Semantic Shift in Italian: A Computational Study

Gaudenzia Genoni
Computational Linguistics Course
Professor Raffaella Bernardi
University of Trento

Abstract
This study examines semantic change in Italian over the 20th century using computational methods inspired by Hamilton et al.'s framework for diachronic embeddings. Using 5-gram data from the Google Ngram dataset, word embeddings were trained for two periods (1900–1959 and 1980–1999) and aligned with the orthogonal Procrustes method. Word meaning shifts were quantified by semantic displacement, and regression models were used to evaluate relationships with word frequency and polysemy. While significant shifts were observed for some words, no statistically significant relationships emerged, likely due to dataset limitations. Future research could address these limitations by expanding the dataset, enriching context, and improving polysemy measurement methods.

1 Motivation

An important topic in computational linguistics and historical semantics is the phenomenon of semantic shift, which studies how words change their meanings over time. In this respect, the Italian language offers a particularly interesting case. Historically a literary and intellectual language, Italian served as a vehicle for high culture while local dialects dominated daily communication. Its widespread use as a spoken language emerged relatively late, following Italy's political unification in the mid-19th century. This delayed transition, combined with its literary heritage, suggests that Italian may exhibit semantic patterns distinct from those observed in other languages.

Accurately quantifying and analyzing these semantic shifts poses a significant challenge due to the lack of robust methodologies and comprehensive diachronic corpora. However, recent advances in natural language processing, particularly the use of word embeddings, offer a promising solution. Embedding-based methods allow for the precise analysis of semantic change across large datasets and time periods. My project, in particular, draws inspiration from Hamilton et al.'s work on diachronic embeddings: by applying their methodologies to the Italian context, I aim to explore how Italian words have evolved over time, focusing on the interplay between word frequency and polysemy.

2 Literature Review

The study of semantic change has gained significant attention in recent years, thanks to the availability of large-scale historical corpora and advances in computational methods. One interesting contribution to this field is the work by Hamilton et al. (2016)[2], which introduces a framework for quantifying semantic change using diachronic word embeddings.

Their study is grounded in the analysis of six extensive historical corpora spanning four languages—English, French, German, and Chinese—covering two centuries of linguistic evolution.
Central to their approach is the construction of word embeddings for individual time periods, using three distinct techniques: PPMI (Positive Pointwise Mutual Information), SVD (Singular Value Decomposition), and SGNS (Skip-gram with Negative Sampling). These methods encode semantic relationships by capturing word co-occurrence patterns, based on the distributional hypothesis that words appearing in similar contexts tend to share meanings. Among the three, SGNS stands out as a neural embedding model optimized to predict word-context pairs, making it particularly effective at capturing nuanced semantic relationships.
To enable meaningful comparisons of word meanings across time periods, Hamilton et al. apply the orthogonal Procrustes method[3], a mathematical alignment technique that minimizes distortions while preserving pairwise similarities between words. Their technique ensures that changes in word embeddings reflect genuine semantic shifts rather than artifacts introduced by the embedding process.
With these aligned embeddings, Hamilton et al. analyze semantic change using two complementary metrics: pairwise similarity and semantic displacement. Pairwise similarity examines the evolving relationships between specific word pairs, providing insights into cultural and linguistic shifts. For instance, they track how the word "gay" transitioned from meaning "cheerful" to its modern association with homosexuality. Semantic displacement, on the other hand, quantifies the degree to which a word's overall meaning shifts over time by measuring changes in its embedding vector.
The authors validate their methodology through detailed analyses of known historical shifts, such as the evolution of "broadcast" (from "casting seeds" to "transmitting signals") and "awful" (from "awe-inspiring" to "terrible"). Their results demonstrate that diachronic embeddings are not only effective at capturing these established changes but also capable of uncovering novel semantic shifts aligned with historical trends.
Building on these findings, Hamilton et al. propose two statistical laws of semantic change. The **law of conformity** posits that frequent words are more resistant to change, a phenomenon consistent with the linguistic stability often observed in high-frequency vocabulary. Conversely, the **law of innovation** highlights that polysemous words are more susceptible to semantic change, even when controlling for frequency. In their study, these laws—supported by regression analyses—explain a significant portion of the variance in semantic change rates across languages and corpora.

3 Research Questions and Proposal

3.1 Research Questions

While Hamilton et al. studied semantic change using corpora from English, French, German, and Chinese, their work did not extend to the Italian language. This raises the question of whether their statistical laws of semantic change—the law of conformity and the law of innovation—are observable in the Italian language. It also invites exploration of how computational methods, such as SGNS embeddings and Procrustes alignment, perform in capturing semantic change in Italian and what these methods might reveal about the relationships between word frequency, polysemy, and semantic drift in the Italian context.

3.2 The Proposal

To address these questions, I propose a computational investigation of semantic change in Italian. For the analysis, I will train SGNS embeddings on an Italian diachronic corpus, and align them with the Procrustes method for temporal comparisons. I will quantify shifts in word meanings via semantic displacement, using t-SNE for visualization. Finally, I will evaluate the relationships between semantic displacement, word frequency, and polysemy with regression models, testing whether frequent and polysemous words are more prone to change.

4 Project Description

4.1 The Dataset

This project serves as an initial experiment to explore the research questions, providing a practical approach to testing the proposed methodology. Given the computational complexity of analyzing an entire Italian diachronic corpus, this study narrows its scope to Italian 5-grams beginning with the letters "ac," extracted from the 2012 Google Ngram dataset[5]. While a comprehensive diachronic analysis would ideally include data from before the 20th century, this study focuses on the period 1900–2000: this temporal range is chosen to ensure computational feasibility, and particularly because the data quality and volume in the Google Ngram dataset (which is entirely derived from digitized books) improve significantly during this time period.
The final dataset comprises 1,260,736 records, each providing details about a 5-gram term, its year of usage, the match count (frequency of occurrence), and the volume count (number of unique books in which the n-gram appears).

4.2 Data Preprocessing

To enable meaningful diachronic analysis, the dataset was divided into two periods: 1900–1959 and 1980–1999, each containing a comparable amount of data. Consequently, the semantic shifts analyzed in this study reflect changes that occurred in Italian between the first six decades and the final two decades of the 20th century.
To ensure data consistency and quality, grammatical tags (automatically included in the dataset) were removed, n-grams containing non-alphabetic characters were excluded, and all n-grams were converted to lowercase. Stopwords were removed using SpaCy's Italian language model[1], and any rows where the n-grams were reduced to a single word after stopword removal were dropped.
Next, a "period" column was added to indicate the temporal range (1900–1959 or 1980–1999) for each entry, and the dataset was aggregated by period and n-gram, summing the match count and volume count for each n-gram.
To normalize linguistic variations, the n-grams were lemmatized using SpaCy, reducing words to their base forms for consistency.
Finally, a filtering step was applied to focus on the most frequently occurring n-grams, minimizing noise: for each period, the 75th quantile of match count was calculated, and only n-grams with a match count at or above this threshold were retained.
After filtering, the dataset for the period 1900–1959 contained 1,815 n-grams with a total match count of 413,325, while the period 1980–1999 included 1,974 n-grams with a total match count of 497,710.

4.3 Embedding Training and Alignment

To model semantic shifts over time, embeddings were trained separately for the two temporal periods (1900–1959 and 1980–1999) and

aligned to enable meaningful comparisons.
First, tokenized datasets were created for each period by expanding each lemmatized n-gram proportionally to its match count. Skip-Gram with Negative Sampling models were then trained for each period using the Word2Vec algorithm. These models encoded semantic relationships as dense vector representations in a 300-dimensional space. The context window size was set to 3, with 5 negative samples applied per positive pair.
To compare embeddings across periods, a shared vocabulary was extracted, consisting of words present in both temporal datasets (in total, 932 words). Embedding matrices were constructed for the common vocabulary, one for each period, and aligned using the orthogonal Procrustes method.
Finally, semantic displacement was calculated for each word in the shared vocabulary by measuring the cosine similarity between its normalized vector in the 1900–1959 matrix and its aligned vector in the 1980–1999 matrix. The displacement score quantifies how much a word's meaning shifted over time, with higher scores indicating greater semantic change.

4.4 Visualization and Evaluation of Semantic Change

The final step involved visualizing semantic change and evaluating its relationship with word frequency and polysemy.
To visualize the alignment of embeddings between the two periods, a dimensionality reduction technique, t-SNE[4], was applied: this method reduces the high-dimensional embedding vectors to two dimensions, allowing them to be plotted and compared.
Ordinary Least Squares (OLS) regression was then used to evaluate the relationship between semantic displacement, word frequency and polysemy. Word frequencies were computed based on their occurrences in the dataset and log-transformed to account for the wide range of values. Polysemy, approximated through context diversity, was measured as the number of unique n-grams in which each word appeared, with these values also log-transformed. Both the log-transformed frequencies and context diversity values were regressed against semantic displacement scores.

5 Results and Discussion

5.1 Semantic Displacement

The analysis of semantic displacement allowed to identify words in the dataset with the most and least significant changes in meaning across the two periods. Here, displacement scores, measured on a scale derived from cosine similarity, range from 0 (no shift, identical embeddings) to 2 (maximum shift, completely divergent embeddings).
The three words with the most significant semantic shifts are "acuto" (displacement score: 1.1738), "dominante" (1.1459), and "accenno" (1.1447). Likely, their shifts reflect evolving contexts or metaphorical uses influenced by cultural, scientific, or societal changes. For instance, "acuto" (sharp/acute) may have experienced semantic broadening due to its application in diverse domains such as music (sharp tones), medicine (acute conditions), and intellectual discourse (acute observations).
Conversely, the three words with the lowest semantic displacement are "socio" (0.8352), "donna" (0.8427), and "padre" (0.8487). Indeed, they represent core social and familial concepts that are less prone to contextual or interpretative changes, therefore maintaining stable meanings across decades.

5.2 t-SNE Visualization

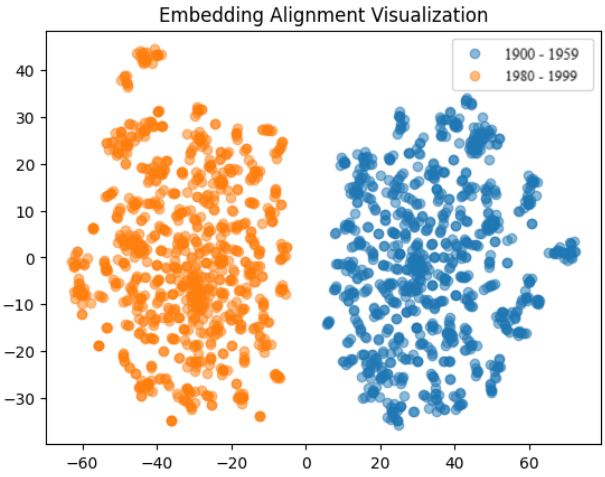

Figure 1: Embedding Alignment Visualization.

The t-SNE plot (Figure 1) reveals two distinct clusters for the periods 1900–1959 and 1980–1999, with no visible overlap between them. This separation, which suggests significant differences in the semantic spaces of the two periods, would indicate a substantial semantic vocabulary shift. However, the interpretation requires careful consideration, as the lack of overlap may be influenced by other factors linked to the exploratory nature of this study. Since the original dataset is composed of 5-grams rather than full sentences, the embeddings might be affected by a limited contextual richness. Additionally, the small shared vocabulary between the two periods (only 932 words) may exaggerate differences in the embedding spaces, as it represents a narrow subset of the overall semantic landscape.

5.3 Regression Analysis

Similarly to the t-SNE visualization, the regression analysis results were not conclusive. The regression between semantic displacement and word frequency produced a near-zero coefficient (0.0006) with a high p-value (0.677), indicating no statistically significant relationship. Likewise, the analysis between semantic displacement and context diversity (used as a proxy for polysemy) returned a coefficient of 0.0002 with a high p-value (0.921), suggesting no meaningful correlation. As such, these findings provide little evidence to support the Law of Conformity or the Law of Innovation in Italian across the 20th century; however, this is likely attributable to the limitations of the dataset discussed earlier.

6 Conclusions

The study explored semantic change in the Italian language throughout the 20th century, employing computational methods inspired by Hamilton et al.'s methodology for diachronic embeddings. To this end, a variety of tools were applied, including the Google Ngram dataset, SpaCy, SGNS with Word2Vec, the orthogonal Procrustes method, t-SNE, and OLS regression. Beyond the technical aspects, this project presented the opportunity to reflect on the representation of words as vectors in multidimensional spaces and to address the challenges of formalizing complex linguistic concepts such as polysemy. Although the results were limited by constraints in the dataset, reflecting the exploratory scope of this analysis, future research on larger corpora could assess the applicability of the laws of conformity and innovation in Italian. Looking ahead, a comprehensive study would span the entire history of the Italian language—from its poetic beginnings in the 13th century to the fast-evolving discourse of contemporary social media—uncovering how its long-standing role as a literary and cultural language and its more recent adoption as a medium of everyday communication have shaped its meanings and expressions over time.

References

[1] Explosion AI. spacy: Industrial-strength natural language processing in python, 2023. Last accessed: 2025-01-11.

[2] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 54:1489–1501, 2016.

[3] Peter H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

[4] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[5] Google Books Ngram Viewer. Google books ngram dataset, 2012. Last accessed: 2025-01-11.