# Automatic Detection of Semantic Shift in Spanish with Context Optimization

Eve Fleisig
Adviser: Christiane Fellbaum

## Abstract

*Many natural language processing tasks, including question answering, topic detection, and historical sentiment analysis, require systems to model the meanings of words even as these meanings continually change. Automatic detection of semantic shift–changes in words' meanings–helps to build models of language that can account for changes in the meanings of words over time. These models are particularly useful for tasks sensitive to changes in words' senses that draw on word sense discovery or disambiguation. This project models 400 years of semantic shift in Spanish words and trains a classifier to detect whether words underwent change between two time periods. In addition, it examines whether the amount of context used to predict a word's embedding can be optimized to improve classification results and analyzes whether the optimal context window size varies for words in different lexical categories. The models created for this research draw on recently developed methods of analyzing semantic shift by creating vector embeddings for words trained via neural networks, after which a support vector classifier is trained to detect semantic shift using a new dataset of words annotated with their historical change. This project finds that these methods of modeling and classification are indeed effective at detecting semantic shift. Further evaluation suggests that trigram models are optimal for semantic shift detection, although words in different lexical categories were best modeled with context windows of different sizes. Closer examination of specific words suggests that the models capture different types of semantic shift, long- and short-term change, and changes in usage outside those captured by traditional lexicography.*

## 1. Introduction

As the meaning of words constantly changes, computational models of language must continue to accurately represent their evolving senses. In particular, natural language processing tasks that

are sensitive to changes in the meaning of words, such as question answering, topic detection, and diachronic[1] sentiment analysis, as well as wider fields such as word sense discovery and disambiguation, are vulnerable to inaccuracies when words undergo semantic shift.[2] Language models, such as vector embeddings of words' meanings, that are trained on data from one time period may not work for another. For example, a question-answering system will have difficulty answering a user's question about "uploading to the cloud" if the system only understands clouds as meteorological phenomena. Thus, the performance of models trained on language from a given time period can significantly decrease when applying the model to language from future years–relevant to maintaining systems in the long term–or past years–relevant to work on historical text. One solution to these problems is to automatically detect the semantic shift of words over time. Recent research has found promising results for the use of historical word embeddings trained via neural networks for detecting semantic shift. However, these methods still lack substantial research regarding languages besides English or refinements for performance optimization.

This paper presents findings in two overlooked areas of semantic shift detection. First, it studies semantic shift in diachronic Spanish documents, which has not previously been addressed. Second, it investigates context optimization for this task: determining the optimal number of context words used to model a given word when determining whether it underwent semantic shift. Using new methods of analyzing semantic shift by modeling words based on vector embeddings trained via neural networks, this project models 400 years of semantic shift in Spanish words and trains a support vector classifier to detect whether words underwent change between two time periods.

Section 2 gives an overview of the context of this problem, while Section 3 describes the motivation for the approach in this work. Section 4 details the process of creating the historical word embeddings and training the classifier. Section 5 concludes by analyzing the overall performance of the models, along with the optimal context size and results of interest on specific lexical categories of words.

[1] Across time.

[2] Changes in words' meanings, including acquiring, losing, or altering a sense.

## 2. Background and Related Work

### 2.1. Background

The semantics of language undergoes continuous change as speakers apply words to new situations. These changes, such as broadening (e.g., *thing* shifting from "public assembly" to "entity"), narrowing (e.g., *meat* shifting from "food" to "animal flesh"), or metaphor (e.g., *root* shifting from a plant structure to "source," such as the root of a word), result in words acquiring, losing, or altering senses over time. These semantic changes may be gradual, such as the narrowing of *wife* from "woman" in Old English to "(female) spouse" in Modern English, or sudden, such as the new sense of *tweet* resulting from the development of Twitter.

Modeling this change is relevant to computational research as well as historical linguistics. Many natural language processing tasks rely on language models to determine the meaning of an input, particularly tasks involving challenges under the broad umbrella of word sense discovery (learning the senses of a word in different contexts) or disambiguation (determining which sense is being used in a given context) [10]. However, these models are sensitive to changes in the meanings of words–both when applying models of modern data to text in future years as words evolve in the present day, and when applying models of modern language to historical data. One remedy for this issue is to automatically detect changes in the meanings of words.

### 2.2. Related Work

Methods for automatically identifying semantic shift have rapidly evolved in recent years. Earlier approaches modeled words in several time periods as vectors based on how frequently they cooccurred and used cosine similarity or related measures of the distance between words' vectors in different time periods to determine the degree of change [3, 7].

Later efforts shifted to graph- and clustering-based models in order to represent the contexts in which words appeared. Wijaya and Yeniterzi (2011) modeled documents as a combination of topics using Latent Dirichlet Allocation, then used k-means clustering to identify English words whose

senses changed clusters over time [16]. Mitra et al. (2014, 2015) expanded on this work, examining English bigrams,[3] by creating a graph of the corpus vocabulary with edge weights corresponding to the frequency of the bigrams and clustering the graph using the Chinese Whispers method [9, 10]. They then identified words that occurred in different clusters over time as having different senses and classified these changes as different types of semantic shift (the birth or death of a word sense, the joining of word senses, and the splitting of word senses). Dubossarsky et al. (2017) used similar approaches to test traditional laws of semantic shift by automatically searching for evidence of semantic shift [4].

New approaches have highlighted the value of using neural networks to improve word embeddings and tested refined metrics for measuring embeddings' similarity. Rosenfeld and Erk (2018) modeled words continuously over time by training a neural network to create vector embeddings given a word and year as input [13]. Using Word2vec's neural word embeddings, Fomin et al. (2019) tested different methods of scoring the change in a word's embedding over time, including Procrustes alignment, global anchors alignment, and Kendall and Jaccard distance scoring. These scores were then used to train a logistic regression classifier that classified words as having shifted or not shifted between time periods on two datasets, one dividing long-term data into three time periods (pre-1918, 1918-1991, and post-1991) and one dividing a 15-year time period into individual years. They found that Procrustes analysis (see Section 4.5) outperformed the other metrics as an input feature to the classifier. The neural approach significantly outperformed earlier efforts at identifying whether semantic shift occurred; however, Fomin et al.'s research on long-term shift only modeled a given word's change between three time periods, raising the question of whether their approach can be refined to examine more granular long-term change by dividing words into shorter time periods.

## 3. Approach

Research on semantic shift has overwhelmingly focused on English: except Fomin et al. (2019)'s work in Russian and Tang et al. (2016)'s work in Chinese, all major work in the past five years has

[3] Sets of two words.

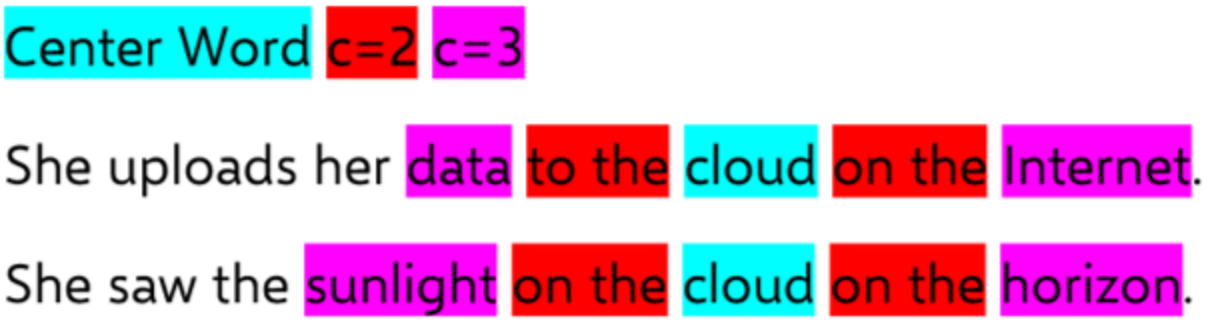



**Figure 1: Example of the effect of context sizes on the embedding of the word *cloud*. A context size of 2 (red) will result in near-identical embeddings for the different senses of the word, whereas a context of 3 (pink) will result in different embeddings.**

used English corpora [15]. The closest research in Spanish includes determining changes in themes across a diachronic corpus of Spanish poetry [11] and examining changes in Spanish participial construction on a small set of verbs [14], neither of which involve examining individual words' semantic change. Thus, there is a gap in the literature regarding research on semantic shift in Spanish.

Furthermore, given the novelty of using neural networks for semantic shift detection, little work has been done on optimizing these approaches. In particular, optimal context window sizes–given a word $w$, the number of words before and after $w$ to consider when building $w$'s embedding–for semantic shift analysis have not yet been studied. Figure 1 gives an example of the importance of context size for the word *cloud*, whose original meteorological meaning differs from the second modern sense of a shared computing resource. An embedding's ability to capture this change in the example sentences depends on the context window size: whereas a context size of 2 will result in near-identical embeddings for the different senses of *cloud*, a context size of 3 allows models to capture the different contextual information provided by "data" and "Internet" compared to "sunlight" and "horizon," resulting in different embeddings that provide more relevant information for semantic shift analysis.

Given the need for research on semantic shift in Spanish, as well as for word context optimization, my approach was to build models to automatically detect semantic shift in Spanish documents, train a classifier to detect semantic shift using these models, and test whether different context

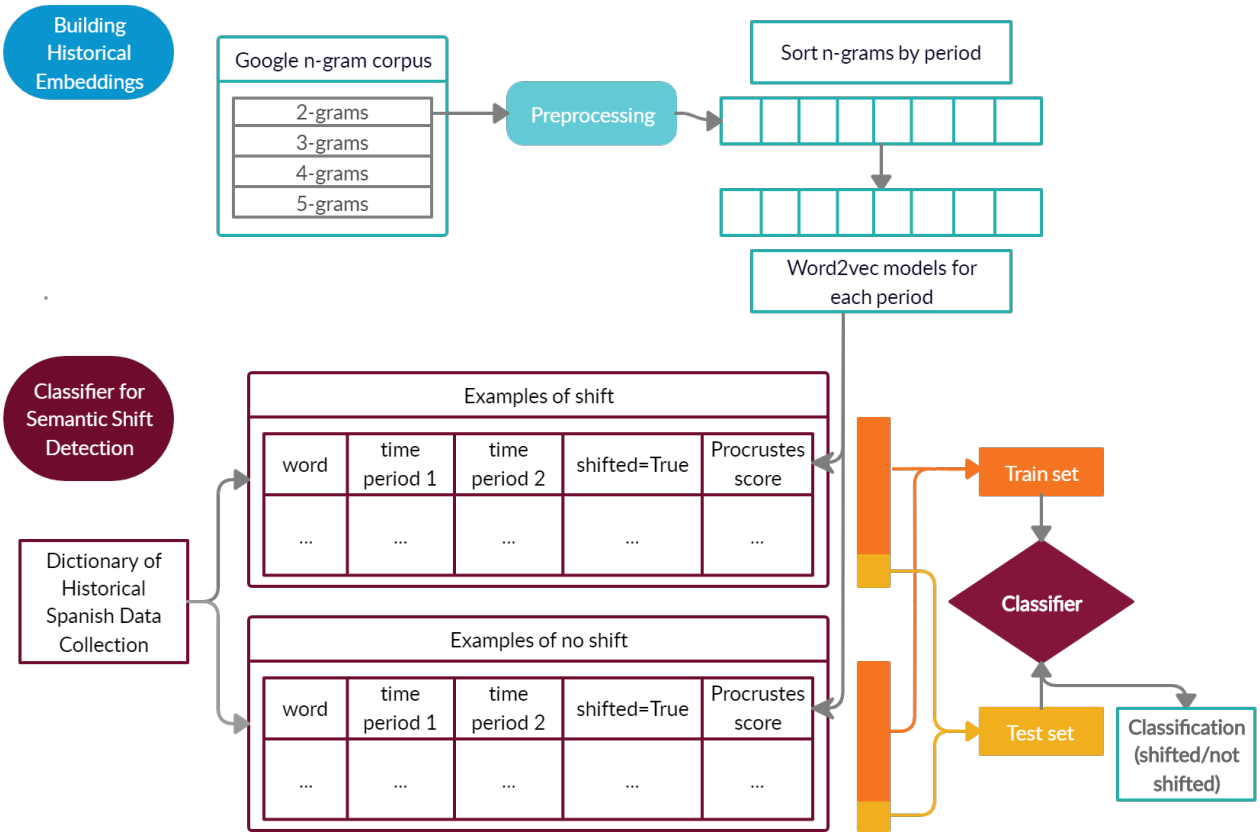



**Figure 2: Workflow for creating (a) historical word embeddings and (b) a classifier for semantic shift detection.**

window sizes improve classifier performance. These models are most similar to that used by Fomin et al. (2019), as the least research has been done on neural word embeddings, and they provided the most promising results. For each context window size studied, I created neural network-based word embeddings for several historical time periods and trained a support vector classifier to detect whether words had shifted between time periods. I hypothesized that larger context windows would increase the accuracy of semantic shift detection, though there would likely be a point after which increasing the context size would no longer improve the embedding.

## 4. Implementation

### 4.1. Overview

The implementation proceeded by first building word embedding models for different historical time periods, then building a classifier to determine whether a word had shifted meanings between two time periods, as shown in Figure 2. To build the word embedding models, separate datasets of bigrams, trigrams, quadgrams, and quintgrams were extracted from the Google n-gram corpus of Spanish. For each n-gram length, the n-grams were preprocessed and sorted into eight time periods. Then, custom word embedding models were created from the n-grams in each time period.

To build the classifier, a dataset was constructed and annotated from the Royal Spanish Academy's Dictionary of Historical Spanish, marking the years in which words were recorded as gaining or losing a sense. Then, for each word $w$ and for each pair of time periods $p_1, p_2$, the entry $(w, p_1, p_2)$ was added to a list of examples of shift if $w$ gained or lost a meaning between those two time periods, and the entry was added to a list of examples of no shift occurring otherwise. Both lists were sampled evenly to create training and test sets with balanced classes. A support vector classifier was then trained to output whether or not a word in a test set entry $(w, p_1, p_2)$ had shifted meaning between periods $p_1$ and $p_2$.

The implementation was first performed on a smaller subset of the corpus as a test case before implementation on the full dataset. The code implementing the models and classifiers was run on the Princeton Adroit computing cluster; all code and models can be found on GitHub at github.com/efleisig/semantic_shift_IW.

### 4.2. Corpus

To create the word embedding models, I selected the Google n-gram corpus of Spanish, which provides the frequency of different n-grams in Spanish from 1522 to 2009. Compared to other Spanish corpora, the Google n-grams corpus provides the most data (45 billion tokens). Furthermore, its division of data into n-grams of length 2 to 5 naturally limits words' context, making it possible to test the optimal amount of context for semantic shift detection.

Because more data is available for more recent years, I followed Mitra et al. (2015)'s approach to account for corpus imbalance. I divided the data into eight time periods containing comparable

amounts of data, with a mean of 361 million n-grams each: 1522-1899, 1900-1949, 1950-1969, 1970-1984, 1985-1994, 1995-1999, 2000-2004, and 2005-2009.

I then built a test case from a subset of the data to examine semantic shift before implementing the model on the larger corpora. The following steps were then implemented on each larger n-gram dataset (bigrams, trigrams, quadgrams, and quintgrams) and performance was evaluated on the classifier for each n-gram size.

### 4.3. Preprocessing and Word Embeddings

For each n-gram size, the corresponding corpus from the Google n-grams dataset, consisting of n-grams in the format [<$w_1, w_2 ... w_n$> <year> <frequency>], first underwent preprocessing. Stopwords and proper nouns were removed from the n-grams, and words were lemmatized using Python's Natural Language Toolkit. The n-grams were then sorted into each of the eight time periods.

For each time period, a custom Word2vec model was trained from the n-grams belonging to that time period. Word2vec models words with vector embeddings using a continuous bag-of-words representation of the n-grams, wherein context words are used to predict a target word, and a neural network adjusts words' feature vectors such that target words can be predicted more accurately [8]. In addition, after aligning two Word2vec models (Section 4.4), the similarity between the vector embeddings of a word under models from two different time periods can be used to quantify the word's semantic change between those time periods.

### 4.4. Data Collection for Classifier Training and Evaluation

In order to train a classifier to label words as shifted or not shifted between two time periods, I collected and annotated a dataset of Spanish words with their historical change. The dictionary of the Royal Spanish Academy is the official resource in hispanophone countries for whether a word has acquired or lost a sense [1]. The Academy's Historical Dictionary of the Spanish Language (Nuevo Diccionario Histórico, or NDH) provides the first and last recorded usage of a particular sense of a word since the 1100s [2].

A notable limitation of the Royal Spanish Academy's work is its history of occasionally pre-

scribing definitions for words that may not reflect their actual usage and its relative conservativism on accepting new usages of words [17]. However, the NDH undergoes less scrutiny, since it does not dictate current usage, and it is by far the most complete source on historical usage of Spanish. Therefore, this work uses it as a source for ground-truth data that is flawed but nonetheless more informative than any alternative.

```
acordeón 1836 1880 1909 1944 1958 1976 1998 2009
agote 1524 1607 1852 2001
ametralladora 1868 2010
```

**Figure 3: Sample entries from the dataset created from the Dictionary of Historical Spanish.**

I randomly selected 400 open-class[4] words from the NDH and annotated them with the dates that, according to the NDH, they acquired or lost a sense (Figure 3). For example, the word *acordeón* ("accordion") was first defined as a musical instrument in 1836, then acquired a variety of colloquial meanings, such as "articulated bus" in 1944 and "cheat sheet" in 1958.

Then, a list $L_{shifted}$ of ($w$, $p_1$, $p_2$) pairs where word $w$ shifted meanings between time periods $p_1$ and $p_2$ and a list $L_{unshifted}$ of ($w$, $p_1$, $p_2$) pairs where $w$ did not shift meanings between $p_1$ and $p_2$ was created as follows:

```
for each word:
    for each pair of time periods p1: (p1_start, p1_end),
    p2: (p2_start, p2_end), where p1 < p2:
        if word lost or gained a sense at date d >=p1_start, d <= p2_end:
            add [word, p1, p2, shifted=1] to L_shifted
        else:
            add [word, p1, p2, shifted=0] to L_unshifted
```

This resulted in a list $L_{shifted}$ with 1284 entries and a list $L_{unshifted}$ with 3149 entries.[5]

[4] Open-class words, such as nouns, verbs, and adjectives, more commonly accept the addition of new words and new word senses.

[5] Later on, slight variation from these numbers for some n-gram sizes is due to the fact that a small number of words occurred too infrequently in some models to be scored for semantic shift.

### 4.5. Semantic Shift Scoring and Classification

For each entry ($w$, $p_1$, $p_2$) in $L_{shifted}$ and $L_{unshifted}$, I scored the semantic shift for word $w$ between time periods $p_1$ and $p_2$ according to the historical word embedding models. Procrustes analysis was used for scoring, as Fomin et al. (2019) found that on long-term data, Procrustes analysis outperformed other scoring metrics used as feature input to a semantic shift classifier, including using multiple metrics as features [5].

Procrustes analysis computes the difference between the vector representations of a word under two models by aligning the models' embedding matrices, then measuring the dissimilarity between a word's vectors under the aligned models. Given the embedding matrices $m_1$ and $m_2$ for the models corresponding to time periods $p_1$ and $p_2$, $m_2$ is aligned with $m_1$:

$$U, S, V^H = SVD(m_2^T \cdot m_1) \text{ where } USV^H = SVD(m) \text{ is the singular value decomposition of } m \quad (1)$$

$$m_2' = U \cdot V^H \quad (2)$$

The final score is the dot product of the vector embeddings of $w$ under $m_1$ and $m_2'$; lower scores indicate more change [6].

After scoring the entries in both lists, I sampled 1284 entries from $L_{unshifted}$ in order to train and test the model on balanced classes of shifted and unshifted data. I then used an 80/20 split to randomly partition the data (1284 shifted entries and 1284 unshifted entries, 2568 in total) into a training set, consisting of 2054 entries, and a test set, consisting of 514 entries.

For each n-gram size, a support vector classifier was trained on the training set, then used to predict whether entries in the test set had undergone semantic shift or not. A support vector classifier was used because these classifiers are often better able to represent complex, nonlinear decision boundaries and avoid overfitting than the logistic regression classifiers used in studies such as Fomin et al. (2019) [5, 12].

## 5. Results and Evaluation

Section 5.1 presents the overall results on the dataset, while Section 5.2 analyzes the results on different lexical categories of words. Section 5.3 presents examples of the models' ability to capture different types of changes in usage.

### 5.1. Performance

After training classifiers on the bigram, trigram, quadgram, and quintgram datasets, the precision, recall, and F1 scores[6] for each classifier were calculated (Figure 4 and Figure 5). For all three metrics, trigrams outperformed the other n-gram sizes (Table 1), with a precision of 0.730, recall of 0.729, and F1 score of 0.729.[7] Although the trigram model outperformed the bigram model, as expected, it is notable that this trend did not continue for quadgrams and quintgrams. Instead, this finding suggests that trigrams provide an optimal context size for semantic shift analysis.

| N-gram length | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| Precision | 0.7110 | **0.7304** | 0.7063 | 0.7114 |
| Recall | 0.7104 | **0.7296** | 0.7062 | 0.7073 |
| F1 | 0.7109 | **0.7293** | 0.7062 | 0.7058 |

**Table 1: Precision, recall, and F1 scores for semantic shift classification.**

[6]Precision:

$$\frac{\text{[words marked shifted that truly shifted]}}{\text{[entries marked shifted that truly shifted]+[entries marked shifted that did not truly shift]}}$$

Recall:

$$\frac{\text{[entries marked shifted that truly shifted]}}{\text{[entries marked shifted that truly shifted]+[entries marked unshifted that truly shifted]}}$$

F1 Score:

$$\frac{\text{2*Precision*Recall}}{\text{Precision + Recall}}$$

[7]Although results from other languages may not be comparable due to linguistic differences and data collection methods, it may be helpful for reference to note that the state-of-the-art F1 score for semantic shift detection, from Fomin et al. (2019)'s work in Russian, is 0.767. Their long-term work only attempted to find semantic shift between three time periods, rather than the eight time periods analyzed in this paper.

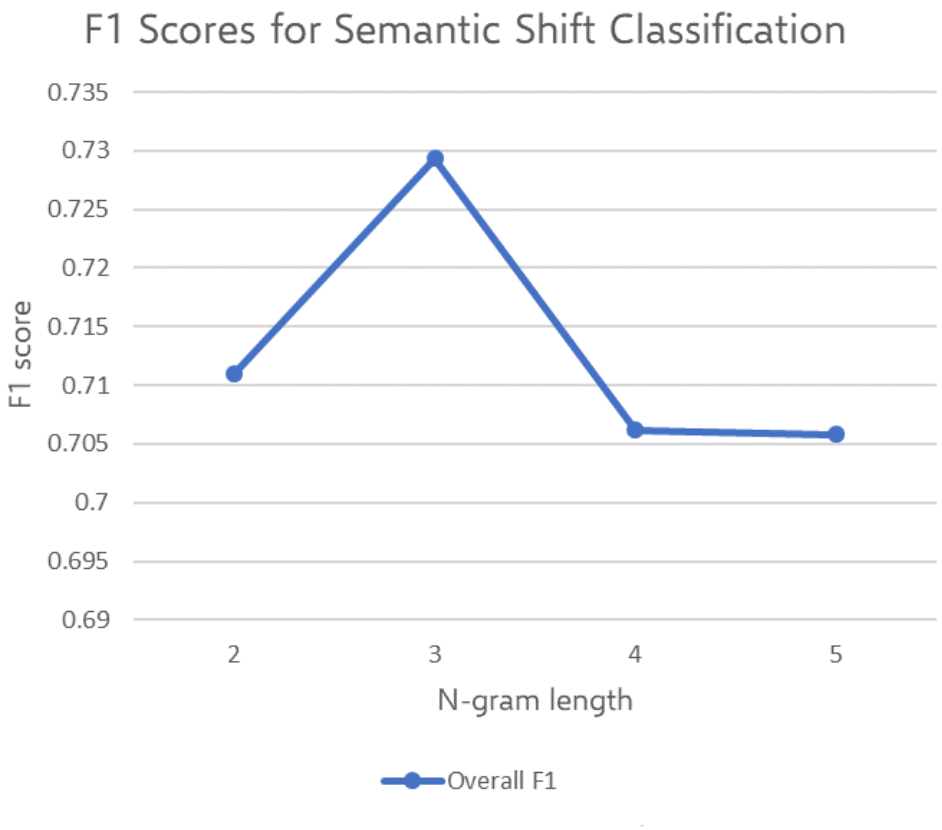



**Figure 4: F1 scores for semantic shift classification.**

### 5.2. Analysis by Lexical Category

One direction in which semantic shift detection models can be improved is by analyzing the classifiers' performance on different types of words. After analyzing the models' overall performance, I investigated how the models with different context sizes performed on different lexical categories. The test set was divided into nouns, adjectives, and verbs,[8] and the performance of the classifiers for each n-gram length was evaluated on these subsets of the test set by calculating their precision, recall, and F1 scores (Tables 2, 3, and 4, with maxima in bold).

The models' performance indeed varied by lexical category. The results on verbs universally surpassed those on nouns and adjectives–by all three metrics and for all context window sizes. In addition, the resulting optimal context size varied by lexical category. Whereas nouns were most

[8]Words in other lexical categories occurred too infrequently in the training and test sets to be analyzed.

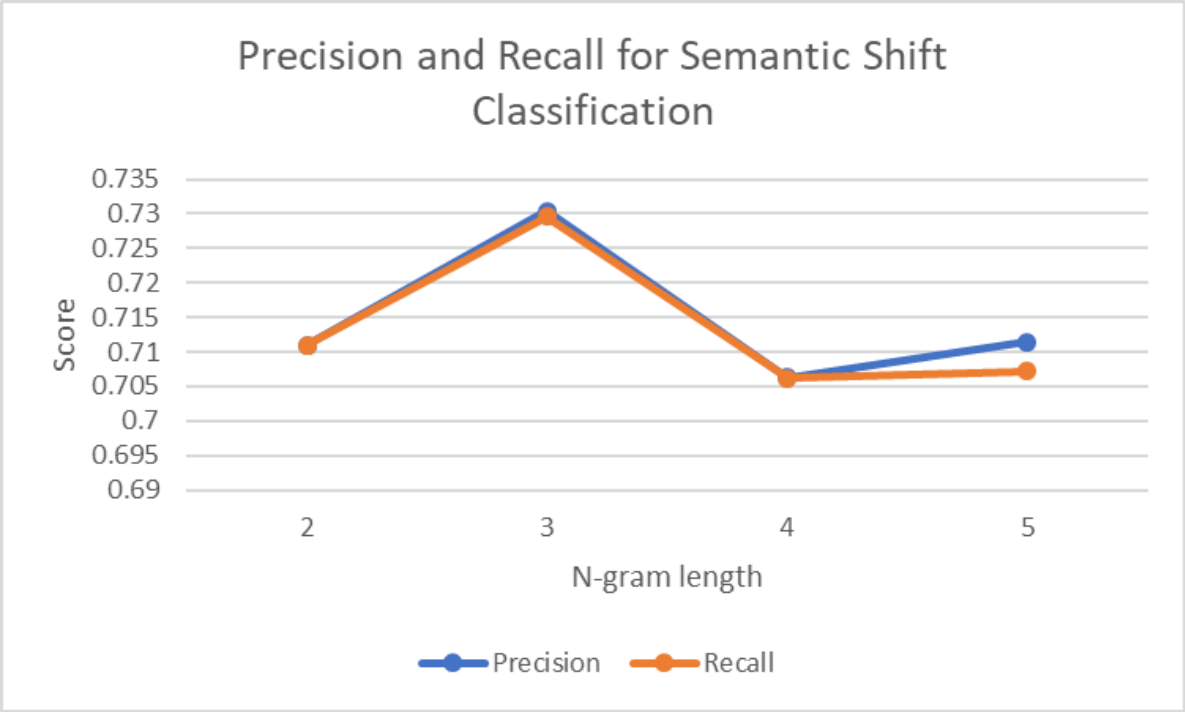



**Figure 5: Precision and recall scores for semantic shift classification.**

**Table 2: Precision scores by lexical category.**

| N-gram length | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| Nouns | 0.6998 | **0.7350** | 0.6984 | 0.7122 |
| Verbs | 0.7590 | 0.8681 | **1** | **1** |
| Adjectives | 0.7589 | 0.6490 | **0.7917** | 0.5399 |
| Overall | 0.7110 | **0.7304** | 0.7063 | 0.7114 |

**Table 3: Recall scores by lexical category.**

| N-gram length | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| Nouns | 0.6998 | **0.7332** | 0.6984 | 0.7044 |
| Verbs | 0.7692 | 0.7692 | **1** | **1** |
| Adjectives | 0.7468 | 0.6078 | **0.750** | 0.5217 |
| Overall | 0.7109 | **0.7296** | 0.7062 | 0.7073 |

**Table 4: F1 scores by lexical category.**

| N-gram length | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| Nouns | 0.6998 | **0.7330** | 0.6984 | 0.7034 |
| Verbs | 0.7588 | 0.7776 | **1** | **1** |
| Adjectives | 0.7486 | 0.6204 | **0.7624** | 0.530 |
| Overall | 0.7109 | **0.7293** | 0.7062 | 0.7058 |

effectively modeled with n-grams of length 3, adjectives were best modeled with those of length 4, and verbs were best modeled with those of length 4 or 5 (Figures 6, 7, and 8).

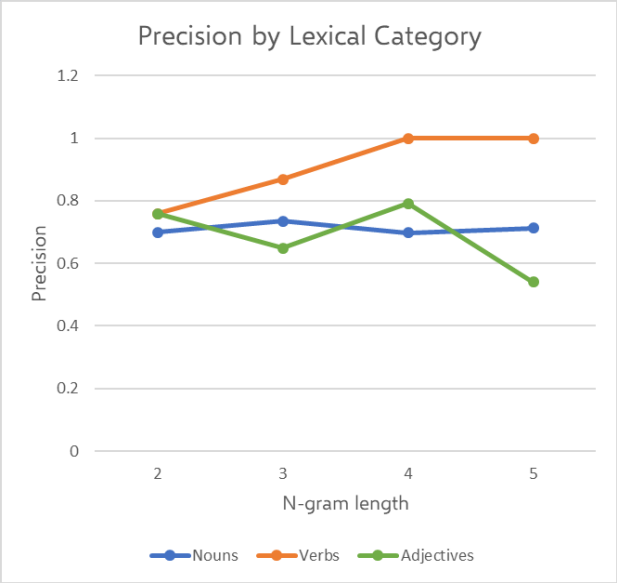



**Figure 6: Precision scores by lexical category for different n-gram lengths.**

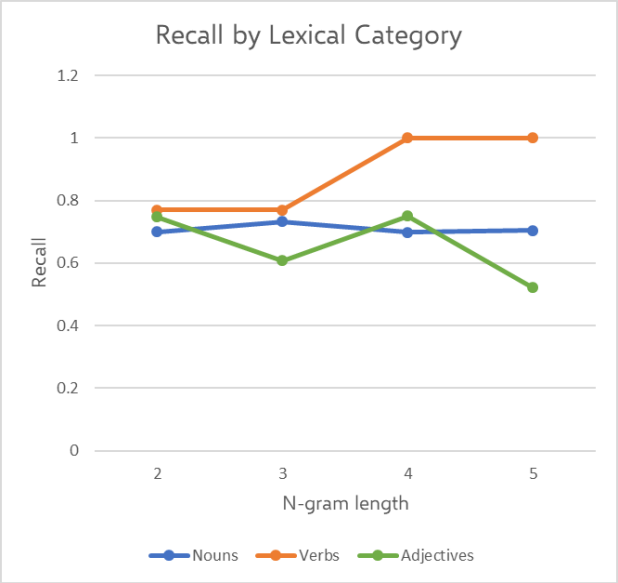



**Figure 7: Recall scores by lexical category for different n-gram lengths.**

The dataset as a whole was most effectively modeled with n-grams of length 3, which is likely due to the fact that nouns dominate the dataset (Table 5 [9]). However, the variation in optimal context size for different lexical categories suggests that for the purposes of semantic shift detection, it may be most effective to model words in different lexical categories using context windows of different sizes.

**Table 5: Train and test set breakdown by lexical category.**

| | Train Set | Test Set |
|---|---|---|
| Nouns | 86.8% | 86.5% |
| Verbs | 2.4% | 2.5% |
| Adjectives | 10.8% | 10.9% |

[9] As in the overall dataset, slight variation from these numbers for some n-gram sizes is due to the fact that a small number of words occurred too infrequently in some models to be scored for semantic shift.

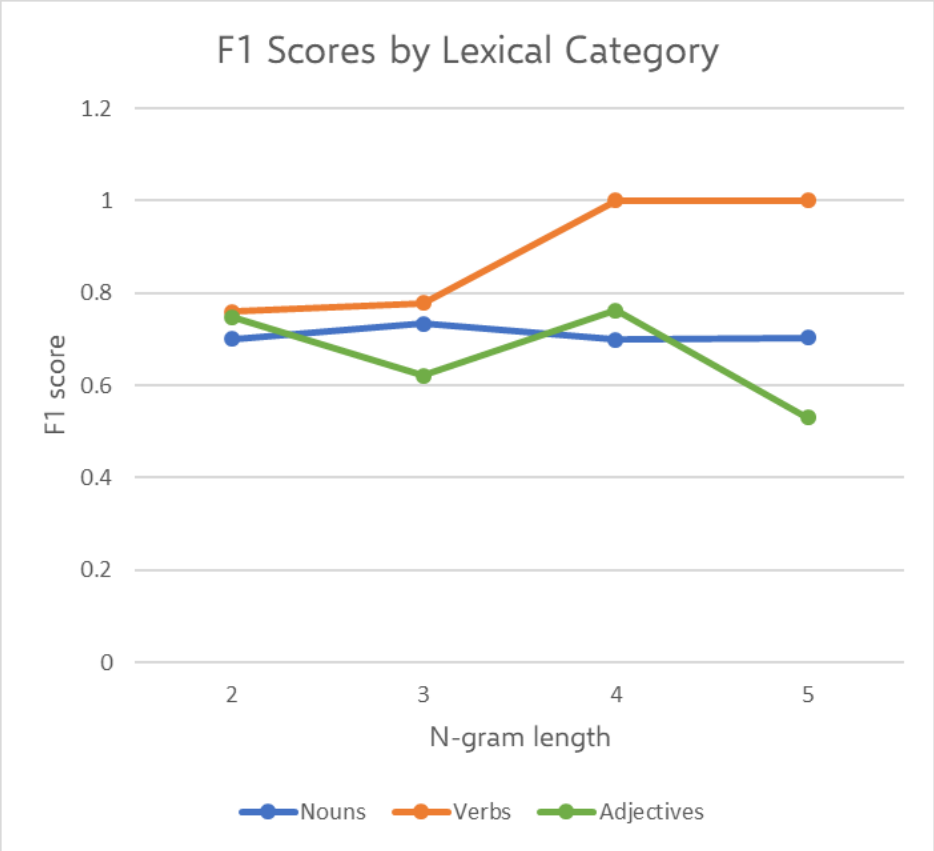



**Figure 8: F1 scores by lexical category for different n-gram lengths.**

### 5.3. Examples and Visualizations

To better understand the changes captured by the models, this section examines some example words and the changes in their embeddings over time. For each word examined, I found the words with the most similar embeddings (by cosine similarity of the embedding vectors) under the trigram models and plotted them using t-distributed Stochastic Neighbor Embedding (t-SNE), which models the similarity between the vectors as the distance between points in two-dimensional space.

These results give some intuition for the types of change captured by the models. For example, the word *guerra* "war," which originally referred to physical warfare, gradually acquired the meaning of a metaphorical struggle. Figures 9 and 10 give the words most similar to *guerra* during the time periods 1522-1900 and 2005-2009. Whereas the words associated with *guerra* in the earlier time period are associated with traditional warfare, such as *machacamiento* "destruction" and *astilleros* "shipyards," the words associated with *guerra* in more recent times, such as *socialistas*

"socialists" and *trastorna* "[civil] disturbance" allude to modern social and political struggles in the hispanophone world. *Guerra* received a Procrustes score of 0.370 (out of 1), indicating lower similarity between these two time periods.

**Figure 9: t-SNE plot of words most similar to *guerra* ("war"), 1522-1900**

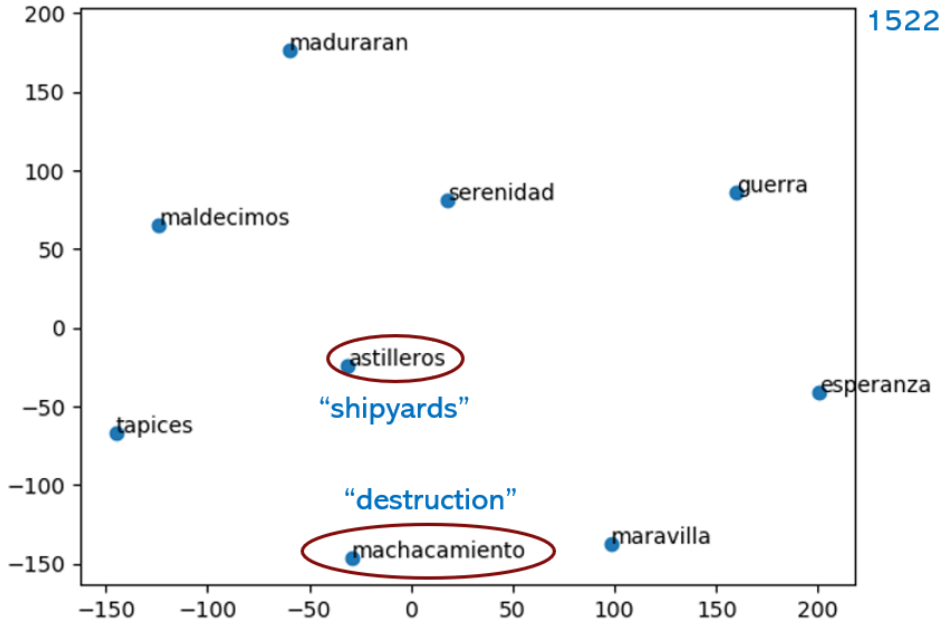
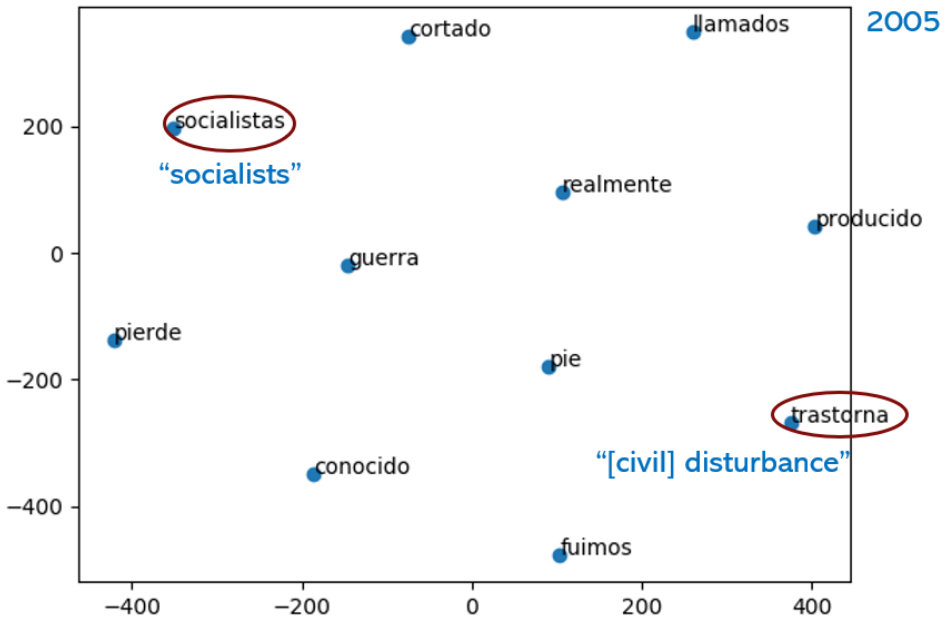



**Figure 10: t-SNE plot of words most similar to *guerra* ("war"), 2005-2009**

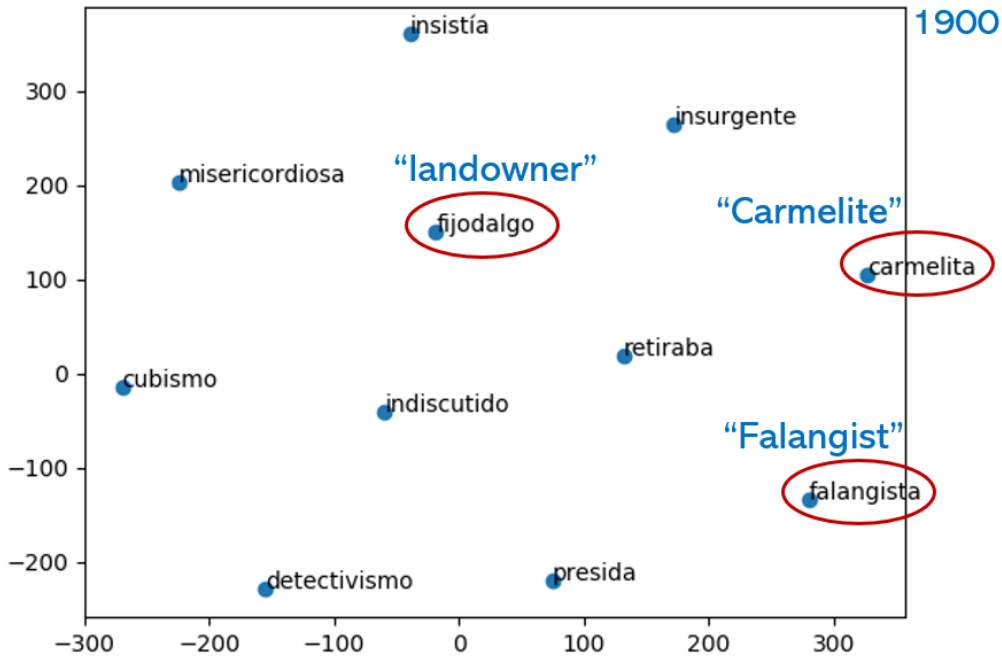



Besides long-term change, the model also appears to capture sudden changes in usage stemming from specific events. For example, the most similar words to *insurgente* "insurgent" in the

period 1900-1950 (Figure 11) include *carmelita* "Carmelite," *falangista* "Falangist," and *fijodalgo* "landowner." The Carmelite nuns and Falangist party were two groups heavily involved in the Spanish Civil War (1936-1939),[10] while *fijodalgo* was a Mexican term for landowners who were the target of multiple revolts in the early 20th century.

One issue may be that these changes do not necessarily indicate change in the senses of a word, but rather in the contexts to which they are applied. This raises a question for future analyses of semantic shift: whether to prioritize detection of changes in the semantic senses of a word, or take into account changes in context, usage, or sentiment regarding a word that affect its usage. Depending on the application used, either of these approaches may be more appropriate for modeling changes in language.

**Figure 11: t-SNE plot of words most similar to *insurgente* ("insurgent"), 1900-1950**




Other changes captured by the model appear to capture changes in linguistic usage that extend outside those captured by traditional lexicography. The Royal Spanish Academy still defines *prensa* "press media" as "periodical publications, particularly newspapers." The words most similar to

[10]The Falangists, a fascist group led by Francisco Franco, took control of Spain and oppressed, among other groups, the Carmelite nuns.

*prensa* in the period 1950-1970, such as *diarismo* "newspaper journalism" and *cartillas* "notepads," align with this definition (Figure 12). However, by the next time period (1970-1985), the model begins to capture the use of *prensa* to include other forms of journalism, including radio and television (Figure 13). *Prensa* received a Procrustes score of 0.333 between these two time periods, suggesting significant change.

Thus, there are discrepancies between the semantic change captured by lexicographers and the actual changes in words' usage captured by the model. Although these differences decrease the accuracy of the classifier due to the limitations of the test set, they do indicate that the model may capture more nuanced or gradual semantic changes than the discrete changes in dictionary definitions.

## 6. Conclusion

### 6.1. Summary

Automatic detection of semantic shift helps to build models of language that can account for changes in the meanings of words over time. These models are particularly useful for tasks sensitive to changes in words' senses, such as diachronic sentiment analysis, word sense discovery, and word sense disambiguation.

This project modeled words in Spanish documents from the past 400 years using neural embeddings based on four datasets with different context sizes (bigrams, trigrams, quadgrams, and quintgrams from the Google n-gram corpus). For each n-gram size, historical word embeddings were created for each of eight time periods; then, a support vector classifier was trained on a dataset of words annotated with when they gained or lost a sense to detect whether words underwent change between two time periods. The optimal context size for semantic shift detection was then evaluated on that dataset. This method was indeed effective at detecting semantic shift, and the best precision, recall, and F1 score were achieved when training models on n-grams of length 3.

Upon analyzing the results by lexical category, the results suggest that for semantic shift detection, words in different lexical categories benefit from being modeled with different context window

**Figure 12: t-SNE plot of words most similar to *prensa* (“press media”), 1950-1970**

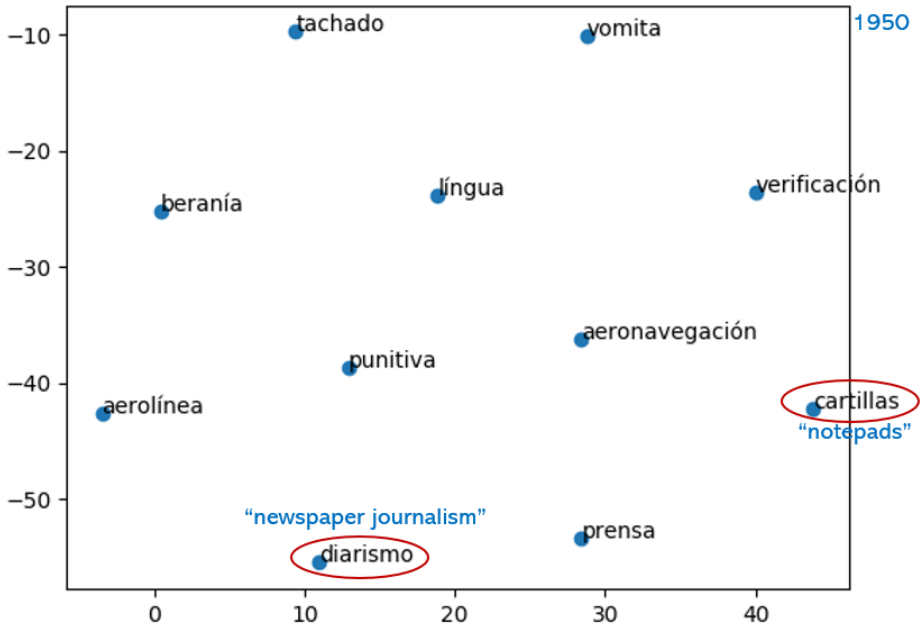
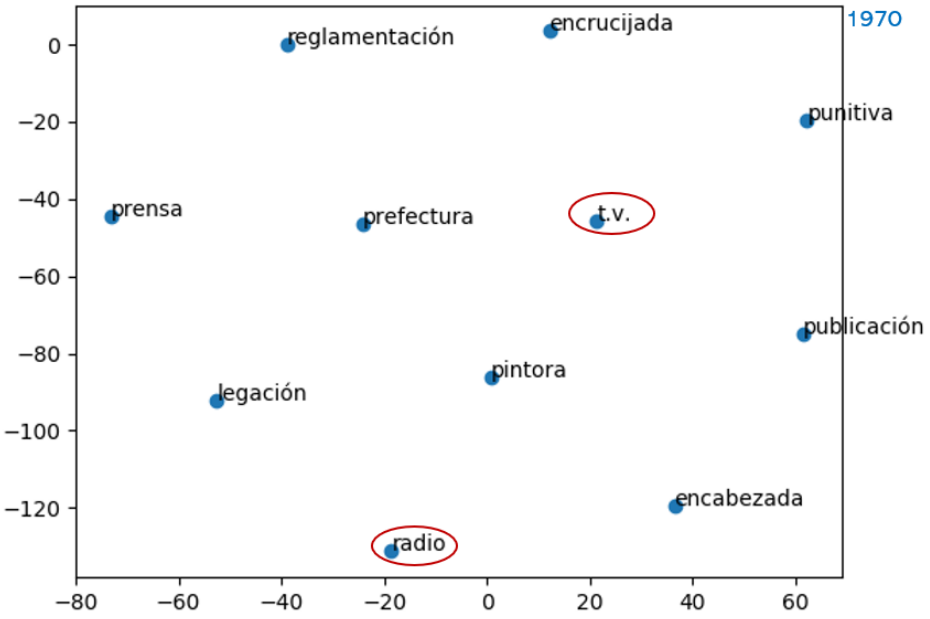



**Figure 13: t-SNE plot of words most similar to *prensa* (“press media”), 1970-1985**




sizes. Whereas nouns were best modeled with the trigram classifier, adjectives were best modeled with quadgrams and verbs with quadgrams or quintgrams. All models performed best at detecting shift in verbs, raising the question of whether different models or classifiers would perform better on different lexical categories. In addition, closer examination of specific words suggests that the models capture different types of semantic shift, long- and short-term change, and changes in usage

outside those captured by traditional lexicography.

A limitation of this work is that it is difficult to set a precise date for when a word acquired or lost a sense, since semantic change is a gradual process; furthermore, the Royal Spanish Academy is conservative about updating definitions, which interferes with the accuracy of the dataset. However, without comparable sources of information on historical usage of Spanish, little can be done to resolve this issue short of forgoing extrinsic evaluation for intrinsic metrics. Another issue is that less research has been done on evaluating the best metrics for quantifying semantic shift, so the classifiers used for this paper relied on only one feature measuring semantic change that earlier work found to be effective [5]. Future research could suggest other effective features to improve classifier performance.

### 6.2. Future Work

The results of the analysis of lexical categories raise several directions for future work. One open question is to examine why the ideal context sizes for semantic shift detection varies among words in different lexical categories, and whether there are semantic or syntactic aspects of Spanish that explain this trend. A related area open for research is determining whether these results hold for languages besides Spanish or for tasks besides semantic shift detection.

Other extensions to this research could increase the accuracy and applicability of semantic shift modeling. Earlier studies attempted to detect specific types of semantic shifts (e.g., the broadening, narrowing, joining, or splitting of senses), with limited results [9]. However, since these studies did not draw on new methods using neural network-based word embeddings, future studies attempting the same task with more robust embeddings may yield better results.

Another possible direction of research is to examine larger-scale or multilingual change, such as tracing the etymology of words or changes in words' meaning between mother and daughter languages. Such studies would be relevant to the work of historical and comparative linguists, including the reconstruction of proto-languages and language families. In addition, the models' ability to capture more gradual change than the discrete changes recorded by lexicographers raises

the possibility that these models could be used as aids for recording and analyzing more precise changes in linguistic usage.

The computational models' ability to model change more fluidly could be further improved by the use of continuous models of words' change over time (i.e., rather than modeling their usage in each of several time periods), an approach pioneered by Rosenfeld and Erk (2018) but not yet applied to languages besides English [13]. The success of new approaches at detecting semantic shift suggests that increasing the nuance and scope of semantic shift studies could benefit a variety of areas in natural language processing and linguistics.

## 7. Acknowledgments

Many thanks to Professor Christiane Fellbaum for her invaluable guidance throughout the semester. Her experience, advice, and enthusiasm made the process of completing this independent work a delight. Thank you as well to the students of the Natural Language Processing seminar for their feedback and support.

Also, special thanks to the Princeton Research Computing group for their support with the Adroit cluster, without which it would have been very difficult to create neural embeddings of millions of words.

## 8. Honor Code

This paper represents my own work in accordance with University regulations.

Eve Fleisig

## References

[1] "Diccionario de la lengua española, versión 23.3 [Dictionary of the Spanish language, version 23.3]," 2019. [Online]. Available: https://dle.rae.es

[2] "Nuevo diccionario histórico de la lengua española [Historical dictionary of the Spanish language]," 2019. [Online]. Available: http://web.frl.es/DH

[3] P. Basile, A. Caputo, R. Luisi, and G. Semeraro, "Diachronic analysis of the italian language exploiting google ngram," *CLiC it*, p. 56, 2016.

[4] H. Dubossarsky, D. Weinshall, and E. Grossman, "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*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 1136–1145.

[5] V. Fomin, D. Bakshandaeva, J. Rodina, and A. Kutuzov, "Tracing cultural diachronic semantic shifts in russian using word embeddings: test sets and baselines," *arXiv preprint arXiv:1905.06837*, 2019.

[6] W. L. Hamilton, J. Leskovec, and D. Jurafsky, "Diachronic word embeddings reveal statistical laws of semantic change," *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016. [Online]. Available: http://dx.doi.org/10.18653/v1/P16-1141

[7] Y. Kim, Y.-I. Chiu, K. Hanaki, D. Hegde, and S. Petrov, "Temporal analysis of language through neural language models," in *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. Baltimore, MD, USA: Association for Computational Linguistics, Jun. 2014, pp. 61–65. [Online]. Available: https://www.aclweb.org/anthology/W14-2517

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.

[9] S. Mitra, R. Mitra, S. K. Maity, M. Riedl, C. Biemann, P. Goyal, and A. Mukherjee, "An automatic approach to identify word sense changes in text media across timescales," *Natural Language Engineering*, vol. 21, no. 5, p. 773–798, 2015.

[10] S. Mitra, R. Mitra, M. Riedl, C. Biemann, A. Mukherjee, and P. Goyal, "That's sick dude!: Automatic identification of word sense change across different timescales," *CoRR*, vol. abs/1405.4392, 2014. [Online]. Available: http://arxiv.org/abs/1405.4392

[11] B. Navarro-Colorado, "On poetic topic modeling: Extracting themes and motifs from a corpus of spanish poetry," *Frontiers in Digital Humanities*, vol. 5, p. 15, 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fdigh.2018.00015

[12] N. Pochet and J. Suykens, "Support vector machines versus logistic regression: Improving prospective performance in clinical decision-making," *Ultrasound in obstetrics gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology*, vol. 27, pp. 607–8, 06 2006.

[13] A. Rosenfeld and K. Erk, "Deep neural models of semantic shift," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 474–484. [Online]. Available: https://www.aclweb.org/anthology/N18-1044

[14] C. Sánchez-Marco, R. Marin, and S. Evert, "Measuring lexical extension: The case of spanish estar + past participle," *Linguistic Evidence*, 2012.

[15] X. Tang, "A state-of-the-art of semantic change computation," *CoRR*, vol. abs/1801.09872, 2018. [Online]. Available: http://arxiv.org/abs/1801.09872

[16] D. T. Wijaya and R. Yeniterzi, "Understanding semantic change of words over centuries," in *Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural diversiTy on the Social Web*, ser. DETECT '11. New York, NY, USA: ACM, 2011, pp. 35–40. [Online]. Available: http://doi.acm.org/10.1145/2064448.2064475

[17] A. C. Zentella, "Limpia, fija y da esplendor: Challenging the symbolic violence of the Royal Spanish Academy."