Is Change the Only Constant? An Inquiry Into Diachronic Semantic Shifts in Italian and Spanish

Matteo Melis[1], Anastasiia Salova[1] and Roberto Zamparelli[1]

[1]Centre for Mind/Brain Sciences, University of Trento, Rovereto, Italy

Abstract

An increasingly prevalent approach to studying the gradual change of word meanings over time involves using distributional semantics, which is based on neighboring words. This study combines methods from Hamilton et al. (2016) [1] and Uban et al. (2019) [2] to analyze deceptive cognate pairs in historical and contemporary Italian and Spanish corpora. By employing fastText word embeddings and various similarity measures, it aims to investigate the change of word meanings and test two laws of regularity proposed by Hamilton et al. (2016) [1], along with a new hypothesized regularity in language change regarding analogy. The findings show a coherent evolution of deceptive cognates across the two languages. However, no meaningful correlation is found regarding the two aforementioned laws. Nevertheless, the results of the hypothesized regularity offer valuable insight into how the context of word usage shifts along with the word.

Keywords

Diachronic semantics, semantic shifts, distributional semantics, similarity measures, deceptive cognates

1. Introduction

1.1. Background

In recent years, there has been a growing interest in studying the shift of word meanings over time, with word embeddings emerging as a valuable tool for this purpose. Hamilton et al. (2016) [1] conducted research focusing on diachronic word embeddings to uncover specific statistical laws associated with semantic change. They examined the law of conformity, which suggests that words tend to change inversely to their frequency. Additionally, they explored the law of innovation, which proposes that words with greater polysemy tend to undergo semantic changes more frequently, regardless of how often they are used. The findings confirmed the hypothesized statistical laws. The study primarily focused on English, aligning word embeddings from different time periods and measuring semantic similarity using cosine similarity.

Dubossarsky et al. (2017) [3] contested the validity of the reported laws of semantic change based on word representation models. Replicating previous studies, they found that the law of conformity and the law of innovation did not withstand the more rigorous standard. The negative correlation between word frequency and meaning change was weaker than previously claimed, and the positive correlation between polysemy and meaning change was largely dependent on word frequency without independent contribution.

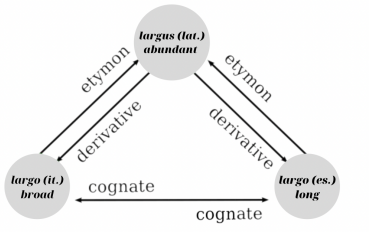

Figure 1: A pair of deceptive cognates in Italian-Spanish, with a shared etymon.

Similarly, to Hamilton et al. (2016) [1], Uban et al. (2019) [2] investigated semantic divergence across languages by examining deceptive cognate sets, which are words with a common origin in different languages. They focused on analyzing modern embeddings to quantify semantic shifts originating from shared etymology, identify false friends (deceptive cognates) in the cognate sets, and measure their score of falseness, namely the dissimilarity between the cognates. The study primarily concentrated on six Romance languages. The authors introduced methodologies such as aligning word embeddings across languages, measuring semantic similarity and divergence between cognate sets, and quantifying the magnitude of semantic changes. Their findings contradict those of Hamilton et al. (2016) [1], who found a negative correlation between frequency and meaning shift. However, they align with their findings regarding the law of innovation.

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

matteo.melis@studenti.unitn.it (M. Melis);
anastasiia.salova@studenti.unitn.it (A. Salova);
roberto.zamparelli@unitn.it (R. Zamparelli)
https://github.com/matteo-mls (M. Melis);
https://github.com/anastasiia-slv (A. Salova)

1.2. Objectives

The primary focus of this study is to investigate the presence of statistical laws governing semantic shifts within the Romance language group, specifically Italian and Spanish. The research questions revolve around exploring the laws of conformity and innovation. It is hypothesized that more frequent words are less likely to undergo semantic shifts, while more polysemous words are more prone to such changes. Additionally, the study introduces a new follow-up analysis on analogy, suggesting that over time periods the meaning of a word which is semantically related to a target (in terms of context-based nearest neighbors), tends to shift in the Euclidean space coherently with the target word.

The study uses distributional semantics as a methodology to explore language change. A crucial part of this research involves analyzing deceptive cognate pairs, which have a similar or the same form in different languages but diverged in meaning over time, unlike true cognates that retain the same meaning. For instance, Figure 1 illustrates how *largo* (broad) in Italian and *largo* (long) in Spanish have diverged in meaning through a semantic shift, despite both words originating from the shared Latin etymon *largo* (abundant). We believe this allows for a robust comparison of semantic changes, especially in related languages, providing illustrative examples and easily interpretable results. Our primary focus is on systematic semantic change that originates from the shared etymon and continues, while also controlling for the random appearance of lexical units in a language. Moreover, this approach would enable cross-language analysis in prospective studies.

Our study aims to expand the current understanding of language change by incorporating cognate comparisons across languages and examining individual changes within specific time periods. To enhance the robustness of our analyses, we introduce various similarity measures.

2. Corpora

2.1. Corpora Selection Criteria

The study uses two different time periods of language usage in its corpora: the 19th and 20th centuries (until 1969) for historical data, and the 21st century for modern data.

To address the size difference between the two datasets, we reduced the modern data to match the historical data's size. This was achieved by counting the number of required tokens and removing the tokens exceeding this number. This allowed for two different training sets for the modern data, enabling comparisons and allowing us to draw conclusions about the minimum amount of data needed for these analyses.

2.1.1. Italian

Four corpora were collected online for this study: Histcorp [4], ChroniclItaly v3.0 [5], Unità corpus [6], and PAISÀ corpus [7]. The first three corpora were merged to form the historical dataset, covering the years 1805-1969, with a total of 545,068,401 tokens. The PAISÀ corpus represented the modern data, containing 1,089,014,748 tokens, while the reduced modern version consisted of 545,106,781 tokens.

2.1.2. Spanish

Similarly, four corpora were collected online for Spanish: Conha19 [8], Impact-es (BVC section) [9], Corpus of Political Speeches [10], and The Large Spanish Corpus [11]. The historical data consists of a merged collection of the first three corpora, covering the period from 1830 to 1969 and containing 204,904,549 tokens. The modern data representation utilizes 'The Large Spanish Corpus' (Wikipedia section), containing 975,251,278 tokens from 2019. Additionally, a reduced version of The Large Spanish Corpus was created, containing 206,900,109 tokens.

2.2. Pre-processing Techniques

The pre-processing for both languages followed the same steps. After collecting the text files for each corpus, we used the NLTK library [12] for tokenization and stopword removal. The files were cleaned by removing URLs, numbers, non-letters, multiple empty spaces, and set to lowercase. For Spanish, diacritic marks were replaced using unicodedata. The spaCy library [13], with its reported accuracy of 0.96 for Spanish and 0.97 for Italian, was employed for lemmatization, and the files were merged into a representative single file for each historical period and language.

2.3. Cognate Dataset

We used an existing resource: an automatically generated multilingual lexicon of false friends [14]. Following the logic that cognate pairs are considered false friends if a word in the second language is closer in meaning to the original word in the shared semantic space than its cognate in that language, a falseness score is provided.

For instance, given the cognate pair (*imbarazzata, embarazada*), where *imbarazzata* (embarassed) is a word in Italian and *embarazada* (pregnant) is a word in Spanish, if there is a word x in Spanish such that for any word w in Spanish the distance (imbarazzata, x) is less than the distance (imbarazzata, w), then the pair is considered

a deceptive cognates pair. Since the Spanish word *avergonzada* (embarassed) exists, the pair (*imbarazzata, embarazada*) constitutes a set of false friends, and their arithmetic difference is the score of falseness, which ranges from 0 to 1. It is lower for false friends that are closer in meaning and higher for more distant false friends.

Given this, we decided to extract the 156 deceptive cognate pairs with a falseness value higher than 0.25. This step was taken to ensure the accuracy of the dataset and account for its limitations in the unsupervised data collection method.

3. Methodology

Methodologically, the study can be divided into the following steps[1]:

3.1. FastText Word Embeddings Retrieval

We trained six fastText models [15] in an unsupervised regime using the six corpora that we obtained and prepared. For each model, we employed the skip-gram algorithm, set the vector dimension to 100, and trained for 5 epochs. These parameters are considered default, and as indicated by Mikolov et al. (2013) [16], the algorithm has been found to work well with small datasets. This resulted in three models for each language, trained on historical data, modern data, and modern reduced data, respectively. This produced a total of 6 different vector spaces.

3.2. Embeddings Overview with RSA

In order to obtain a comprehensive overview of the vector spaces and as the initial step of our analysis, we computed Representational Similarity Analysis (RSA) between dissimilarity matrices of 156 deceptive cognate words from the dataset by Uban and Dinu (2020) [14]. These matrices were created by extracting vectors for specific cognates from the common vector spaces obtained in the previous step. The aim was to assess general similarity patterns within the word embeddings. Based on the results thus obtained we chose to exclusively use the model trained on the full modern data and discard the one trained on the reduced modern data to ensure higher-quality word embeddings in later steps. Detailed results of this analysis will be discussed later.

[1]All the code can be found at https://github.com/matteo-mls/diachronic-semantic-shift.

3.3. K-Nearest Neighbors Retrieval Using a Similarity Measure

To obtain more qualitative data, the fastText library [15] was used to retrieve embeddings closest to the target cognate in Euclidean space. The retrieval process utilized the K-Nearest Neighbors (K-NN) function, where the cosine similarity measure was employed to compare two vectors. The number of nearest neighbors to retrieve (k) was predetermined and set to 5, 10, 20, and 50 for comparative analysis purposes.

3.4. Semantic Shift Calculation within Each Language

After retrieving the nearest neighbors for cognates, we calculated the overlap between the sets of nearest neighbors in each language. This overlap was measured using the Jaccard similarity coefficient, which determines the similarity between two sets. The semantic shift was then computed as the difference in overlap between the sets of nearest neighbors over time. Finally, by using the Pearson correlation measure to assess the shifts between the two languages, Italian and Spanish, we were able to draw conclusions.

3.5. Word Frequency and Semantic Divergence Analysis

For the frequency analysis, we followed the following steps:

1. We applied Procrustes alignment [17] to the two vector spaces (historical to modern for each language) to ensure that similar vectors represented the same concepts across different embedding spaces. This alignment was necessary as the embeddings were trained on different corpora in different languages.
2. We calculated the cosine similarity for the cognates in different time periods.
3. We counted the occurrences of each cognate word from both the historical and modern corpora in Italian and Spanish.
4. We normalized the occurrences of cognate words by dividing each value by the maximum value, which is the sum of all values. This normalization resulted in a total of 1, effectively replacing the actual frequency values.

Using the NumPy library [18], we computed the correlation coefficient and linear regression coefficients of the frequency and semantic shift across time. In this analysis, we incorporated polysemy covariance, considering the correlation between polysemy and frequency.

3.6. Word Polysemy and Semantic Divergence Analysis

After conducting the frequency and semantic divergence analysis, we proceeded to measure the polysemy of words. To accomplish this, we utilized the WordNet library [19], specifically leveraging the functionality provided by the "nltk.corpus.wordnet" module. Polysemy was quantified as the number of synsets associated with a word in WordNet, following the methodology described by Uban et al. (2019) [2].

Subsequently, we investigated the correlation between the cosine similarity over time, which indicates the degree of semantic shifting, and the number of meanings a word can have according to WordNet. In this analysis, we took into account the co-variance with frequency, similarly to our previous approach.

3.7. Word Analogy and Semantic Divergence Analysis

In addition to the previous analyses, we further examined how the cosine similarity changes over time for the K-Nearest Neighbors (K-NN) that exhibit overlap between the two different time periods. For each cognate word, we employed a K-NN approach with varying values of K (5, 10, 20, 50). We examined the overlapping nearest neighbors (NN) in both the historical and modern lists of NN. For each overlapping NN, we calculated the cosine similarity and measured the difference in the shift, determining whether the NN moved closer to or further from the target cognate word.

By calculating the ratio of positive (closer) or negative (further) shifts, we could now assess the coherence (the consistency of neighbors' movement relative to the target cognate) of the shift in the K-NN of that specific target cognate word. To identify significant coherent shifts, we set a threshold (>0.75). This threshold was chosen to be substantially higher than chance, ensuring a rigorous approach. If this ratio is crossed, it implies a major coherent shift in the K-NN of the target cognate word.

In carrying out this analysis for all the cognates in the list we removed those that had 0 or 1 NN, since they do not provide informative results.

4. Results

4.1. Representational Similarity Analysis

As shown in the Appendix A (Figure 4), the reduced Italian modern embedding space exhibits a lower correlation compared to the complete Italian modern embedding space, with a difference of 0.0322 (a). This suggests that the improved embedding obtained by using more data in unsupervised word embedding contributes to this outcome. Furthermore, when comparing the reduced historical Spanish embedding space with the modern embedding space, a difference of 0.0956 is observed (b). Therefore, while the results for Italian remain consistent between the full and reduced spaces, reducing the Spanish modern space to match the historical space produces different outcomes compared to using the full modern space. Given the choice between data quality and balance, we have opted for better data quality by discarding the models trained with reduced datasets.

4.2. Calculation of Semantic Shifts

4.2.1. Within-Language Comparison: K-NN with Jaccard Distance

In reference to the selection of K Nearest Neighbors (KNN) values at 5, 10, 20, and 50, the obtained results are presented in the tables provided in the Appendices B and C (Tables 3 to 10). These tables display the average number of overlapping nearest neighbors in the cognate list, the ratio of overlapping nearest neighbors considering the extracted KNN, and the Jaccard distance. Please refer to the Appendix for a detailed representation of these values.

4.2.2. Inter-Language Comparison: K-NN with Jaccard Distance

The values in Appendix D (Tables 11 and 12) represent dissimilarity scores, specifically semantic shifts, calculated using the Jaccard distance (1-Jaccard index). The Pearson correlation score of 0.999 indicates a strong correlation between the shifts for Italian and Spanish as the particular K value increases. Overall, the scores show compatible semantic shifts. However, in this analysis, we can only infer the magnitude of the shifts and not the patterns, which will be explored in later analyses.

4.3. Law of Conformity

Figure 2 (upper) showcases the correlation results for the law of conformity in both Italian and Spanish. The obtained correlation coefficients demonstrate a moderate positive correlation, with a coefficient of 0.408 for Italian and 0.470 for Spanish. However, when accounting for the influence of polysemy through partial correlation analysis, the coefficients decrease to 0.261 for Italian and 0.3 for Spanish. These values are generally considered weak. While these findings provide only weak evidence for the law of conformity, they are at least consistent in their trend with the results reported by Hamilton et al. (2016) [1].

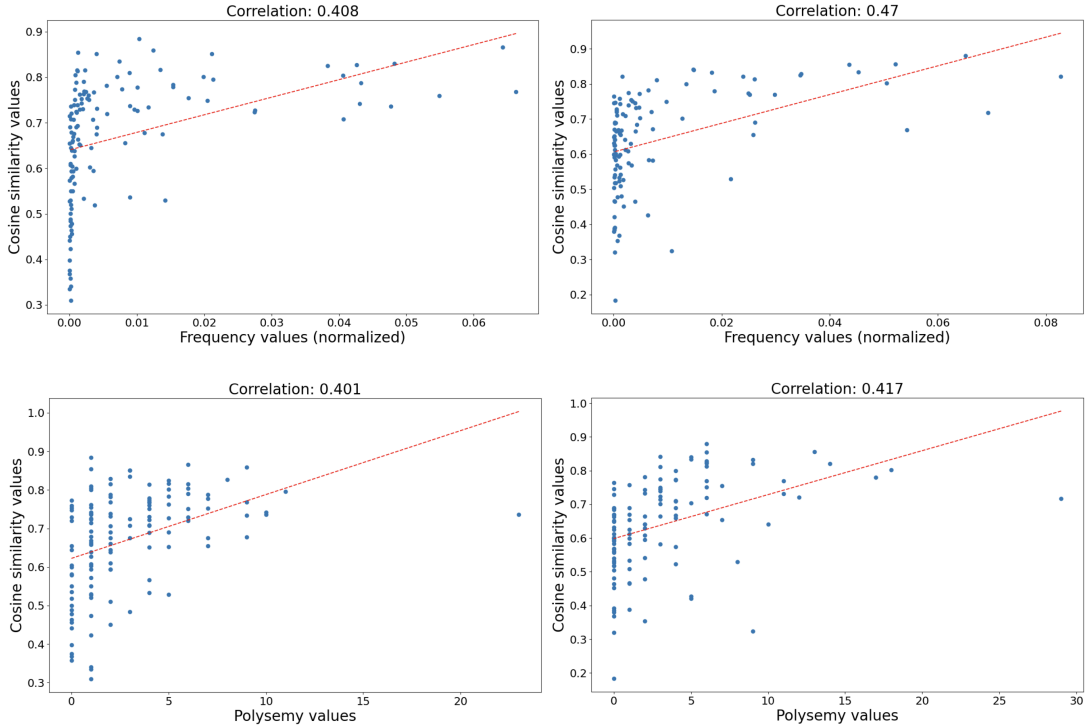

Figure 2: Law of conformity (upper) and law of innovation (lower) visualized for Italian (left) and Spanish (right).

4.4. Law of Innovation

Conversely, in our study the results for the law of innovation (more polysemy = greater shift), depicted in Figure 2 (lower), differ from those reported by Hamilton et al. (2016) [1] and Uban et al. (2019) [2]. While we observed a moderate positive trend, similar to that of the law of conformity, with correlation scores of 0.401 for Italian and 0.417 for Spanish, the partial correlation, which accounts for the frequency compound, reveals weaker values of 0.249 for Italian and 0.188 for Spanish. These findings suggest that the data does not provide strong support for the existence of the law of innovation in Romance languages. However, due to the weak partial correlations observed, it is challenging to draw definitive conclusions.

4.5. Law of Analogy

One trend that emerges from our study is that semantically related words (as indicated by contextual nearest neighbors) tend to shift coherently closer or farther to the target word. Table 1 and Table 2 provide supporting evidence for this observation: as the number of nearest neighbors (K-NNs) increases, the ratio of coherent shifts tends to decrease. This aligns with the intuition that with more K-NNs, the distances between the neighbors and their target cognate increase, leading to less consistent shifts. To provide a visual representation, Figure 3 displays an example visualization for a single cognate pair.

Table 1
Analogy analysis for Italian

K-NN	N° of Cognates	Coherent shift	%
5	53	36	67.92
10	83	51	61.45
20	104	52	50
50	121	64	52.89

Table 2
Analogy analysis for Spanish

K-NN	N° of Cognates	Coherent shift	%
5	48	35	72.92
10	67	46	68.66
20	88	59	67.04
50	102	68	63.72

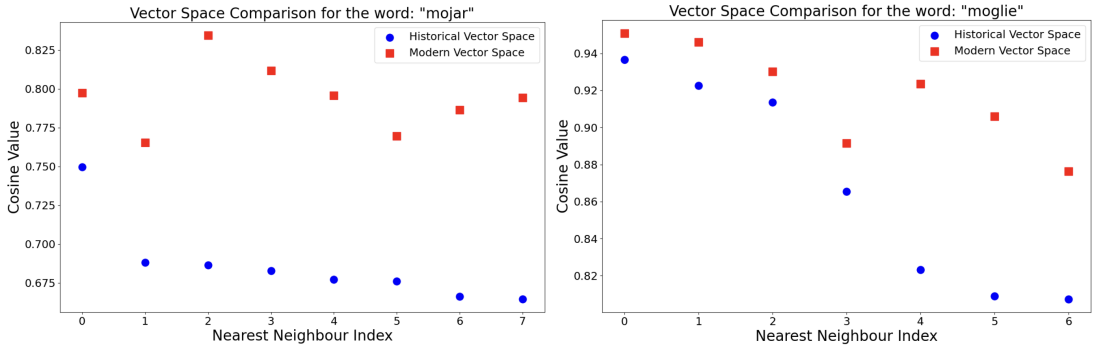

Figure 3: An example of the analysis of the law of analogy visualized for Italian (right) and Spanish (left) using the cognate pair ”mojar”/”moglie”.

5. Discussion

The hypothesized regularity regarding analogy, a follow-up analysis in this study, has provided intriguing insights into semantic shifts. However, it is important to note that further research into this topic is necessary to validate and expand upon these initial findings.

On the other hand, the analyses conducted in this study do not yield definitive results supporting the statistical laws of semantic shifts. Firstly, the RSA evaluation of the embedding spaces revealed that the scarcity of data significantly impacted the quality of the embeddings. Furthermore, while the law of conformity agrees with previous literature in a general trend, such as Hamilton et al. (2016) [1], our study identified a contrasting trend for the law of innovation. This discrepancy in findings may be attributed to the limitation of our study, namely the scarcity of data resulting from the use of relatively short time periods.

An additional factor is the relatively short temporal distance between the historic (as recent as 1969) and the modern corpora. Increasing this span is likely to lead to greater shifts, but also to greater data sparsity. Last but not least, the alignment technique employed for matching the embedding spaces could have contributed to the divergent outcomes in the analysis of the law of conformity and the law of innovation.

It is noteworthy that both the laws of conformity and the law of innovation conform to the findings of Dubossarsky et al, (2017) [3]. Their study revealed that the suggested positive correlation between meaning change and polysemy was primarily influenced by word frequency, and the correlation between word frequency and meaning change is indeed weaker. Here, after conducting partial correlation analysis, a weak correlation was observed. Furthermore, we noticed a high compatibility between frequency and polysemy, indicating an inherent dependence, despite our efforts to disentangle them using partial correlation.

Utilizing the fastText model, known for its improved performance on non-English languages, and preprocessing freely available data, the results still highlight poor quality embeddings. This underscores the need for ongoing research and development of word embedding models, alongside the creation of larger, well-curated diachronic corpora. Improving data quality and quantity can enhance the accuracy and reliability of future studies in the field.

It is important to note that due to the limitations of the embeddings used in this study, the shifts observed in the inter-language Jaccard distance analysis are relatively small and close to each other. This leads to an extremely high correlation coefficient between the languages being analyzed, which should be interpreted with caution.

In addition to the aforementioned directions, other potential areas of research include expanding further in time and broader in the scope of languages. For instance, this could involve going beyond the Romance or even the Indo-European language family to conduct a more comprehensive investigation into language change.

Acknowledgements

We would like to express our gratitude to Dr. Raffaella Bernardi for her support and feedback throughout this project, which has been helpful in shaping our research. We also appreciate her encouragement regarding the conference submission.

We also extend our gratitude to Dr. Lorella Viola for her generous assistance in providing a portion of the corpus used in our analysis.

References

[1] W. L. Hamilton, J. Leskovec, D. Jurafsky, Diachronic word embeddings reveal statistical laws of semantic change, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1489–1501. URL: https://aclanthology.org/P16-1141. doi:10.18653/v1/P16-1141.

[2] A. S. Uban, A. M. Ciobanu, L. P. Dinu, Studying laws of semantic divergence across languages using cognate sets, in: Proceedings of the Workshop, 2019, pp. 161–166. doi:10.18653/v1/W19-4720.

[3] H. Dubossarsky, D. Weinshall, 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, Association for Computational Linguistics, Copenhagen, Denmark, 2017. URL: https://aclanthology.org/D17-1118. doi:10.18653/v1/D17-1118.

[4] E. Pettersson, B. Megyesi, The histcorp collection of historical corpora and resources, in: Digital Humanities in the Nordic Countries Conference, 2018. URL: https://api.semanticscholar.org/CorpusID:19243754.

[5] L. Viola, A. M. Fiscarelli, Chroniclitaly 3.0. a deep-learning, contextually enriched digital heritage collection of italian immigrant newspapers published in the usa 1898-1936, in: Proceedings of the Conference, 2021. doi:10.5281/zenodo.4596345.

[6] P. Basile, A. Caputo, T. Caselli, P. Cassotti, R. Varvara, A diachronic italian corpus based on "l'unità", in: CLiC-it 2020 Italian Conference on Computational Linguistics 2020, volume 2769, CEUR Workshop Proceedings (CEUR-WS.org), 2020.

[7] V. Lyding, E. Stemle, C. Borghetti, M. Brunello, S. Castagnoli, F. Dell'Orletta, H. Dittmann, A. Lenci, V. Pirrelli, PAISÀ corpus of italian web text, 2013. URL: http://hdl.handle.net/20.500.12124/3, eurac Research CLARIN Centre.

[8] U. Henny-Krahmer, Corpus de novelas hispanoamericanas del siglo xix (conha19) version 1.0.1, in: Proceedings of the Conference, 2021. doi:10.5281/zenodo.4781947.

[9] F. Sánchez-Martínez, I. Martínez-Sempere, X. Ivars-Ribes, R. C. Carrasco, An open diachronic corpus of historical spanish, Language Resources and Evaluation 47 (2013) 1327–1342.

[10] E. Álvarez-Mellado, A corpus of Spanish political speeches from 1937 to 2019, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 928–932. URL: https://aclanthology.org/2020.lrec-1.116.

[11] J. Cañete, Compilation of large spanish unannotated corpora, Zenodo, 2019.

[12] S. Bird, E. Klein, E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit, O'Reilly Media, Inc., 2009.

[13] M. Honnibal, I. Montani, spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, 2017.

[14] A. S. Uban, L. P. Dinu, Automatically building a multilingual lexicon of false friends with no supervision, in: International Conference on Language Resources and Evaluation, 2020. URL: https://api.semanticscholar.org/CorpusID:218973843.

[15] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146.

[16] T. Mikolov, K. Chen, G. S. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: International Conference on Learning Representations, 2013. URL: https://api.semanticscholar.org/CorpusID:5959482.

[17] J. Gower, Generalized procrustes analysis, Psychometrika 40 (1975) 33–51. URL: https://EconPapers.repec.org/RePEc:spr:psycho:v:40:y:1975:i:1:p:33-51.

[18] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. Fernández del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T. E. Oliphant, Array programming with NumPy, Nature 585 (2020) 357–362. doi:10.1038/s41586-020-2649-2.

[19] C. Fellbaum, WordNet: An Electronic Lexical Database, Bradford Books, 1998. URL: https://mitpress.mit.edu/9780262561167/.

Appendix

A. RSA Correlation of Italian and Spanish

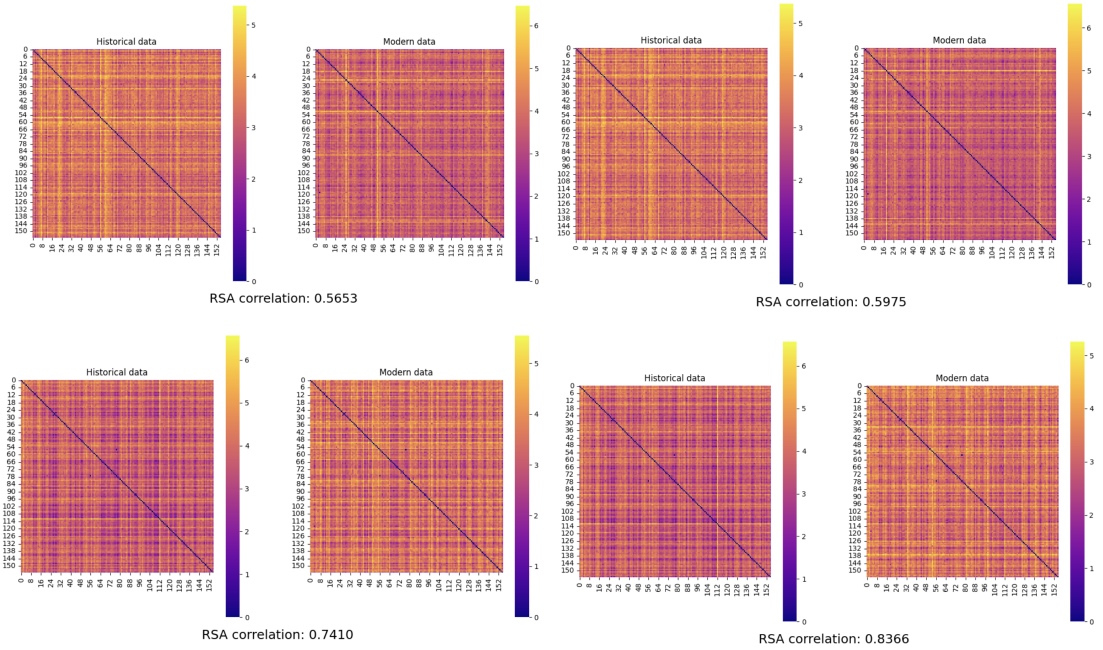

Figure 4: RSA correlation of Italian (upper) and Spanish (lower). Comparison between historical and modern (left) and historical and modern reduced (right)

B. Italian K-Nearest Neighbour

Table 3
Italian, K = 5 NN Overlap

Italian - K = 5	Word	N° of overlap
1	Fiaccola	4
2	Maggio	4
3	Ottimo	4
...	...	...
94	Verso	1
95	Voluta	1
96	Vendicare	1
Average	171/96	**1.7812**
Jaccard Distance	1 - J	**0.7833**

Table 4
Italian, K = 10 NN Overlap

Italian - K = 10	Word	N° of overlap
1	Maggio	9
2	Cardinale	7
3	Mantello	6
...	...	...
112	Servo	1
113	Via	1
114	Vigile	1
Average	294/114	**2.5789**
Jaccard Distance	1 - J	**0.8520**

Table 5
Italian, K = 20 NN Overlap

Italian - K = 20	Word	N° of overlap
1	Maggio	12
2	Cardinale	11
3	Decima	10
...	...	...
124	Venia	1
125	Tonno	1
126	Servo	1
Average	488/126	**3.8730**
Jaccard Distance	1 - J	**0.8928**

Table 6
Italian, K = 50 NN Overlap

Italian - K = 50	Word	N° of overlap
1	Impadronirsi	27
2	Cardinale	26
3	Giudicare	25
...	...	...
132	Oste	1
133	Sotto	1
134	Vado	1
Average	1005/134	**7.5000**
Jaccard Distance	1 - J	**0.9189**

C. Spanish K-Nearest Neighbour

Table 7
Spanish, K = 5 NN Overlap

Spanish - K = 5	Word	N° of overlap
1	Ardor	4
2	Diverso	4
3	Imaginario	4
...	...	...
82	Derrame	1
83	Verso	1
84	Vivir	1
Average	153/84	**1.8214**
Jaccard Distance	1 - J	**0.7773**

Table 8
Spanish, K = 10 NN Overlap

Spanish - K = 10	Word	N° of overlap
1	Cometer	6
2	Importar	6
3	Muerto	6
...	...	...
101	Derrame	1
102	Verso	1
103	Decir	1
Average	261/103	**2.5340**
Jaccard Distance	1 - J	**0.8549**

Table 9
Spanish, K = 20 NN Overlap

Spanish - K = 20	Word	N° of overlap
1	Cometer	13
2	Prender	11
3	Importar	10
...	...	...
114	Ensear	1
115	Tata	1
116	Tenia	1
Average	448/116	**3.8620**
Jaccard Distance	1 - J	**0.8931**

Table 10
Spanish, K = 50 NN Overlap

Spanish - K = 50	Word	N° of overlap
1	Cometer	25
2	Importar	20
3	Jurar	19
...	...	...
124	Patrón	1
125	Radio	1
126	Tenia	1
Average	920/126	**7.3016**
Jaccard Distance	1 - J	**0.9212**

D. Cosine Similarity

Table 11
Italian, Cosine Similarity

ITALIAN	Word	N° of overlap
1	Moglie	0.8845485
2	Ancora	0.8659243
3	Finire	0.8588681
...	...	...
146	Venia	0.3331086
147	Così	0.31215054
148	Caudale	0.30994532
8 cognates not found	**Average**	**0.6655**

Table 12
Spanish, Cosine Similarity

SPANISH	Word	N° of overlap
1	Querer	0.88015264
2	Decir	0.8567517
3	Pueblo	0.8563638
...	...	...
124	Radio	0.3236405
125	Das	0.3200544
126	Craso	0.18371347
30 cognates not found	**Average**	**0.6470**