Historical Ink: Semantic Shift Detection for 19th Century Spanish

Tony Montes[1] **Laura Manrique-Gómez**[2] **Rubén Manrique**[1]
[1] Systems and Computing Engineering Department, Universidad de los Andes
[2] History and Geography Department, Universidad de los Andes
Bogotá D.C.
{t.montes, l.manriqueg, rf.manrique}@uniandes.edu.co

Abstract

This paper explores the evolution of word meanings in 19th-century Spanish texts, with an emphasis on Latin American Spanish, using computational linguistics techniques. It addresses the Semantic Shift Detection (SSD) task, which is crucial for understanding linguistic evolution, particularly in historical contexts. The study focuses on analyzing a set of Spanish target words. To achieve this, a 19th-century Spanish corpus is constructed, and a customizable pipeline for SSD tasks is developed. This pipeline helps find the senses of a word and measure their semantic change between two corpora using fine-tuned BERT-like models with old Spanish texts for both Latin American and general Spanish cases. The results provide valuable insights into the cultural and societal shifts reflected in language changes over time[1].

1 Introduction

The study of how word meanings evolve over time, influenced by social, historical, and political factors, is a fundamental pursuit within linguistics and natural language processing. This evolution poses challenges in detection and interpretation, often addressed through Semantic Shift Detection (SSD) task, also known as Lexical Semantic Change Detection task (LSCD) (Montanelli and Periti, 2023; Hu et al., 2021). Traditionally reliant on manual methods such as discourse analysis, recent computational linguistics advancements have revolutionized this field. These approaches streamline analysis and open doors to interdisciplinary research applications spanning sociology, history, and beyond, offering invaluable insights into cultural and societal shifts using digitized corpora.

In 2013, static word embeddings, also known as word vector representations, were first introduced by Mikolov et al. (2013) using the bag-of-words and skip-gram architectures. These embeddings represent words as static vectors that remain unchanged and are based on their surrounding words. Hamilton et al. (2016) first proposed using these embeddings for the SSD task by employing diachronic word2vec static embeddings to measure word meaning changes across consecutive decades. Various approaches have been explored to automate this task effectively. Montanelli and Periti (2023) proposed using contextual embeddings instead to capture multiple meanings assigned to the same word due to polysemy and homonymy, which static embeddings cannot achieve. This was accomplished by comparing multiple BERT-like Language Models (Devlin et al., 2018) such as XLM-RoBERTa.

In this paper, we focus on two things: crafting a 19th-century Spanish corpus (C_{old}) from sources spanning 1800 to 1914 and creating a customizable pipeline for assessing the SSD task. Utilizing this pipeline, we analyze the semantic changes of a set of target words, for both the global context and the specific Latin-American context. We explore a variety of known and novel solutions for the SSD task by comparing the 19th-century Spanish corpus with the Spanish portion of the "EUBookShop" corpus as the modern corpus (C_{new}) (Cañete, 2019)[2].

2 Related Work

Recent advances in Semantic Shift Detection have leveraged many computational approaches based on natural language processing techniques. Contextual embeddings, capable of capturing multiple-word usages and meanings, have been used in most of the state-of-the-art solutions, summarized by Montanelli and Periti (2023) who defines a classification framework based on three dimensions

[1]The pipeline and code can be found at https://github.com/historicalink/SSD-Old-Spanish

[2]This portion was taken from the large Spanish corpus available at https://huggingface.co/datasets/josecannete/large_spanish_corpus

Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change, pages 29–41

of analysis: meaning representation (*form-* and *sense-oriented* approaches), time awareness (*time-oblivious* and *-aware*) and learning modality (*supervised* and *unsupervised*, referencing to the injection of external knowledge support like a dictionary), useful for *Contextualized Semantic Shif Detection*.

Martinc et al. (2020) and Giulianelli et al. (2020) explore transformer-based BERT models for detecting semantic change. Martinc et al. use contextualized embeddings to capture shifts in word usage over time, outperforming traditional techniques like Word2Vec and Glove by leveraging BERT's dynamic word representations. Giulianelli et al. (2020) adopt an unsupervised approach, obtaining and clustering word representations to measure change over time, aligning with human judgments. Both studies underscore the effectiveness of BERT-based models in identifying and analyzing diachronic linguistic changes.

Although most of the research in the field of semantic change has been done on a wide scope of languages, Spanish hasn't played such an important role in this field, except for some research, like LSCDiscovery in Spanish, a task presented by Zamora-Reina et al. (2022). This task has facilitated the development and evaluation of SSD systems in this language, accompanied by an unannotated Spanish corpus for both modern and old texts, which has a size of $22M$ and $13M$ tokens respectively. Additionally, the task paper highlighted effective techniques and approaches within the solutions. The most successful solution for the LCSDiscovery task was GlossReader, developed by Rachinskiy and Arefyev (2022), which involved fine-tuning XLM-RoBERTa, a Language Model trained on more than 100 languages, with old English datasets and employing the model zero-shot cross-lingual transferability of the model to build contextualized embeddings for Spanish, and using this fine-tuned model for SSD tasks. This approach has demonstrated good performance, especially in avoiding issues associated with word form bias and labor-intensive annotation requirements. These advancements underscore the increasing significance and potential of computational methodologies in enhancing our comprehension and automation of semantic shifts in multiple languages.

Also, Hu et al. (2021) present a set of methodological considerations for low-resource languages such as 15th-century Spanish, where a lower amount of data is available, and the data is not as clean as in other high-resource languages such as English and Mandarin Chinese, stating that common SSD techniques are also useful for these cases, but must be used carefully, under a set of considerations.

3 Data

Selecting the data is a crucial step for the reliability of the results. The LSCDiscovery shared task provides a useful corpus for old Spanish texts within the years 1810-1906, with a size of $13M$ tokens (Zamora-Reina et al., 2022). However, this paper aims to construct a larger old Spanish corpus, also adding more presence from Latin-American countries. The main sources selected and filtered for this corpus were **Project Gutenberg**[3] which was filtered by language and by the given date ranges (1800-1914), **The British Library books**[4] (portion from 1800-1899) which was also filtered by language (British Library Labs, 2021), and the **LatamXIX**[5] dataset from the *Historical Ink* project which contains Latin American texts from newspapers within years 1845-1899 (Manrique-Gómez et al., 2024).

3.1 Cleaning

The cleaning step is essential for The British Library and Project Gutenberg datasets since some texts from these sources consisted solely of chapter, book, or newspaper titles, or were filled with numbers and other characters that added noise to the dataset. In the case of the LatamXIX dataset, these noisy rows were already filtered and complemented with an LLM OCR correction process that corrected many OCR errors within the corpus, making it cleaner and more fittable for the SSD task, as it preserves better semantic meaning for words and less noise.

For The British Library books, an initial filter was applied using word confidence information to retain only those books with a mean OCR word confidence higher than 0.5. This experimental threshold was set to balance data loss (2.26% of rows) and text quality. After conducting several revisions with different examples, it was observed that this threshold maintained a high standard of text quality. Therefore, it was selected as the optimal balance

[3]Available at https://www.gutenberg.org/browse/languages/es

[4]Available at https://huggingface.co/datasets/TheBritishLibrary/blbooks

[5]Available at https://huggingface.co/datasets/Flaglab/latam-xix

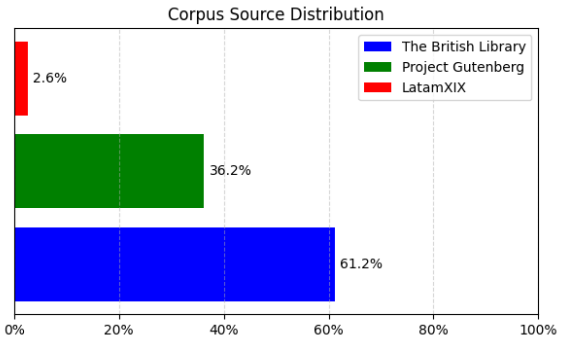

Figure 1: Final corpus distribution by source. The percentage is computed over the total number of rows of the whole C_{old} chunked corpus

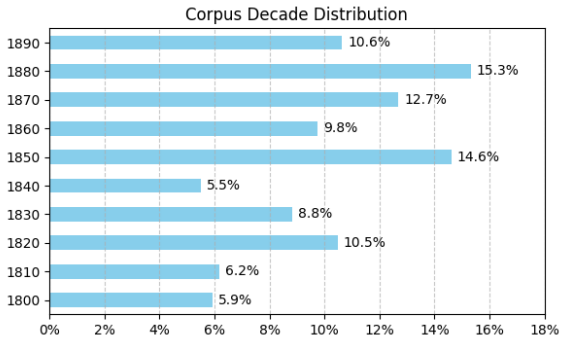

Figure 2: Final corpus distribution by decade. The percentage is computed over the total number of rows of the whole C_{old} chunked corpus

Feature	Value
Size	$\sim 865MB$
Rows	$1,141,490$
Words	$\sim 125M$
Tokens	$\sim 160M$
Years Range	1800 - 1914

Table 1: Final C_{old} chunked corpus information

Feature	Value
Size	$\sim 27MB$
Rows	29.972
Words	$\sim 4.5M$
Tokens	$\sim 5.7M$
Years Range	1845 - 1899

Table 2: Final C_{old} Latin-American portion chunked corpus information

between data retention and textual accuracy.

Same as in Manrique-Gómez et al. (2024), the cleaning steps to perform were:

1. Remove duplicates and empty rows within the whole dataset. 6.94% of rows were removed.

2. Filter out rows where over 50% of the characters are non-alphabetic, including spaces. 0.92% of rows were removed.

3. Remove the rows with 4 or fewer tokens. Samely, a new tokenizer was trained with a vocabulary size of 52,000, trained from the BETO pre-trained tokenizer (Cañete et al., 2020). 0.50% of rows were removed.

These filters were applied to minimize the risk of compromising the results due to noise in the dataset.

3.2 Chunking

As the historical texts from the corpus come from books and newspapers, many are very large, or some are very short with an average of $\sim$ 110 words and $\sim$ 140 tokens per text. For BERT-like models, the maximum sequence length consists of 512 tokens, which is not enough for very large texts like the current corpus texts.

Because of this, it's necessary to chunk the large texts within the dataset in a number shorter than 512 tokens. A much lower number was selected to make the chunked corpus fit for many different Language Models (LMs), for instance, a maximum of 256 tokens per text chunk, where a token was measured by training a new tokenizer over the cleaned version of the corpus[6].

During this step, over 67.6% of the rows were chunked, adding 460,543 new rows. Each row was transformed into a part of a paragraph or left as a whole paragraph (no chunking) with no more than 256 tokens while preserving as much semantic meaning as possible. The preservation of semantic meaning in the chunked segments was achieved by splitting through punctuation marks and common paragraph-sentence separators. The rows distribution and corpus information can be found in Figures 1, 2, and Table 1 respectively. Also, the information on the Latin-American portion of the corpus can be found in Table 2.

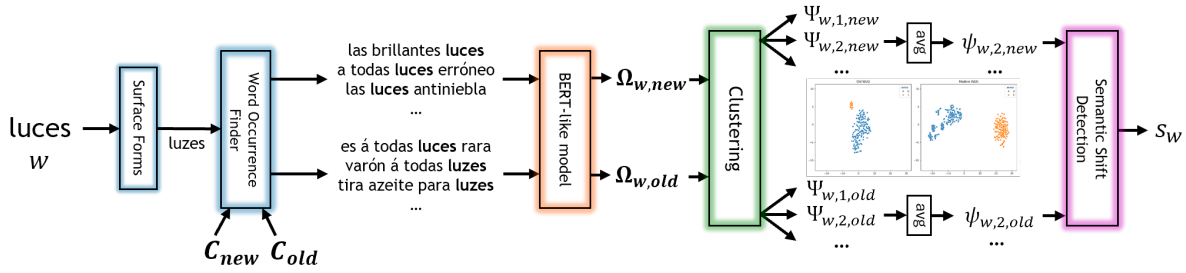

Figure 3: Historical Ink SSD Pipeline Architecture

4 Methodology

To achieve effectively the desired task, and be able to perform a quality analysis of the results, we have defined the pipeline observed in Figure 3, with the following steps:

1. Find the occurrences of a given word w in C_{old} and C_{new} corpora.
2. Retrieve the word embeddings in the found occurrences, using a BERT-like language model.
3. Cluster the word usage by its meaning (sense), and average to get the centroids of the clusters.
4. Perform the SSD task to identify lost/gained senses and measure the semantic change of the word (s_w).

It's important to note that the pipeline was designed as a flexible and reusable solution for various contexts and configurable stages. Beyond analyzing the specific case of 19th-century Spanish, we propose a modular, plug-and-play pipeline with numerous adjustable stages. Each component of the pipeline can be used independently and configured for different use cases, ensuring versatility and adaptability for further research or applications.

4.1 Find the Occurrences

Given corpora C_{old} and C_{new}, finding all texts where a word w is used is straightforward when looking for *exact occurrences*. However, this task becomes more complex with inflectional variations typical of languages like Spanish. For example, the word "crear" (to create) may appear as "creaste" (you created) or "creado" (created). Stemming can help by extracting the base form of the word, but it may lose some contextual meanings.

Also, in old Spanish, language rules have changed significantly, as noted by Montgomery (1966). These changes are detectable using the semi-automated framework presented as part of the Historical Ink project (Manrique-Gómez et al., 2024), which extracts useful lists of *surface forms* (i.e. specific appearance of a word in a given context) for words that underwent orthographic changes in 19th-century Latin-American Spanish (e.g., "luces" historically written as "luzes").

To address these challenges, we propose a method to find occurrences of a given word w in diachronic corpora C_{old} and C_{new}. This method organizes all word's expected usages and tokenizes both the word and the searching text, searching for each subword within a list of different orthographic forms of writing a given word.

For example, the word "gente" would be searched in C_{old} as "gente", "jente" (surface form), "gent", or "jent" in that order. This method relies heavily on the tokenizer, so using one trained in the specific language is recommended for better performance.

4.2 Word Embeddings

For the SSD task, contextual embeddings are very useful as they can capture the evolving meaning of words over time. By considering the surrounding context of a word within a sentence or document, contextual embeddings can provide an enhanced representation of its semantics, enabling the detection of particular shifts in meaning. In particular, there are some BERT-like LMs trained on Spanish corpora. Some of the most representative are BETO: Spanish Bert[7] in both uncased and cased versions (Cañete et al., 2020), Multilingual BERT[8] in both uncased and cased versions, which has an

[6]The final corpus can be found at https://huggingface.co/datasets/Flaglab/spanish-corpus-xix in all its three versions: "original", "cleaned", and "chunked"

[7]Available at https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased

[8]Available at https://huggingface.co/google-bert/bert-base-multilingual-cased

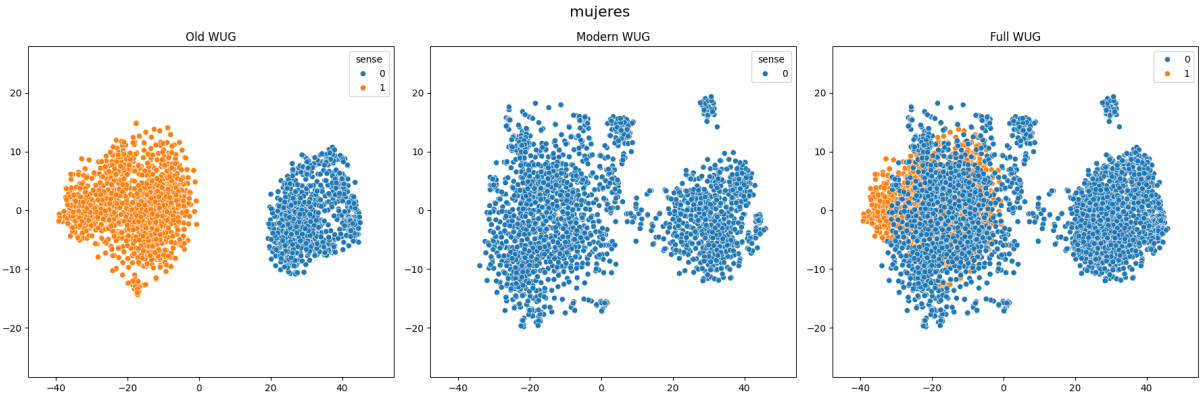

Figure 4: DWUG of the word "mujeres" (women), using the whole corpus fine-tuned model embeddings, the T-SNE dimensionality reduction algorithm, and the KMeans clustering algorithm (with the silhouette metric). Each color represents a meaning (cluster) of the word. The color changes between the left (old corpus) and center (modern corpus) images illustrate the overall semantic change between the two diachronic corpora.

important portion of training in Spanish (Devlin et al., 2018), and AlBERT Spanish version[9]. All these models are BERT-based and have the same maximum sequence length of 512 tokens, BERT has an embedding size of 768, while ALBERT has a more compact embedding size of 128.

For this paper, we performed the SSD task using the mentioned LMs. Some were trained with the whole 19th-century Spanish corpus, while others were trained only with the Latin-American portion of the dataset. We fine-tuned these models using the specific corpus for each case, employing the Masked Language Modeling (MLM) task. In this task, 15% of the corpus tokens were randomly masked, and the model learned to predict the masked tokens based on their context. This approach ensured that the model learned the unique linguistic style of each corpus, enabling it to generate word embeddings that accurately reflect the corpus' linguistic patterns, which is essential for detecting semantic shifts.

During the training phase, an Adam optimizer with a learning rate of 2×10^{-5} was employed, and the training proceeded with a batch size of 32, during a total of 5 epochs. Due to the low number of epochs, no Early Stopping was required, and the chosen parameters led to good resource utilization. The training time with the given batch size depended on the model but was on average 47 hours for the whole C_{old} corpus, and 1 hour and 20 minutes for the Latin-American portion. The training was performed on an A40 GPU.

[9]Available at https://huggingface.co/dccuchile/albert-base-spanish

4.3 Clustering

We applied a joint clustering approach, combining both corpora within the same set of embeddings before clustering. Given two corpus C_{old} and C_{new}, and a particular word w, the sets $\Omega_{w,old}$ and $\Omega_{w,new}$ are defined as the set of word embeddings generated in each corpus respectively, for the word w.

The clustering algorithm is meant to find the different meanings of a word within a given period, and overall the whole timespan of both C_{old} and C_{new} periods. This generates a well-known Diachronic Word Usage Graph (DWUG) for the word in both periods (Schlechtweg et al., 2021), allowing to perform the semantic shift detection and change measurement between *old* and *new* periods, as seen in Figure 4, where each color refers to a word meaning.

The particular algorithms used were Affinity Propagation and KMeans with an automatic K finder under a certain score function such as silhouette score or inertia. The main problem with KMeans are words with a single meaning across the whole timespan. As common KMeans K-evaluation metrics are not fittable for one-cluster evaluation, so it wouldn't be possible to validate if the best number of clusters should be just one. As this occurs for many of the target words selected for analysis, a very good alternative for it is the Affinity Propagation (AP) clustering algorithm with a damping parameter of 0.975; this parameter was selected through a test with different values and a manually-driven evaluation of the number of clusters automatically selected by the algorithm. Selecting a high damping value for the AP algo-

rithm leads to a more stable selection of the number of clusters as the requirements for new cluster creation are more strict, which is expected for this case.

The T-SNE dimensionality reduction algorithm was used to plot the DWUGs shown in this paper, with a perplexity of 50, which proved the best for better cluster space separation. For words with a lower number of found occurrences in the dataset, a lower perplexity was employed for its representation.

4.4 Semantic Shift Measurement

Once clustering is performed, the measurement for Semantic Shift is straightforward. There are two main divisions of the SSD task which are Binary Change Detection (BCD) and Graded Change Detection (GCD) (Zamora-Reina et al., 2022), where Graded Change Detection is the most common and useful, but also the most challenging task for change classification, which consists of ranking a list of target words based on their degree of change (Periti and Tahmasebi, 2024).

The consolidation of techniques for measuring semantic shift detection has been a high-growth area, with the proposal of many different techniques, some of them comparing sets of embeddings (e.g. the clusters), and others comparing individual embeddings (e.g. the centroids). Montanelli and Periti (2023) present a survey that compiles many of the most used state-of-the-art techniques for grading the semantic shift of a word between two temporal-different corpora, classifying them between form- and sense-based approaches.

Given m number of clusters (senses) for the word w, returned by the clustering algorithm, we define $\Psi_{w,s,t}$ as the cluster with the sense s for the word w in the period t, such that all the senses compound the whole set of embeddings.

$$\Omega_{w,t} = \bigcup_{s=1}^{m} \Psi_{w,s,t} \quad \forall t = \{new, old\} \tag{1}$$

For these clusters, a centroid embedding is computed as the average:

$$\psi_{w,s,t} = \text{avg}(\Psi_{w,s,t}) \quad \forall t = \{new, old\} \tag{2}$$

Finally, two different formulas were taken from Montanelli and Periti (2023) to measure the semantic shift f, based on the cosine similarity function (CS). With this shift, for each word, we would have as many semantic shifts f as the number of clusters given by the algorithm (m), so we could determine which senses have had a diachronic shift and which haven't, for each word.

Cosine Distance (CD):

$$f_{CD}(w, s) = 1 - CS(\psi_{w,s,old}, \psi_{w,s,new}) \tag{3}$$

Inverted similarity over Word Prototype (PRT):

$$f_{PRT}(w, s) = \frac{1}{CS(\psi_{w,s,old}, \psi_{w,s,new})} \tag{4}$$

It should be noted that if a sense is not present within a period, whether old or new period, f_{CD} should be 1.0, meaning a complete change of the given sense. If the sense is absent from the embeddings of the old period ($\Psi_{w,s,old} = \emptyset$), it means that the sense was gained in modern Spanish; otherwise, if the sense only exists in the embeddings of the old period ($\Psi_{w,s,new} = \emptyset$), it means that the sense was lost in modern Spanish, as seen in Figure 4 where the sense 1 (orange color) is not present in the modern WUG.

For this task, it is crucial to consider the frequency of points per cluster within each period. If a cluster has significantly fewer points in a period, specifically less than 10% of the total, we classify these points as either misclassifications or obsolete words. This allows us to treat the cluster as a gained or lost sense. We chose this threshold based on testing with few known examples, where it provided the best performance in detecting gained and lost senses.

5 Evaluation and Model Selection

As mentioned, several pre-trained Language Models (LMs) are available for large Spanish corpora. We needed an evaluation method to select the best model for our analysis. The LSCDiscovery shared task (Zamora-Reina et al., 2022) provides over 65,000 annotated examples for 100 target words using the DURel framework proposed by Schlechtweg et al. (2018). This annotated corpus is highly useful for evaluating the LMs, as its time period is within the 19th century. Even though the LSCDiscovery task differs from the one in this paper, it offers a valuable benchmark. Our task focuses on detecting the different meanings of a word in a diachronic corpora and measuring their

LM		Average		Clustering			Non-Clustering	
#	Name	Clustering	All	AP	KM inertia	KM silhouette	CD	PRT
1	BETO cased FT	0.5799	0.6017	**0.6124**	0.5598	0.5676	0.6285	0.6402
2	**BETO cased LFT**	**0.5872 (1)**	0.6064	0.5853	0.5815	0.5947	0.6307	0.6396
3	BETO cased	0.5832	0.6041	0.5600	0.5790	**0.6107**	0.6302	0.6405
4	BETO uncased FT	0.5578	0.5837	0.5442	0.5579	0.5714	0.6224	0.6227
5	BETO uncased LFT	0.5658	0.5890	0.5594	0.5676	0.5703	0.6223	0.6255
6	BETO uncased	**0.5862 (3)**	0.6043	0.5916	0.5819	0.5850	0.6167	0.6463
7	mBERT cased LFT	0.5806	0.5951	0.5692	0.5788	0.5939	0.6163	0.6172
8	mBERT cased	0.5782	0.5949	0.5675	0.5808	0.5863	0.6100	0.6297
9	mBERT uncased LFT	0.5593	0.5929	0.553	0.5633	0.5615	0.6405	0.6464
10	mBERT uncased	0.5762	0.6065	0.5523	0.5924	0.5839	**0.6457**	0.6581
11	AlBERT LFT	0.5717	0.5928	0.5731	0.5796	0.5624	0.6160	0.6328
12	AlBERT	**0.5869 (2)**	**0.6132**	0.5758	**0.5992**	0.5857	0.6373	**0.6682**

Table 3: LM benchmark through the LSCDiscovery (Zamora-Reina et al., 2022) F1 of the Binary Change Detection task. Each model was fine-tuned for the Latin-American corpus (LFT) and both BETO-cased and uncased models were also fine-tuned for the whole corpus (FT), comparing also with non-fine-tuned versions.

semantic shift over time. By comparing with the LSCDiscovery task, we ensure a rigorous evaluation, confirming that the models are robust and effective across various contexts and not overly tailored to a single specific task.

The task's corpus includes pairs of sentences rated from 1 to 4, where 1 indicates identical word usage and 4 indicates completely different usage (Schlechtweg et al., 2018). To evaluate the models, we converted this numerical assessment into a binary evaluation: ratings 1-2 indicated no semantic change, while ratings 3-4 indicated a semantic change. We then defined five specific methods to classify a pair of word uses as either semantic change (1) or no change (0). Among these, two methods — *cosine distance* (CD) and *inverted similarity over word prototype* (PRT) — were tested purely for task purposes. However, the methods of primary importance for this paper are those related to sense clustering.

The three clustering-based evaluation methods consist of grouping all the embeddings of the occurrences of a word, as mentioned in the SSD section. Then, given two uses, if they do not belong to the same cluster, a semantic change is indicated (1); otherwise, no semantic change is indicated (0). This was evaluated using Affinity Propagation and KMeans (with silhouette and inertia metrics) methods. Finally, the model with the best average results across the three clustering methods was selected. The benchmark results can be seen in Table 3.

While the results provide valuable insights into the models' capabilities, they should not be directly compared to those from the LSCDiscovery leaderboard (Zamora-Reina et al., 2022). Instead, they serve as an effective benchmark for assessing how well the LMs perform in detecting semantic changes within our specific historical context. The differences in tasks and the method approaches for our study reflect that direct comparisons with LSCDiscovery scores are not applicable.

Given the results, the best-performing model was BETO fine-tuned on the Latin American dataset[10]. A possible explanation for this is that the Latin American portion of the corpus underwent an additional step of LLM OCR correction, which removed OCR-related errors and produced cleaner text. This likely reduced noise and improved the quality of fine-tuning. Additionally, BETO was trained solely in Spanish, unlike multilingual BERT, which was trained in many different languages. According to (Cañete et al., 2020), this single-language focus tends to result in better performance compared to multilingual models. This model was the one used for evaluating the target words and creating the DWUGs presented in appendix C.

[10]Fine-tuned model was uploaded to HuggingFace and is available at https://huggingface.co/Flaglab/beto-cased-finetuned-xix-latam

6 Results

The results of the trained model focus on a specific group of 255 target words[11] selected for their historical significance and relevance to generate hypotheses about potential semantic shifts over time, confirming the consistency of the results. Some examples of the DWUGs analyzed in this section are available in Appendix C for both AP and KMeans.

One of the main results of this research was to highlight the success and failure cases for both AP and KMeans clustering algorithms, as both were used to compute the senses of all 255 words. Affinity Propagation (AP) performed poorly in many cases where it couldn't detect multiple usages of a word, such as "grave" (serious/bass), or detected many different senses for other words, such as "honor" (honour), as shown in Figure C2. However, it effectively detected single-sense words, a task that KMeans wasn't capable of due to metrics used to choose the best K. However, KMeans performed very well in most cases, effectively detecting and clustering the senses of multi-meaning words over time.

As displayed in Figure C2, some words like "rey" (king) and "usurer" (usurer) present neither polysemy nor notable historical changes. However, the term "mujeres" (women), as shown in Figure 4, shows a change in modern usage. This finding is particularly interesting in the context of both historical discourse analysis in gender studies and historical linguistics studies, as it is an example of computational verification.

The semantic transformation of the word women, as plotted in Figure 4 and in Appendix B, primarily pertains to the antiquated use of "mujeres" designating a particular group of female individuals. In 19th-century Spanish, lexical tradition mandated the rigorous use of masculine forms of nouns and adjectives as the universal form, encompassing both genders (feminine and masculine) (Porto-Dapena, 1975). Thus, the word "hombres" (men) could be used as a synonym for humanity, while the use of "mujeres" (women) was more likely to be reserved for describing a private group of women. Twentieth-century gender studies introduced a unified meaning to the word "mujeres". Joan W. Scott famously stated that "-Women's experience- or -women's culture- exists only as the expression of female particularity in contrast to male universality" (Scott, 1988). This idea explains the rupture in the modern usage of the word women towards the relational concept of gender in the 20th century (Lux and Pérez, 2020).

Consequently, the term "mujeres" evolved from a specific designation to a broader and more inclusive reference, reflecting significant social and cultural shifts in gender discourse. As we have observed, the contemporary usage of "mujeres" tends to encompass all women more generically, since it was not until the 20th century that historical consideration began to differentiate "women" as a collective separate from "men". In the past, the term was used to refer to a distinct group of women, thereby distinguishing women from other plural nouns such as men, children, or even animals. Modern usage of "women" almost exclusively serves to differentiate women from men.

Other insightful results demonstrate both how the polysemy of words changes over time, as seen in examples in Appendix A, and the particularities of word semantics diachronically used in Latin American Spanish. Historical linguistics studies acknowledge "El español de América" as a main Spanish variant, for which corpus studies are yet to be conducted. Newspapers are recognized as a legitimate source for exploring the particularities of linguistic variants (Gutiérrez Maté and Diez del Corral Areta, 2023). Hence, the LatamXIX dataset we used to model the quantitative experiments might initiate a triangulation with new regional research. For example, we have observed how the term "infancia" (infancy/childhood), as depicted in Figure C1, was predominantly used in the 19th century as an abstract reference to the nascent phase of objects, entities, or people. This suggests a metaphorical use of the word, indicative of a broader, symbolic interpretation of "infancy" or "early development" during this era.

Newly formed Latin American nations in the 19th century viewed themselves as children recently independent from their mother, metropolitan Spain. Consequently, the term "infancia de la patria" (infancy of the nation) described the contradictory and highly unstable political and social times experienced in Latin America during that era. These old meanings have largely been supplanted by the modern understanding of "childhood", which specifically refers to the population segment of children. These results align with the

[11] From all 255 words, only 233 had enough occurrences in the modern corpus. The DWUGs and SSD for both AP and KMeans algorithms are available in the notebook https://colab.research.google.com/drive/1eaULQocxyuCNX0ftBvDJwe8nfpEi5s6i

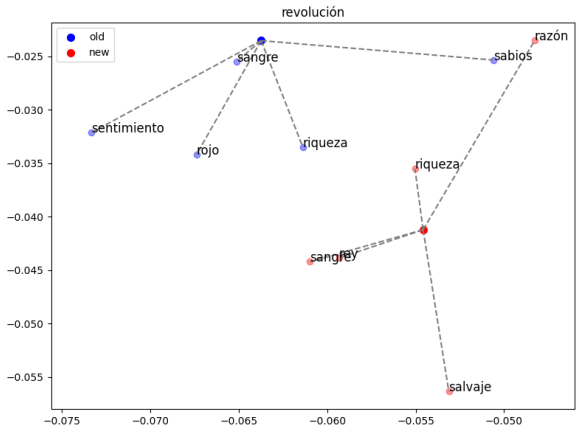

Figure 5: Diachronic comparison of word "revolución" (revolution) and its related words, between the old and the modern period using PCA dimensionality reduction algorithm.

second wave of human rights in the 20th century, which expanded the 19th century's initial civil rights to include specific rights for various western population groups, such as children and women.

Words like "sentimiento" (sentiment) have lost one of their historical meanings, as illustrated in Figure C1. In contemporary usage, "sentimiento" serves as a synonym for the feelings experienced by an individual or group of people. However, one of its older meanings has almost disappeared. In the 19th century, "sentimiento" was used to describe the expression of a person's correctness, effectively acting as a synonym for morality, or even referring to someone's elevated religious or artistic spirit. On the other hand, the term "sublime" (sublime or elevated) has largely fallen out of use and is scarcely found in the modern dataset, as depicted in Figure C1. Appendix B contains examples of the 255 words' semantic shift detection outputs, including other examples such as "luces" (ideas/lights) and "servidores" (servants/servers).

Finally, word comparison also proves highly valuable for numerous diachronic analyses. In each period, the most representative sense of a word is determined based on its frequency dominance among other senses. Then, its sense cluster centroid is computed to allow comparison between words. Within the set of 255 words, the 5 words exhibiting the highest cosine similarity to this centroid are selected, indicating their related usage contexts. For example, as observed in Figure 5, the word "revolución" (revolution) historically exhibited close associations with the words blood, richness, feeling, wise, and red. In contemporary contexts, however, the term "revolución" is linked to terms like king, reason, and savage, and it remains related to blood and richness in different proportions, with blood now more distant and richness closer.

This study provides significant insights into the SSD of 19th-century Spanish words, utilizing computational linguistics to uncover shifts in word meanings relevant to both global and Latin American contexts. By developing a specialized corpus and employing methods such as fine-tuning BERT-like models and diachronic word embeddings, we achieved a nuanced analysis of historical semantic changes. Our examination of selected words reveals the relation between societal, cultural, and political events and the shift of words' semantic meaning over time.

The application of SSD and modern computational techniques highlights the evolution of linguistic analysis from manual to systematic approaches, enhancing the accuracy of semantic shift detection and deepening our understanding of language as a dynamic entity. This study's interdisciplinary implications are notable, offering potential benefits to fields like history, sociology, and digital humanities, where these insights can provide deeper context to historical cultural shifts.

Looking ahead, the methodologies and findings of this project can serve as a framework for future research in other languages and periods, suggesting a scalable approach to historical linguistics and semantic analysis. The flexible and reusable pipeline developed here can be adapted for various contexts and stages. Future research could apply this pipeline with modified parameters or data for different use cases or languages, to prove its performance on different contexts.

However, an evaluation of the selected models for the Latin-American corpus, particularly for clustering, is still needed. An annotated dataset similar to the given in the AXOLOTL-24 shared task (Fedorova et al., 2024), but for Latin-American Spanish, would be highly beneficial. Such a dataset, with examples of specific word usages, their periods, and a gold standard for word senses, would enable a more focused assessment of the models beyond the task evaluation presented in Table 3.

7 Acknowledgements

We would like to thank the three anonymous reviewers from the ACL 2024 LChange'24 conference for their helpful feedback and suggestions.

References

British Library Labs. 2021. Digitised books. c. 1510 - c. 1900. JSONL (OCR derived text + metadata).

José Cañete. 2019. Compilation of large spanish unannotated corpora.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Mariia Fedorova, Timothee Mickus, Niko Tapio Partanen, Janine Siewert, Elena Spaziani, and Andrey Kutuzov. 2024. AXOLOTL'24 shared task on multilingual explainable semantic change modeling. In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, Bangkok. Association for Computational Linguistics.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Miguel Gutiérrez Maté and Elena Diez del Corral Areta. 2023. El español en américa (III): de las independencias a nuestros días. variedades andinas y caribeñas. In *Lingüística histórica del español / The Routledge Handbook of Spanish Historical Linguistics*, pages 539–545. Routledge, London.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Cultural shift or linguistic drift? comparing two computational measures of semantic change. *CoRR*, abs/1606.02821.

Hai Hu, Patrícia Amaral, and Sandra Kübler. 2021. Word embeddings and semantic shifts in historical spanish: Methodological considerations. *Digital Scholarship in the Humanities*, 37(2):441–461.

Martha Lux and María Cristina Pérez Pérez. 2020. Los estudios de historia y género en américa latina. *Historia Crítica*, 1.

Laura Manrique-Gómez, Tony Montes, and Rubén Manrique. 2024. Historical Ink: 19th century Latin American spanish corpus with LLM OCR correction.

Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020. Capturing evolution in word usage: Just add more clusters? In *Companion Proceedings of the Web Conference 2020*, WWW '20. ACM.

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

Stefano Montanelli and Francesco Periti. 2023. A survey on contextualised semantic shift detection. *Preprint*, arXiv:2304.01666.

Thomas Montgomery. 1966. On the development of spanish y from "et". *Romance Notes*, 8(1):137–142.

Francesco Periti and Nina Tahmasebi. 2024. A systematic comparison of contextualized word embeddings for lexical semantic change. *Preprint*, arXiv:2402.12011.

José A. Porto-Dapena. 1975. En torno a las entradas del "diccionario" de rufino josé cuervo. *Boletín del Instituto Caro y Cuervo*, 30(1).

Maxim Rachinskiy and Nikolay Arefyev. 2022. GlossReader at LSCDiscovery: Train to select a proper gloss in English – discover lexical semantic change in Spanish.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 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, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021. DWUG: A large resource of diachronic word usage graphs in four languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joan Scott. 1988. *Gender and the politics of history*. Columbia University Press, New York, NY.

Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. LSCDiscovery: A shared task on semantic change discovery and detection in spanish.

A Usage Examples per Sense

Infancia: The word has presented a semantic shift as shown in Figure C1

Sense 0 in New-"Adopción derechos del niño, protección de la infancia, tráfico de personas".

Sense 1 in Old-"Los pueblos, como los hombres, tienen su infancia, embrion todavía entre nosotros, período delicado y peligroso, en el que todo exceso é indiscreción trastorna el organismo é impide el desarrollo, si es que no lo destruye." "Escamilla, por

ejemplo, se casó desde la infancia con una matrona llamada Portería del Congreso de Escamilla: lleva dos apellidos, esta señora, no porque sea bigama, (pues no ha tenido mas que un solo marido) sino porque su papà es el señor Congreso, un viejo, mui necio."

Sentimiento: The word has presented a semantic shift as shown in Figure C1

Sense 0 in New- "65% de la personas que expresan un sentimiento personal de temor o esperanza". "Reforzar entre los europeos el sentimiento de pertenencia a una misma Comunidad".

Sense 1 in Old- "Será un gran artista de mucho sentimiento, posee una rica voz, si la educa, y tiene mucho aplomo en las tablas, es feo, pero simpático". "Una forma de expresión nueva, en la que brillaban un profundo sentimiento poético y una suerte de ingenuidad". "Que el divino arte de la música, lenguaje de la inteligencia y del sentimiento, ejerce sobre todos los hombres una influencia poderosa, que al mismo tiempo que atempera las pasiones, despierta las ideas de moralidad y de sociabilidad". "Republicano de ideas y de sentimiento, ha sabido armonizar sus opiniones políticas con sus creencias".

Sublime: The word presented polysemy in the past but is no longer in use as shown in Figure C1

Sense 0 in Old-"Hé aquí un epitafio sublime; la madre que busca al hijo bajo la sombra de los laureles, en la soledad de la muerte como dos almas inseparables, siempre unidas, siempre amantes".

Sense 1 in Old- "Bolívar, el del genio sublime que todo lo abarcó, que todo lo comprendió, y á quien debieron su existencia y su gloria, en menos de un cuarto de siglo, la mayor parte de las nacionalidades del Nuevo Mundo".

Sense 2 in Old- "á veces las leyes naturales- puede sí ejercer el sublime ministerio de aliviar (obra divina, según Hipócrates) y consolar á los que sufren."

Servidores: The word gained a new sense as shown in Figure C1

Sense 0 in Old-"Era allí donde se alojaba el Cacique, su familia y sus principales servidores". "a depositar- sus votos en favor de los buenos y leales servidores de la causa".

Sense 0 in New- "si la joven no está en un convento, rodearla de servidores que la acompañen por todas partes". "La Comisión y nosotros somos los servidores de los ciudadanos de nuestros Estados miembros".

Sense 1 in New- "la adquisición o el alquiler de ordenadores personales, servidores y microordenadores". "operación de los sistemas y de la red, y servidores para bases de datos, la Web, el FTP".

B SSD Examples

Some of the SSD results chosen were selected from Affinity Propagation algorithm clusterization, particularly those with only one sense such as "rey" and "usurero".

	Word	Sense	CD	PRT	gained/lost Sense
AP	Rey	0	0.005	1.005	
	Usurero	0	1.0	∞	**lost**
KMeans	Luces	0	0.012	1.012	
	Luces	1	0.012	1.013	
	Infancia	0	0.017	1.017	
	Infancia	**1**	1.0	∞	**gained**
	Sentimiento	**0**	1.0	∞	**gained**
	Sentimiento	1	0.003	1.003	
	Sublime	**0**	1.0	∞	**lost**
	Sublime	**1**	1.0	∞	**lost**
	Sublime	**2**	1.0	∞	**lost**
	Servidores	0	0.043	1.045	
	Servidores	**1**	1.0	∞	**gained**

Table B1: SSD for some of the 255 target words; the ones mentioned in the paper, and others added in the appendix DWUGs.

C DWUGs Examples

Figure C1: DWUG using the Latin American portion of the corpus fine-tuned model embeddings, the T-SNE dimensionality reduction algorithm, and the **KMeans** clustering algorithm (with the silhouette metric). All words are correctly clustered.

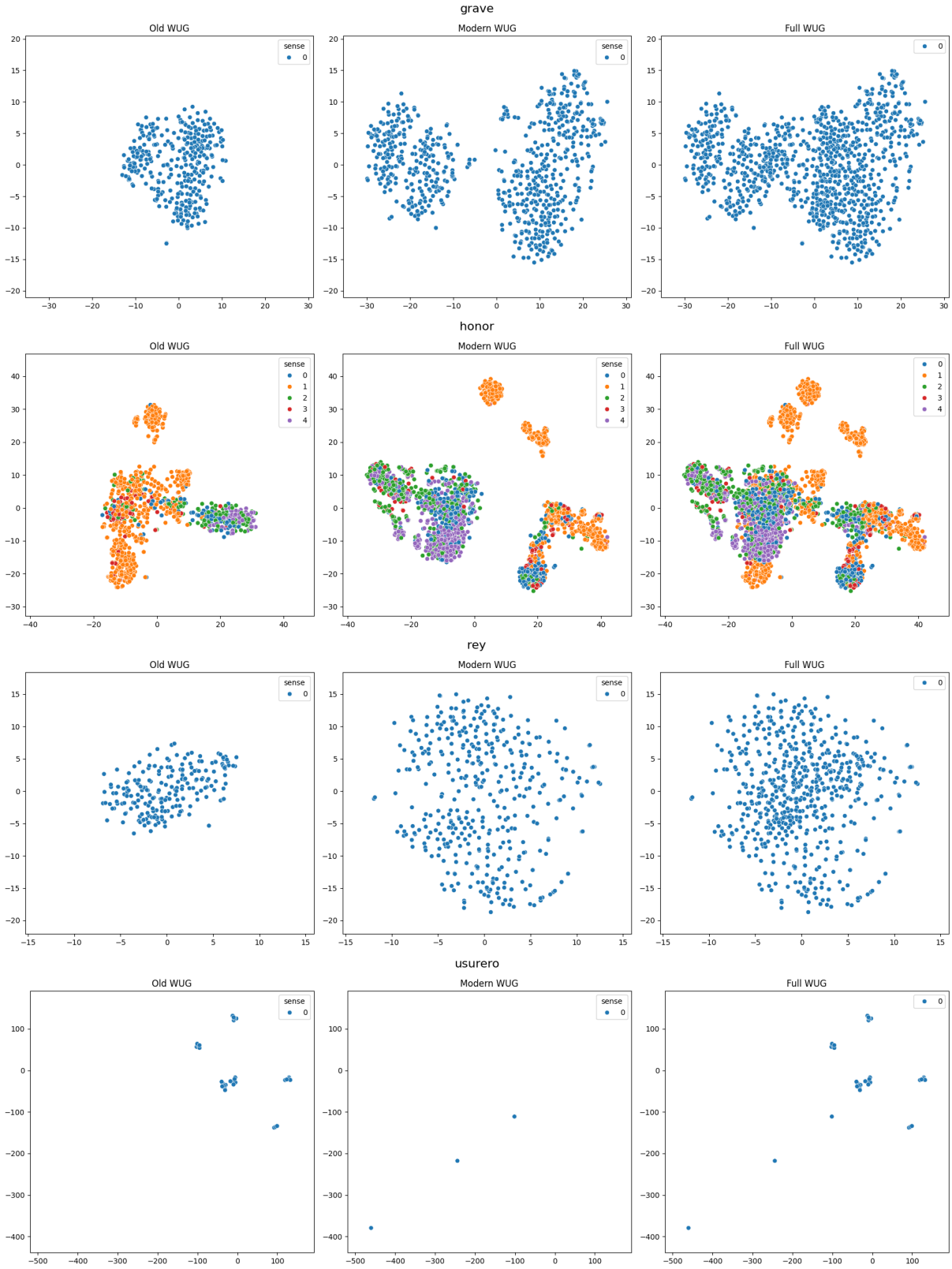

Figure C2: DWUG using the Latin American portion of the corpus fine-tuned model embeddings, the T-SNE dimensionality reduction algorithm, and the **Affinity Propagation** clustering algorithm. Words "grave" and "honor" are wrong clustered, and words "rey" and "usurero" are correctly clustered.