “noise” from the results. For instance, *abenagit* is chosen by the second model of SGNS in Table 2 as the second closest word to *algo* in the Chronicles corpus, but does not occur among the 10 most similar words in the

¹⁶When there are ties, which occur occasionally, we consider these words to have the same mean rank, and treat them equally.

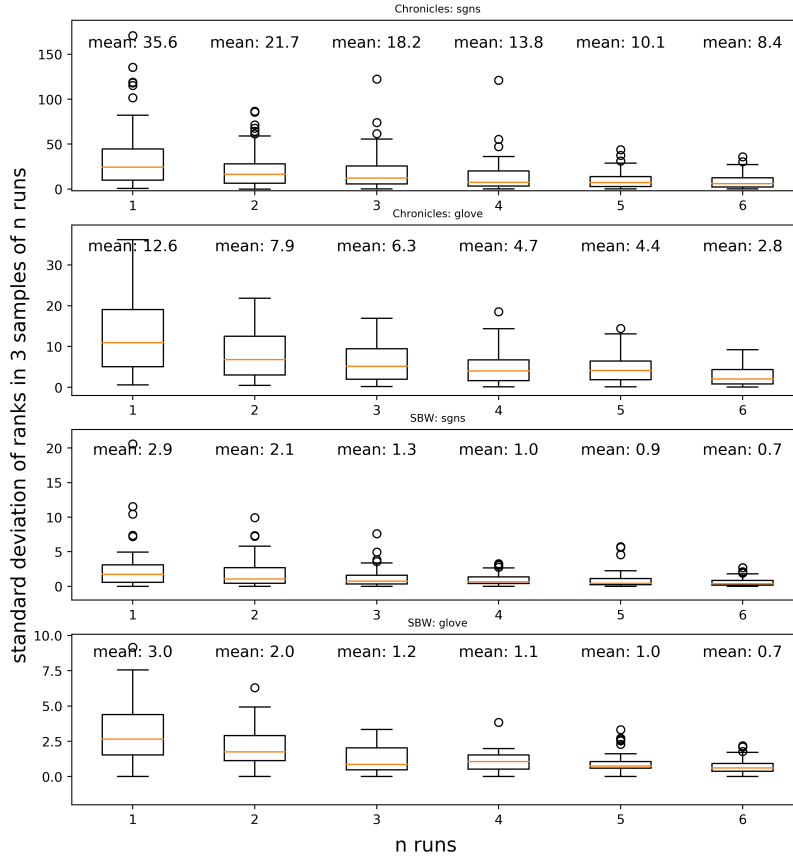


Figure 4: Standard deviation of mean ranks as a function of n runs. Models trained on the small Chronicles corpus (top 2) have more variance than models trained on SBW (bottom 2). The mean of the standard deviation in the three samples is listed in each plot, where a mean close to zero would indicate that the three samples have exactly the same rankings for the words.

other two runs (see Table 2). *Abenagit* is a proper name, the name of a king mentioned in some of the chronicles. It is chosen because a common collocation in medieval Spanish is *fijos de algo* (lit.: ‘children of possessions / riches’; cf. Eng.: noble men), which is similar to the expression found in the Chronicles texts, *los hijos de Abenagit* ‘the children of Abenagit’. By averaging across the three runs, we obtain the average ranks for the word (see Table 3), which gives us a more reliable indication of the position of the word in terms of similarity to *algo*. The second advantage is that averaging makes the results more robust because for a word to have a high average rank, it has to be high ranking in all the models.

Finally, this method can also be applied across algorithms. That is, instead of choosing between different algorithms for our semantic analysis, we can compute the *Rank* across all the nine runs (i.e., the 3 runs of SGNS, 3 runs of GloVe, and 3 runs of SVD). This ranking is more informative than the individual results because it takes into account the information from 3 runs of each model, i.e., it reflects the aggregated information from different algorithms (see the last column of Table 4). For instance, the word *abenagit* does not appear in the last column of Table 4.

Table 4 shows the results of averaging over 3 runs each for SGNS, GloVe, and SVD for the SBW and the Chronicles corpus, along with the results for averaging the three runs of all three algorithms. These results show that the averages over the individual algorithms are fairly stable for the contemporary corpus, as indicated by the low ranks. Averaging over the three algorithms results in considerably higher ranks: 5

SGNS		GloVe		SVD		Three algorithms	
word	<i>Rank</i>	word	<i>Rank</i>	word	<i>Rank</i>	word	<i>Rank</i>
nada	1	nada	1	nada	1	nada	1
realmente	2	eso	2	eso	2	eso	2.57
eso	3.33	parece	3	mucho	3	mucho	4.14
mucho	3.67	mucho	5	pensar	4	realmente	6.57
bastante	5	cosa	6	realmente	5	bastante	8.14
cosa	6	poco	6.67	imaginarlo	6	parece	10.14
alguien	7	bastante	6.67	bueno	7	porque	11.43
porque	8	porque	8	quizá	8	alguien	13.43
aspecto	10	alguien	10	parece	9	quizás	13.71
tan	10	cierto	11	quizás	10	tan	13.86

(a) *Rank* of words trained on the **SBW** corpus.

SGNS		GloVe		SVD		Three algorithms	
word	<i>Rank</i>	word	<i>Rank</i>	word	<i>Rank</i>	word	<i>Rank</i>
aueres	1	auer	3	nada	1	demas	4
pro	4	demas	3.33	mester	2	dones	6.14
demas	4	quanto	3.67	quanto	3	nada	11
dones	5.67	dones	3.67	dar	4	aueres	12.57
heredades	6.67	nada	6	ello	5	auer	15.71
soldadas	7	ganar	6	demas	6	sabor	29.43
criar	7	dar	8.67	gelos	7	dioles	32.57
ualiesse	8.67	sabor	8.67	gelo	8	precio	44.86
abenagit	9.33	farie	9	recabdo	9	dineros	46.14
mayordomos	10.67	delo	13.33	rentas	10	pro	46.57

(b) *Rank* of words trained on the **Chronicles** corpus.

Table 4: *Rank* of words trained on two corpora, averaging across 3 runs for SGNS and GloVe.

words have an average rank over 10, showing the higher differences between the individual algorithms. For the Chronicles corpus, we see similar trends, but more pronounced. Averaging over the three algorithms here results in 4 words having an average over 30, thus indicating major differences between algorithms.

5.2 Determining the Accuracy of Embeddings

Our next question concerns how to decide between the different embeddings. Ultimately, we need to know which type of word embedding gives us the best results for investigating semantic change. However, since this is an open research question, we need to determine if we can use other, existing methods for determining the accuracy of the embeddings. Currently, there exist two methods: 1) intrinsic evaluation based on analogy and similarity tasks, and 2) extrinsic evaluation in downstream NLP tasks such as Named Entity Recognition (Mundra and Socher, 2015). Since in our setting, there is no downstream task with quantifiable evaluation measures, we use a battery of analogy tasks (Mikolov et al., 2013a) to determine the quality of the embeddings. These tasks make sense for our purposes since they are different from our task but still meaning-related. The similarity task would be too similar to our own investigation, so we decided not to use it. We assume that if a type of embeddings reaches a high quality on these tasks, this translates into applicability for investigating semantic shifts. This hypothesis will be further tested in the next question.

Before we can start determining quality, however, we need to decide on the hyper-parameter settings for the embeddings algorithms since it is a well known fact that differences in hyper-parameters can influence results considerably (see e.g., [Hutter et al. \(2014\)](#)). Since both tasks are intrinsically related, we will first discuss hyper-parameter tuning (which uses the analogy test to determine the best setting), and then describe the analogy test and the performance of the individual embeddings (given the final hyper-parameters).

5.2.1 Hyper-Parameter Tuning

The first question that we need to decide is which hyper-parameters we should use for generating the embeddings. We can 1) use the ones suggested in prior work ([Levy et al., 2015](#); [Hamilton et al., 2016b](#)), or 2) we can perform tests to determine the optimal settings. Since our data set is smaller by an order of magnitude than the one used by [Hamilton et al. \(2016b\)](#), we assume that the former approach will give misleading answers. Thus we have decided to use the latter approach. To evaluate the quality of the resulting embeddings, we use the analogy test described below. However, we will first review one previous study on tuning the hyper-parameters on very large corpora, and then turn to our experiments.

[Levy et al. \(2015\)](#) performed a comprehensive study of the effect of hyper-parameter choices on the results of intrinsic evaluation (word similarity tests and word analogy tests). They experimented with a multitude of hyper-parameters – window size, dynamic context window, context distribution smoothing, subsampling, etc. – on four embeddings algorithms: PPMI, SVD, SGNS and GloVe. Their results show that having context distribution smoothing ($c_{ds} = 0.75$) is the only setting that always gives better performance while other hyper-parameter choices depend on the specific embeddings algorithm and the evaluation task used in the experiment.

In our investigation, we experimented with the following hyper-parameters: `window size = {3, 5, 7}`, `vector dimension = {50, 100, 200}`, `context window smoothing (cds) = {0.75, 1}` where 0.75 is the smoothed and 1 is the unsmoothed context window. This context window smoothing is first introduced by [Mikolov et al. \(2013b\)](#), who raise the counts of context words in the unigram distribution from which the negative examples are sampled to the power of 0.75, to increase the probabilities of sampling rare words in SGNS. [Levy et al. \(2015\)](#) demonstrate that there is an analog in PPMI-based algorithms for the same hyper-parameter. One can also raise all the counts of context words to 0.75 and perform the rest of the computation as in the PPMI algorithm. Thus we tune c_{ds} in the SGNS and SVD algorithms.

We chose c_{ds} rather than other hyper-parameters because it is the only one that gives a performance boost in all cases reported by [Levy et al. \(2015\)](#). We leave a more thorough exploration of all the hyper-parameters in small corpora and low-resource scenarios to future work.

Our results corroborate our assumption: Settings that produce higher performance for large corpora, as shown by [Levy et al. \(2015\)](#), may not lead to better results in our low-resource scenario. Our experiments show that smoothing for SVD is detrimental to the results of the analogy test. More specifically, using context distribution smoothing yields 20.1% accuracy on our tailored analogy test, compared to 23.8% without smoothing. The same trend is observed for SGNS. And while [Levy et al.](#) found that SVD prefers shorter context windows, we did not see any clear trend for context windows. It is impossible to say whether this is due to the corpus size or the choices in the analogy tests. Thus for studies using small corpora, the hyper-parameter recommendations derived from training with much larger corpora may need to be re-examined to see if they produce the optimal results.

5.2.2 Spanish Analogy Test

We now need to determine the quality of the word embeddings produced by the algorithms described above. The most intuitive approach for a semanticist would be to evaluate the embeddings model based on what we expect from prior work on semantic change. However, this means that we would evaluate how many of

the already *known* contexts we can find with a model. One of the reasons for using word embeddings is to obtain new insights. If we select the computational method based on what we already know, we may curtail new insights.

Hence, as mentioned above, we decided to perform an evaluation based on an analogy test following Mikolov et al. (2013a). In this type of test, we need word pairs that stand in a certain relationship, for example *big* and *biggest*. Then we can ask which word is in the same relationship to *small*, expecting the word *smallest* as the correct answer. In order to find this word, we perform algebraic operations on the word vectors to model the following reasoning: We start with *biggest*, then take away the adjective *big*, which should leave us with a representation of the superlative. Then we add the new adjective *small*, which should result in *smallest*. Performing these operations on embeddings means using addition and subtraction: $X = \text{embedding}(\textit{biggest}) - \text{embedding}(\textit{big}) + \text{embedding}(\textit{small})$. Then we retrieve the word in the vector space that is closest to X . If this word corresponds to the superlative of *small*, i.e., *smallest*, we count this as correct. Thus, in order for this method to work, we need sets of word pairs, based on different lexical relationships.

There is a Spanish analogy test set (Cardellino, 2016) created by translating the English set by Mikolov et al. (2013a) into Spanish (to which we refer as MTS). Since the coverage of this test set on our Chronicle corpus was minimal (see next section), we used this test set as our basis, then filtered out all the word pairs not covered, and extended it to cover other relationships. For the extension of the test set, we followed the approach by Li et al. (2018), who adapted the set by Mikolov et al. (2013a) to Chinese morphological patterns.

We could not use the following tests from the analogies in MTS (Cardellino, 2016): capitals of common countries, capitals of the world, city in state, currency, and nationality adjectives. These are not used in the Chronicles corpus.

We used the family set from the original analogy set but filtered out some words because they were not used in the corpora (e.g., *mamá*, *papá*). We also filtered out the test “adjective to adverbs”, which maps an adjective to its corresponding adverb in *-mente*. This test could not be used due to the variability of forms of adverbs in *-mente* in the time period of the Chronicles corpus. During this period, this suffix had several allomorphs (e.g. *-mente*, *-miente*, *-mentre*, *-mientre*), and it could be orthographically represented either attached to or detached from the previous noun, due to categorial variation (Company and Flores Dávila, 2014; Barrio de la Rosa, 2016).

In order to extend the test’s coverage of verbal morphology, we created the following new tests: “infinitive to participle”, e.g., *saber sabido* : *amar amado* ‘to know known : to love loved’. The set of plural nouns in MTS was not appropriate because of the vocabulary, consequently we adapted it to the nouns in the Chronicles corpus. Given the importance of inflectional morphology in Spanish, we also added tests for other types of inflection (gender and number in adjectives). The morphology tests were generated by using vocabulary based on the frequency counts from the Chronicles corpus.

As for semantic tests, we kept the set of semantic opposites formed by adjective antonyms and added one test set using antonyms based on the Chronicles corpus vocabulary (e.g., *cerca lejos* : *bien mal* ‘nearby far : well badly’).

The structure of our analogy test is summarized in Table 5.

5.3 Accuracy on Modern Data

Now we can test the accuracy of our embeddings on the contemporary SBW corpus, assuming that this is the easier corpus to model. The results are shown in Table 6. For this evaluation, a question is considered valid if all four words are attested in the corpus. The percentage of correctly answered questions is calculated based on valid questions only.

The table shows that most of the questions of the two data sets, the translations of the original English data and our own data set, occur in the SBW corpus (see # valid). Thus, the coverage of both data sets is very

Source	Category	Example	#Questions
MTS	Morphology nouns: kinship terms	padre madre : hijo hija	506
	Morphology verbs: third person singular	comer come : ir va	650
	Morphology verbs: infinitive to participle	saber sabido : tomar tomado	1190
	Morphology verbs: gerund to participle	sabiendo sabido : tomando tomado	1190
ours	Morphology adj.: singular to plural	negra negras : rica ricas	992
	Morphology adj.: singular to plural	negro negros : rico ricos	992
	Morphology adj.: masc to fem	negro negra : negros negras	992
	Morphology adj.: masc to fem	negros negras : ricos ricas	992
	Morphology nouns : singular to plural	casa casas: capilla capillas	1332
	Morphology/Semantics: antonyms	feliz infeliz : posible imposible	42
	Semantics: antonyms	cerca lejos : bien mal	342
Total			9220

Table 5: Structure of the analogy test following the model of Table 1 by Li et al. (2018). MTS denotes the analogy test from Mikolov et al. (2013a), translated into Spanish.

Analogy test	# total	# valid	# correct			% correct		
			SVD	SGNS	GloVe	SVD	SGNS	GloVe
MTS	14 764	14 672	4725	7 424	7 805	32.2	50.6	53.2
MTS+ours	9 220	8 146	2038	3 720	3 486	25.0	45.7	42.8

Table 6: Number of valid and correct questions of the two analogy test sets on the **SBW** corpus. Results are based on seed 1 for SGNS and GloVe; MTS: analogy test from Mikolov et al. (2013a) translated into Spanish; MTS+ours: combination of applicable questions combining MTS and our test.

high. If we compare the embeddings based on the three algorithms, we see that both data sets lead to the same trends: SVD has the lowest accuracy, SGNS the highest, and GloVe reaches a middle ground between the two. We note that the results for SVD are considerably lower than SGNS and GloVe. It may be possible to obtain better results for this algorithm, but this would require a very extensive search of hyper-parameters. Thus while SVD is commended for yielding stable results across different runs, the model is not entirely satisfactory for our task.

We can also look at the overall quality of the embeddings, even though this is not completely relevant in our situation since we are mostly interested in choosing between one of the models. The accuracies range between 14% and 41% for the MTS test and between 18% and 37% for our test. This means that the embeddings represent the analogies with moderate accuracy. It also means that our test is slightly more challenging than the MTS test.

5.4 Accuracy on Historical Data

We now focus on the historical data to see how the analogy tests work on the Chronicles corpus. The results are shown in Table 7. In contrast to the results on the contemporary corpus, we see that the original MTS test has very little coverage on the historical corpus: Only 331 questions are valid. For our test, more than half of the questions (4 950) are valid. Given the small number of valid questions in the MTS test, it is likely that the accuracies of the different methods are not very meaningful. It is also obvious that the accuracies overall are considerably lower, the highest one reaching 27.3%, for SGNS. This result is comparable to SBW, where SGNS also reaches the highest accuracy. However while GloVe fares better than SVD on the modern data, it reaches the lowest accuracy, 19.9%, on the historical data.

Analogy test	# total	# valid	# correct			% correct		
			SVD	SGNS	GloVe	SVD	SGNS	GloVe
MTS	14 764	331	47	47	31	14.2	14.2	9.4
MTS+ours	9 220	4 950	1 178	1 351	985	23.8	27.3	19.9

Table 7: Number of valid and correct questions of the two analogy test sets on the **Chronicles** corpus. Results are based on seed 1 for SGNS and GloVe; MTS: analogy test from Mikolov et al. (2013a) translated into Spanish; MTS+ours: combination of applicable MTS and our test.

Overall, we need to stress the importance of having a set of questions that reach a good coverage on all corpora. I.e., if we need to decide on one algorithm for embeddings, we need to take into account both the contemporary and the historical corpus. The results show that across both corpora, SGNS provides the embeddings that reach the highest accuracy.

5.5 Applicability of Embeddings

As the final question, we need to decide if the accuracy based on the analogy tests corresponds to the applicability of the word lists for determining semantic shifts, based on the semanticist’s expertise. If this is the case, we can use the analogy test in the future to determine the best type of embeddings, and the best hyper-parameter settings. If the results of the test do not agree with the semanticist’s expertise, we are back to evaluating the embeddings manually. Since we are trying to determine the validity of the evaluation method, we need to have a comparison with an existing study (in our case, Amaral (2016)) but once the methods have proven to be useful, our goal is to detect new semantic shifts and gain new insights that would otherwise not be available through an intuition-based analysis of the corpus.

In SBW the results of the different models are rather similar (see Table 4a). For this reason, we mainly focus on the historical Chronicles corpus, where the three algorithms provide very different word lists. This is reflected in the last column of Table 4b, which shows that there is only one word in common among the top-10 words in the word lists based on the three algorithms, *demás*. Most words in the last column of Table 4b rank very high for only one model, but low for the other two. This demonstrates that relying on just one algorithm may give an incomplete picture of the distribution of a word over time.

In the SGNS models, 6 out of 10 of the words ranked most similar to *algo* in Table 4b are nouns with meanings in the domain of property and value. These are consistent with previous work showing that *algo* was a noun meaning ‘possessions’ in the medieval period (Amaral, 2016), besides being an indefinite pronoun ‘something’. However, the results of the GloVe and SVD models are quite different. For example, in the SVD model only one word is a noun from the semantic domain of property and value (*rentas*). Hence, if we had exclusively used the SVD algorithm, we would have little evidence for the syntactic and semantic change involving the medieval noun.

The last column in Table 4b shows the importance of averaging across algorithms. Comparing the word list from SGNS with the word list in the last column of Table 4b, we see that in both lists, 6 out of the 10 words are nouns from the semantic domain of property and value (i.e., ranking across algorithms confirms the pattern from the algorithm that yields the best results). Both lists provide information attesting to the original value of *algo* as a noun meaning ‘possessions’. Also, the word list resulting from ranking across algorithms does not include the word *abenagit*, which is noise, as mentioned above. Third, the noun *sabor* ‘taste’ is in the output of the ranking across algorithm. According to Amaral (2016), co-occurrence with abstract nouns (like “taste”) in pseudo-partitive constructions plays a role in the change undergone by *algo*, so the presence of this word can provide insight into the types of context favoring semantic change. In addition, the last column provides aggregated information about the position of each word in the models, hence giving a more accurate picture of the weight of the neighbors across all three algorithms.

On the basis of the findings of [Amaral \(2016\)](#), the best model for the neighbors of *algo* in the Chronicles is SGNS because it best represents the range of meanings of the word found in the medieval data. This corresponds to the accuracy scores in the analogy tests. Thus, we can conclude that the test scores correlate with the reliability of the models for most similar words. From our findings reported in [Table 2](#), the choice of model seems particularly relevant when we have a small data set. However, the list of words most interesting to the semanticist is the result of averaging across all three algorithms.

6 Conclusion

The work presented here extends previous research on distributional approaches to semantic change by showing that word embeddings can detect semantic and syntactic change even when trained on a relatively small data set (7 million tokens in our case).

Our systematic comparison of three state-of-the-art models demonstrates that word embeddings are applicable for semantic change detection, but the choice of algorithm and hyper-parameters matters: If we had relied on SVD alone, based on prior work ([Hellrich, 2019](#)), we would not have identified the change of *algo* from a noun in the older data. As previously pointed out by [Hellrich \(2019\)](#), SGNS and GloVe lack replicability. Hence, it is important to report the design decisions because they affect the results. These findings are important for historical linguistic research in the digital humanities since available corpus data is often scarce, and it is not possible to find more data.

In addition, we contribute to the research on word embeddings and their applicability in tracing semantic change by showing a potential solution to addressing the lack of stability of embeddings: Using rankings on the most similar words produced by averaging over several runs of the algorithms ensures more stable and robust results regarding neighboring words. The ranking measure makes it possible to filter out possible noise generated by the algorithms.

With respect to the evaluation of the models, we have shown that we cannot use established analogy tests for smaller data sets like ours. Instead, we need to design new tests adapted to the data. This is especially important for data sets where the small size is generally accompanied by a wide range of spelling variations, exacerbating the problem. We have also shown that the accuracy results of the analogy test correspond to an expert’s judgment, thus allowing us to choose the best embeddings model automatically once we have established the test. It is important to find robust ways of evaluating word embeddings, not only from the perspective of NLP tests, but because the results of embeddings models can provide insight on semantic shifts and hence shed light on theories of language change.

Our main focus in the current work was on establishing a methodology and best practices for using word embeddings for modeling semantic change. While our results in terms of investigating semantic change are limited to a case study using a small set of data, they nonetheless highlight the potential of word embeddings on a wider empirical ground, hence expanding their contribution to the field of diachronic semantics.

Acknowledgment

We thank our two anonymous reviewers for their helpful comments and suggestions.

Funding

This project was partially supported by the Indiana University Institute for Digital Arts and Humanities (IDAH) and the New Frontiers in the Arts and Humanities Program through two fellowships awarded to Patrícia Amaral.

References

- Amaral, P. (2016). When *something* becomes a *bit*. *Diachronica*, 33(2):151–186.
- Antoniak, M. and Mimno, D. (2018). Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119.
- Bamler, R. and Mandt, S. (2017). Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning*, pages 380–389, Sydney, Australia.
- Barrio de la Rosa, F. d. (2016). La distribución de las variantes *-mente*, *-miente* y *-mientras* en el CODEA (1221-1420): Espacio variacional y cambio lingüístico. *Scriptum Digital*, 5:85–102.
- Benito Moreno, C. (2019). Los corpus del español desde la perspectiva del usuario lingüista. *Scriptum Digital*, 8:1–21.
- Boleda, G. (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6:13.1–13.22.
- Boukhaled, M. A., Fagard, B., and Poibeau, T. (2019). Modelling the semantic change dynamics using diachronic word embedding. In *11th International Conference on Agents and Artificial Intelligence (NLPinAI Special Session)*, pages 944–951, Prague, Czech Republic.
- Bréal, M. (1897). *Essai de Semantique*. Hachette, Paris.
- Bybee, J., Perkins, R., and Pagliuca, W. (1994). *The Evolution of Grammar*. Chicago University Press.
- Cardellino, C. (2016). Spanish Billion Words Corpus and embeddings. Online at <https://crscardellino.github.io/SBWCE/>; retrieved August 2019.
- Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.
- Company, C. and Flores Dávila, R. (2014). Adverbios en mente. In Company, C., editor, *Sintaxis histórica de la lengua española. Tercera parte: Adverbios, preposiciones y conjunciones. Relaciones interoracionales*, pages 1195–1340. Fondo de Cultura Económica y Universidad Nacional Autónoma de México.
- Davies, M. (2001). Corpus del Español. Online at <http://www.corpusdelespanol.org>; retrieved August 2019.
- Del Tredici, M., Fernández, R., and Boleda, G. (2019). Short-term meaning shift: A distributional exploration. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 2069–2075, Minneapolis, MN.
- Deo, A. (2015). Diachronic semantics. *Annual Review of Linguistics*, 1:179–197.
- Dubossarsky, H., Hengchen, S., Tahmasebi, N., and Schlechtweg, D. (2019). Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470, Florence, Italy. Association for Computational Linguistics.
- Dubossarsky, H., Weinshall, D., and Grossman, E. (2017). Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, Copenhagen, Denmark.

- Eckardt, R. (2006). *Meaning Change in Grammaticalization. An Enquiry into Semantic Reanalysis*. Oxford University Press.
- Firth, J. R. (1957). *Papers in Linguistics*. Oxford University Press, London.
- Gago Jover, F. (2011). *Spanish Chronicle Texts. Digital Library of Old Spanish Texts*. Hispanic Seminary of Medieval Studies. Online at <http://www.hispanicseminary.org/t&c/ac/index.htm>; retrieved August 2019.
- GITHE (2015). Corpus de Documentos Españoles Anteriores a 1800. Online at <http://corpuscodea.es/>; retrieved August 2019. Universidad de Alcalá.
- Golub, G. H. and Reinsch, C. (1971). Singular value decomposition and least squares solutions. In Bauer, F., Householder, A., Olver, F., Rutishauser, H., Samelson, K., and Stiefel, E., editors, *Handbook for Automatic Computation*, pages 134–151. Springer. Volume II: Linear Algebra.
- Golordava, K. and Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *GEMS Workshop on Geometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, UK.
- Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016a). Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2116–2121, Austin, TX.
- Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016b). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1489–1501, Berlin, Germany.
- Harris, Z. (1954). Distributional structure. *Word*, 10(2-3):146–162.
- Hellrich, J. (2019). *Word Embeddings: Reliability and Semantic Change*. PhD thesis, Jena University Language and Information Engineering Lab.
- Hellrich, J. and Hahn, U. (2017). Don't get fooled by word embeddings – better watch their neighborhood. In *Proceedings of Digital Humanities*, pages 250–252, Montreal, Canada.
- Hengchen, S. (2017). *When Does it Mean? Detecting Semantic Change in Historical Texts*. PhD thesis, Université Libre de Bruxelles.
- Hock, H. H. and Joseph, B. D. (2009). *Language History, Language Change, and Language Relationship: An Introduction to Historical and Comparative Linguistics*. Mouton de Gruyter, Berlin, 2nd edition.
- Hopper, P. J. and Traugott, E. C. (2003). *Grammaticalization*. Cambridge University Press, Cambridge.
- Hutter, F., Hoos, H., and Leyton-Brown, K. (2014). An efficient approach for assessing hyperparameter importance. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 754–762, Beijing, China.
- Jurafsky, D. and Martin, J. H. (2019). *Speech and Language Processing*. 3rd edition. Online at <https://web.stanford.edu/~jurafsky/slp3/>; retrieved April 2020.
- Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., and Petrov, S. (2014). Temporal analysis of language through neural language models. In *Proceedings of the ACL Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD.

- Kutuzov, A., Øvrelid, L., Szymanski, T., and Velldal, E. (2018). Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, NM.
- Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284.
- Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, 4:151–171.
- Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.
- Li, S., Zhao, Z., Hu, R., Li, W., Liu, T., and Du, X. (2018). Analogical reasoning on Chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 138–143, Melbourne, Australia.
- Luo, Y., Jurafsky, D., and Levin, B. (2019). From insanely jealous to insanely delicious: Computational models for the semantic bleaching of English intensifiers. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 1–13, Florence, Italy.
- Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605.
- McGillivray, B., Hengchen, S., Lähteenoja, V., Palma, M., and Vatri, A. (2019). A computational approach to lexical polysemy in Ancient Greek. *Digital Scholarship in the Humanities*, 34(4):893–907.
- Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations (ICLR)*, Scottsdale, AZ.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3111–3119, Lake Tahoe, NV.
- Mundra, R. and Socher, R. (2015). Lecture notes for CS 224d: Deep learning for NLP. Online at: https://cs224d.stanford.edu/lecture_notes/LectureNotes2.pdf; retrieved August 2019.
- Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar.
- Perrone, V., Palma, M., Hengchen, S., Vatri, A., Smith, J. Q., and McGillivray, B. (2019). GASC: Genre-aware semantic change for Ancient Greek. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 56–66.
- Real Academia Española (n.d.). Corpus Diacrónico del Español. Online at <http://corpus.rae.es/cordenet.html>; retrieved August 2019.
- Rodda, M., Senaldi, M., and Lenci, A. (2017). Panta Rei: Tracking semantic change with distributional semantics in Ancient Greek. *Italian Journal of Computational Linguistics*, 3(1):11–24.

- Sagi, E., Kaufmann, S., and Clark, B. (2012). Tracing semantic change with latent semantic analysis. In Allan, K. and Robinson, J., editors, *Current Methods in Historical Semantics*, pages 161–183. Mouton de Gruyter, Berlin.
- Sánchez-Martínez, F., Martínez-Sempere, I., Ivars-Ribes, X., and Carrasco, R. C. (2013). An open diachronic corpus of historical Spanish. *Language Resources and Evaluation*, 47(4):1327–1342.
- Schlechtweg, D., HäTTY, A., Del Tredici, M., and Schulte im Walde, S. (2019). A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy.
- Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., and Tahmasebi, N. (2020). SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online).
- Schlechtweg, D., Schulte im Walde, S., and Eckmann, S. (2018). Diachronic usage relatedness (DUREl): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 169–174, New Orleans, LA.
- Sweetser, E. (1990). *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge University Press.
- Tahmasebi, N., Borin, L., and Jatowt, A. (2018). Survey of computational approaches to lexical semantic change. arXiv:1811.06278.
- Tang, X. (2018). A state-of-the-art of semantic change computation. *Natural Language Engineering*, 24(5):649–676.
- Traugott, E. C. (1982). From propositional to textual and expressive meanings: Some semantic–pragmatic aspects of grammaticalization. In Lehman, W. P. and Malkiel, Y., editors, *Perspectives on Historical Linguistics*, pages 245–271. John Benjamins, Amsterdam.
- Traugott, E. C. and Dasher, R. (2002). *Regularity in Semantic Change*. Cambridge University Press, Cambridge.
- Ullmann, S. (1959). *The Principles of Semantics*. Blackwell.
- Yao, Z., Sun, Y., Ding, W., Rao, N., and Xiong, H. (2018). Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 673–681, Marina Del Rey, CA.
- Zhang, Y., Jatowt, A., Bhowmick, S., and Tanaka, K. (2015). Omnia mutantur, nihil interit: Connecting past with present by finding corresponding terms across time. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 645–655, Beijing, China.
- Zhang, Y., Jatowt, A., Bhowmick, S. S., and Tanaka, K. (2016). The past is not a foreign country: Detecting semantically similar terms across time. *IEEE Transactions on Knowledge and Data Engineering*, 28(10):2793–2807.