Universität Zürich[UZH]

Bachelorarbeit
zur Erlangung des akademischen Grades
Bachelor of Arts
der Philosophischen Fakultät der Universität Zürich

Multilingual Semantic Change Detection with Word Embeddings in 19th Century Novels

Verfasserin: Seraina Betschart
Matrikel-Nr: 17-960-162

Referent: Prof. Dr. Martin Volk
Institut für Computerlinguistik

Abgabedatum: 01.12.2024

Abstract

Recent years have seen a rise in using computational methods to detect semantic change in words. This thesis explores the use of static word embeddings to measure semantic change and compare it cross-linguistically. I propose a method for identifying English words of Latin origin and manually curate a small dataset of cognates in French, Spanish and English. I measure semantic change using cosine similarity and find that English words of Latin origin undergo a 3% larger semantic change compared to words without Latin roots. Further, I examine the curated cognate sets and report a strong correlation between the measured semantic change of French and Spanish words. Finally, I show that parts-of-speech exhibit similar patterns of change across the three examined languages, with function words being more stable than content words, and proper nouns exhibiting the largest measured change.

Zusammenfassung

Die Verwendung von rechnerbasierten Methoden zur Erkennung semantischer Veränderungen in Wörtern hat in den letzten Jahren zugenommen. Die vorliegende Arbeit untersucht die Bedeutungsänderungen von Wörtern unter Einsatz von statischen «word embeddings» und deren vergleichende Analyse über verschiedene Sprachen hinweg. Ich entwickle eine Methode, um englische Wörter mit lateinischem Ursprung zu identifizieren, und stelle eine Sammlung von Wörtern gleichen Ursprunges in Französisch, Spanisch und Englisch zusammen. Mittels der Kosinus-Ähnlichkeit wird der Grad der semantischen Veränderung gemessen. Dabei stelle ich fest, dass englische Wörter lateinischen Ursprunges eine um 3% grössere semantische Veränderung durchlaufen haben, als diejenigen ohne lateinische Wurzeln. Zusätzlich untersuche ich die Sammlung an Wörtern mit gleichem Ursprung und ermittle eine starke Korrelation zwischen der gemessenen semantischen Veränderung der französischen und spanischen Wörtenr. Schliesslich zeige ich, dass Wortarten in den drei untersuchten Sprachen ähnliche Veränderungsmuster aufweisen, wobei Funktionswörter grundsätzlich stabiler sind als Inhaltswörter und dass Eigennamen die grösste gemessene Veränderung aufweisen.

Acknowledgment

I thank my supervisor Martin Volk for his consistent support throughout the entire process and for encouraging me to explore my own path.

I extend my gratitude to Julian Hofer for proofreading the text.

Last but not least, I owe thanks to my family and friends for helping me stay sane and reminding me to go outside every once in a while. Without all of you, I would probably still be typing away at my computer.

Contents

List of Figures

List of Tables

Appendix

List of Acronyms

ELTeC	European Literary Text Collection
LLM	Large Language Models
NLP	Natural Language Processing
OCR	Optical Character Recognition
OED	Oxford English Dictionary
POS	Part-Of-Speech
PPMI	Positive Pointwise Mutual Information
SGNS	Skip-Gram with Negative Sampling
SVD	Singular Value Decomposition

1 Introduction

1.1 Motivation

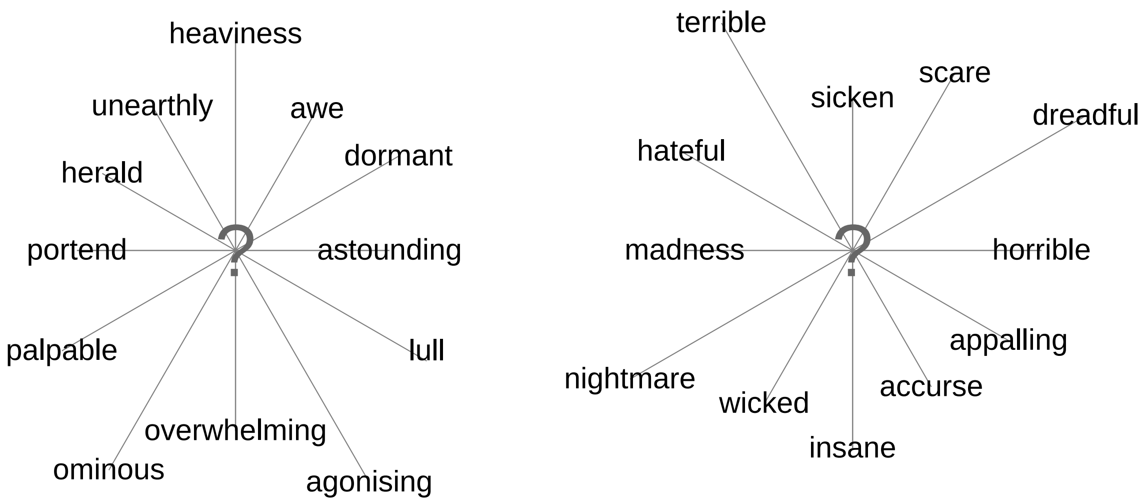


Figure 1: Nearest neighbours of two unknown words

For years, word embeddings have been used to analyse a word's meaning. As Firth famously stated in 1957: "You shall know a word by the company it keeps" (Firth, 1957).

Why not start with a guessing game: The two illustrations in Figure 1 depict a selection of words whose embeddings are closest to that of the unknown word in the middle. Which words might belong in the middle in each of the two examples?

So, have you made your guess as to what the unknown words might be?

The answer may surprise you: Both figures represent the vector space of one and the same word – *awful* – just at different points in time. The illustration on the left is based on a word embedding model trained on data from 1840 to 1875. The right, however, on data from 1885 to 1920.

As has been extensively demonstrated, word embeddings are very effective at capturing a word's meaning (Mikolov et al., 2013; Grave et al., 2018). By training separate embedding models on different time periods, one can thus capture the meaning of a word during a specific period and compare the different embeddings. Words with little meaning change during the observed period are expected to have consistent embeddings. However, when words undergo semantic shifts, their vector embeddings can illustrate such changes, as seen in Figure 1.

Using word embeddings to detect semantic change holds tremendous potential, as it is possible to compare and evaluate historical data on an unprecedented scale. Recent years have seen a rise of studies employing word embeddings for semantic change detection (Tahmasebi and Dubossarsky, 2023). Most of these studies have experimented with static embeddings, using algorithms such as PPMI or SGNS to generate a single embedding for each word in the vocabulary.

While many studies have shown the ability of such models to successfully detect semantic change (compare Hamilton et al. (2016); Schlechtweg et al. (2019); Shoemark et al. (2019)), the field still lacks a standardized evaluation framework. Currently, most evaluations are conducted manually, relying on a qualitative review of selected examples.

Although several studies have worked with different languages individually, the potential for cross-lingual comparisons remains underexplored. Uban et al. (2021) used pre-trained embeddings to compare cognate sets across Romance languages and English in modern-day use.

In this thesis, I dive deeper into the exploration of Romance cognates. I construct cognate sets adapted to a multilingual corpus for the languages English (which boasts a significant number of Latin words), French and Spanish. I train embedding models to investigate semantic change patterns in English words of Latin origin compared to those without Latin origin, as well as French and Spanish. Additionally, I explore the semantic change of parts-of-speech in each of the three languages.

1.2 Research Questions

The goal of this thesis is to investigate words with Latin roots across different languages. I investigate English words, compute their semantic change across several decades, and compare the results with those of their cognates in French and Spanish. Thereby the following research questions are posed:

1. Do English words of Latin origin undergo a larger semantic change during the 19th century than English words without Latin roots?
2. How does the change in meaning of English words with Latin roots compare

to the semantic change of their French and Spanish counterparts?

3. Which part-of-speech categories exhibit the most change in meaning and can cross-lingual differences be observed?

Examining these questions allows me to explore the potential of word embeddings for comparing related words in different languages.

1.3 Thesis Structure

I begin with a brief background on the study of semantic change and an overview of some computational methods for detecting it. Chapter 3 describes the data used to train the embedding models, the pre-processing steps taken, and the process of how determining the etymology of a word. In Chapter 4, I go into detail about the methods I employed, including the algorithms for model training, hyperparameter tuning, and evaluation of the trained models. In Chapter 5, I present the experiments conducted and their results. Finally, in the discussion, I address the research questions outlined earlier.

All code is available in the following GitHub repository:

https://github.com/sebets1/lsc-detection-with-embeddings

2 Background

2.1 Linguistic Background

2.1.1 Semantic Change in Historical Linguistics

Semantic change is the change of meanings of individual words (Uban et al., 2021). It happens through an innovative use of a word through one speaker or a small group. If this innovation is picked up by more speakers, it can lead to lexicalization of this innovation as an additional meaning of the word (Blank, 2001). Over time, a word can acquire and lose meanings several times through different processes such as broadening, narrowing and amelioration. Figure 2, based on an example from Blank (2001, translated from German), shows how the Latin word *tutari*, meaning "protect", has ended up taking the form *tuer* in French with an entirely different meaning as "kill".

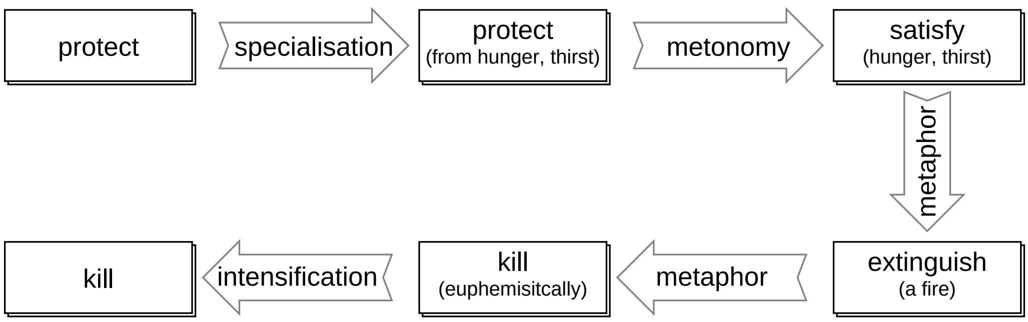


Figure 2: How "protect" became "kill": An example of semantic change from the meaning "protect, ward off" in Latin *tutari* to "kill" in today's French *tuer* (adapted from Blank (2001)

2.1.2 The Latin Influence on English Vocabulary

In this thesis, I am going to concentrate on semantic change in English words which can be traced back to Latin. While English is a Germanic language, a big part of the vocabulary can be traced back to Latin (Green (2014) estimates that over 60% of English words have Latin or Greek roots).

The profound influence of Latin on English vocabulary is due to several historical and cultural factors. Two pivotal reasons are the Norman Conquest and the high prestige Latin held for centuries.
The Norman conquest of England in 1066 introduced Old French as the language of the court, culture and literature for several centuries. This transition brought a wave of Latin-derived words into English through French (Solodow, 2010).
Additionally, many words in the English vocabulary were directly adapted from Latin itself. Latin held a high status in Europe for centuries, serving as the language of science and clergy. English writers often “enriched” their texts with Latin terms, many of which eventually integrated into everyday speech (Solodow, 2010).
These circumstances led to instances where the same Latin word entered the English vocabulary twice - once directly from Latin and once through French; such as *legal* and *loyal* (from Latin *lex*) or *fabric* and *forge* (from Latin *fabrica*) (examples from Solodow (2010)).

Today, many Latin-derived words have become part of the vocabulary of several Romance languages, as well as English. Some of these words, although still very similar in form, have not kept the same meanings across languages.
One example is the English word *actual* (“real, existing”) going back to Latin *actualis* (”active, pertaining to action”)[1]. The Spanish *actual*, derived from the same root, now means “current, up-to-date”. Meanwhile, the French word has slightly changed form and is now spelled *actuel* but has shifted in meaning to the same as the Spanish actual.
In the meantime, other words are still so similar in form and meaning that it comes as no surprise to learn that they share a common ancestor. For instance, English *different*, French *différent*, and Spanish *diferente* all trace back to Latin *differentem* (accusative case of *differens*). Similarly, English *silence* originates in Latin *silentium*, the same as French *silence* and Spanish *silencio*.

Words from related languages which go back to the same etymon of a common proto-language are called **cognates** (Campbell, 1998), such as the example in Figure 3. A **cognate set** is a set of cognates across different languages. English words with Latin origin are loanwords, borrowed into the language, as English itself is not a Romance language. For the purpose of this work, I follow Uban et al. (2021) by including English loanwords in the Romance cognate sets. In this sense, even though

[1] Etymology and definitions are taken from Wiktionary (https://en.wiktionary.org/) and Etymonline (https://www.etymonline.com/).

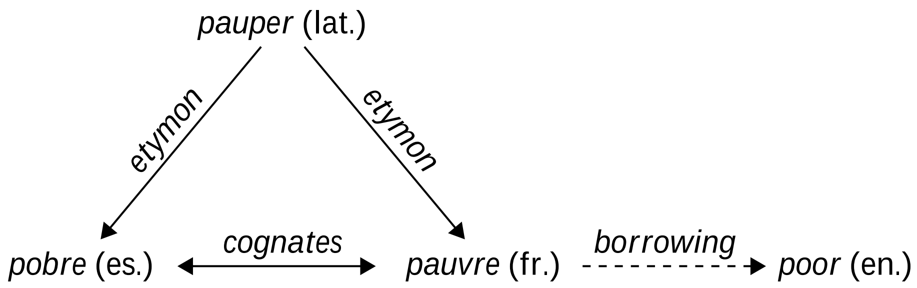


Figure 3: Cognates example

English *different* and *silence* have been imported into the language through French, I will still refer to them as cognates of their Spanish and French counterparts as they are eventually derived from the same ancestor.

2.2 Computational Methods to Detect Semantic Change

With the development of large language models, new paths have opened for historical linguistics. Recent years have seen a rise in computational methods aimed at discovering semantic shift and semantic change (Kutuzov et al., 2018). These methods include clustering techniques, topic models, count-based approaches and distributional semantic methods (Tahmasebi and Dubossarsky, 2023). In the following sections, I focus specifically on the research around training word embeddings to detect semantic change.

Through word embeddings, it is possible to derive much about a word's meaning based on the contexts in which it appears. This approach relies on the distributional hypothesis, which posits that "we can infer the meaning of words by systematically analyzing the context in which they are used" (Tahmasebi and Dubossarsky, 2023).

So, what if we were to train two separate embedding models with the same prerequisites, algorithm and hyperparameter settings but on data from different time periods? For example, we could use one corpus based on 18th-century texts and another on 20th-century texts. We would likely observe words appearing in different contexts, reflecting changes in their meanings over time. This is the intuition behind using word embeddings to detect semantic change. While various methods have been tested, no universally accepted "gold standard" approach has yet emerged (Tahmasebi and Dubossarsky, 2023). Still, most studies on the topic now follow a pattern involving these three steps:

1. Determine a corpus and split it into equal time bins as a basis for training embedding models.
2. Choose a training algorithm and apply it to each time bin of the corpus separately.
3. Select an evaluation metric to measure the distance between two vectors representing the same word at different points in time.

In the following sections, I briefly outline each of the three steps and mention the most common approaches. I also present some evaluation strategies.

2.2.1 Corpus Considerations

A common approach is to use two corpora of the same genre from different time periods to assess semantic shifts over time (Kutuzov et al., 2018). The duration of each time span within a corpus and the gap between the two corpora can be varied depending on the research focus. Early studies using word embeddings tended to employ wider time gaps (Kutuzov et al., 2018) but more recent research, such as studies by Shoemark et al. (2019) and Wang and Choi (2023), has shown that semantic shift can be detected over short periods of only a few years between two corpora.

A repeatedly used corpus for evaluation is the English corpus for the SemEval-2020 shared task (see also Section 2.2.4). Along with an evaluation set, Schlechtweg et al. (2020) published two training corpora, the first one spanning from 1810 to 1860 (6.5 million words) and the second one from 1960 to 2010 (6.7 million words). Not only were these openly available corpora useful for the evaluation of my models (see 4.2), but the processing steps undertaken for these corpora influenced my own process too.

While most studies focus on detecting semantic change over time, researchers have also used embeddings trained on different corpora to explore semantic variation across domains and languages. For instance, Schlechtweg et al. (2019) examined semantic change detection across domains building a corpus of cooking-related data. They showed that domain-specific changes of word senses compared to general-language use can be detected with embedding models.

Interesting for this work was a study by Uban et al. (2021), who measured semantic shift among cognates in Romance languages and English. The study aimed at investigating whether words derived from the same Latin root were used in similar contexts across Romance languages today and whether Latin-derived English words show distinct patterns of semantic shift depending on the time they came into the English language. Instead of putting together comparable corpora in various Romance languages, they used the pre-trained embeddings.

Ehrenworth and Keith (2023) proposed "intertextual semantic change detection", using individual books as corpora. This method produced a ranked list of words according to the degree of semantic change between two books. Their approach is still very interesting under the consideration that the ranked list should not be seen as definitive, but as a starting point for manual linguistic investigation. Such a list can produce surprising words which might not have been considered to be investigated otherwise (Ehrenworth and Keith, 2023).

With their study, Ehrenworth and Keith (2023) raised the question of the minimum corpus size needed to produce reliable embeddings. While corpus size is an important factor, several others also influence the embeddings, such as the type of text, OCR errors or high amounts of non-standard spelling (Tahmasebi and Dubossarsky, 2023).

2.2.2 Types of Embedding Models

There have been various studies on semantic change detection with embedding models. These different model architectures can be loosely categorized into **static** and **contextualized embeddings** (Schlechtweg et al., 2020). Models that produce static embeddings generate a single embedding for each word in the vocabulary, regardless of the fact that many words have multiple meanings. **Static embeddings** are "used to detect how a word changes in its dominant sense, without considering the possible additional, subordinate senses/meanings that the word can have" (Periti and Montanelli, 2024). In contrast, **contextualized embeddings** produce varied embeddings for a word depending on the specific context in which it appears (Periti and Montanelli, 2024).

In recent years, contextualized embeddings have become increasingly popular for semantic shift detection. However, in the SemEval-2020 task, contextualized embeddings were outperformed by static embeddings (Schlechtweg et al., 2020). The authors reason that this may have been due to a "lack of proper usage conventions", as contextualized embeddings were still a relatively new technology at the time. Furthermore, contextualized embeddings are pretrained and thus already carry information which might not have been relevant for the specific data in that evaluation (Schlechtweg et al., 2020). While Periti and Montanelli (2024) confirm in their survey paper that "form-based approaches" (i.e. static embeddings) generally outperform "sense-based approaches" (i.e. contextual embeddings), they attribute this result to the structure of most test sets, which typically provide only one semantic change score per target word. This setup could place contextual models at a disadvantage by failing to capture the nuanced changes that sense-based embeddings are designed to detect.

In this research, I did not focus further on contextual embeddings but instead worked with static embeddings. This decision allowed me to use my own data and training set-up instead of large pre-trained models. I hoped that this would better enable me to adapt the models to the specific needs of my research questions and build models adapted for cross-lingual comparison. Furthermore, static embeddings for semantic change detection have been more thoroughly explored, providing more resources and studies for comparison.

2.2.3 Static Embedding Models

Past studies on semantic change detection have most commonly compared the following three types of static embeddings:

- Positive Pointwise Mutual Information (PPMI)
- Singular Value Decomposition (SVD)
- Skip-Gram with Negative Sampling (SGNS)

Positive Pointwise Mutual Information (PPMI)

The simplest form of static embeddings are **count-based embeddings**. Count-based embeddings are high-dimensional and sparse. Each word is represented by a vector whose number of dimensions matches the total number of words in the vocabulary. Each dimension of the vector reflects how often another word in the vocabulary co-occurs with the target word. Sparse vectors typically have many dimensions with values of 0.

To enhance the performance of raw co-occurrence counts, additional transformations are often applied (Schlechtweg et al., 2019). A common method is **Positive Pointwise Mutual Information (PPMI)**, which measures "how often two events x and y occur, compared with what we would expect if they were independent" (Jurafsky and Martin, 2024). PPMI is an easy and computationally inexpensive method to apply, also embeddings of a word from different models can be directly compared without alignment. But its performance in semantic change detection cannot keep up with more complex methods (Schlechtweg et al., 2019; Hamilton et al., 2016; Wang and Choi, 2023).

Singular Value Decomposition (SVD)

Another popular approach is **Singular Value Decomposition (SVD)**. SVD is applied to PPMI embeddings reducing the dimensions of the vectors by trying to find the most important dimensions. SVD can significantly improve the embedding quality and compared to PPMI seems to be more robust (Hamilton et al., 2016). It has been frequently employed for semantic change detection with overall superior performance over PPMI (Hamilton et al., 2016; Schlechtweg et al., 2019). This method was not chosen for this thesis and will therefore not be examined further.

Skip-Gram with Negative Sampling (SGNS)

One of the most common and best-performing approaches using static embeddings for semantic change detection is **Skip-Gram with Negative Sampling (SGNS)**. SGNS is one of two algorithms proposed by Mikolov et al. (2013) and implemented in the word2vec software. It is a training approach producing dense vectors, which means that each word gets a vector representation of only a few hundred dimensions. These dimensions do not correspond to specific words (unlike count-based vectors) but rather encode complex, hard-to-interpret semantic and syntactic patterns. Typically, vectors with 50 to 300 dimensions are trained on SGNS models for semantic change detection (Tahmasebi and Dubossarsky, 2023).

SGNS has repeatedly been shown in studies to be the best-performing embedding model for semantic change detection. Schlechtweg et al. (2019) evaluated various count-based and static embedding models, finding that Skip-Gram with orthogonal alignment and cosine distance performed best overall. Furthermore, approaches using SGNS produced the best results in the Sem-Eval2020 Shared Task (Schlechtweg et al., 2020).

Generally, SGNS for semantic change detection has already been extensively studied and well documented, with several studies trying the technique on a corpus of similar size or smaller to the one I was working with (see Chapter 3). I thus decided to opt for this popular approach too.

The standard approach to measure semantic shift is filtering out the vector representations of a word at different points in time and comparing their distance, most commonly by measuring the cosine distance (Tahmasebi and Dubossarsky, 2023). While methods like PPMI allow for direct comparison of vectors across corpora, dense vector models require an additional alignment step. A common aligning technique uses orthogonal Procrustes in order to transform word vectors from one model as best as possible into the vector space of the other model while maintaining the relative distances between vectors (Hamilton et al., 2016; Tahmasebi and Dubossarsky, 2023).

2.2.4 Evaluation Strategies

A key challenge in semantic change detection is the absence of a standardized evaluation framework, which complicates the comparison of findings across studies. According to a 2023 survey (Periti and Montanelli, 2024), there is still no "consolidated and widely-accepted classification framework." Tahmasebi and Dubossarsky (2023) identifies three primary strategies for evaluating semantic change detection models:

- Evaluate on a test set previously evaluated by hand
- Evaluate the outcome of an experiment
- Evaluate using control settings

Evaluate on a test set previously evaluated by hand

When using manually annotated test sets, a common process is choosing a set of a few dozen target words which meet a certain frequency threshold in the evaluated data. Several annotators will then assess the level of semantic change, either in binary terms (the word's meaning changed or did not change) or on a scale (e.g. 1–5, with 1 representing no change and 5 indicating complete change). These human judgements then serve as the gold standard against which the model outcome is assessed (Tahmasebi and Dubossarsky, 2023; Kutuzov et al., 2018).

In 2020, a shared task was introduced offering a standardized approach to testing semantic change embedding models. For the SemEval-2020 Task 1, Schlechtweg et al. (2020) created test sets in English, Swedish, German and Latin with a set of target words for each language, each target word receiving an annotated semantic change score. They furthermore provided two pre-processed corpora for each language (i.e. the English corpora spanning the time from 1810-1860 and 1960-2010).

Creating evaluation datasets in this way requires a lot of resources and is not a doable approach for every study (Tahmasebi and Dubossarsky, 2023). However, using pre-existing standardized test sets makes it possible to automatically evaluate different model architectures and setups under uniform conditions. While newer studies often rely on these pre-made test sets (i.e., Ehrenworth, 2023; Wang, 2023), it is important to note that these sets were created for specific corpora and may not be fully adaptable to other data sources.

Evaluate the outcome of an experiment

Another common strategy is directly evaluating the outcome of the models. The majority of studies on semantic change detection have opted for this kind of evaluation (Tahmasebi and Dubossarsky, 2023).

Hamilton et al. (2016) employed a hybrid approach, making two evaluations. In the first one, they put together a list of words with known meaning shifts or newly acquired meanings taken from previous works on semantic change and as well as from a dictionary. After training their models, Hamilton et al. (2016) evaluated if these could correctly detect the shifts in meaning of the target words. Subsequently, they investigated the ten words showing the largest semantic change according to their models and assessed by hand with the help of etymological dictionaries if these words actually showed semantic change.

Ehrenworth and Keith (2023) worked with very small corpora of approximately 150'000 tokens each, comparing two individual novels. They recreated several top-performing models to study their performance on small subsets of SemEval-2020 Task 1 data showing that there was a steep drop in performance on smaller datasets. Subsequently, they used one of the models to analyse their own corpora and examined the top ten most-changed words. They concluded that models were needed which could perform well on small corpora but added that even the existing models could already be interesting to use "in an exploratory manner to aid literary critics" (Ehrenworth and Keith, 2023).

Direct evaluation of experimental outcomes comes with a few downsides. First of all, each study may yield different results due to varying methods, making a comparison difficult. Additionally, every modification to a model will change the results as well, which must then be reevaluated manually (Tahmasebi and Dubossarsky, 2023).

However, this approach also has advantages. If we go beyond evaluation with predetermined test sets, we might encounter interesting phenomena which we could have missed otherwise. Even if the models are not capable yet of providing a perfectly accurate number of semantic change that a word underwent, their output can be a valuable starting point for discussion and analysis by hand (Ehrenworth and Keith, 2023; Tahmasebi and Dubossarsky, 2023).

Evaluate using control settings

The third evaluation strategy is less explored. When using control settings, the goal is not to evaluate the semantic change of single words but to determine whether the models actually detect semantic change or something else.

Shoemark et al. (2019) proposed synthetic evaluation data to simulate semantic change in words. They inserted made-up words into their corpora in a controlled setting to model different patterns of model behaviour that they either aimed to observe or avoid. For example, they would replace a word with a fixed probability and a word whose probability increased over time with the same pseudoword. This should show a pattern where a word would be modelled that initially starts with the meaning of the first word and over time acquires a new meaning related to the second word (Shoemark et al., 2019).

Another such control mechanism involves training a model with data from several time periods, then using the same data but shuffling the sentences and training the model again. Ideally, comparing these two models should not result in detecting significant semantic change. If this were the case, we could assume that something different, such as fluctuations in the amount of data over time, is modelled (Tahmasebi and Dubossarsky, 2023).

2.2.5 Semantic Change Detection with Text Generation Models

A method tried recently in semantic shift detection is prompting text generation models. Wang and Choi (2023) compared PPMI and SGNS trained on a small dataset of tweets (3'287 sentences) with embeddings from a BERT model. The models assessed the semantic change of 34 target words over short time spans (years 2019, 2020 and 2021). Furthermore, they experimented with prompting a GPT-4 model by providing a sentence from one year and a sentence from another year containing the same target word, then asking the model to return a binary result (indicating whether a meaning change had occurred). In this study, GPT prompting outperformed the other methods. In a study of Periti et al. (2024) however, GPT performed significantly worse then BERT in recognizes long-term and especially short-term changes in words.

While experimenting with text generation models was out of the scope of this study, they seem to provide much potential and further investigation would be interesting.

3 Corpus Building and Etymology Detection

In this chapter, I outline which data I used to train embedding models, how the corpora were put together, how I pre-processed the data and created a list of cognates in French, Spanish and English.

3.1 Data Collection

For this thesis, I aimed to compare the semantic change of English words with Latin roots to their cognates in Spanish and French. To accomplish this, I used six corpora in total – one for each language at two different time points. In order to ensure comparability of the corpora, the data collected for each language should be from the same time period, evenly distributed and within a similar domain. I chose to focus on literary texts from before 1900, as each of the three languages has a rich tradition in producing literary fiction, making it feasible to gather sufficient data across similar genres. Ultimately, I chose to work with the **European Literary Text Collection (ELTeC corpus)**.

3.1.1 European Literary Text Collection

The ELTeC corpus[2] (Schöch et al., 2021) provides a curated collection of literary texts published between 1840 and 1920 across several European languages. The completed language collections, to which Spanish, French and English belong, contain 100 novels each, evenly distributed over four 20-year time sections. The texts are all in the public domain. The collections were built according to a set of rules (e.g. even distribution across time, a minimum of female authors if possible, even distribution of lengths of novels), attempting to make the collections as comparable as possible across time and domain.

[2] ELTeC corpus: https://www.distant-reading.net/eltec/

With several million words for each language (see Table 1), the collections should suffice in size in order to get decent semantic change embeddings. However, the number of documents is on the smaller side with between 30 and 47 novels in each corpus. This increases the risk that a single novel which might use a certain term frequently could skew the trained embedding of that term.

3.1.2 Splitting the Corpora

ELTeC documents consisted of pre-formatted XML files, one for each novel. Following some experimentation with different time cut-offs, I decided on the period from 1840–1875 for Corpus 1 (C_1) and 1885–1920 for Corpus 2 (C_2). This ensured balanced corpus sizes across languages. As shown in Table 1, each corpus contains between 2 and 5 million words. Finally, I saved the documents, ordered by time period, into raw text files.

	English	French	Spanish
Older Epoch C_1 (1840-1875)	5’942’305 words 34 novels	3’473’785 words 39 novels	2’662’163 words 30 novels
Newer Epoch C_2 (1885 -1920)	3’999’995 words 47 novels	3’688’348 words 44 novels	3’388’583 words 45 novels

Table 1: Word and document counts of each corpus

3.2 Pre-processing

Before I could use the raw text files for model training, they needed to be pre-processed. Following established procedures (Tahmasebi and Dubossarsky, 2023; Schlechtweg et al., 2019), I performed word and sentence tokenization, removed non-alphabetical characters and stopwords to focus on meaningful content.

I used the NLP toolkit SpaCy (Montani et al., 2023) for lemmatization and part-of-speech tagging. The following SpaCy pipelines were used; *fr_core_news_sm* for French, *es_core_news_sm* for Spanish and *en_core_web_sm* for English. I added the part-of-speech tag to every lemma in order to later train the models on the lemmas with part-of-speech tags (e.g. “she hoped” – “she_det”, “hope_verb”). This would allow not only for more analysis options but also ensure that different words with the same lemma, which are common especially in English, would be trained separately.

SpaCy assigns UPOS tags from the Universal Dependencies Project[3]. The UPOS tags in table3.2 are sorted into open and closed class (following the categorization of Jurafsky and Martin (2024). Most closed class words are function words, which are important for structuring and grammar. Meanwhile, content words contain semantic content and are usually open class words (Jurafsky and Martin, 2024).

Open Class		Closed Class	
ADJ	Adjective	ADP	Adposition
ADV	Adverb	AUX	Auxiliary
NOUN	Noun	CCONJ	Coordinating conjunction
VERB	Verb	DET	Determiner
PROPN	Proper noun	NUM	Numeral
INTJ	Interjection	PART	Particle
		PRON	Pronoun
		SCONJ	Subordinating conjunction

Table 2: UPOS tags sorted by open class and closed class words

Following common practice, I randomly shuffled the sentences within each corpus. I saved the output as sentence-delimited JSON files.

3.3 Creation of Cognate List

The following section describes how I determined which words I use for my analysis. First, I compiled a list of English words with Latin origins (either taken into the English language directly from Latin or through French as an intermediary). Next, I identified corresponding Spanish and French descendants based on the Latin roots of these English words. Finally, I reduced the list to include only those cognates that were among the 500 most frequent words in each of the six corpora to ensure reliable analysis.

3.3.1 Identifying English Words with Latin Origin

I determined the following requirements for English words to become a possible target word for the cognate list:

- **Frequency**: The word needs to appear frequently in both English corpora to ensure stable embeddings.

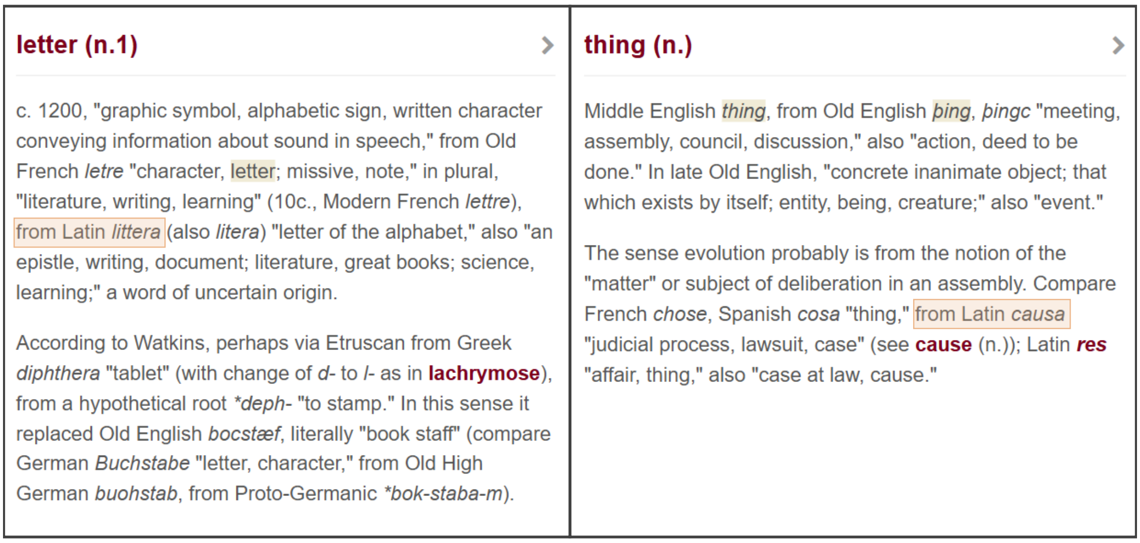
[3] Universal POS tags: https://universaldependencies.org/u/pos/

- **Content Words**: The word has to be a content word. As they carry more meaning than function words, they seem more interesting for semantic change studies.
- **Latin Origin**: The word needs to descend from a Latin ancestor.

After preprocessing the English corpora, I filtered the top 500 most occurring content words from each corpus, filtering by POS tags for open class words (as seen in Table 3.2). I then combined both outputs retaining only words that were among the top frequent words in both corpora. This produced a list of 391 English content words.

For each of the remaining 391 words I made an automatic call to the Online Etymology Dictionary[4], which is openly available. The dictionary would then return the entry found for that word. I searched each entry for the phrase "from Latin," which reliably indicated Latin-derived words. This method captured the word following "from Latin" as the original Latin form, which was then stored alongside the English word and the dictionary entry (for later manual verification).

To test the reliability of this method, I sampled a subset of 100 words from the top English words and manually evaluated it. The accuracy of 90% indicated that the structure of the online dictionary is in fact very consistent and the filtering worked well (full evaluation in Appendix).



letter (n.1)

c. 1200, "graphic symbol, alphabetic sign, written character conveying information about sound in speech," from Old French *letre* "character, letter; missive, note," in plural, "literature, writing, learning" (10c., Modern French *lettre*), from Latin *littera* (also *litera*) "letter of the alphabet," also "an epistle, writing, document; literature, great books; science, learning;" a word of uncertain origin.

According to Watkins, perhaps via Etruscan from Greek *diphthera* "tablet" (with change of *d-* to *l-* as in **lachrymose**), from a hypothetical root **deph-* "to stamp." In this sense it replaced Old English *bocstæf*, literally "book staff" (compare German *Buchstabe* "letter, character," from Old High German *buohstab*, from Proto-Germanic **bok-staba-m*).

thing (n.)

Middle English *thing*, from Old English *þing*, *þingc* "meeting, assembly, council, discussion," also "action, deed to be done." In late Old English, "concrete inanimate object; that which exists by itself; entity, being, creature;" also "event."

The sense evolution probably is from the notion of the "matter" or subject of deliberation in an assembly. Compare French *chose*, Spanish *cosa* "thing," from Latin *causa* "judicial process, lawsuit, case" (see **cause** (n.)); Latin ***res*** "affair, thing," also "case at law, cause."

Figure 4: Two example entries from the Online Etymology Dictionary

Among the sampled 100 words were four words with Latin origin falsely assigned. This was because dictionary entries of non-Latin words would sometimes contain information about related words, synonyms or comparisons to other languages. So,

[4] https://www.etymonline.com/

some originally Germanic or Slavic words would also appear on the filtered Latin word list together with a Latin word that might mean something similar but is not their predecessor.

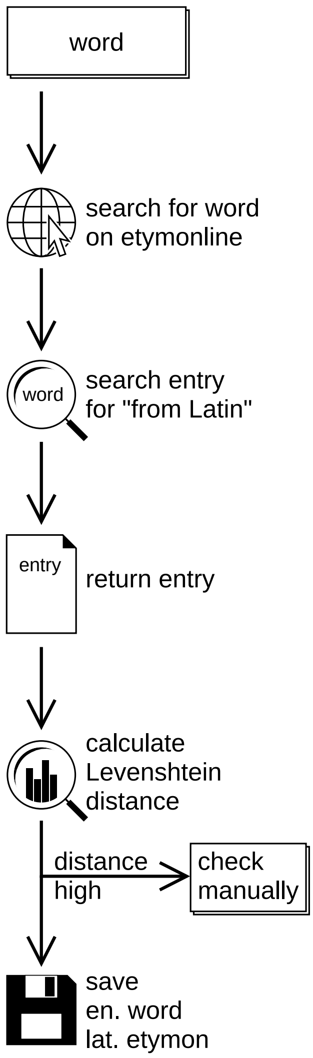


Figure 5: Flow diagram of etymology identification process

In Figure 4 there are two examples taken from the online dictionary. As can be seen in the first one, in the entry about the word *letter*, "from Latin" was found with the following word being the Latin word *littera*. Another word being filtered was *thing*, which, in actuality is of Germanic origin. This happened because its sense evolution was discussed by the author of the entry who compared it to French and Spanish words going back to *causa*. So, the filter returned the English word thing with its claimed Latin root *causa*.

To improve precision, I implemented an additional correction step. First, the Levenshtein distance between the found word and its supposed Latin predecessor was computed. If the Levenshtein distance exceeded my set threshold (the length of the English word minus 1), a notification was displayed (this was also the case for the example of Figure 4: *thing* – *causa*). These words I would manually check and delete from the list if they turned out to not be of Latin descent. I set the Levenshtein distance threshold rather strict, which meant I would have to check more words but it minimized false positives in the list. Not surprisingly, some words scoring a high Levenshtein distance were of Latin descent and had just developed quite far from the original Latin word. (e.g. *people* – *populus*). Those were left in the list. This additional step increased precision greatly. All four false positives from the sample evaluation were filtered out.

Figure 5 outlines my process of etymology detection. These steps produced a list of totally 116 English words with their Latin origin.

3.3.2 Spanish and French Descendants of Latin Words

For the 116 words of the English-Latin list, I identified cognates in Spanish and French. Initial attempts of automatization turned out to be too error-prone and given the small set of target words, I opted to manually curate the French and Spanish lists.

English	Latin	French	Spanish
moment_noun	momentum	moment_noun	momento_noun
people_noun	populus	peuple_noun	pueblo_noun
view_noun	videre	vue_noun	vista_noun

Table 3: Excerpt of curated cognate list

I used Wiktionary[5], a huge open multilingual dictionary, to look up all Latin words from the list. Wiktionary entries offer a list of descendants. From there, I would add the Spanish and French descendant respectively, which was closest in form and meaning to the English descendant from the list. Table 3 shows a few examples of the curated cognate list (full list in Appendix).

I reduced the list further by assessing if every word in a cognate set was among the top 500 most frequent words of the corpora in their language. This final list consisted of 32 cognate sets (see Chapter 5).

[5] https://en.wiktionary.org/

4 Training and Evaluation of Embedding Models

In the previous section, I outlined the creation of my six corpora; an old corpus C_1 (covering 1840 to 1875) and a new corpus C_2 (covering 1885-1920) for each of the three languages: English, French and Spanish. In this chapter, I first explain the architecture I used to train embedding models (Section 4.1), then I show how I evaluated the models and performed hyperparameter tuning (Section 4.2).

4.1 Model Architecture

To train and compare static embeddings, I used skip-gram with negative sampling (SGNS) (as described in Section 2.2.3), aligned the vector spaces with orthogonal Procrustes and measured the semantic change using the cosine similarity. With this procedure, I followed several previous studies (i.e. Hamilton et al. (2016); Schlechtweg et al. (2019); Ehrenworth and Keith (2023).

I trained six embedding models, one for each corpus, using the Gensim library[6] for Python to implement the word2vec SGNS approach. The hyperparameter settings were consistent across models. Below are some of the key hyperparameters:

- **Vector dimensions**: While the number of dimensions of models with sparse vectors equals the length of the vocabulary, dense vectors have much fewer dimensions (see Section 2.2.3), whose number can be set with this hyperparameter. Tahmasebi and Dubossarsky (2023) suggests using between 50 and 300 dimensions.
  While I initially tried using 200 vector dimensions, I found out during hyperparameter tuning, which is described in Section 4.2, that 300 dimensions produced better, more stable results.
- **Window size**: This hyperparameter controls the context which is taken into

[6] https://pypi.org/project/gensim/

account when training a word's embedding. A window size of 3, for example, includes the three preceding and three following words of a target word in each instance.
Word2vec's default window size is 5, but after experimenting with different hyperparameter settings I chose a larger window size of 10. This increase provides more context per word occurrence, although at the same time enlarging the risk of having less precise relationships, as distant co-occurrences are also included.

- **Training epochs**: The number of epochs represents how many times the model cycles through the corpus during training. A higher epoch helps the model learn nuanced relationships but can also increase the risk of overfitting, especially with smaller datasets. A lower epoch number however, may result in insufficient learning and not be enough for the model to capture semantic relationships (Tahmasebi and Dubossarsky, 2023)).
  I applied 5 epochs based on the recommendations of Schlechtweg et al. (2019) and Hamilton et al. (2016).

- **Minimum frequency**: Another common hyperparameter is minimum frequency. This threshold sets the minimum number of occurrences a word must have to be included in the vocabulary. The embeddings of high-frequency words are updated more often in training than those of low-frequency words. If words appear very rarely, they may not have enough training material to produce reliable embeddings, thus it makes sense to exclude them (Tahmasebi and Dubossarsky, 2023).
  I tested different frequency thresholds to analyze their impact on vocabulary size to choose possible hyperparameters and tested these during hyperparameter tuning, finding that models using a minimum occurrence of 5 or 10 performed best. To filter out very rare words but also ensure sufficient training material, I chose a minimum frequency of 5.

After training I aligned the models trained on the corpora of the same language using orthogonal Procrustes, as outlined in Hamilton et al. (2016). For the specific implementation, I used a code snippet[7], which was directly adapted from the original code used in Hamilton et al. (2016) to work with gensim models. This option worked nicely and has been shown multiple times to perform well.

Finally, I calculated semantic change by measuring the cosine similarity between the vectors of the same word in the aligned models.

[7] Original code from Hamilton et al. (2016): https://github.com/williamleif/histwords
Adapted code: https://gist.github.com/zhicongchen/9e23d5c3f1e5b1293b16133485cd17d8

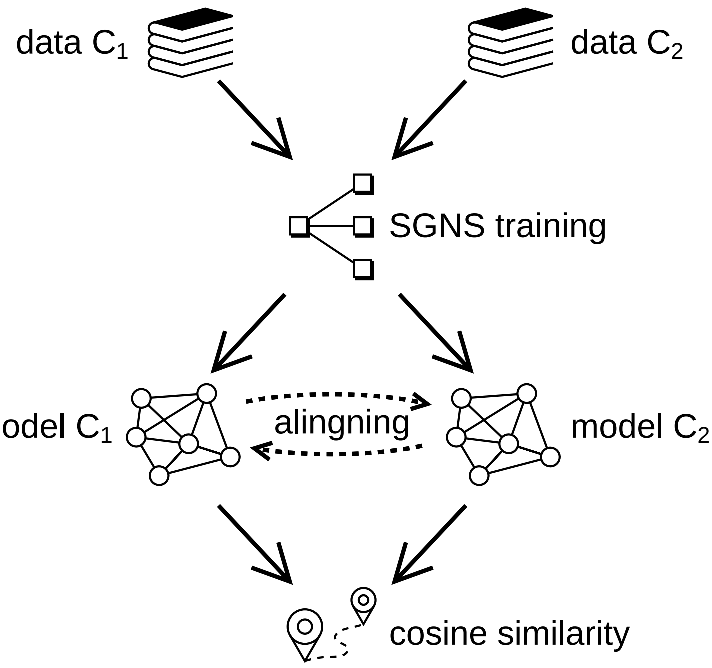


Figure 6: Illustration of model training process

4.2 Evaluation Setup

In the last section, I described my model architecture and the most relevant hyperparameters.

In the following section, I explain how I evaluated my models using a hybrid approach (as suggested by Tahmasebi and Dubossarsky (2023)): I first used an existing dataset and test set to perform hyperparameter tuning and then trained my six corpora with the most successful model configuration, which I evaluated manually.

4.2.1 Automatic Evaluation and Hyperparameter Tuning

For the automatic evaluation, I worked with the English test set provided by Schlechtweg et al. (2020) for the SemEval-2020 shared task 1 (for more details about the test set see Section 2.2.4). I followed the procedure suggested by Schlechtweg et al. (2020): For every word in the test set, I extracted the vectors of the English C_1 and C_2 models and measured the cosine similarity between them. Then, I compared these results with the test set scores provided Schlechtweg et al. (2020) using Spearman's rank correlation. Spearman correlation is a ranked metric which assigns a value between 0 (no correlation) and 1 (identical order). It does not take into account the exact differences in scores but only the rankings themselves. If we had a Spearman correlation of 1 for example, the list with measured cosine distances and the test set would have the words in the exact same order from most to least semantically changed.

Initial evaluation of models trained on ELTeC data

Data	Vector dimension	Window size	Training epochs	Minimum frequency	Spearman correlation (p-value)
ELTeC data (English)	200	5	5	5	0.102 (0.551)

Table 4: Training run with ELTeC data

Initially, I had trained two models on my pre-processed English corpora based on the ELTeC data, as outlined in Section 3.1 (refer to Table 4.2.1 for hyperparameter settings). I performed a Spearman evaluation on these models, as explained above.

This first evaluation, comparing the test set with the models trained on the ELTeC data, led to inconclusive results. This comes as no surprise as the ELTeC corpora the models were trained on are not from the same time period as the corpora provided by the SemEval-2020 Task 1. Thus, the scores in the test set were based on semantic change during a different time period than what the models could capture. Furthermore, the words in the test set were selected based on the datasets of the SemEval-2020 task and many of these words were not common in the ELTeC data.

Evaluation of model architecture when trained on SemEval-2020 Task 1 data

Data	Vector dimension	Window size	Training epochs	Minimum frequency	Spearman correlation (p-value)
SemEval-2020 dataset	200	5	5	5	0.334 (0.018)

Table 5: Training run with SemEval-2020 data

I trained and aligned two new models, this time based on the datasets provided for the SemEval-2020 shared task. With around six million tokens for C_1 and C_2 each, they were of comparable size to my datasets. I repeated the Spearman evaluation between these models and the test set.

Though this approach could not confirm if the training data I used was sufficient, it was still helpful to see that the training setup functioned correctly. Using the same training set-up as in the last section (compare Tables 4.2.1 and 4.2.1), the models reached a Spearman correlation of 0.33 with the test set, meaning a small but significant correlation.

Hyperparameter tuning

Continuing, I trained models with different hyperparameter settings. I tried these settings using recommendations by Tahmasebi and Dubossarsky (2023) and Schlechtweg et al. (2019), as well as following my own intuition:

- Vector dimension: 100, 200, 300
- Window size: 2, 3, 5 and 10
- Minimum frequency: 2, 3, 5, 10

This led to trying out 48 different combinations. I trained every combination twice, one on the older and one on the newer time period of the SemEval-2020 datasets, aligned each pair of models and evaluated them.

Vector dimensions	Context window	Minimum frequency	Spearman correlation with test set	p-value
Best models				
300	10	5	-0.423	0.009
200	5	10	-0.394	0.016
200	10	10	-0.372	0.023
200	5	3	-0.366	0.026
200	10	3	-0.353	0.032
Worst models				
200	2	10	-0.262	0.117
100	10	2	-0.258	0.124
100	2	2	-0.255	0.128
100	10	10	-0.254	0.13
100	3	2	-0.247	0.141

Table 6: Best and worst performing hyperparameter settings

Some hyperparameter settings and their evaluation results can be seen in Table 6 shows the hyperparameter choices of the models with the best and worst performance (full table in Appendix). Training epochs were always 5. The negative correlation is due to the fact higher SemEval test scores indicate larger semantic change while a higher cosine similarity indicates more stability. Thus, we hope for a negative correlation here. Note that the low-scoring models all had a very high p-value increasing the possibility that the little correlation that was measured could be due to chance. In the evaluation, all models with vector dimensions set to 100 performed poorly. Also, a small context window of 2 or 3 impaired performance.

	200, 5, 10	300, 10, 5	200, 10, 10
Run1	-0.394	-0.423	-0.372
Run2	-0.444	-0.436	-0.405
Run3	-0.377	-0.359	-0.315
Run4	-0.334	-0.417	-0.334
Run5	-0.306	-0.390	-0.324
average	-0.371	-0.405	-0.350

Table 7: Evaluation of five training runs with the three best hyperparameter settings

The training algorithm includes random elements that result in different vector configurations in each training run even with identical parameters. As a measure to counteract that, I set the random seed hyperparameter to a fixed number for all trainings. Still, some randomness persisted. Before deciding on my final model settings, I thus made five training iterations with each of the three hyperparameter settings which had previously produced the best results. As can be seen in Table 4.2.1, the Spearman correlation scores vary with around 0.1 difference for every setting from weakest to strongest run. However, all three settings could consistently achieve a weak to moderate correlation with the test set. The hyperparameter settings with the best average performance were: **300 vector dimensions, window size of 10, and a minimum frequency of 5**. I used these settings to train the pre-processed ELTeC corpora and perform several experiments which are described in Section 5.

4.2.2 Manual Screening of Results

In the previous section, I explained how I selected the hyperparameter settings. I used the final settings to train the pre-processed ELTeC corpora. While I did not have a test set specifically adapted to the corpora I was working with, I conducted a small screening of the trained embeddings, an excerpt of which is presented here. From the list of the 500 most frequent words appearing in both corpora, I extracted the ten words with the lowest cosine similarity, representing the largest measured semantic changes. I first checked whether the words were just frequent in very few novels which might have impacted the embeddings strongly but in fact all ten words appeared regularly in more than 70I looked up the ten English words in the Oxford English Dictionary (OED)[8], which not only provides an overview of all known meanings for a word but also estimates for when a meaning first appeared,

[8] https://www.oed.com/

respectively when a meaning became obsolete. For every word, I assessed how many new and obsolete meanings the OED mentions during the time period covered by my corpora (between 1840 and 1920). These observations served as indicators of whether a word could have experienced a semantic shift. Table 8 summarizes this evaluation and indicates whether I deem a semantic change during the measured time period probable or not.

Semantic change probable	unclear	Semantic change improbable
uncle_noun (3 new meanings, 1 obsolete)	*continue_verb* (2 new meanings)	*sometimes_adv*
party_noun (3 new meanings, 4 obsolete)	*companion_noun* (1 new meaning)	*thus_adv*
real_adj (3 new meanings, 1 obsolete)		*certain_adj*
		half_adj
		also_adv

Table 8: Manual analysis of English words with largest measured semantic shift

Sometimes looking at the words with the closest embeddings to a target word can be an indicator of meaning change. Looking thus at the nearest neighbours sharing the same POS tag, both words for which I assessed “unclear” semantic change appear to have gained a more negative connotation between model C_1 and model C_2 (see Table 9).

Interestingly, two words for which semantic change was assessed as “improbable” - *also* and *sometimes* - were among the ten most polysemous words identified in a study by Hamilton et al. (2016). This raises the question of whether highly polysemous words tend to exhibit lower cosine similarity due to their occurrence in more varied contexts during model training.

Word	Nearest neighbours in model C_1	Nearest neighbours in model C_2
continue_verb	add, answer, say, resume, enquire	upbraid, afflict, displease, interpose, query
companion_noun	diversion, friend, playmate, associate, schoolfellow, protégé, betrothed	questioning, interlocutor, aversion, calmness, chaperon, discomposure

Table 9: Example words and nearest neighbours with same part-of-speech

4.2.3 Limitations

The previously described setup and evaluation have several limitations that should be noted.

As documented in Section 4.2, the hyperparameter tuning and evaluation were performed automatically. While evaluation on the gold standard yielded promising results, this only demonstrates that the model setup performs as expected when trained on the English SemEval-2020 data. It does not guarantee similar performance when applied to my pre-processed ELTeC data. One significant difference between the SemEval-2020 training set and the ELTeC data is the time scope: the SemEval dataset includes a time gap of 100 years between the older and newer corpus and only the time span of the older partially overlaps with the ELTeC data.

Moreover, the model was only automatically evaluated on English data, not on French or Spanish. This was primarily due to the difficulty of obtaining accessible gold standard and training sets which fit the scope of this work.
Manual evaluation was also conducted more extensively on English words. Again, one factor was available resources, the OED providing an impressive service for historical linguistics, for which I could find no comparable tool in Spanish or French. Another factor was certainly my own limitations: assessing semantic changes in words from 200 years ago over a relatively short time span proved to be a challenging task for which I could not identify an adequate solution.

Lastly, the stochastic nature of model training introduces variability. As shown in Table 4.2.1, even with identical initial conditions, different hyperparameter settings emerged as optimal across several training runs.

5 Experiments and Results

In the previous two chapters, I showed how I assembled datasets, identified the words I wanted to evaluate and trained models. In the following section, I describe the experiments conducted and their results, using these models and data.

Experiment 1: Comparing change in English words with Latin roots to those without Latin roots

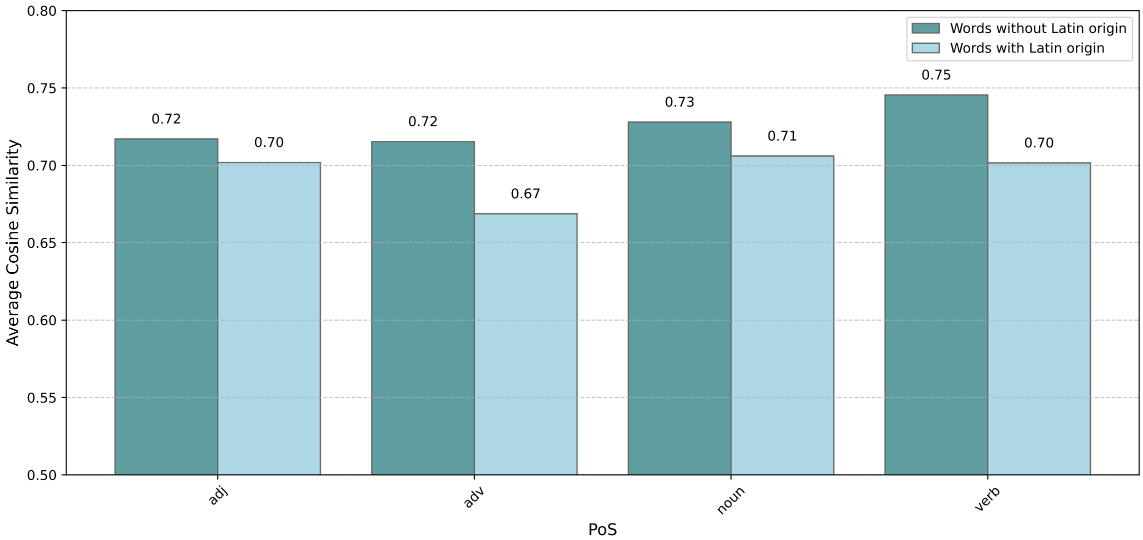


Figure 7: Average cosine similarity of open class parts-of-speech

In the first experiment, I investigated how much the vector embeddings of English words with Latin roots changed from model C_1 to model C_2 compared to English words without Latin roots. I worked with the list containing the most frequent words among both English corpora (comprising a total of 391 words). Of this list, 116 English words had previously been identified to contain Latin roots, as presented in Section 3.3.1. Proper nouns were excluded from this experiment as they exhibited very low cosine similarity and seemed strongly influenced by the novels they occurred in. Additionally, words represented in fewer than 60% of all novels in either corpus C_1 or corpus C_2 were excluded as well to ensure that the embeddings were not overly skewed by frequent appearances in a small subset of novels. I then calculated the cosine similarity for each word.

Overall, words with Latin origin achieved a cosine similarity of 0.70, which is 0.03 lower than the average of words without Latin origin (0.73).

I further examined the ten most and ten least stable English words with and without Latin origin (full table in Appendix). While nine out of ten of the most stable words without Latin origin turned out to be verbs, there was only one verb among the most stable words with Latin origin. In Figure 7, I compiled the mean cosine similarities by part-of-speech, a higher mean indicating more stable meaning across the two time periods.

Experiment 2: Measuring change across cognate sets

lat word	en word	en cos	fr word	fr cos	es word	es cos
continuare	continue_verb	0.595	continuer_verb	0.726	continuar_verb	0.643
repetere	repeat_verb	0.644	répéter_verb	0.690	repetir_verb	0.750
familia	family_noun	0.645	famille_noun	0.727	familia_noun	0.677
spiritus	spirit_noun	0.650	esprit_noun	0.748	espíritu_noun	0.741
aer	air_noun	0.651	air_noun	0.797	aire_noun	0.697
ordinem	order_noun	0.666	ordre_noun	0.680	orden_noun	0.599
persona	person_noun	0.668	personne_noun	0.820	persona_noun	0.736
praesentia	presence_noun	0.671	présence_noun	0.731	presencia_noun	0.672
effectus	effect_noun	0.682	effet_noun	0.700	efecto_noun	0.696
status	state_noun	0.684	état_noun	0.751	estado_noun	0.702
videre	view_noun	0.691	vue_noun	0.685	vista_noun	0.703
idea	idea_noun	0.698	idée_noun	0.738	idea_noun	0.720
partem	part_noun	0.705	part_noun	0.750	parte_noun	0.807
pungo	point_noun	0.708	point_noun	0.752	punto_noun	0.771
offerre	offer_verb	0.710	offrir_verb	0.730	ofrecer_verb	0.708
recipere	receive_verb	0.713	recevoir_verb	0.785	recibir_verb	0.766
possibilis	possible_adj	0.715	possible_adj	0.748	posible_adj	0.709
passus	pass_verb	0.720	passer_verb	0.855	pasar_verb	0.831
intrare	enter_verb	0.728	entrer_verb	0.817	entrar_verb	0.775
dubitare	doubt_noun	0.728	papier_noun	0.772	duda_noun	0.741
papyrus	paper_noun	0.729	doute_noun	0.761	papel_noun	0.704
rationem	reason_noun	0.734	raison_noun	0.739	razón_noun	0.696
securus	sure_adj	0.737	sûr_adj	0.737	seguro_adj	0.722
quiritare	cry_verb	0.748	crier_verb	0.749	gritar_verb	0.776
historia	story_noun	0.753	histoire_noun	0.756	historia_noun	0.744
sortem	sort_noun	0.758	sorte_noun	0.664	suerte_noun	0.623
silentium	silence_noun	0.769	silence_noun	0.795	silencio_noun	0.803
stare	stay_verb	0.775	être_verb	0.828	estar_verb	0.823
tonus	tone_noun	0.779	ton_noun	0.752	tono_noun	0.703
vocem	voice_noun	0.784	voix_noun	0.848	voz_noun	0.811
momentum	moment_noun	0.788	moment_noun	0.839	momento_noun	0.798
hora	hour_noun	0.795	heure_noun	0.879	hora_noun	0.776

Table 10: List of 32 cognate sets and their cosine similarity, color-coded by rank of change in each language

In the second experiment, I worked with the list of 32 cognates, where each word appeared frequently in the corpus of their respective language (see Section 3.3.2). I measured the cosine similarity between the embeddings of the two models for each word. Note that these results in Table 5 are not directly comparable because the models have not been aligned cross-lingually.

Language pairs	Spearman's correlation coefficient	p-value
fr - en	0.488	0.0053
en - es	0.469	0.0077
es - fr	0.708	$8.21 \cdot 10^{-6}$

Table 11: Spearman Correlation between language pairs based on Table 10

To enable a meaningful comparison between the languages I used Spearman's rank correlation for each language pair. While the cosine similarity scores cannot be compared directly across languages, the Spearman Correlation reveals the extent to which cognates in various languages have undergone similar changes. The Spearman correlation scores and p-values in Table 5 indicate a significant correlation between all language pairs. Notably, the measurements for French and Spanish show the highest Spearman correlation.

Experiment 3: Evaluating the average change across parts-of-speech

For the evaluation of part-of-speech categories, I compiled a new list comprising the intersection of the 1000 most frequent words in both corpora for each language. This time, all parts-of-speech were included.

I computed the cosine similarities for all words and calculated the mean cosine similarity for every part-of-speech. This process was repeated for all three languages. Table 12 provides an overview of the number of words analyzed for each part-of-speech and their corresponding average cosine similarities. It is evident that content words are much more numerous, with 81% of the most frequent words in English tagged as open class parts of speech. While there are much fewer function words in a language, they tend to occur much more frequently; for instance, 52% of all word occurrences in the two English corpora were classified as closed class words. Figures 8, 9 and 10 illustrate the results for each language, with colour separation for open class (primarily content) words and closed class (most function) words. On average, closed class words are more stable across the two corpora in each language: 7.6% more stable in English, 7.2% more stable in Spanish, and 8.5% more stable in French.

	English		Spanish		French	
	Word types	Ø cosine	Word types	Ø cosine	Word types	Ø cosine
Open class words						
adj	110	0.699	84	0.705	69	0.737
adv	82	0.709	49	0.720	79	0.759
intj	5	0.725	5	0.693	1	0.785
noun	253	0.703	280	0.722	341	0.734
propn	26	55	14	0.602	10	0.607
verb	182	0.721	185	0.734	240	0.754
total	658	0.684	617	0.696	740	0.729
Closed class words						
adp	38	0.761	21	0.785	30	0.815
aux	14	0.8	7	0.820	4	0.843
cconj	7	0.741	9	0.773	7	0.783
det	18	0.767	28	0.766	17	0.834
num	10	0.754	11	0.745	10	0.790
part	4	0.755	1	0.712		
pron	44	0.765	29	0.765	38	0.823
sconj	24	0.74	9	0.775	8	0.816
total	159	0.760	115	0.717	114	0.815

Table 12: Individual word counts and average cosine similarity among the top 1000 most frequent words for each part-of-speech

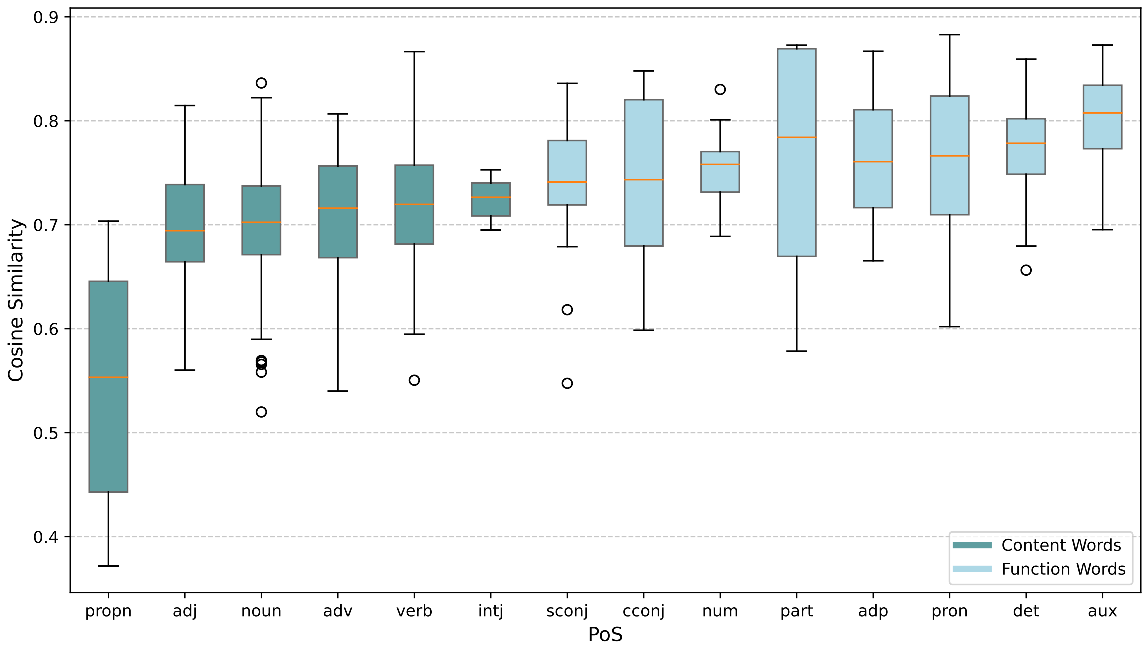


Figure 8: Part-of-speech tags cosine similarity distribution: English

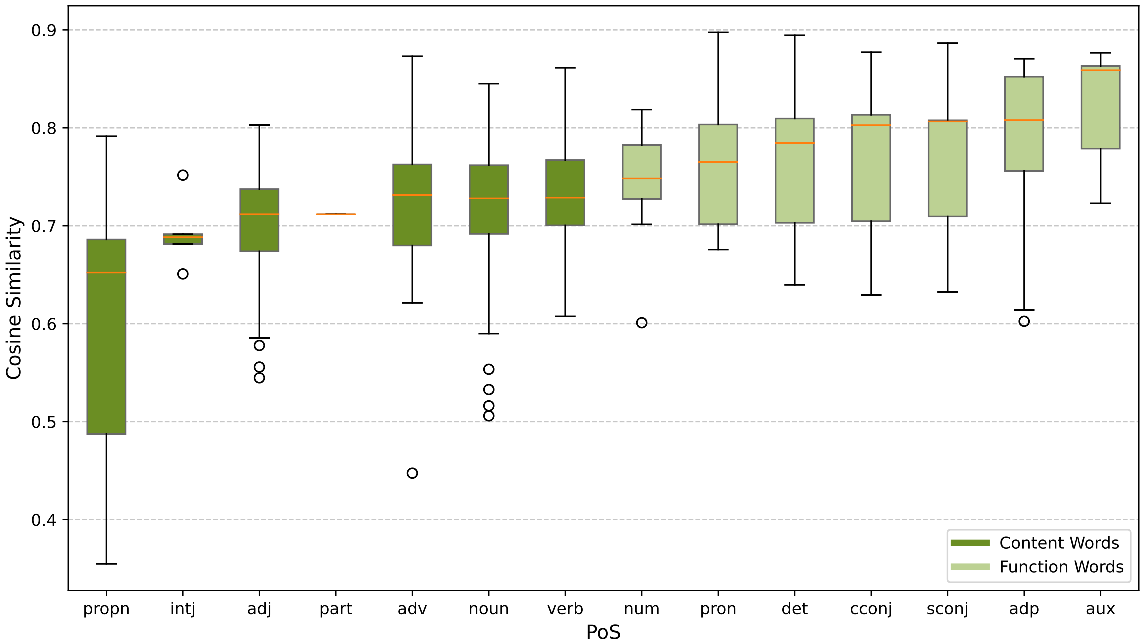


Figure 9: Part-of-speech tags cosine similarity distribution: Spanish

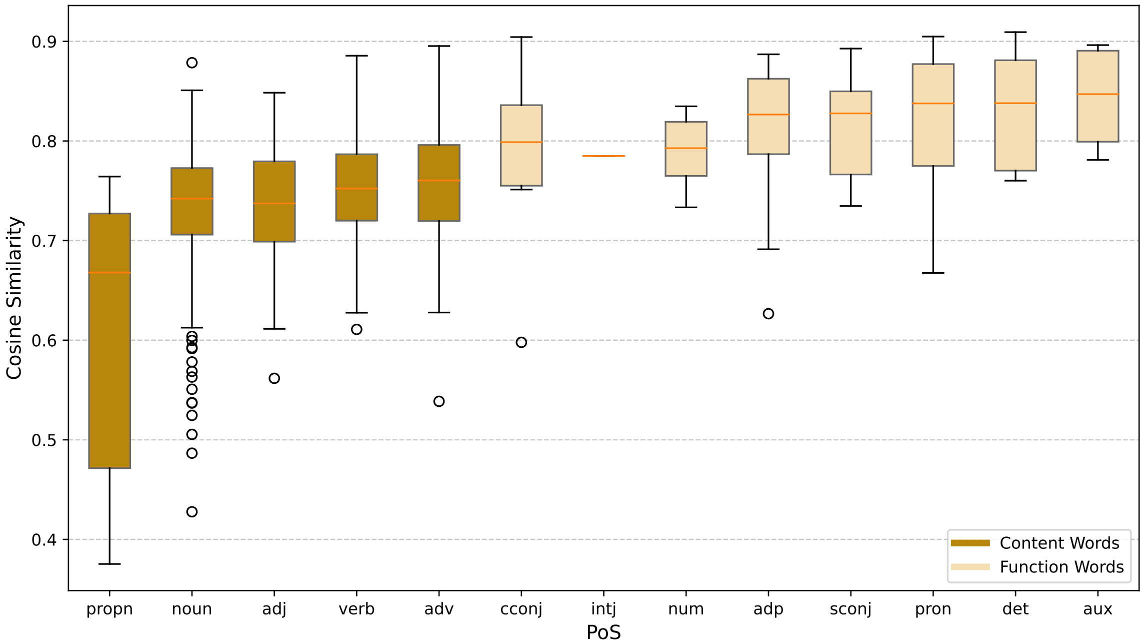


Figure 10: Part-of-speech tags cosine similarity distribution: French

6 Discussion

At the beginning of this thesis, three questions were posed:

- Do English words of Latin origin undergo a larger semantic change during the 19th century than English words without Latin roots?
- How does the change in meaning of English words with Latin roots compare to the semantic change of their French and Spanish counterparts?
- Which part-of-speech categories exhibit the largest change in meaning and can cross-lingual differences be observed?

Subsequently, I discuss these questions based on the results in Chapter 5.

Do English words of Latin origin undergo a larger semantic change during the 19th century than English words without Latin roots?)

Results from the experiment revealed that English words of Latin origin exhibited a 3% lower average cosine similarity than other English words, suggesting that they underwent slightly greater semantic change during the measured time period. While this percentage is small, a t-test confirmed high statistical significance (*p-value* < 0.05). The difference was particularly pronounced in verbs, with a 5% larger change on average (though keep in mind the small sample size of $n = 26$ for verbs with Latin origin). Surprisingly, among the words with Latin roots, verbs exhibited larger change than nouns and adjectives, even though verbs overall demonstrated the most stability among all content words (see Table 12).

An important question remains: are the measured embedding distances reliable indicators of semantic change? As explained in Section 4.2, conducting a large-scale evaluation was beyond the scope of this work. However, in Section 4.2.2, I analyzed the ten English words with the largest measured change. Of these, six were of Latin origin, and four were not.
For the four English words without Latin roots that showed the largest measured changes, the evaluation indicated no apparent semantic change. Three of these words were adverbs, and two were among the most polysemous words identified in

the Hamilton et al. (2016) study. Thus, it is possible that polysemous words have skewed the results in this study.
Among the six English words of Latin origin with the largest measured changes, three demonstrated strong indicators of semantic change, two were unclear, and for one word, semantic change was deemed improbable.

How does the change in meaning of English words with Latin roots compare to the semantic change of their French and Spanish counterparts?

Due to limitations of corpus size and the requirement for each word in a cognate set to be frequent among the corpora of the same language, an analysis of this question could only be conducted on a small scale.

Experiment 2 showed that the measured cosine similarities of the 32 analysed cognates exhibited a similar change pattern in French and Spanish with a high Spearman correlation. This leads to the conclusion, that in this experiment, French and Spanish words have shown to change in a more similar direction than their English cognates. However, it is not possible to definitively infer whether the English words underwent more or less change in absolute terms. Since the models were not aligned cross-lingually, comparing cognate similarity scores directly is not possible.

An interesting aspect of this experiment was the closer examination of certain cognates, particularly those where one word in a cognate pair showed a significantly different change compared to the others.

For example, in the colour-coded Table 5, the English word *sort* stands out as more stable compared to its French and Spanish counterparts. Now, interestingly from previous examination I had expected Spanish *suerte* to be an outlier. The English word sort seems to have retained its meaning of ”order” and frequently appears in the old and new corpora in the phrase sort of (as “in a manner of speaking, quasi”). Meanwhile, Spanish *suerte* in the new corpus frequently appears in the context of “luck”, which is its today’s primary meaning (compare example in Table 6. It having received a low similarity score could thus have been expected. For French *sorte* though, I could not find anything among example sentences of the corpus, closest embeddings or in the dictionary that could confirm the measured semantic shift.

Another English word which caught my eye is *spirit*. But this time because it was measured to be much less stable than its Romance cognates. This word has a wide range of meanings documented in the OED, with four of them becoming obsolete during the time of my corpora. Standing out were the multiple appearances of *spirit*

word	example C_1 (1840-1875)	example C_2 (1885-1925)
spirit (en)		'Dr Burrows won't listen to me': 'I tell him how I dislike the taste of spirits, but he says they are absolutely necessary for my constitution': 'my medical man insists on something at bedtime'; that's the style.
suerte (es)	De esta suerte me explico que mi padre se descuide, y no recele que, hasta a pesar mío, pudiera tener un rival en mí.	"Ánimo, chico, que hoy te va a sonreír la suerte."

Table 13: Example of semantic shift in word *suerte* and newly appearing meaning in word *spirit*

in the context of liquor in the new corpus (such as the example in Table 6. While not a new meaning per se I, could not find a mention of *spirit* in this context in the older corpus.

Which part-of-speech categories exhibit the largest change in meaning and can cross-lingual differences be observed?

Across all three languages, closed class part-of-speech showed more stable embeddings than open class part-of-speech. This finding affirms the already established notion that function words tend to be more stable than content words (compare Section 3.2).

Among the parts of speech, proper nouns showed the largest measured change in meaning across all three languages. Notably, about a third of all proper nouns in this evaluation were names of characters. As every *Julia* and every *Pierre* will have different characteristics in each novel they appear in, it is expected that their embeddings will differ accordingly. And in effect, as shown in Table 6, the most changed proper nouns in each language were character names. However, the most stable proper nouns were among the most stable overall nouns in their respective languages, leading to a large divergence within this part of speech. There also appear to be a few words wrongly tagged as proper nouns, especially in the French and Spanish results (such as the word *oui*, which achieved the second highest stability among French proper nouns).

No significant differences were found across the examined languages. In this study on 19th century novels, the parts-of-speech show a similar change pattern across all three languages, French, Spanish and English.

Spanish		French		English	
Proper nouns with lowest cosine similarity					
miguel_propn	0.355	louis_propn	0.375	esther_propn	0.372
alfonso_propn	0.447	jacques_propn	0.440	julia_propn	0.388
luis_propn	0.456	pierre_propn	0.470	margaret_propn	0.394
Proper nouns with highest cosine similarity					
dios_propn	0.791	dieu_propn	0.762	london_propn	0.703
don_propn	0.711	oui_propn	0.729	god_propn	0.681
españa_propn	0.699	paris_propn	0.722	england_propn	0.672

Table 14: Proper nouns with highest and lowest cosine similarity among most frequent words

7 Conclusion

7.1 Summary

Semantic change detection with word embeddings has been an expanding field of study in recent years. However, many studies in this field primarily focus on English data.

In this work, I employed a Skip Gram with Negative Sampling model, aligned using Orthogonal Procrustes and applied cosine similarity as a metric, following methodologies established in previous studies (Hamilton et al. (2016); Schlechtweg et al. (2019)). The hyperparameters for the model were set to 300 vector dimensions, 5 training epochs, a window size of 10, and a minimum word frequency of 5, after performing hyperparameter tuning.

The experiments in this study introduced a novel approach by comparing semantic change over time and across languages. I trained six word embedding models for three languages – English, French and Spanish – across two time periods (1840-1875 and 1885-1920). While I followed standard approaches, lemmatizing the words and shuffling the sentences in each corpus, I also trained the words together with their part-of-speech to ensure separate embeddings for words sharing the same lemma (e.g. *start_noun* and *start_verb*). To identify English words with Latin origins, I used an automated approach that achieved 90% accuracy. Moreover, I manually curated a list of cognate sets across the three languages.

The results indicated that English words with Latin roots experienced a 3% larger change compared to non-Latin words. However, examining individual words did not fully confirm these measured changes.
Among the examined cognate sets, cosine similarity measurements revealed a strong Spearman correlation (0.71) between French and Spanish compared to lower correlations between English–Spanish (0.47) and English–French (0.49). This suggests that French and Spanish cognates have undergone more similar semantic changes than their English counterparts. While the results were statistically significant, they were based on a small dataset of only 32 cognate sets.

The part-of-speech evaluation revealed that part-of-speech categories show similar changes across all three measured languages. Function words were more stable than content words and proper nouns exhibited the largest measured change.

The models often produced results in line with previous studies, as could be notably seen in the part-of-speech evaluation. The results are encouraging to dive deeper into the topic and replicate the study on different datasets.

7.2 Outlook

This study explored how word embeddings could be used for cross-lingual semantic change evaluations. While the results provided some interesting insights, it is clear that further research is needed. Below, I address some points which could be considered in the future.

- **Larger Time Gaps Between Corpora**: A clearer and more definitive manual evaluation could be achieved by choosing corpora with a larger time gap. Many previous studies have used corpora separated by one or two centuries, which may provide a more distinct comparison of semantic change over time.
- **Using More Recent Data**: Instead of extending the time gap, another possibility would be to work with corpora that contain more recent data. Not only is much more data available, but semantic change may be easier to interpret with current world knowledge if it happened in recent years.
- **Additional Time Bins**: In a future study, I would try out more time bins. This could make it possible to quantify the rate of semantic change for different words by looking at the displacement between consecutive time points. However, more data would be needed than in this study, ensuring sufficient training material for each time bin.
- **Expanding the Range of Cognates and Borrowings**: In future research, it would be valuable to include more Romance cognates and borrowings in other languages being added to the evaluation, as well as to explore cognate sets across different language families.
- **Contextual Models and LLM Prompting**: As mentioned in Section Background, recent advancements in contextual models and LLM prompting hold significant potential for improving semantic change detection. It would be interesting to repeat the experiments of this study with more recently developed models.

While computational approaches to semantic change detection show great promise, further exploration is needed to grasp their full potential. Meanwhile, there is still a lack of standardized testing methodologies that can be adapted to a wide range of corpora. This limitation makes it difficult to directly compare different methods and evaluate their relative effectiveness.

Nonetheless, the potential of these computational tools seems vast. They offer the ability to investigate language at an unprecedented scale, opening up new possibilities for linguistic research and helping us explore the dynamic nature of language over time.

References

Andreas Blank. 2001. *Einführung in die lexikalische Semantik für Romanisten.* Niemeyer, Tübingen.

Lyle Campbell. 1998. *Historical Linguistics: An Introduction.* Edinburgh University Press.

Jackson Ehrenworth and Katherine Keith. 2023. Literary intertextual semantic change detection: Application and motivation for evaluating models on small corpora. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 1–14, Singapore. Association for Computational Linguistics.

JR Firth. 1957. A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis, Special Volume/Blackwell.*

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Tamara M. Green. 2014. *The Greek & Latin Roots of English.* Rowman & Littlefield Publishers.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. 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)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2024. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released August 20, 2024.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013.

Ines Montani, Matthew Honnibal, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. explosion/spaCy: v3.7.2: Fixes for APIs and requirements.

Francesco Periti, Haim Dubossarsky, and Nina Tahmasebi. 2024. (chat)GPT v BERT dawn of justice for semantic change detection. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 420–436, St. Julian's, Malta. Association for Computational Linguistics.

Francesco Periti and Stefano Montanelli. 2024. Lexical semantic change through large language models: a survey. *ACM Comput. Surv.*, 56(11).

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Christof Schöch, Roxana Patras, Tomaž Erjavec, and Diana Santos. 2021. Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives. *Modern Languages Open*, (1):25.

Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. 2019. Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*

(EMNLP-IJCNLP), pages 66–76, Hong Kong, China. Association for Computational Linguistics.

Joseph B. Solodow. 2010. *Latin Alive: The Survival of Latin in English and the Romance Languages*. Cambridge University Press.

Nina Tahmasebi and Haim Dubossarsky. 2023. Computational modeling of semantic change. *Preprint*, arXiv:2304.06337.

Ana Sabina Uban, Alina Maria Cristea, Anca Dinu, Liviu P. Dinu, Simona Georgescu, and Laurentiu Zoicas. 2021. Tracking semantic change in cognate sets for English and Romance languages. In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 64–74, Online. Association for Computational Linguistics.

Ruiyu Wang and Matthew Choi. 2023. Large language models on lexical semantic change detection: An evaluation. *Preprint*, arXiv:2312.06002.

A Additional Tables

Example Sentences from Corpora

Working with these huge corpora consisting of millions of words, it is sometimes easy to only see numbers and percentages and forget about the treasure of content hidden among the data. Here is a collection of some nuggets I stumbled upon during my investigations, which made me appreciate 19th century literature a little more.

- It isn't worth extinguishing oneself in order to make a world for those others, anyhow.
- "It is sad to suppose that places may ever be more important than people," continued Margaret.
- No soy poeta, continuó Medina sonriéndose; pero he tenido tambien mis dias de gloria, y sé estimar en lo que valen los ensueños del dia despues.
- Il ressentait peu à peu une sorte de reconnaissance pour l'aimable femme qui sacrifiait si simplement sa réputation aux yeux du seul homme à l'opinion duquel elle tenait au monde.
- "By them," continued Corbulo, disregarding the interruption, "we must shape our course, and this true shaping of our course, and not drifting with tides, or blown hither and thither by winds—this is the seamanship of life."
- My education, it is just to say, was neglected by no one but myself.
- I never say what I am afraid of having repeated.

And some fun ones...

- "I'll make you a proposal, if you like," continued the Marquis, grinning like an antiquated goat.
- "You may report to your Government that the British youth of the present day, hot from the university, are very often prigs."
- He had lived nearly seven years at Petershof, and, like many others was obliged to continue staying there if he wished to continue staying in this planet.
- Un homme en spencer, en 1844, c'est, voyez-vous, comme si Napoléon eût daigné ressusciter pour deux heures.
- I regard the whole university system as a wretched sham.

find_verb	TN	word_noun	FP	chance_noun	FN
whole_adj	TN	course_noun	TP	idea_noun	TP
offer_verb	TP	remain_verb	TP	fall_verb	TN
know_verb	TN	glad_adj	TN	one_noun	TN
able_adj	TP	enough_adv	TN	lie_verb	TN
happen_verb	TN	lip_noun	TN	fine_adj	TP
presence_noun	TP	work_noun	TN	step_noun	TN
grow_verb	TN	water_noun	TN	world_noun	TN
lord_propn	TN	fact_noun	TP	mind_noun	TN
beautiful_adj	X	suppose_verb	TP	learn_verb	TN
present_adj	TP	soon_adv	TN	want_verb	TN
reason_noun	TP	break_verb	TN	true_adj	TN
house_noun	TN	short_adj	TN	miss_propn	TN
truth_noun	TN	look_verb	TN	much_adv	TN
large_adj	TP	home_adv	TN	begin_verb	TN
open_adj	TN	’_verb	FP	look_noun	TN
evening_noun	TN	even_adv	TN	hand_noun	TN
age_noun	FN	son_noun	TN	deal_noun	TN
keep_verb	TN	make_verb	TN	servant_noun	FN
life_noun	TN	effect_noun	TP	give_verb	TN
lady_noun	TN	sort_noun	TP	possible_adj	TP
first_adj	TN	change_verb	FN	real_adj	TP
back_adv	TN	question_noun	TP	arm_noun	TN
glass_noun	TN	bear_verb	TN	god_propn	TN
alone_adv	TN	trouble_noun	TP	girl_noun	FP
wonder_verb	TN	right_adj	X	week_noun	TN
thought_noun	TN	pay_verb	TP	stand_verb	FP
perhaps_adv	TN	leave_verb	TN	understand_verb	TN
party_noun	TP	marriage_noun	TP	pass_verb	TP
consider_verb	TP	sit_verb	TN	repeat_verb	TP
bed_noun	TN	come_verb	TN	letter_noun	TP
day_noun	TN	window_noun	TN	reply_verb	FN
certain_adj	FN	door_noun	TN	death_noun	TN
live_verb	TN				

True positives: 25
True negatives: 63
False positives: 4
False negatives: 6
Precision: 0.862
Recall: 0.806
Accuracy: 0.898
f-Measure: 0.833

Table 15: Evaluation of sampled 100 English words to assess automatic etymology identification.

face_noun	facies	faciès_noun	haz_noun
people_noun	populus	peuple_noun	pueblo_noun
turn_verb	tornare	tourner_verb	tornar_verb
moment_noun	momentum	moment_noun	momento_noun
voice_noun	vocem	voix_noun	voz_noun
place_noun	platea	place_noun	plaza_noun
pass_verb	passus	passer_verb	pasar_verb
hour_noun	hora	heure_noun	hora_noun
cry_verb	quiritare	crier_verb	gritar_verb
sort_noun	sortem	sorte_noun	suerte_noun
poor_adj	pauper	pauvre_adj	pobre_adj
suppose_verb	supponere	supposer_verb	suponer_verb
course_noun	cursus	course_noun	curso_noun
remember_verb	rememorari	remémorer_verb	rememorar_verb
table_noun	tabula	table_noun	tabla_noun
letter_noun	littera	lettre_noun	letra_noun
part_noun	partem	part_noun	parte_noun
fact_noun	factum	fait_noun	hecho_noun
sure_adj	securus	sûr_adj	seguro_adj
matter_noun	materia	matière_noun	madera_noun
sense_noun	sensus	sense_noun	sentido_noun
air_noun	aer	air_noun	aire_noun
question_noun	quaestionem	question_noun	cuestión_noun
marry_verb	marītāre	marier_verb	maridar_verb
money_noun	moneta	monnaie_noun	moneda_noun
move_verb	movere	mouvoir_verb	mover_verb
minute_noun	minuta	minute_noun	minuto_noun
idea_noun	idea	idée_noun	idea_noun
strange_adj	extraneus	étrange_adj	extraño_adj
large_adj	largus	large_adj	largo_adj
appear_verb	apparere	appareiller_verb	aparar_verb
person_noun	persona	personne_noun	persona_noun
chair_noun	cathedra	chaise_noun	cadera_noun
catch_verb	captare	capter_verb	captar_verb
silence_noun	silentium	silence_noun	silencio_noun
case_noun	casus	cas_noun	caso_noun
expect_verb	expectare/exspectare	expectative_adj	expectar_verb
point_noun	pungo	point_noun	punto_noun
add_verb	addere	!	añadir_verb
reason_noun	rationem	raison_noun	razón_noun
carry_verb	carrum	charger_verb	cargar_verb
wall_noun	vallum	!	valla_noun
tone_noun	tonus	ton_noun	tono_noun
stay_verb	stare	être_verb	estar_verb
possible_adj	possibilis	possible_adj	posible_adj

story_noun	historia	histoire_noun	historia_noun
paper_noun	papyrus	papier_noun	papel_noun
nature_noun	natura	nature_noun	natura_noun
human_adj	humanus	humain_adj	humano_adj
family_noun	familia	famille_noun	familia_noun
country_noun	contra	contrée_noun	contra_noun
enter_verb	intrare	entrer_verb	entrar_verb
remain_verb	remanere	remanoir_verb	remanir_verb
continue_verb	continuare	continuer_verb	continuar_verb
clear_adj	clarus	clair_adj	claro_adj
different_adj	differentem	différent_adj	diferente_adj
dinner_noun	dis+ieiuno	dîner_noun	!
power_noun	potis	pouvoir_noun	poder_noun
spirit_noun	spiritus	esprit_noun	espíritu_noun
pay_verb	pacare	payer_verb	pagar_verb
able_adj	habilem	habile_adj	hábil_adj
position_noun	positionem	position_noun	posición_noun
allow_verb	allocare	allouer_verb	alocar_verb
round_adv	rotundus	rond_adv	redondo_adv
silent_adj	silentium	*silence_noun	*silencio_noun
real_adj	res	réel_adj	real_adj
exclaim_verb	exclamare	exclamer_verb	exclamar !
conversation_noun	conversationem	conversation_noun	conversación_noun
pleasure_noun	placere	plaisir_noun	placer_noun
impossible_adj	impossibilis	impossible_adj	imposible_adj
order_noun	ordinem	ordre_noun	orden_noun
save_verb	salvus	sauver_verb	salvar_verb
repeat_verb	repetere	répéter_verb	repetir_verb
view_noun	videre	vue_noun	vista_noun
fine_adj	finis	fin_adj	fino_adj
uncle_noun	avunculus	oncle_noun	!
note_noun	notare	note_noun	nota_noun
subject_noun	subjectum	sujet_noun	sujeto_noun
receive_verb	recipere	recevoir_verb	recibir_verb
usual_adj	usus+alis	usuel_adj	usual_adj
marriage_noun	maritatus	mariage_noun	maridaje_noun
trouble_noun	turba	troubler_verb	!
companion_noun	com+panis	compagnon_noun	compañero_noun
party_noun	partire/partiri	partir_verb	partir_verb
present_adj	praesentem	présent_adj	presente_adj
beauty_noun	bellus	beauté_noun	belleza_noun
consider_verb	considerare	considérer_verb	considerar_verb
presence_noun	praesentia	présence_noun	presencia_noun
quiet_adj	quies	quiet_adj	quieto_adj
offer_verb	offerre	offrir_verb	ofrecer_verb

duty_noun	debitus	!	!
form_noun	forma	forme_noun	forma_noun
scene_noun	scaena	scène_noun	escena_noun
natural_adj	naturalis	naturel_adj	natural_adj
character_noun	character	caractère_noun	carácter_noun
effect_noun	effectus	effet_noun	efecto_noun
pain_noun	poena	peine_noun	pena_noun
state_noun	status	état_noun	estado_noun
doubt_noun	dubitare	doute_noun	duda_noun
change_verb	cambire	changer_verb	cambiar_verb
age_noun	aetatem	âge_noun	edad_noun
chance_noun	cadens	chance_noun	cadencia_noun
servant_noun	servus	servant_noun	siervo_noun
reply_verb	*replicare	replier_verb	replicar_verb
certain_adj	certus	certain_adj	cierto_adj

Table 16: Curated list of 116 cognate sets based on the most frequent English words of Latin origin.'!' markers indicate that no fitting descendant for French or Spanish was found based on the Latin etymon and the English descendant.

Vector dimension	Context window	Minimum frequency	Spearman correlation with test set	p-value
300	10	5	-0.423	0.009
200	5	10	-0.394	0.016
200	10	10	-0.372	0.023
200	5	3	-0.366	0.026
200	10	3	-0.353	0.032
300	5	10	-0.325	0.049
100	3	10	-0.321	0.053
100	5	10	-0.317	0.056
200	10	5	-0.317	0.056
300	10	10	-0.314	0.059
300	2	5	-0.310	0.062
100	2	3	-0.309	0.063
100	5	3	-0.308	0.063
300	2	10	-0.308	0.064
100	3	3	-0.305	0.031
100	3	5	-0.304	0.067
200	5	5	-0.300	0.071
100	10	3	-0.297	0.074
200	2	3	-0.297	0.074
200	2	5	-0.290	0.082
300	5	5	-0.289	0.082
200	3	10	-0.288	0.084
200	3	5	-0.285	0.087
300	3	10	-0.283	0.089
100	5	5	-0.280	0.093
200	3	3	-0.280	0.093
100	2	5	-0.280	0.039
300	3	5	-0.269	0.107
100	5	2	-0.268	0.063
200	2	10	-0.262	0.117
100	10	2	-0.258	0.124
100	2	2	-0.255	0.128
100	10	10	-0.254	0.130
100	3	2	-0.247	0.141
100	10	5	-0.227	0.177

Table 17: Overview of training runs with different hyperparameter settings

English words with Latin origin			
Words with Lowest Cosine Similarity (largest change)		Words with Highest Cosine Similarity (smallest change)	
uncle_noun	0.520	hour_noun	0.795
continue_verb	0.595	dinner_noun	0.788
party_noun	0.601	moment_noun	0.788
companion_noun	0.605	voice_noun	0.784
certain_adj	0.606	chair_noun	0.781
real_adj	0.606	tone_noun	0.779
exclaim_verb	0.624	stay_verb	0.775
age_noun	0.628	people_noun	0.769
form_nou	0.630	silence_noun	0.769
natural_adj	0.642	minute_noun	0.767
Average cosine similarity across all words: 0.704			

English words without Latin origin			
Words with Lowest Cosine Similarity (largest change)		Words with Highest Cosine Similarity (smallest change)	
sometimes_adv	0.567	say_verb	0.867
thus_adv	0.596	know_verb	0.857
half_adj	0.614	come_verb	0.843
also_adv	0.617	take_verb	0.841
instead_adv	0.618	door_noun	0.836
choose_verb	0.621	think_verb	0.834
body_noun	0.625	make_verb	0.832
short_adj	0.631	look_verb	0.831
forward_adv	0.631	leave_verb	0.830
start_verb	0.634	tell_verb	0.829
Average cosine similarity across all words: 0.730			

Table 18: Words with largest and smallest measured semantic change among 500 most frequent English words

Declaration of Independent Authorship

Original work

I expressly declare that the written work I submitted to the University of Zurich in the spring/autumn semester of 2024 with the title

Multilingual Semantic Change Detection with Word Embeddings in 19th Century Novels

is an original work written by myself, in my own words, and without unauthorized assistance. If it is a work by several authors, I confirm that the relevant parts of the work are correctly and clearly marked and can be clearly assigned to the respective author.

I also confirm that the work has not been submitted in whole or in part to receive credit for another module at the University of Zurich or another educational institution, nor will it be submitted in the future.

Use of sources

I expressly declare that I have identified all references to external sources (including tables, graphics etc.) contained in the above work as such. In particular, I confirm that, without exception and to the best of my knowledge, I have indicated the authorship both for verbatim statements (citations) and for statements by other authors reproduced in my own words (paraphrases).

Use of text generation models

I expressly declare that I have not only identified existing external sources, but also any automatically generated text that is contained in the above work. I have used the same citation style as if the text had been generated by a human to indicate the source of the automatically generated text. If the contribution of text generation models cannot be linked to specific text passages (see the associated guidelines), I have included a chapter describing the contributions of the text generation model. I acknowledge that no explicit citation is necessary where text generation models are merely used correctively (to improve grammar or idiomaticity of my own words).

Sanctions

I acknowledge that a thesis that is used to acquire credit and proves to be plagiarism with the meaning of the document *Erläuterung des Begriffs „Plagiat“* leads to a grade deduction in minor cases, a grade 1 (one) in more severe cases, without the possibility of revision, and in very severe cases can have the corresponding legal and disciplinary consequences according to §§ 7ff of the “Disziplinarordnung der Universität Zürich” and § 36 of the “Rahmenordnung für das Studium in den Bachelor- und Master-Studiengängen der Philosophischen Fakultät der Universität Zürich”.

I confirm with my signature that this information is correct:

Name: Betschart First Name: Seraina

Matriculation number: 17-960-162

Date: 01.12.2024 Signature: S. Betschart