Is Change the Only Constant? An Inquiry Into Diachronic Semantic Shifts in Romance Languages

Matteo Melis and Anastasiia Salova

CIMeC, University of Trento

Semantic shift is a multifaceted phenomenon encompassing the gradual evolution of word meanings over time. While significant computational research has focused on lexical semantic change, a novel approach to investigating meaning shifts involves using distributional semantics to quantify the semantic change of a word based on its neighboring words. Moreover, it is crucial to recognize that languages do not evolve in isolation, prompting exploration of semantic shift patterns across related languages to yield valuable insights. This study employs fasttext word embeddings to analyze historical and contemporary corpora from both Italian and Spanish languages, aiming to examine semantic shifts in language pairs through similarity measures. The study further seeks to confirm or challenge whether these two Romance languages adhere to statistical laws governing meaning change, such as the law of conformity, the law of innovation, and the law of analogy.

diachronic semantics | distributional semantics | semantic shift

Introduction

Background. In recent years, there has been a growing interest in studying the evolution of word meanings over time, with word embeddings emerging as a valuable tool for this purpose. Hamilton et al. (2016) [5] 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 with similar semantic characteristics change in a coordinated manner over time, aligning with prevailing usage patterns. Additionally, they explored the law of innovation, which proposes that certain words undergo unique or idiosyncratic semantic changes that deviate from dominant usage patterns. The study primarily focused on English, aligning word embeddings from different time periods and measuring semantic similarity using cosine similarity.

Dubossarsky et al. (2017) [4] contested the validity of reported laws of semantic change based on word representation models and emphasized the need for a stricter standard of proof. 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.

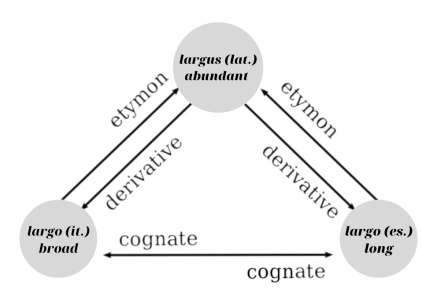


Fig. 1. A pair of false cognates in Italian-Spanish, with a shared etymon.

Similarly to Hamilton et al. (2016) [5], Uban et al. (2019) [13] investigated semantic divergence across languages by examining 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 and measure the score of falseness. 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), who found a negative correlation between frequency and meaning shift. However, their results align with the original finding regarding the law of innovation.

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, innovation, and analogy in relation to semantic shifts. 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 hypothesis concerning analogy, suggesting that the meaning of a word may shift towards other words that share similar form or meaning.

The study employs distributional semantics as a methodology to address fundamental inquiries regarding language evolution. A key aspect of this research involves the analysis of deceptive cognate pairs.

By building upon previous research, our study aims to expand the current understanding of language evolution by

incorporating cognate comparisons across languages and examining individual changes within well-defined time periods. To enhance the robustness of our analyses, we introduce a variety of similarity measures.

Corpora

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 account for the size difference between the two data sets, the modern data was reduced to a similar amount as the historical data by counting the number of tokens and then removing the "excessive" tokens. This allowed for two different training models for the modern data, enabling comparisons and conclusions about necessary data quantity.

Italian. Four corpora were collected online: Histcorp [11], ChroniclItaly v3.0 [15], Unità corpus [1], and PAISÀ corpus [7]. The first three were merged to represent historical data, and the PAISÀ corpus represented modern data.

Table 1. Corpora Used for Italian

Type	Corpora	Years	Tokens
Historical	Histcorp ChroniclItaly v3.0 Unità corpus	1805-1969	545,068,401
Modern	PAISÀ corpus	2010	1,089,014,748
Modern reduced	PAISÀ corpus	2010	545,106,781

Spanish. Four corpora were collected online: Conha19 [6], Impact-es (BVC section) [12], Corpus of Political Speeches [8], and The Large Spanish Corpus [2]. Similarly to Italian, we merged the first three to represent the historical data, while we used 'The Large Spanish Corpus' (Wikipedia section) to represent the modern data.

Table 2. Corpora Used for Spanish

Type	Corpora	Years	Tokens
Historical	Conha19 Impact-es (BVC section) Corpus of Political Speeches	1830-1969	204,904,549
Modern	The Large Spanish Corpus	2019	975,251,278
Modern reduced	The Large Spanish Corpus	2019	206,900,109

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 for tokenization and stop-word 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 was used for lemmatization, and the files were merged into a representative single file for each historical period and language.

Cognate Dataset. We used an existing resource: an automatically generated lexicon of false friends described in Uban and Dinu (2020) [14]. By setting a threshold of 0.25 in falseness, we extracted 156 cognate pairs from the provided CSV file to ensure the accuracy of our dataset. Although we acknowledge the possibility of true cognates being included due to limitations in the data collection method, we believe the chosen threshold reduces this risk.

Methodology

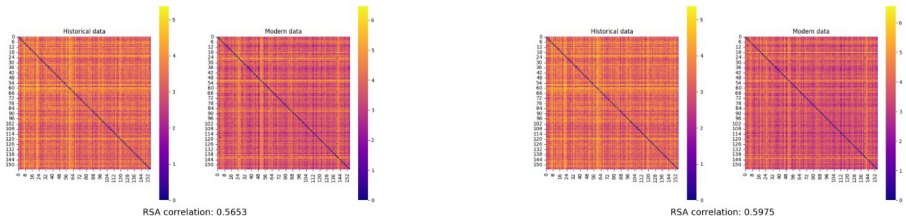
Methodologically, the study can be divided into the following steps[1]:

FastText Word Embeddings Retrieval. We trained six Fasttext models 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) [9], 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.

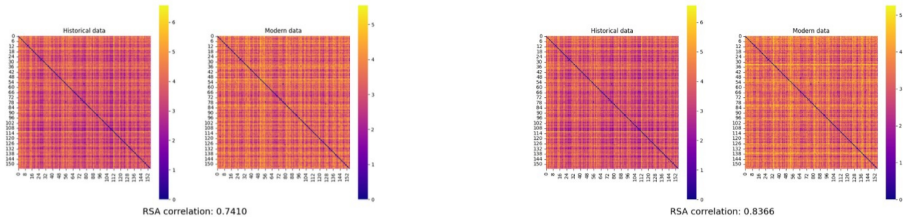
Embeddings Overview with RSA. In order to obtain a comprehensive overview of the vector spaces and as the initial step of our analysis, we opted to compute Representational Similarity Analysis (RSA) between dissimilarity matrices of the cognate words extracted from the trained models. The intention was to evaluate the general similarity patterns within the word embeddings. It is important to note that we made the decision to exclusively utilize the model trained on the full modern data and discard the one trained on the reduced modern data for subsequent analyses. This choice was made to ensure a higher quality of the word embeddings. Detailed results of this analysis will be discussed in the subsequent results section.

K-Nearest Neighbors Retrieval Using a Similarity Measure. To obtain more qualitative data, the FastText library was used to retrieve embeddings that are most similar to the target or cognate embeddings. The retrieval process utilized the K-Nearest Neighbors (KNN) 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.

[1] All the code can be found at GitHub.



(a) RSA on Italian hist. and full modern embeddings (left) and hist. and reduced modern embeddings (right).



(b) RSA on Spanish hist. and full modern embeddings (left) and hist. and reduced modern embeddings (right).

Fig. 2. Results of RSA performed on the 6 embedding spaces.

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.

Word Frequency and Semantic Divergence Analysis. For the frequency analysis, we followed the following steps:

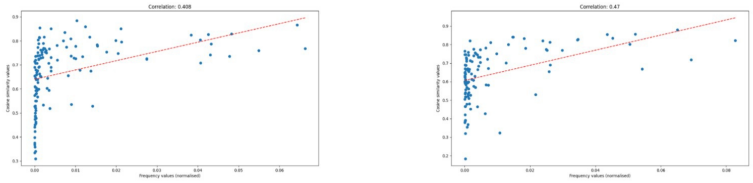
1. We applied Procrustes alignment to the two vector spaces (historical to modern for each language) to ensure that similar vectors represented the same concepts across different embedding spaces. This alignment was necessary as the embeddings were trained on different corpora in different languages.
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, 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. Additionally, we generated a graphical representation of the correlation using Matplotlib.

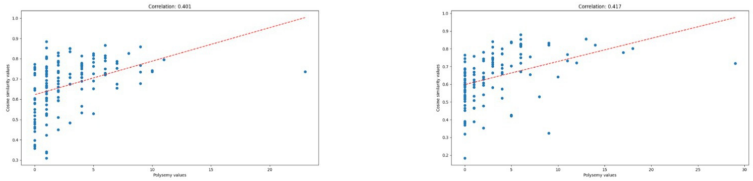
Word Polysemy and Semantic Divergence Analysis. After conducting the frequency and semantic divergence analysis, we proceeded to measure the polysemy of words. To accomplish this, we utilized the WordNet library, specifically leveraging the functionality provided by the "nltk.corpus.wordnet" module. Polysemy was quantified as the number of synsets associated with a word in WordNet, following the methodology described by Uban et al. (2019) [13].

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. Likewise in the previous analysis, we generated a graphical representation of the correlation using Matplotlib.

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 employ a K-NN approach with varying values of K (5, 10, 20, 50). We examine the overlapping nearest neighbors (NN) in both the historical and modern lists of NN. For each overlapping NN, we calculate the cosine similarity and measure the difference in the shift, determining whether the NN moved closer to or further from the target cognate word.



(a) The law of conformity visualized for Italian (left) and Spanish (right).



(b) The law of innovation visualized for Italian (left) and Spanish (right).

Fig. 3. Results of the analysis for the statistical laws of semantic changes: the law of conformity and the law of innovation.

By calculating the ratio of positive (closer) or negative (further) shifts, we can assess the coherence 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.

Following this analysis for all the cognates in the list, we remove those that have 0 or 1 NN because they do not provide informative results.

Results

Representational Similarity Analysis. As shown in Fig. 2, the reduced Italian modern embedding space shows 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). Consequently, while the results for Italian remain consistent across the full and reduced spaces, the variation in the Spanish modern spaces leads to different outcomes in the analyses and a higher cosine similarity. This difference suggests an insufficient amount of data in the historical Spanish embedding space. As a result, we have decided not to utilize the reduced models.

Calculation of Semantic Shifts.

Within-Language Comparison: KNN 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 Appendix section. 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 section for a detailed representation of these values.

Inter-Language Comparison: KNN with Jaccard Distance. The values presented in Table 3 correspond to dissimilarity scores, specifically semantic shifts, calculated using the Jaccard distance (1-Jaccard index) measure. The Pearson correlation score of 0.999, indicative of the overall correlation between the shifts for Italian and Spanish, irrespective of the particular K value, suggests a strong alignment in the observed semantic shifts in both languages. This high correlation implies the presence of similarities in the evolutionary patterns of language or semantic changes between Italian and Spanish.

Table 3. Correlation between the shifts for Italian and Spanish

K	Italian	Spanish
5	0.7833	0.7773
10	0.8520	0.8549
20	0.8928	0.8931
50	0.9189	0.9212
Correlation: 0.999		

Law of Conformity. Figure 3(a) 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 consistent in their trend with the results reported by Hamilton et al. (2016) [5].

Law of Innovation. Conversely, the results for the law of innovation in our study, as depicted in Figure 3(b), differ from

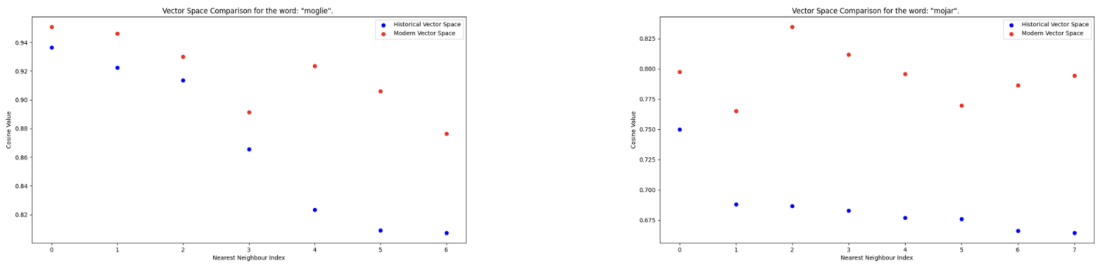


Fig. 4. An example of the analysis of the law of analogy visualized for Italian (left) and Spanish (right) using the cognate pair "moglie"/"mojar."

those reported by Hamilton et al. (2016) [5] and Uban et al. (2019) [13]. 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.

Law of Analogy. Here, we observe a trend, indicating that concepts that are closer to each other in the Euclidean space tend to exhibit shifts towards each other. Table 4 and Table 5 provide supporting evidence for this trend, as we can see that as the number of nearest neighbors (K-NNs) increases, the ratio of coherent shifts decreases. This aligns with the intuition that with more K-NNs, the distances between neighbors and their target cognate increase, leading to less consistent shifts. Furthermore, to provide visual representation, Fig. 4 displays an example visualization for a single cognate pair. These results provide a consistent pattern that supports the existence of a trend in the semantic shifts observed.

Table 4. 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 5. 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

Discussion

Indeed, the law of analogy, a new law examined in this study, has provided intriguing insights into semantic shifts. By considering the relationships between words and their analogies, we can gain an additional understanding of the underlying mechanisms driving these shifts. However, it is important to note that further investigation and research into this topic are necessary to validate and expand upon these initial findings.

However, 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 aligns with previous literature in a general trend, such as Hamilton et al. (2016) [5], 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 narrowly defined small time periods. Additionally, the alignment technique employed for aligning the embedding spaces could have contributed to the divergent outcomes in the analysis of the law of conformity and law of innovation. It is noteworthy that both the law of conformity and the law of innovation align with the findings of Dubossarsky et al. (2017) [4]. Their study revealed that the previously 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. In our analysis, after conducting partial correlation analysis, we also observed a weak correlation. Furthermore, we noticed a high compatibility between frequency and polysemy, indicating an inherent dependence, despite our efforts to disentangle them using partial correlation.

While utilizing the fasttext model, which is known for its improved performance on non-English languages, and compiling and pre-processing freely available data, the results still highlight the poor quality of embeddings obtained. This emphasizes the need for continued research and development of word embedding models, as well as the creation of larger and well-curated diachronic corpora. Improving the quality and quantity of data can potentially 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. As a result, this lead to an extremely high correlation coefficient between the languages being analyzed. This high correlation coefficient 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 evolution and to examine the universality claim. However, as mentioned previously, it is important to note that many languages still lack sufficient resources, and bridging the gap between these languages and their representations will require a significant effort.

Acknowledgements

We would like to express our gratitude to Dr. Raffaella Bernardi for her support throughout this project. Her feedback has been helpful in shaping our research.

We would also like to thank Dr. Lorella Viola for generously providing us with the corpus used in our analysis.

Bibliography

[1] P. Basile et al. "A diachronic Italian corpus based on "L'Unita"". In: *Proceedings of the Conference*. 2020, Page range. DOI: 10.4000/books.aaccademia.8245.

[2] José Cañete. *Compilation of Large Spanish Unannotated Corpora*. Zenodo. 2019.

[3] G. Collell and M. F. Moens. "Do neural network cross-modal mappings really bridge modalities?" In: *Journal Name* Volume.Number (2018), Page range. DOI: 10.18653/v1/P18-2074.

[4] Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. "Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017. DOI: 10.18653/v1/D17-1118. URL: https://aclanthology.org/D17-1118.

[5] W. L. Hamilton, J. Leskovec, and D. Jurafsky. "Diachronic word embeddings reveal statistical laws of semantic change". In: *Journal Name* Volume.Number (2016), Page range. DOI: 10.18653/v1/P16-1141.

[6] Ulrike 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.

[7] V. Lyding et al. "The PAISA corpus of Italian web texts". In: *Proceedings of the Workshop*. 2014. DOI: 10.3115/v1/W14-0406.

[8] E. Á. Mellado. "A Corpus of Spanish Political Speeches from 1937 to 2019". In: *Proceedings of the Conference*. 2020, pp. 928–932.

[9] T. Mikolov et al. "Efficient estimation of word representations in vector space". In: *Journal of Machine Learning Research* 13.Oct (2013), pp. 3071–3018. DOI: arXiv:1301.3781.

[10] S. Montariol. "Models of diachronic semantic change using word embeddings". PhD thesis. Université Paris-Saclay, 2021.

[11] E. Pettersson and B. Megyesi. "The histcorp collection of historical corpora and resources". In: *Proceedings of the Conference*. 2018, pp. 306–320.

[12] F. Sánchez-Martínez et al. "An open diachronic corpus of historical Spanish". In: *Language Resources and Evaluation* 47 (2013), pp. 1327–1342.

[13] A. Uban, A. M. Ciobanu, and 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.

[14] A. S. Uban and L. P. Dinu. "Automatically Building a Multilingual Lexicon of False Friends With No Supervision". In: *Proceedings of the Conference*. 2020, pp. 3001–3007.

[15] L. Viola and A. M. Fiscarelli. "ChronicItaly 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.

Appendix

Italian K-NN.

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 6. Italian, K = 5 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 7. Italian, K = 10 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 8. Italian, K = 20 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**

Table 9. Italian, K = 50 NN Overlap

Spanish K-NN.

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 10. Spanish, K = 5 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 11. Spanish, K = 10 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 12. Spanish, K = 20 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**

Table 13. Spanish, K = 50 NN Overlap

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 14. Italian, 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**

Table 15. Spanish, Cosine Similarity