Where in Semantic Space Do Spanish Neologisms Emerge?

Bianca Delgado, Shira Wein
Amherst College
Amherst, MA, United States
{bdelgado27, swein}@amherst.edu

Abstract

English neologisms, or newly coined words, have previously been shown to emerge in sparser semantic neighborhoods (filling semantic gaps) and near other neologisms (in growing semantic areas). In this work, we investigate where in semantic space Spanish neologisms emerge, and whether this mirrors English neologism development. We find that Spanish neologisms, in comparison to non-neologisms, do indeed appear both nearer to other neologisms and further from non-neologisms. We additionally investigate the prevalence of loanwords from other languages through time in Spanish neologism production and manually assess the topics that appear as loanwords at four years: 1810, 1900, 1950, and 1990. Our findings show that on average, the Spanish neologisms in our dataset have fewer neighboring words in semantic space compared to non-neologisms and tend to cluster more tightly in the semantic space, indicating that patterns of neologism emergence span languages. This suggests that novel methods for neologism detection may be cross-lingually applicable, with these features serving as multilingual predictors of neologism emergence.

Keywords: vector semantics, Spanish, neologisms

1. Introduction

Neologisms are newly coined words that have been accepted in speech communities (Picone, 1996), and emerge across the world's languages. Ryskina et al. (2020) investigate how English neologisms relate to other words in semantic space, finding that English neologisms are more likely to (1) appear in sparser semantic neighborhoods (filling semantic gaps), and (2) emerge near other neologisms (in growing semantic areas). In this work, we investigate how typological diversity impacts how neologisms emerge, specifically examining whether these two findings by Ryskina et al. (2020) apply for Spanish neologisms. We perform this analysis through a temporal lens, performing this investigation on Spanish texts at various points in time and with multiple thresholds of "closeness" for semantic similarity (cosine similarities of at least 0.35, 0.45, and 0.55).

To address these research questions, we produce static embeddings of Spanish words appearing in the Google Ngram Viewer corpus (Michel et al., 2011), which contains frequencies of Spanish words used in books from 1500-2019. We classify these unigrams as neologisms or non-neologisms based on their proportion of usage before and after the year being analyzed. We then compare the embeddings of the Spanish words for each year via cosine similarity, counting how many words are at least as similar as the specified similarity lower bound. Then, we can use these counts to evaluate whether neologisms are more likely to emerge in (1) sparser areas, i.e. have fewer close semantic neighbors, and (2) growing semantic neighborhoods, i.e. have a higher proportion of neologisms as close semantic neighbors and tend to be grouped together in semantic space.

Next, we inspect clusters of Spanish neologisms to identify the topics that have grown in popularity over each of the various years, and assess the role that language contact has played in the creation of new words across time, by identifying the source language of the neologisms for each year.

We find that new Spanish words, like new English words, do indeed emerge in sparser and growing neighborhoods. We also find that language contact and globalization tend to impact the loanwords that appear in Spanish text over time. These findings indicate that neologisms tend to emerge in similar areas of semantic space across languages, given that these patterns appear to be consistent multilingually. This result opens up new avenues of multilingual neologism detection, which corresponds with one persistent challenge for large language models (LLMs) in the modeling of contemporary speech: unknown token handling.

2. Related Work

Neologisms reflect cultural, technological, and societal change. Consequently, it is important to be able to accurately detect and infer the meaning of neologisms from limited context in order to keep language models current and effective for real-world applications. In fact, Zheng et al. (2024) introduce NEO-Bench, a benchmark that serves to evaluate LLMs' robustness in handling neologisms, and conclude that LLMs are not yet fit to generalize on neologisms.

Proceedings of NEOLOGY AND LARGE LANGUAGE MODELS, pages 47–52
16 MAY 2026.

Prior work related to neologisms has largely focused on detecting their presence in corpora, in particular for English (Würschinger et al., 2016; McCrae, 2019; Zalmout et al., 2019). Kulkarni et al. (2018) propose leveraging the appearance of neologisms to help estimate when a document was written by tracking their appearance and frequency. In a non-English setting, prior work has investigated detecting neologisms in Persian (Megerdoomian and Hadjarian, 2010), Mandarin (Liu et al., 2013), French (Falk et al., 2014; Lejeune and Cartier, 2017), Russian (Lejeune and Cartier, 2017), and Japanese (Breen et al., 2018). Mizrahi et al. (2020) introduce a model which, rather than detecting or analyzing existing Hebrew neologisms, is designed to create new words with the goal of reducing reliance on loanwords.

In this work, we focus on characterizing the semantic qualities of Spanish neologisms, in particular for nouns. On the other hand, Rello and Basterrechea (2010) present the first system able to identify and conjugate Spanish verb neologisms, and Wein (2020) categorizes utterances in a Spanish language learner corpus as being neologisms, loanwords, or errors.

As discussed in Section 1, Ryskina et al. (2020) propose two hypotheses surrounding neologism emergence, in order to analyze neologisms through the lens of distributional semantics. Separate Word2Vec embeddings are trained on the Corpus of Historical American English (COHA; Davies, 2010) and the Corpus of Contemporary American English (COCA; Davies, 2008), and then aligned. Neologisms are identified as nouns that occur at least 20 times more frequently in the contemporary corpus, following Ryskina et al. (2020). Semantic density is measured by counting words that fall within certain cosine similarity thresholds ranging from 0.35 to 0.55, and frequency growth is calculated by averaging the change in frequency of a given word's neighbors over time. In concluding that both semantic sparsity and frequency growth serve as strong predictors, with frequency growth outperforming, this study offers valuable insight into neologism emergence for English words. These findings motivate our work on Spanish neologisms.

3. Methods

To test our hypotheses, we represent words using word embeddings and measure the semantic similarity between them. In doing so, we are able to define neighborhoods around each word via a similarity threshold in order to measure the density of the neighborhood as well as the presence of nearby neologisms.

3.1. Data

We utilize the unigram data from the 2020 Spanish-language Google Ngram Viewer corpus (Michel et al., 2011), which contains frequency counts of words in Spanish books published between 1500 and 2019. Each entry within this dataset contains a word and its frequency for each year.

Following Ryskina et al. (2020), before classifying neologisms, we filter the data to include only nouns using the part-of-speech tagger from the SpaCy package (Honnibal et al., 2020). Additionally, we filter out words beginning with capital letters as well as words with special characters or numbers. This preprocessing limits our dataset to just nouns, because they are an open-class part-of-speech, and helps reduce some of the noise from our large dataset by filtering out some named entities.

We identify neologisms at four cutoff years: 1810, 1900, 1950, and 1990. We select these four years given the amount of data available in each of those four intervals and the cultural shifts that occurred between those times. We then calculate the amount of times each word is used before and after each cutoff and their proportion of modern usage, which we define as the ratio of occurrences after the cutoff to occurrences before. We determine which words to label as neologisms based on their proportion of modern usage, ultimately labeling the 1,000 words with the highest proportion at each year as neologisms (we discuss examples of these neologisms in Section 4). We select the top 1,000, which follows Ryskina et al. (2020) and is affirmed by our qualitative assessment of the point at which the modern usage proportion drastically declines. The average ratios for these 1,000 highest-proportion items are as follows: 76,649 for 1810, 12,716 for 1900, 3,877 for 1950, and 564 for 1990. The decrease in these proportions over the years is an expected trend that likely reflects the amount of time each word had to accumulate usage after each cutoff year. For example, words that emerged around 1810 had over 200 years to become widely used, while words that emerged around 1990 had only a few decades until the end of the dataset in 2020.

3.2. Approach

In order to represent our words as vectors, we use static embeddings from fastText (Bojanowski et al., 2017),[1] as our dataset consists of isolated words and frequencies.

To address the first hypothesis, which is that neologisms tend to appear in less dense areas of the

[1] Specifically, we use the embeddings produced by the cc.es.300.bin Spanish-language model, which is pretrained on Common Crawl and Wikipedia data.

Language	Neologism	Non-neologism
Spanish	403	293429
English	198	95880
Portuguese	69	61794
Italian	61	52669
French	31	35193

Table 1: Source languages identified for neologisms and non-neologisms for words in the year 1990.

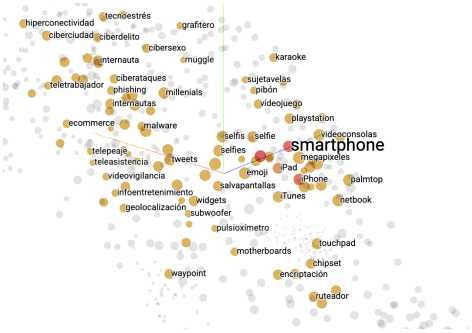

Figure 1: Technology Cluster: 1990 Cutoff. Words that are redder (darker) are closer in semantic space to smartphone.

semantic space, we calculate the cosine similarity between the each word and all other words in the dataset. Then, for each word, we count the number of similar words above a certain threshold. We choose to analyze several lower bound thresholds in order to explore varying degrees of similarity: 0.35, 0.45, and 0.55 following Ryskina et al. (2020); these thresholds are defined as the minimum cosine similarity that is required for a word to be considered "similar" to a given word, i.e. to be a part of that word's semantic neighborhood. This allows us to determine how many words are similar to the original word, and thereby ascertain the density of a given word's semantic space.

We operationalize the second hypothesis, which is that neologisms tend to arise in growing semantic neighborhoods, in terms of proximity of neologisms to each other. Accordingly, we perform the same calculation as for the first hypothesis, but now only focus on how many neologisms are near the word, indicating whether the area is growing.

Finally, to investigate the question of language contact's role in neologism creation, we perform language identification of the neologisms for each year. We leverage fastText's pretrained model for language identification, which supports 176 languages (Joulin et al., 2017).[2] We count the amount of words identified as belonging to each language by the model.

[2]Specifically, we use the lid.176.bin fastText model.

4. Results

First, we perform a qualitative analysis of the neologisms that emerge in the dataset at each year. We are able to notice clear trends in the types of words classified as neologisms. For the interval starting at 1810, we notice words like *fotografía* (photography), *vodka* (vodka), and *fútbol* (soccer/football) being labeled as neologisms. All of these words not only give insight into what terms are growing in popularity at the time, but are also examples of globalization's impact on language, as all of these terms are either loanwords or adapted from other languages—mostly from English, but with *vodka* being a Russian loanword. In the 1900-1950 data split, we see words like *telenovelas* (television soap opera), *ecologismo* (environmentalism), and *neurociencia* (neuroscience) reflecting a growing focus on scientific, environmental and media-related domains for the time period. For 1950-1990, we see *feminicidio* (femicide), *rockero* (rocker), *hippie* (hippie), *interculturalidad* (interculturality), and even *píxeles* (pixels), which reflects more modern developments. Finally, in the data split with cutoff year 1990 (which spans 1990 to 2020), we notice more contemporary concepts that are emblematic of 21st century advancements, such as *smartphone* (smartphone), *bitcoin* (bitcoin), and *ibuprofeno* (ibuprofen). Overall, this progression underscores the influence of globalization, technological innovation, and societal shifts on the Spanish language, revealing predictable patterns for neologism emergence. When visualizing the neologisms into cluster via T-SNE (Figure 1), we are able to observe similar clusters developing, including technologically- and scientifically-focused terms clustering together for each year.

When examining the amount of neologisms identified as belonging to each language for each cutoff year, we observe an increase in the proportion of neologisms that are non-Spanish words over time, with Portuguese and English appearing most frequently. In 1810, Portuguese is the most common non-Spanish source of neologisms, with English behind it. However, in 1900, English not only surpasses Portuguese, but continues to increase and ultimately makes up more than double the amount of Portuguese neologisms by 1990 (Table 1). This shift suggests a rise of English influence on Spanish vernacular.

Addressing the first hypothesis, our results (Table 2) show that on average, neologisms have fewer neighboring words in the semantic space compared to non-neologisms. Noticeably, the absolute counts are large, which is due both to the size of the dataset and to the use of a relatively low similarity threshold, such as 0.35, permit the inclusion of many loosely related words. For this

		Lower Bound		
Year	**Type**	**0.35**	**0.45**	**0.55**
1810	Neologism	497,530	73,111	10,147
1810	Non-neologism	693,953	569,252	324,618
1900	Neologism	473,241	104,741	17,723
1900	Non-neologism	693,985	569,210	324,608
1950	Neologism	461,496	139,012	28,781
1950	Non-neologism	694,001	569,165	324,593
1990	Neologism	488,120	233,502	102,150
1990	Non-neologism	693,965	569,039	324,495

Table 2: Average number of similar *words* within the threshold for neologisms and non-neologisms, at each lower bound. For example: for the year 1810, for all 1,000 neologisms the average number of similar words from the dataset in each neologism's neighborhood is 497,530 with a cosine similarity of $0.35 \leq x \leq 1$.

		Lower Bound		
Year	**Type**	**0.35**	**0.45**	**0.55**
1810	Neologism	37.4	9.25	2.59
1810	Non-neologism	3.44	0.309	0.045
1900	Neologism	34.0	8.53	2.53
1900	Non-neologism	7.19	1.18	0.190
1950	Neologism	32.7	8.60	2.55
1950	Non-neologism	15.8	2.87	0.336
1990	Neologism	57.4	21.6	6.00
1990	Non-neologism	53.6	13.3	1.87

Table 3: Average number of similar *neologisms* for neologisms and non-neologisms within each threshold. For example: for the year 1810, for all 1,000 neologisms the average number of similar neologisms in a given word's neighborhood is 37.376, for a cosine similarity of $0.35 \leq x \leq 1$.

reason, the values are best interpreted in comparison to one another rather than in isolation. The fact that neologisms have fewer neighboring words in the semantic space compared to non-neologisms indicates that neologisms are indeed more likely to emerge in sparser areas of the semantic space, suggesting that Spanish, like English, follows this supply-driven pattern. This is indicative of a cross-linguistic trend of conceptual gaps that exist in semantically sparse regions, which neologisms fill.

Conversely, when assessing the proximity of neologisms to each other (Table 3), the average similarity count is higher than that of non-neologisms across all years and thresholds. This suggests that new words tend to cluster more tightly in semantic space, in line with the second hypothesis, which outlines the demand-driven theory that neologisms emerge in areas of growing popularity. Our findings suggest that Spanish neologisms are also subject to such cultural trends and shifts, and thus that neologisms are not just filling gaps in the semantic space, but responding to increased demand in culturally relevant spaces.

These findings and the similarity of Spanish neologism emergence to English neologism emergence indicates that both semantic sparsity and growing popularity serves as a multilingual predictor of future neologism emergence.

5. Conclusion

In this work, we explore where in semantic space Spanish neologisms emerge in relation to other words. We specifically investigate whether Spanish neologisms are more likely to fill semantic gaps (thus appearing in sparser neighborhoods and having fewer close neighbors than non-neologisms do) and emerge in areas of growing popularity (thus being closer to other neologisms). As Ryskina et al. (2020) find for English, we find that Spanish neologisms do emerge in both sparser and growing semantic neighborhoods, suggesting that these phenomena carry across languages. Our qualitative analysis reveals that words related to technology and global politics regularly emerge as neologisms in our data. We additionally investigate the role of loanwords, finding that more English loanwords appear with increasing frequency over the centuries. Our findings motivate future work detecting multilingual neologisms given their relationships with other words and known neologisms. In particular, given that handling unknown tokens (such as neologisms) is a persistent challenge for LLMs, this work provides critical insight into how we may detect new words across languages, which would prove useful for enhancing performance of multilingual LLMs.

6. Limitations

We select the Michel et al. (2011) n-gram dataset because of its size, historical scope, and well-documented temporality metadata. The words appear as individual unigrams, and thus we are not able to leverage the context that the words appear in for our analysis (or use contextualized embeddings). A limitation of our approach is that semantic neighborhoods are computed in a modern embedding space. Because the dataset only contains isolated word usage per year and no text sequences, we cannot train embeddings that reflect the historical usage of each neologism. Future work using time-stamped corpora with contextual information and embeddings over time could address these issues, allowing analyses of historical semantic structure and the dynamics of emerging words, following the approach of Ryskina et al. (2020).

Further, our static embeddings do not allow us to determine whether semantically sparse regions existed prior to the emergence of neologisms or appear sparse because neologisms are newly introduced.

Additionally, we focus on Spanish nouns in particular, removing nouns with capital letters or special characters. While we filter out all nouns that begin with capital letters, some proper nouns remain in the dataset (such as “iPhone”).

Acknowledgments

We thank anonymous reviewers and members of the Amherst College NLP lab for their feedback. This work is supported by the Amherst College HPC, which is funded by NSF Award 2117377.

7. Bibliographical References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

James Breen, Timothy Baldwin, and Francis Bond. 2018. The company they keep: Extracting japanese neologisms using language patterns. In *Proceedings of the 9th Global Wordnet Conference*, page 163–171, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

Mark Davies. 2008. The corpus of contemporary american english (COCA).

Mark Davies. 2010. The corpus of historical american english (COHA).

Ingrid Falk, Delphine Bernhard, and Christophe Gérard. 2014. From non word to new word: Automatically identifying neologisms in French newspapers. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC‘14)*, pages 4337–4344, Reykjavik, Iceland. European Language Resources Association (ELRA).

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Vivek Kulkarni, Yingtao Tian, Parth Dandiwala, and Steve Skiena. 2018. Simple neologism based domain independent models to predict year of authorship. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 202–212, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Gaël Lejeune and Emmanuel Cartier. 2017. Character based pattern mining for neology detection. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 25–30, Copenhagen, Denmark. Association for Computational Linguistics.

Tsun-Jui Liu, Shu-Kai Hsieh, and Laurent Prevot. 2013. Observing features of PTT neologisms: A corpus-driven study with n-gram model. In *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013)*, pages 250–259, Kaohsiung, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

John Philip McCrae. 2019. Identification of adjective-noun neologisms using pretrained language models. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 135–141, Florence, Italy. Association for Computational Linguistics.

Karine Megerdoomian and Ali Hadjarian. 2010. Mining and classification of neologisms in Persian blogs. In *Proceedings of the NAACL HLT 2010 Second Workshop on Computational Approaches to Linguistic Creativity*, pages 6–13,

Los Angeles, California. Association for Computational Linguistics.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Moran Mizrahi, Stav Yardeni Seelig, and Dafna Shahaf. 2020. Coming to Terms: Automatic Formation of Neologisms in Hebrew. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4918–4929, Online. Association for Computational Linguistics.

Michael Picone. 1996. *Anglicisms, Neologisms, and Dynamic French.* John Benjamins B.V.

Luz Rello and Eduardo Basterrechea. 2010. Automatic conjugation and identification of regular and irregular verb neologisms in Spanish. In *Proceedings of the NAACL HLT 2010 Second Workshop on Computational Approaches to Linguistic Creativity*, pages 1–5, Los Angeles, California. Association for Computational Linguistics.

Maria Ryskina, Ella Rabinovich, Taylor Berg-Kirkpatrick, David Mortensen, and Yulia Tsvetkov. 2020. Where new words are born: Distributional semantic analysis of neologisms and their semantic neighborhoods. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 367–376, New York, New York. Association for Computational Linguistics.

Shira Wein. 2020. Classification and analysis of neologisms produced by learners of spanish: Effects of proficiency and task. In *Proceedings of the Fourth Widening Natural Language Processing Workshop*, page 88–91, Seattle, USA. Association for Computational Linguistics.

Quirin Würschinger, Mohammad Fazleh Elahi, Desislava Zhekova, and Hans-Jörg Schmid. 2016. Using the web and social media as corpora for monitoring the spread of neologisms. the case of 'rapefugee', 'rapeugee', and 'rapugee'. In *Proceedings of the 10th Web as Corpus Workshop*, pages 35–43, Berlin. Association for Computational Linguistics.

Nasser Zalmout, Kapil Thadani, and Aasish Pappu. 2019. Unsupervised neologism normalization using embedding space mapping. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 425–430, Hong Kong, China. Association for Computational Linguistics.

Jonathan Zheng, Alan Ritter, and Wei Xu. 2024. Neo-bench: Evaluating robustness of large language models with neologisms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 13885–13906, Bangkok, Thailand. Association for Computational Linguistics.

8. Language Resource References