Treball de fi de màster de Recerca

Using Context Variation Indexes for the Detection of Semantic Neologisms

Nom i Cognoms Laura Moset Estruch

Màster: Lingüística Teòrica i Aplicada

Edició: 2023-2024

Directors: Dra. Núria Bel Dra. Judit Freixa

Any de defensa: 2024

Col·lecció: Treballs de fi de màster

Departament de Traducció i Ciències del Llenguatge

Contents

Abstract

Semantic neologism detection is a complex issue in the field. SNs do not instil a feeling of novelty like formal neologisms because they already exist in the language, their change is only at a semantic level. This difficulty is prevalent when automating the identification process. The literature on the topic defends that the term's context is the best tool for detection, as the term changes meaning, so do the words it collocates with. This study presents a new method of automatic detection that uses context variation indexes calculated employing diachronic corpora. This approach is much simpler than other state-of-the-art procedures because it depends less on other computational tools and requires less software maintenance. The method is tested on 100 Spanish and Catalan terms. All tests demonstrated consistent performance. The results highlight the importance of term frequency and usage frequency of the neological sense as key factors for its success.

Keywords: semantic neologism; automatic detection; Kullback-Leibler divergence; context variation indexes;

1 Introduction

It is a well-known fact that language changes constantly; some words stop being used and new words keep appearing. For the latter, even if one is unaware of the standard name, we are all acquainted with the concept of neologisms. For example, if you have not unplugged from social media and find in 2024 the Spanish verb *gepetear* you may easily single it out because it does exude a feeling of novelty. If you are a speaker of the language, you can sense that the term is new and was not used one or two years ago. This is an example of a formal neologism and, as we will see through this study, not all types of neologisms are that easy to detect. This dissertation is centred around a different type of neologisms: semantic neologisms. For example, the verb *unplug* used to mean "to remove (a plug, such as an electric plug) from a socket or receptacle" (Merriam-Webster, n.d.) but can now be applied to people too to refer to taking a break from digital devices or the internet. A term that was also used in the same sentence as *gepetear* and which you may have skipped over as a possible neologism because you had seen the word before and it did not look new to your eyes.

Thus, we understand semantic neologisms (henceforth, SN) as a preexisting word in a language that suffers a new semantic manifestation —a change in meaning— and this shift can be to content that is totally new conceptually or to content that already existed but that can now be also expressed using this other signifier (Adelstein, 2022; Sánchez Manzanares, 2009). This is what differentiates SNs from other types such as formal neologisms whose change is observable —the addition of a prefix or suffix— or a completely new word, like *gepetear* before. For SNs, the term is already an entry in the language's dictionary; the only change is on a semantic level as it received a new acceptation. This is in part what makes the identification and classification of SNs so difficult to automatize.

Manual detection of SNs requires a lot of effort but it is very possible. In the same way that as a speaker one can recognize the novelty of the verb *gepetear*, one can also notice that the subject of *unplug* is not what it is supposed to be and that it has gone through some kind of metaphoric process. But computers do not have this speaker's feeling and this becomes a problem when one attempts to automate this detection task.

What linguists rely on for the identification of SNs —both manually and automatically— is context. There is this famous quote by Firth (1957) that has been applied to the fields of semantics, pragmatics, and distributional semantics, among others, which will also be completely relevant for SNs and our attempt at their identification: "You shall know a word by the company it keeps" (p. 11). Following this statement, a word usually collocates with similar groups of words from a specific domain. Therefore, a shift in a word's meaning should translate into a change in the words it usually collocates with or the general topic of the text it is found in. Therefore, it is not a surprise that the attempts at automatization until now have focused on one or more of these elements.

1.1 Motivation

While some successful methods have been found for the automatization of SN detection, as a field that is still in its early stages, it is good to explore and try different methodologies. The method presented in this thesis tries to achieve this identification but uses a "simpler" and more stripped-down version than the ones created by other authors. One that does not use other computational tools like topic detection or using word embeddings, but instead context and its relative frequency. As Renouf (2014) mentions, one of the biggest issues with automatic SN detection is that it requires a lot of software maintenance. If a good performance is achieved, this issue could be reduced significantly with the here proposed method. Moreover, while deep-learning tools can be very helpful, they could also result in biases, thus, it is interesting to explore methods that do not depend so heavily on them.

1.2 Objectives

This dissertation aims to analyze a new way in which the semantic change of words could be automatically detected. The idea is to test the Kullback-Leibler Divergence formula on two diachronic concordance files from a target word and use context variation indexes as an identification method. This method will be tested on 26 Spanish terms to establish a detection threshold and after on 50 more words to further observe the performance. For further evaluation, the code will also be tested on 26 Catalan terms. The goal of this other test is not to compare the behaviour of SNs in Spanish and Catalan but instead to look at whether this same method could work for both languages.

Based on the conclusions reached by other authors about the change in a term's context after undergoing a semantic shift, the author expects that there will be a big enough divergence between the concordance files to identify a potential SN. Because this method uses probability divergence as the base for detection and not semantic meaning, no specific type of SN is expected to perform higher than the other.

1.3 Structure of the Report

This present study is organized into five main chapters. After this introduction, there will be a literature review of topics relevant to this thesis divided into two sections. First, there is a look at some current classifications of Semantic Neologisms. Then, another section reviews the general complexities of semantic neologism detection and a subsection focuses on the state of the art on automatic detection. In Chapter 3, Methodology, there will be four sections describing the steps taken to complete this work: starting with data extraction, an explanation of the Kullback-Leibler Divergence formula and how it was applied, a rundown of the code to look at how it works and the output it gives, and finally a justification of how the threshold to detect SN was chosen. Chapter 4 combines the exposition and discussion of the results obtained. Lastly, Chapter 5 concludes the study by presenting a summary of the observations and conclusions reached throughout the work, accompanied by a reflection on the project's limitations and some future work.

2 Theoretical Background

In Chapter 1, Introduction, we saw a definition of SN. However, while authors tend to agree on the general description of what accounts for a neologism of meaning, few concur with the classification of what creates one. The attempt at classification is important, as we will also see in this chapter's section 2, Semantic Neologism Detection, not all SNs act the same way and thus, the way they are detected could vary. For those reasons, literature on both these topics is exposed below.

2.1 Neologism Classification

The classification of SNs usually goes in two ways, either focusing on the process the word went through to receive a change or by directly looking at the change in meaning they suffered.

The creation process of SNs is highly debated as some of the processes are, depending on the author, often attributed to other types of neologisms. As Díaz Hormigo (2007) describes, generally, the process of *semantic transference* —changing from one field to another— and *metaphoric uses* are assigned unanimously as processes to create SN. While *syntagmation* — combining words to create new meanings— and *syntactic or grammatical conversion* —changing grammatical categories— are understood by some as processes of semantic neology creation but as pertaining to formal neology by others.

A perhaps more defined classification is that of Cabré (2006). This classification seeks to be clear and consistent, by adopting the speakers' perspective and basing itself on the principle that words can be multifaceted and have different meanings (Cabré, Domènech-Bagaria, & Solivellas, 2021). Cabré (2006) proposes a classification of SNs that is based on the process of formation of the word and focuses on how the new meaning differs from the original one. As can be seen in Table 1 which is only an extract of her whole classification, within the type of neologisms that went through what she calls *resematization* or *semantic change* she recognizes three subtypes: *reduction*, *expansion*, and *change of meaning*. The criteria for classification looks at whether the word has started being used for fewer or larger contexts or if it obtained a completely different meaning from the one it had before.

Table 1: Extract of Cabré (2006)'s Classification of Neologisms as represented in Cabré et al. (2021)

<table>
<tr><td rowspan="6">Procédés d'innovation lexicale</td><td rowspan="6">Formation</td><td rowspan="6">Changement</td><td rowspan="3">Changement grammatical</td><td>Changement de catégorie</td></tr>
<tr><td>Changement de sous-catégorie</td></tr>
<tr><td>Changement de sous-catégorisation</td></tr>
<tr><td rowspan="3">Changement sémantique</td><td>Extension de sens</td></tr>
<tr><td>Réduction de sens</td></tr>
<tr><td>Changement de sens</td></tr>
</table>

Cabré et al. (2021) also mention the classification produced by Sablayrolles in 2015, Table 2. It is based on the formation process of words but specifically on the mechanisms behind the creation, such as different figures of speech. What both authors have in common is that they initially differentiate semantic or sense change with grammatical and syntactic change. However, when it comes to classifying SN they focus on very different aspects. While authors like Sánchez Manzanares (2009) will deviate towards Sablaroylles' approach of using the figures that provoked the change as descriptors, others such as Adelstein (2022) will choose an approach closer to Cabré (2006)'s and describe the type of change that has happened in the term.

Adelstein (2022) comes up with a classification based on the change of meaning instead of the process. However, she finds that there is some link between the two as she especially focuses on the semantic distance between the newer meaning and the original

Table 2: Extract of Sablayrolles (2015)'s Classification of Neologisms as represented in Cabré et al. (2021)

Matrices internes	Syntactico-sémantiques	Changement de fonction	Conversion
			Conversion verticale
			Déflexivation
			Combinatoire syntax / lexicale
		Changement de sense	Métaphore
			Métonymie
			Autres Figures

one is linked to the creation process of neologisms. She concludes that "[los] tipos con mayor distancia semántica se asocian con procesos metafóricos y metonímicos, mientras que los de menor distancia, con la generalización y la especificación" (p.333). Adelstein then creates four categories: *complete homonymous senses*, those that do not share any information between meanings and require extra-linguistic information to be decoded; *complete polysemous senses*, when the distance between meanings is also significant and extra information is needed, but there is a polysemous relationship between the two senses; *microsenses*, which have minimal semantic modification; and *contextual modulations*, are abstract senses that arise from considering information from the original descriptive meaning.

As Lamaczová (2018) mentions, these complications in reaching an agreement are derived from the base that meaning in its lexical sense is already very flexible and suffers from continuous change with the intent to adjust to speakers' communicative intentions. This results in changes in meaning happening rapidly, being difficult to foresee and being irregular. Cook, Lau, Rundell, McCarthy, and Baldwin (2013) contemplate the difficulty of describing word-sense, "[there is] no general agreement about what constitutes a discrete meaning of a word. [...] word meanings are unstable entities, often with shifting boundaries" (p. 2). While other types of neologisms have more visible elements that can be used to identify and categorize them, SNs display no change in the form of the word, only meaning. And, when meaning and word-sense are the only elements guiding us, classification, as seen above, can become complex.

However, returning to the quote mentioned in the Introduction by Firth (1957): "You shall know a word by the company it keeps" (p. 11); hence, the meaning of a word can be inferred by its context. Applied to the case of SNs, the signifier may not show any change but the words it collocates with do change. Making the context in which the SN can be found the most useful method to identify them. Specifically by observing which words one term is usually associated with and looking for a shift in them (Martínez Linares, 2015).

To summarize, classifying SNs is not an easy feat. There is no agreement on a specific method or on what categories to include. Mainly this is a result of the change happening on an already existing signifier and the flexibility of lexical meaning. These difficulties translate into its detection, and while context makes it possible, it does not come without effort. However, there have been attempts at creating identification methods both manual and automatic.

2.2 Semantic Neologism Detection

Detecting semantic neologisms has long been one of the main difficulties of the field. Unlike formal neologisms that undergo an observable change, semantic neologisms only change on a semantic level which is not perceptible just from a simple glance at the word. Because of this, speakers often find them to be harder to identify and describe and, as a result, they appear a lot less in dictionaries and their classification is less agreed upon and exhaustive than with other types of neologisms (Adelstein, 2022).

One element that authors concur on is the importance of context when looking at semantic neologisms. Very early in the literature on this phenomenon, Bastuji (1974) already mentions how, generally, it is this variation in the context of a word that is significant enough to provoke a change in meaning and the interpretation of a term. He used the examples of the terms *réimporter* and *exportateur*, a verb and a noun that used to be followed by a noun phrase that made reference to an object or thing and which had changed to also include nouns related to people or their characteristics like 'retired', for example, in "nous sommes devenus exportateurs de gens instruits et nous réimportons des retraités" (p. 9).

When talking about possible classifications, Sánchez Manzanares (2009) suggests that "los neologismos de sentido se activan por el contexto, bien en un contexto sintagmático, bien en un contexto discursivo" (p.142). The author concludes in her article that the rupture in words, which would usually appear together, that occurs when a semantic neologism emerges in a context which is not its expected one is essential for its identification.

In her corpus-based study, Lamaczová (2018)) examines if it is conceivable to confirm a change in the meaning of two terms, 'viral' and 'bizarro', by looking at a possible contextual change using the ABC newspaper's corpus between the years 1980 and 2015. Their conclusions confirm that the context of these semantic neologisms reflects the change. Because as the term's sense changes, the nouns and adjectives that occur around it also do. Moreover, they derive that these transformations are incremental and they can reach a point where they surpass the frequency of use of the original meaning.

Lamaczová (2018)'s analysis is mostly done manually, they checked each occurrence and classified it. Some authors such as Cook et al. (2013), Wijaya and Yeniterzi (2011), Torres Rivera (2019) and Torres Rivera and Torres-Moreno (2019) have tried to find methods that required less human effort by proposing semi-automatic solutions.

2.2.1 Automatic Detection of Semantic Neologisms: State of the Art

Renouf (2014) when describing the possible automatization of SN detection enumerates five different parameters that could be used to identify new word senses: word frequency, distributional semantics, lexical priming, heuristics, and onomasiological strategies. To look at word frequency would be to look at significant changes in the usage of a word, for example, if it goes from having a relative low use and to suddenly experiencing a big increment. Distributional semantics would focus on lexical collocations or context of the word. Lexical priming would look for changes of the word position on a text or domain changes. Heuristics works by looking at markers of newness such as italics or quotation marks or by comparing dictionaries, however, this works best for formal neologisms but not for SN. Finally, onomasiological strategies analyse how different expressions or words can represent the same concept.

Cook et al. (2013)'s method introduces Word-Sense Induction (WSI) to identify word-senses that are new with respect of one corpus made of newer texts (Focus Corpus) with another with older texts (Reference Corpus). What WSI does is to group similar usages of a target word in a corpus, so that all usages of a particular sense of a word would be in the same group. Their process consists on taking from the concordances of a target lemma a three-sentence context: the sentence containing the lemma, the previous one and the one after. Then, from these sentences extract a bag of words and its positions in reference to the target lemma which will be used to apply the WSI and group the specific word-sense of the lemma with others in the corpus. Finally, they "calculate the 'novelty' of the induced sense in the Focus Corpus as the ratio of its relative frequency in the Focus and Reference Corpora" and rank lemmas according to their novelty score making the highest-scoring sense for a lemma its novel sense. Cook, et al. claim that this method can be particularly useful for their main goal which was to facilitate the process of updating dictionaries, as it would provide editors with newer senses but also indicate senses that have stopped being used.

Wijaya and Yeniterzi (2011) propose a method that focuses not so much on the detection of possible SNs but on the confirmation of a change and identifying when it happened. Their process consists of counting the words that co-occur with the term over time, so they utilize a match count using frequency and co-occurrences. Then, identifying the topic of the top words, and grouping documents in clusters. They do this with a tool called a TOT (Topics-Over-Time) model and k-means clustering. Finally, each year will be assigned to a cluster and therefore, they suggest that there is a change in meaning when two consecutive years are assigned to different clusters. Their method greatly succeeds at predicting when the change in meaning happened.

Another automaticized identification method is the one presented by Torres Rivera and Torres-Moreno (2019) who introduced a program they called DENISE, that uses Natural Language Processing methods and computational models. They also use the context of the term for neologism detection by topic modelling it -meaning determining the themes of the text fragment–, extracting keywords from it, and comparing the vector embedding of these keywords to the already existing embedding of the SN obtained from deep learning models like Word2Vec, Sense2Vec and FastText. If the embeddings are very different, they suggest that there could be a change in the meaning of the word. DENISE obtained positive results. Sense2Vec had the best results even though the other embeddings also had their advantages. However, the authors remark that their system does not rely on one single method to analyze semantic change.

Now, most of these methods use one or more tools for their identification. Apart from tokenization and lemmatization tools, Cook et al. (2013) used WSI, Wijaya and Yeniterzi (2011) TOT models and k-means clustering, and Torres Rivera and Torres-Moreno (2019) topic modelling and models like Word2Vec. The more steps there are, the more probable it is one of them fails or misses information, especially using deep learning models which are trained on a very large scale. However, we can see this was not a problem as all the studies mentioned before obtained very favourable results. However, these resources used mean more time and computational power, and one can only wonder if there is a more straightforward and simple way for SN detection.

It will be interesting to see with the method employed in this dissertation what affects the identification the most from the elements mentioned before; the creation process, the change in meaning, or frequency. For example, Lamaczová (2018) mentions the differences in the manual detection of *viral* and *bizarro*. While with the former it was easy to determine when it belonged to the neologic meaning or the original one as the thematic contexts were entirely new, this did not happen with *bizarro* whose contexts remained somewhat consistent. They conclude that there is a correlation between the easiness of the terms' identification and the semantic distance between the original sense and the neologic one. If we take into account (Adelstein, 2022)'s ideas about semantic distance, we could assume that words which have a higher distance between meanings will also have a more different context and thus, its detection using the latter would be more clear. While Wijaya and Yeniterzi (2011) put a lot of emphasis on the frequency of appearance of the new meaning for identification.

3 Methodology

Applying all that has been mentioned so far, a new potential approach for the detection of semantic neologism has been thought of and will be presented in this thesis. The method presented here uses the term's context variation indexes for a possible identification. It relies on the previously mentioned corpus linguistics and it introduces the novelty of using the Kullback–Leibler divergence as a detection method.

3.1 Data

For this thesis, three different corpora were used to obtain word concordances. The first two were CREA and CORPES, two Spanish corpora created by the Real Academia Española. The last corpus used was in Catalan, it is called CTILC. From these corpora, the concordances of a total of 102 terms were extracted.

The first corpus was CREA (Corpus de referencia del Español actual)(Real Academia Española, n.d.-b), version 1.0, which was last updated in December 2023. It is compiled using more than 111.000 documents produced between 1975 and 2000 which constitute more than 122.500.000 tokens. These documents could come from any Spanish-speaking country and have no specific domain or topic, thus, they include fiction and non-fiction, news from many areas, etc. The same annotating method developed for the corpus CORPES was applied in this corpus, allowing for searches using lemmas, forms or grammatical categories. This corpus can be accessed online via their website (Real Academia Española, n.d.-b) and is available for unlimited queries. The results are obtained in real-time.

The other Spanish corpus used was CORPES (Corpus del Español del Siglo XXI) (Real Academia Española, n.d.-a) version 1.1, which includes more than 410 million tokens obtained from more than 380.000 documents but more are being added with each version. This corpus includes both written texts and oral transcriptions. The documents were written or produced as early as 2000 and the latest version includes texts from 2024. Similarly to CREA, the texts originate from different Spanish-speaking countries and include many different genres and domains. Both CREA and CORPES use the same annotating tags and system which explanation can be found on their website.

For the test of Spanish SNs, CREA and CORPES were chosen as complementary diachronic corpora, the former going from 1975 to 2000 and the latter from 2000 to 2024. They were chosen because of the variety of texts they include. Moreover, they allowed the download of .txt files with some of the concordances obtained. At the time this is written, only 1000 results can be downloaded at once. Therefore, for this research the oldest concordances were prioritized from CREA and the newest from CORPES, meaning the results were downloaded from oldest to newest in the former and newest to oldest in the latter.

The corpus used for obtaining the concordances of Catalan words was the CTILC (Corpus textual informatitzat de la llengua catalana) which has almost 114 million tokens and keeps getting updated as time passes (Institut d'Estudis Catalans, n.d.). It is created by documents on various genres and topics, although literature accounts for almost half of it. The texts included extend from 1832 to 2020, although the biggest volume of texts comes from the most recent years. The website does not mention the tagging method used but the corpus is annotated and it is possible to search using lemma and part of speech. This corpus was useful because of its size and the large amount of years it includes. Moreover, the fact it allowed for search queries using lemma was important as we wanted to obtain the biggest context of the words as possible. It is important to note that the download of the `.xlsx` file used for this research is not available to the public and was obtained by contacting the corpus coordinators.

For Spanish, 26 terms were employed to create a threshold for detection and 50 more terms were later used for testing and confirmation. These terms are either confirmed SNs or words without any recent semantic change. For the first test, 18 out of 26 words were SNs, while for the second it was 30 out of 50. To create the list of confirmed SNs the author consulted the Neologism database BOBNEO (Observatori de Neologia, 1989–) which is available in both Catalan and Spanish. In BOBNEO, the 'advanced search' tool was used to find terms that had a semantic change by filtering using the 'Tipus de Neologisme > Canvi > Semàntic' (Type of neologism > Change > Semantic) option. Then, from the results obtained different terms were chosen at random, the only condition being that they had to have their first occurrence after 2000. For the list of non-SNs, the author asked the tool chatGPT 3.5 to generate a list of different words at random that included nouns, verbs and adjectives. From that list, the author selected words from different fields and grammatical categories, while also avoiding words generated that could have undergone a semantic change in recent years.

The test for Catalan was only applied to a list of 26 words to observe whether this method could be used on other languages and whether different corpus would affect the results. Still, the method to obtain the 26 words used on the test was the same used for Spanish.

3.2 The Kullback–Leibler Divergence

The novelty of this thesis is introducing the Kullback-Leibler Divergence, also known as relative entropy introduced by Kullback and Leibler (1951), to obtain a relative guess on the probability of a term having had a change in its context and thus being a SN.

KL divergence, often denoted as D KL(p(x)||q(x)), is used to compare two probability distributions (*p(x)* and *q(x)*) over the same variable x. University of Illinois Urbana-Champaign (n.d.) define it as: "a measure of the information lost when *q(x)* is used to approximate *p(x)*. Let *p(x)* and *q(x)* are two probability distributions of a discrete random variable x. That is, both $p(x)$ and $q(x)$ sum up to 1, and $p(x) > 0$ and $q(x) > 0$ for any x in X."

The discrete form of the KL divergence equation can be seen below:

$$D_{\mathrm{KL}}(p(x) \parallel q(x)) = \sum_{x \in X} p(x) \ln \left(\frac{p(x)}{q(x)} \right)$$

It is important to note that while KL is a distance metric, it does not calculate the distance between the two probability distributions but, instead, the steps that would need to be taken to go from one to another (University of Illinois Urbana-Champaign, n.d.). That is because it is not symmetrical, the KL divergence going from $q(x)$ to $p(x)$ will generally be different from the one from $p(x)$ to $q(x)$.

Applied to this thesis, the KL will be used to detect semantic change using the relative frequency of all the nouns in the context of our possible SN or target word in two different periods. To calculate the KL we need two probability distributions and relative frequency is exactly that. By analyzing the co-occurring words in two time periods, noticing a significant change could indicate a possible change in the meaning of the target word. While terms without a shift should occur in the context of similar words in both periods and thus, return a low KL score.

This method is conceptually similar to others mentioned in Section 2.2, Semantic Neologism Detection, as it relies heavily on the words occurring in the context of the target words. However, the main difference is that the main effort is on data extraction but after that, it does not depend on other software. As Renouf (2014) mentions a big issue of automatic neologism detection is data and software maintenance. While the KL method still requires progressively updated data, it is based on a mathematical formula and the only software needed is the Spacy library, lowering the software maintenance significantly compared to other detection methods.

3.3 The Code

The code's [1] main section automatically detects the possibility of a target word being a semantic neologism. Then, various options exist to manually confirm the previous findings by looking at top co-occurring words per year, top co-occurring words per period, 2D plot representation using top word embeddings, and target word frequency per year.

The automatic detection of potential SNs is done the following way. In the first place, the data files are imported. For this thesis, the data used was obtained from the corpora mentioned in Section 3.1 and is imported as `.txt` or `.xlsx` files. In the case of the CREA/CORPES corpus data, the occurrences obtained are already separated into two files, one from the corpus containing older data (CREA) and the other with the newest one (CORPES). For the CTILC, all occurrences are included in a single file.

Then, the data is pre-processed for a better analysis: two data frames are created from the files, separated diachronically. Because we are looking for a recent semantic change that has had time to become established, the separation is made in 2000. Thus, the older data frame, henceforth referred to as the *Reference* concordance files, contains all of the target word's concordances before 2000. While the newer one, the *Current* concordance files, includes everything from 2000 onwards. In the case of Spanish, the corpus went as early as 1975. For Catalan, the same year is used as a limit, so that the time span between the Reference and Current periods is similar.

For each data frame, the information is organized into two columns: `df["FECHA"]` contains the year the occurrence took place and `df["CONCURENCIA"]` has the text string with the concordance. The last step is cleaning the concordance data, which is transformed into lowercase and tokenized, all words that are not stop-words are lemmatized using the Spacy's library models: `ca_core_news_lg` for Catalan and `es_core_news_lg` for Spanish. For better results, only nouns were kept after the lemmatization as they are the part of speech with the most semantic information (Moset Estruch, 2024).

[1] The code can be accessed on the author's Github or the following link: https://github.com/momolaurs/TFM_SN_detection

Once the data is ready, two separate lists are created, one list contains every lemma that appears in the Reference concordance files and the other does the same for the Current one. Then, the absolute frequency of each word in that specific file is counted. After, the relative frequency of each word is calculated. In both cases, the count is calculated relative to the total number of words in the concordance file they belong to —the Reference or the Current—. The relative frequency of the word is what is used to calculate the Kullback-Leibler divergence between the concordance files. To make the calculation possible, first, the words that appear in one of the files but not on the other are added but with a very low relative frequency of 0,0001. Once we obtain two lists that have the same words on them, we can calculate the KL divergence between the files. The number obtained will tell if the concordance files imported are very different from one another. A higher number would mean that the context of the target word has changed a lot from one period to the other. Now we have the context index variation but for the code to classify the target word into a potential SN or not, we need to establish a threshold for detection.

The last part of the code is able to create scatter plot visualizations with the top words that co-occur with the target word in each concordance file. To do this, we use word embeddings —a vector that captures the meaning of the word into numbers so that the computer can "understand" it—, in this case, we use the ones already created by FastText. To be able to display it in a 2D plot, for each word, a PCA transformation finds the most important directions and its dimensions are reduced to two which will equate to the X and Y axis.

3.3.1 Calculating the Threshold

Table 3 shows the results obtained from testing the code using the Kullback-Leibler Divergence on 26 different Spanish words, including known SNs and words without any recent semantic change. The idea was to analyze the KL score SNs would get, compare it to the one obtained by non-SN, and create a threshold to automatically classify them. The Table has five columns. Each row starts with the target word whose concordances were obtained from the corpora. Then, the KL Score column contains the score obtained from comparing the Reference and the Current contexts using the Kullback-Leibler formula. The third column indicates whether the word is considered a SN by BOBNEO.

Based on the results from the second and third columns the threshold was decided. For it, evaluation metrics were performed at potential thresholds 0.35 and 0.4. These metrics will consist of the Precision, Recall and F1 scores calculated with the code's predictions. The F1 will be a particularly useful measure as it provides the test's accuracy of binary classification problems, such as this one. Moreover, F1 works well when it is applied to imbalanced classes, as is the case with this test which has more SNs than non-SNs terms.

Table 4 shows the results for the 0.35 threshold, while Table 5 displays the performance metrics for the 0.4 threshold. While threshold 0.4 obtained a better F1 score of 0.8, 0.35 was able to identify a higher number of SNs correctly, as shown by the True Positive (TP) metric. Even though, the 0.35 threshold did so by also detecting some more non-SNs as False Positives (FP), because the final score is not that different from one another we consider that detecting more SNs should be a priority. Thus, the threshold of 0.35 was chosen for the final evaluation.

In Table 3 shown before, the columns *Code Prediction* and *Performance* show a breakdown of the final results obtained with the 0.35 threshold. The fourth column reports whether the code predicted the word to be a SN (True) or not (False). Finally, the last column indicates if the result is a True Positive (TP) if it was a SN that was correctly identified; False Negative (FN) is a SN that was not identified; True Negative (TN) is a non-SN that was not identified as SN; and False Positive (FP) is a non-SN that was incorrectly categorized as one.

A similar test was applied to 26 Catalan terms, and the results can be observed in Table 6 with the same structure as the one described for the Spanish test. A glance at the KL scores obtained is enough to discover that the same threshold of 0.35 would not be of any use for the classification of the Catalan terms, as most terms, including non-SNs obtained a score over 0.35. However, it is possible to choose a personalized threshold for

Table 3: The KL method applied to 26 Spanish terms, with columns *Code Prediction* and *Performance* showing the results after establishing the detection threshold at 0.35

Word	**KL Score**	**Semantic Neologism?**	**Code Prediction**	**Performance**
agua	0.3204	No	False	TN
alérgico	0.3582	Yes	True	TP
barbie	1.1154	Yes	True	TP
bizarro	1.0831	Yes	True	TP
blanquear	0.4406	Yes	True	TP
camisa	0.2300	No	False	TN
casa	0.3583	No	True	FP
cuenta	0.4340	Yes	True	TP
hibernar	1.6326	Yes	True	TP
ingrediente	0.2799	Yes	False	FN
lámpara	0.2006	No	False	TN
muro	0.2196	Yes	False	FN
nube	0.5276	Yes	True	TP
país	0.3444	No	False	TN
perfil	0.2855	Yes	False	FN
pingüino	0.8850	Yes	True	TP
plataforma	0.4444	Yes	True	TP
seguidor	0.2189	Yes	False	FN
sofá	0.2641	No	False	TN
televisión	0.3930	No	True	FP
tóxico	0.8290	Yes	True	TP
tsunami	1.6077	Yes	True	TP
unicornio	0.9817	Yes	True	TP
ventana	0.2107	Yes	False	FN
viral	0.9542	Yes	True	TP
vaso	0.3021	No	False	TN

this corpus and obtain similar performance results.

The same evaluation metrics as before were performed, in this case with thresholds 0.6

Table 4: Evaluation Metrics for the Spanish test at threshold 0.35

	Predicted Positive	**Predicted Negative**
Actual Positive	13 (TP)	5 (FN)
Actual Negative	2 (FP)	6 (TN)
Evaluation Metrics		
Precision	$\frac{TP}{TP+FP} = \frac{13}{13+2} = 0.8666$	
Recall	$\frac{TP}{TP+FN} = \frac{13}{13+5} = 0.7222$	
F1 Score	$\frac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} = 0.7878$	

Table 5: Evaluation Metrics for the Spanish test at threshold 0.4

	Predicted Positive	**Predicted Negative**
Actual Positive	12 (TP)	6 (FN)
Actual Negative	0 (FP)	8 (TN)
Evaluation Metrics		
Precision	$\frac{TP}{TP+FP} = \frac{12}{12+0} = 1$	
Recall	$\frac{TP}{TP+FN} = \frac{12}{12+6} = 0.6666$	
F1 Score	$\frac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} = 0.8$	

(Table 7) and 0.7 (Table 8). Threshold 0.7 had a very high number of false negatives, meaning that it missed a lot of SN, and its final F1 score was lower than the 0.6 threshold at 0.6451. Therefore, the most optimal threshold was decided to be 0.6, with an F1 score of 0.7058, a performance a bit lower than that of the Spanish test.

It is interesting to note how the thresholds vary significantly from one language to another, as the average KL score obtained diverted greatly. Some possible explanations could lie in the way the data was extracted and therefore the different sizes of the concordance files. As both Spanish corpora only allowed a maximum of 1000 occurrences to be downloaded at a time, meanwhile the Catalan one can reach up to 40,000 concordances on occasion. A higher number of concordances should result in a bigger number of words and thus a bigger divergence and higher KL score, this is possibly what happened. However, there were some terms like ‘unicorn’ and ‘viral’ that had a very low count of concordances,

Table 6: The KL method applied to 26 Catalan terms, with columns *Code Prediction* and *Performance* showing the results after establishing the detection threshold at 0.6

Word	**D KL Score**	**Semantic Neologism?**	**Code Prediction**	**Performance**
actiu	0.3283	Yes	False	FN
aigua	0.9102	No	True	FP
altaveu	1.4023	Yes	True	TP
blanquejar	2.6651	Yes	True	TP
camisa	0.2805	No	False	TN
cartell	0.5024	Yes	False	FN
casa	1.5617	No	True	FP
got	0.4097	No	False	TN
imant	1.8529	Yes	True	TP
ingredient	0.9424	Yes	True	TP
llibre	0.9854	No	True	FP
maneta	1.6394	Yes	True	TP
maquillar	2.5714	Yes	True	TP
marea	1.0468	Yes	True	TP
núvol	0.1994	Yes	False	FN
onze	0.3871	Yes	False	FN
país	0.7528	No	True	FP
plataforma	0.5270	Yes	False	FN
procés	0.6844	Yes	True	TP
sabata	0.3046	No	False	TN
seguidor	0.6949	Yes	True	TP
sofà	0.6545	No	True	FP
televisió	0.2933	No	False	TN
tòxic	0.9708	Yes	True	TP
unicorn	3.6384	Yes	True	TP
viral	3.9805	Yes	True	TP

making the difference between the terms' scores a bit skewed.

Another explanation could lie in the data itself. A quick count of the mean words per

Table 7: Evaluation Metrics for the Catalan test at threshold 0.6

	Predicted Positive	**Predicted Negative**
Actual Positive	12 (TP)	5 (FN)
Actual Negative	5 (FP)	4 (TN)
Evaluation Metrics		
Precision	$\frac{TP}{TP+FP} = \frac{12}{12+5} = 0.7058$	
Recall	$\frac{TP}{TP+FN} = \frac{12}{12+5} = 0.7058$	
F1 Score	$\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}} = 0.7058$	

Table 8: Evaluation Metrics for the Catalan test at threshold 0.7

	Predicted Positive	**Predicted Negative**
Actual Positive	10 (TP)	7 (FN)
Actual Negative	4 (FP)	5 (TN)
Evaluation Metrics		
Precision	$\frac{TP}{TP+FP} = \frac{10}{10+4} = 0.7142$	
Recall	$\frac{TP}{TP+FN} = \frac{10}{10+7} = 0.5882$	
F1 Score	$\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}} = 0.6451$	

concordance of both corpora shows that there are on average 64.1 words in each CREA and CORPES concordance against the 24.8 average words in the Catalan one. This would imply that despite there being fewer concordances, the Spanish context is much larger and therefore there is a higher number of nouns found and compared between concordance files. Because this difference between corpora is not the focus of this thesis, no further analysis has been done, however, it is an intriguing topic that could be explored further in the future.

4 Results and Discussion

This section will present the analysis of the results obtained from testing the KL detection method on the Spanish terms. The data discussed includes the 26 terms checked in the previous section as well as the results from evaluating 50 new terms. This new test was executed once the 0.35 threshold had been established. The CREA and CORPES corpora were used again as data extraction tools. From these 50 new terms, 30 were SNs according to BOBNEO and the other 20 words were not SN. The results from this test are broken down in Table 9 which has the same structure as Table 3 described in the previous section. The performance metrics of this extended test can be found in Table 10 below. The final results report numbers very similar to those of the first test meaning that the performance of the method is consistent and, with an F1 score of 0.7857, it has a relatively high accuracy.

Table 9: Results of the KL method applied to 50 Spanish terms

Word	**KL Score**	**Semantic Neologism?**	**Code Prediction**	**Performance**
abridor	0.4828	Yes	True	TP
abrir	0.3087	No	False	TN
apagón	0.4381	Yes	True	TP
árbol	0.3332	No	False	TN

Continued on next page

Table 9 – continued from previous page

Word	KL Score	Semantic Neologism?	Code Prediction	Performance
arena	0.2533	Yes	False	FN
barato	0.2048	No	False	TN
barracón	0.5661	Yes	True	TP
breve	0.2303	No	False	TN
burbuja	0.4199	Yes	True	TP
caballo	0.3062	No	False	TN
campamento	0.2897	Yes	False	FN
cielo	0.2961	No	False	TN
clonar	0.8673	Yes	True	TP
construir	0.2132	No	False	TN
contento	0.2677	No	False	TN
decir	0.8305	No	True	FP
desahogar	0.4106	Yes	True	TP
descomprimir	1.5149	Yes	True	TP
enlace	0.4822	Yes	True	TP
etiqueta	0.1853	Yes	False	FN
flor	0.3110	No	False	TN
galáctico	0.9087	Yes	True	TP
gatillar	1.3990	Yes	True	TP
generoso	0.2027	No	False	TN
grande	0.4984	No	True	FP
húmedo	0.1891	No	False	TN
icono	0.9258	Yes	True	TP
instalación	0.3157	Yes	False	FN
java	0.7640	Yes	True	TP
levantar	0.3807	Yes	True	TP

Continued on next page

Table 9 – continued from previous page

Word	KL Score	Semantic Neologism?	Code Prediction	Performance
liberar	0.1860	Yes	False	FN
llave	0.2338	Yes	False	FN
máscara	0.3091	Yes	False	FN
metal	0.3654	Yes	True	TP
nuevo	0.4995	No	True	FP
paisa	1.0635	Yes	True	TP
peligroso	0.2524	No	False	TN
pintar	0.1991	Yes	False	FN
pitufo	2.5527	Yes	True	TP
playa	0.2379	No	False	TN
postear	5.0668	Yes	True	TP
rearmar	1.1736	Yes	True	TP
remolcar	0.5547	Yes	True	TP
seguir	0.3948	Yes	True	TP
sentir	0.3400	No	False	TN
sesión	0.4330	Yes	True	TP
sol	0.2644	No	False	TN
termómetro	0.4372	Yes	True	TP
vaca	0.2453	No	False	TN
virtual	0.2440	Yes	False	FN

Similarly to the first evaluation, most of the misclassification consisted of False Negatives —a total of 9 in the 50 terms test and 5 in the 26—, while there were 2 False Positives in the first test and 3 in the second one. This implies that the method with this threshold is more prone to missing SNs rather than incorrectly identifying them. However, this is a result of some of the SNs returning a fairly low KL score which we will soon analyze.

Table 10: Evaluation Metrics for the Spanish test at threshold 0.35

	Predicted Positive	**Predicted Negative**
Actual Positive	22 (TP)	9 (FN)
Actual Negative	3 (FP)	16 (TN)
Evaluation Metrics		
Precision	$\frac{TP}{TP+FP} = \frac{22}{22+3} = 0.88$	
Recall	$\frac{TP}{TP+FN} = \frac{22}{22+9} = 0.7096$	
F1 Score	$\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = 0.7857$	

For this analysis, it will be interesting to look at terms from both evaluations that the code failed to categorize accordingly, as well as, some of the words that were a clear success. Most of these words can be put into groups because a similar reasoning can be found behind their results, however, there are also a few intriguing outliers.

One of the first groups to notice is the terms that were misclassified because, despite BOBNEO indicating that the first instance in their bank is from a year after 2000, they may have been present in the Reference concordance files. For example, 'ventana' with a KL divergence of 0.2107. This term was already used in its neologic sense as early as the 1970s but was consolidated with the commercial release of Apple Lisa in 1983 and Microsoft's Windows 1.0 in 1985. Thus, the code which looks for recent changes that happened after the year 2000 struggles to identify it. A similar case is 'virtual', for which the same events also initiated the shift. In some concordances of the last years of the Reference files, you can already notice the neologic meaning being used. While it is a very low frequency, it can be argued that it still interferes with the score and lowers it.

Something similar could be the reason 'muro', 'perfil', 'seguidor' and 'etiqueta' failed to be identified, as social media where they are used started appearing in 1997, such as the social network *SixDigrees.com.* However, that year is pretty close to 2000 and considering the time it would take for the terms to become established the change in context should not be that noticeable anyway. Thus, another possible hypothesis for the detection's failure would be that sometimes the cognates of words are used to achieve specific purposes. Especially in publicity, the English equivalent of a term is sometimes preferred when talking about technology because the language has connotations of future and innovation but also of social status (Vellón, 2009). For example, in Spanish, the SN 'etiqueta' is very often replaced by the English term 'hashtag' or instead of 'seguidor' one could use 'follower' to seem cooler. This would mean that the frequency of use of the neologic sense of these words decreases as the English equivalent is favoured by the speakers.

This brings us to the topic of the frequency of use of the neological sense. For the words above the frequency is low because cognate words are preferred for stylistic choices. However, there are words where the frequency of use of the neologic sense is low because they are an extension/generalization or reduction of the original sense. That is to say, it is only a small amplification or a niche use relative to the original sense. For example, Cabré (2006) would classify 'perfil' as a SN that is a result of *reduction of meaning*. Using Adelstein (2022)'s classification, the term could belong to the *microsense* category, the original semantic structure remains the same (a representation or description of a person or object), however, it has adapted to refer to a specific context, in this case, social media. 'seguidor' would also belong to this category. These terms still have a similar meaning but in a new and specific context. This is represented in the scatter plot of the top words that co-occur with 'perfil' seen in Figure 1. The words at the top of both periods are a bit bolder. The Reference words in yellow are mostly descriptive of parts of the face. In blue are the Current concordance file words, they are mostly grouped and centred, meaning that they do not deviate much from the original meaning. Thus, we can conclude that the neological sense is present as it appears in this plot but the change is still not significant, a possible reason why it does not go over the threshold.

Speaking of frequency of the neologic use, we could find two other groups: the first includes 'ingrediente', and 'arena'; and the other is composed of 'campamento', 'liberar', 'llave', 'pintar', and 'máscara'. For the former, a native speaker of Spanish can notice that all the previous words are relatively common but their neologic meaning usage is

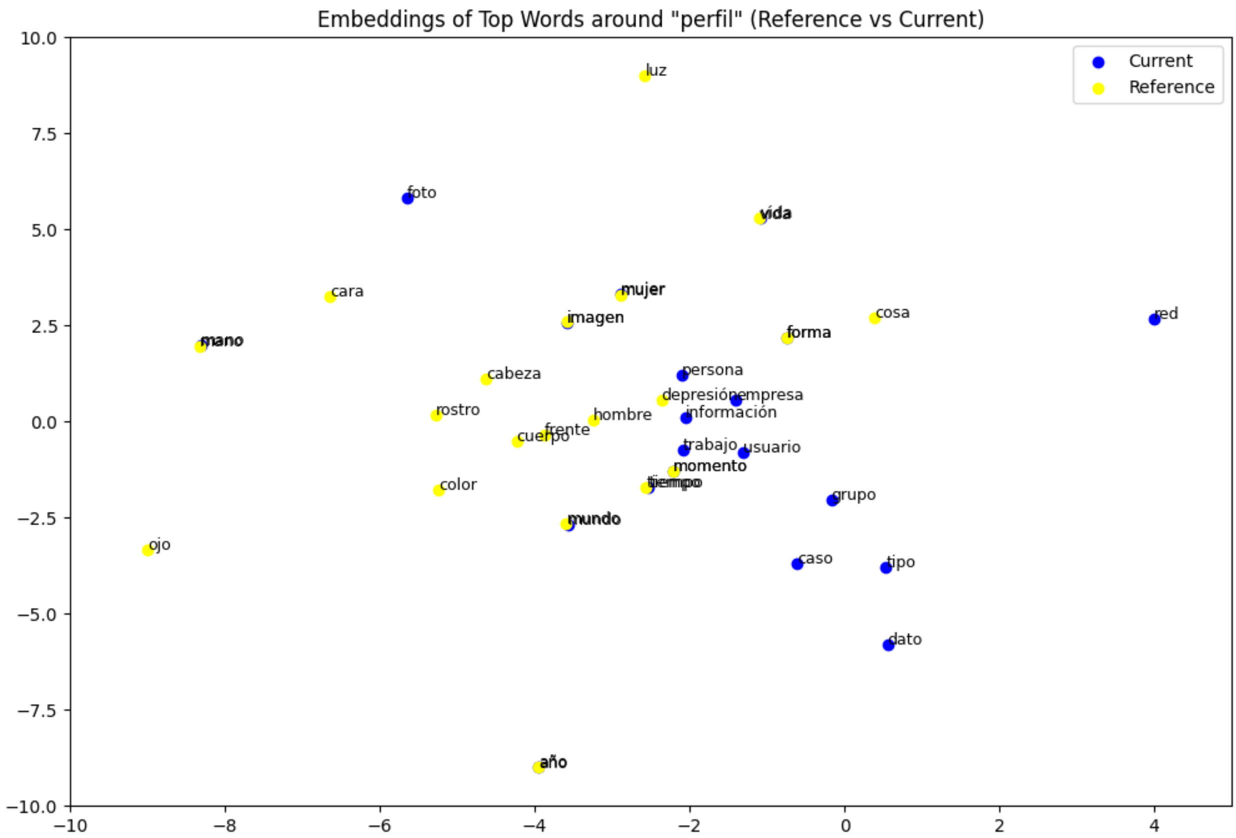

Figure 1: 2D plot of embeddings of Top Words around the term 'perfil'

minimal compared to the original sense. Most of its use is in news articles as metaphors and not so much in our daily lives. Thus, only checking from the occurrences of the term in a corpus where news articles are just a small part may be a reason for their low KL scores. For example, 'ingrediente' will have a lot of occurrences in recipes, and its metaphoric use in the news will pale in comparison.

The second group's failure to be detected could lie in the data used. The terms for the test were randomly chosen from BOBNEO, this neologism bank includes some dialectal examples from different varieties of Spanish and some ended up being selected. For example, a quick look at the occurrences from the term 'máscara' in the BOBNEO website reports only instances from Cuba, while 'campamento' has 12 examples and all are from Chile. However, as seen in Figure 2, a big part of the concordances collected in CORPES are from Spain while Chile occupies a much lower percentage. This is a graph that repeats for most words in the corpus indicating that the corpus used does not have an equal representation of all Spanish varieties. Even if the corpus was better balanced, when using a corpus including all varieties the fact that they are dialectal SNs already reduces largely the amount of nouns in the context that will relate to the neological meaning, making detection using KL difficult. Probably what happened with the terms ‘campamento’, ‘liberar’, ‘llave’, ‘pintar’, and ‘máscara’. Because of these two reasons, we can conclude that the relative frequency of the words from the newer context was not high enough for the code to detect a noticeable change.

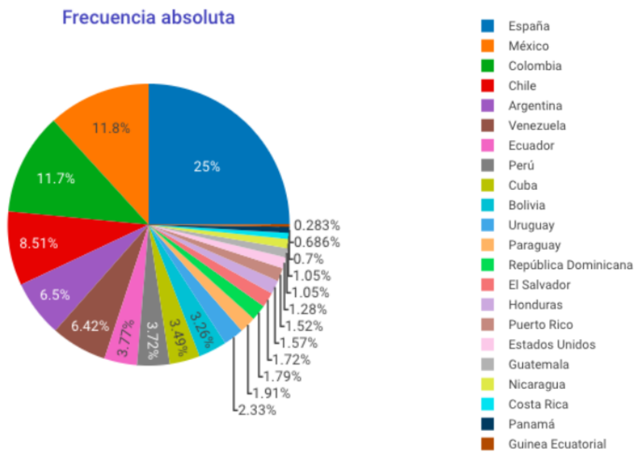

Figure 2: Absolute Frequency of the lemma ‘campamento’ in CORPES separated by country

In opposition, there were some non-SNs that the code classified as potential SNs, some of which with a relatively high KL score. For these we find two different groups: ‘casa’ and ‘televisión’; and ‘decir’, ‘grande’ and ‘nuevo’. Looked at it differently: on the one side are nouns; and on the other side are verbs and adjectives. The easiest one to identify is the latter. The verbs and adjectives here are very general and abstract, an ability which allows them to precede or follow terms from most semantic fields. This versatility allows them to adapt to different situations, contributing to their applicability to a wide range of contexts. This is reflected in both the Reference and Current concordance files and after in their KL score. It can be seen more clearly in Figure 3, a visualization showing the top words that co-occur with ‘decir’, where there are almost no repeated top words and all terms appear to be spread out around the plot. However, the same does not happen with other verbs —like ‘sentir’— and adjectives —like ‘generoso’ or ‘peligroso’— that have more defined contexts and fields where they occur. For example, the scatter plot of 'peligroso' in Figure 4 shows a good representation of a non-SN where the context stays mostly the same. Most words are yellow because a lot of them overlap on both Reference and Current top words.

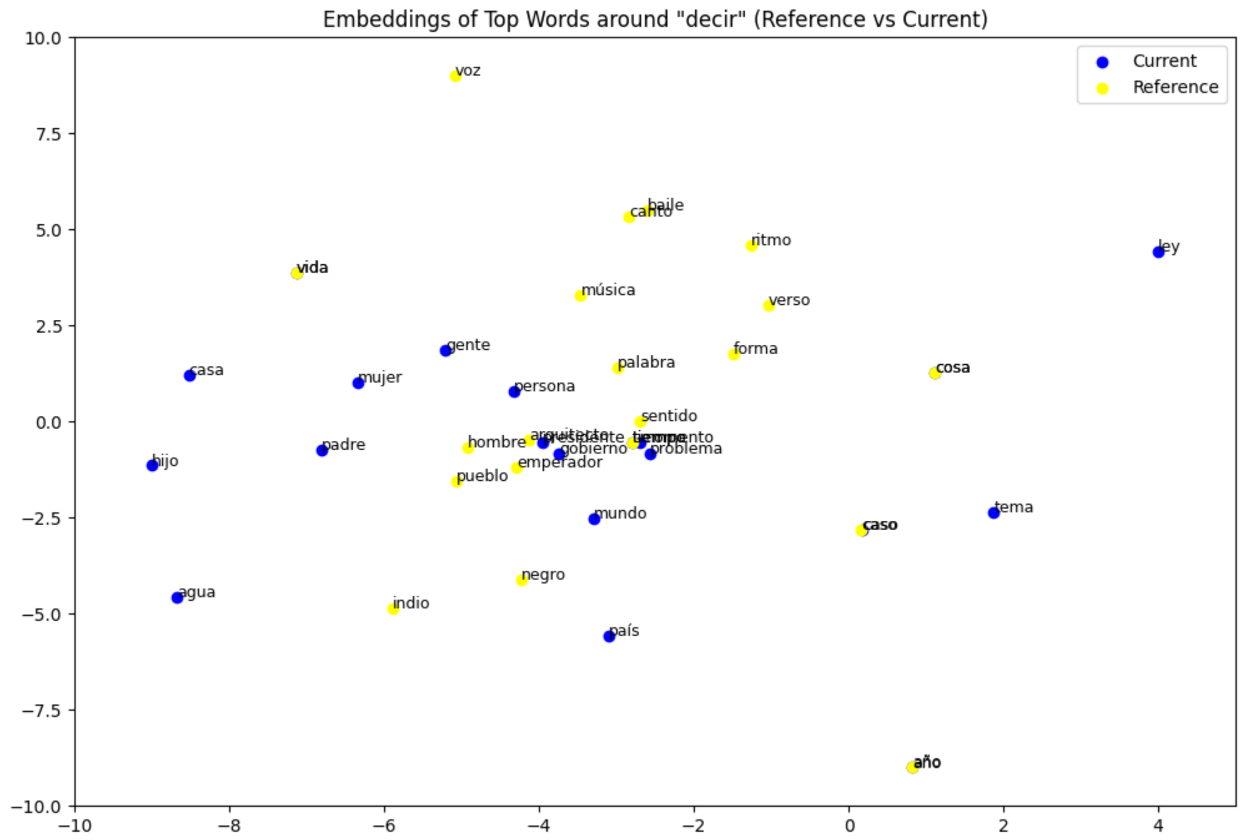

Figure 3: 2D plot of embeddings of Top Words around of the term 'decir'

In the case of the nouns, the reason is more specific to each term. For example, looking at the term 'televisión'. Once again there is an event around the term resulting in a contextual change, for televisions is a big technological advance. So even if 'television' has not changed meanings, it has shifted in how they look, what they can do, and how they are used. This has been reflected in their context and identified by the code. From these words, we derive that there is no need for a semantic change in a word for it to suffer contextual change and this could trick this method on occasion.

However, we should not only observe the terms that the code failed to detect but also some of the successes. Despite the mentioned SNs that the code missed for the above reasons, we could claim that the method still succeeded at detecting all SN that have a high frequency in BOBNEO. Moreover, it managed to detect all types of SN, not limited

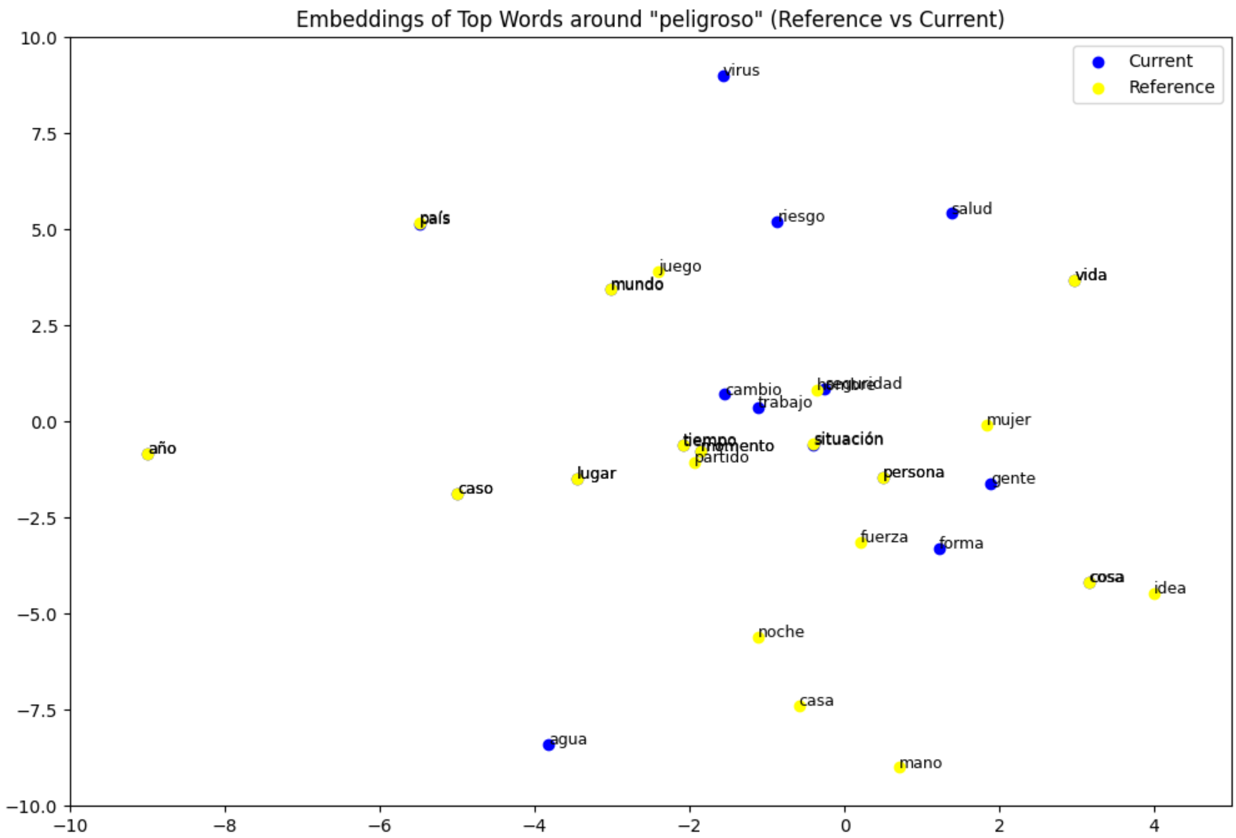

Figure 4: 2D plot of embeddings of Top Words around the term 'peligroso'

by type or specific fields. When comparing it to other previous studies, the KL method managed to automatically detect the SNs 'viral' and 'bizarro', which Lamaczová (2018) also examined manually for their paper and it had a similar F1 score to (Torres Rivera & Torres-Moreno, 2019)'s 0.84.

As seen in the results Tables 3 and 10 shown before, there were many terms where the KL divergence was 0.6 or higher. For this, one important factor has been identified: a change in the general frequency of use. For most of these terms, with some exceptions, the number of occurrences in the Reference concordance file did not reach 1000, while there were significantly many more in the Current file. For example, this can be observed in Figure 5 showing the number of occurrences of the term 'unicornio' per year, and where we can observe a big shift from 2001 onwards.

This is interesting, as Renouf (2014) mentioned one parameter for SN identification would be to look for changes in word frequency. Lamaczová (2018) also highlighted the incre-

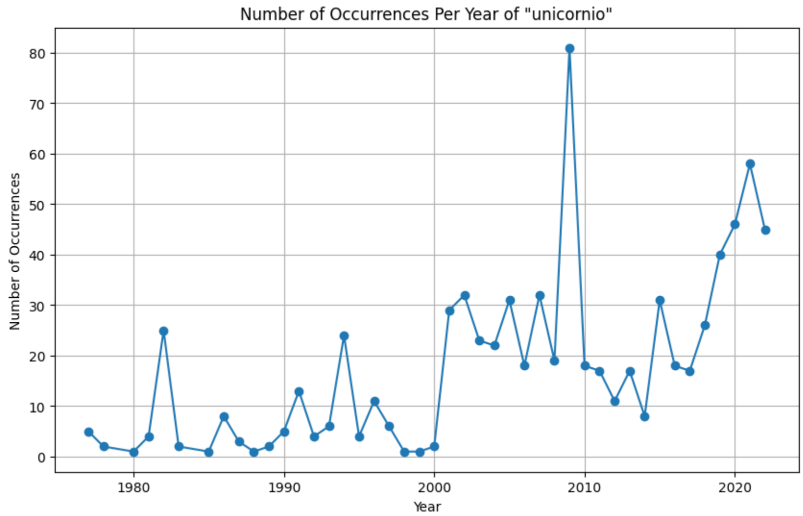

Figure 5: Number of Occurrences per year of the term 'unicornio'

ments in the general frequency as a tale-tell factor during her research of 'viral' and 'bizarro'. If a word which was previously rarely used suddenly has a big increase in occurrences, it can tip us towards hypothesizing that something has happened to that term. Concerning the KL method, more occurrences leave more space for more words to appear in the context and thus create a bigger divergence.

Nevertheless, this difference in frequency can not be attributed as the sole factor for the high KL score of these words. Evidence of this is the SN 'blanquear' which also has double the concordances in the Current file than it does on the Reference one but its KL score is still not that far over the threshold. This means that for the terms with a very high KL D, the words that appeared in the context of each concordance file were also very different from one another. Once more the scatter plot of embeddings will help visualize this. Figure 6 shows the plot for the term 'icono'. Where we can see a separation between the Current and Reference top words, with new contexts related to 'artista', 'música' and 'moda' appearing. This separation is surprising, as 'icono' could also be classified as a *microsense* of the original meaning of "religious pictorial representation" (Real Academia Española & Asociación de Academias de la Lengua Española, s.f.), because both senses have a shared semantic base (icon as a symbol, something representative) but in this neologic case, it is referencing a person. This plot is so different from the other microsense seen before in Figure 1 because 'icono' had another semantic change in the 1990s to refer to icons in computers. The plot still can be divided into three meanings: the left side corresponds to icons on phones or computers, the centre bottom refers to religious icons ('figura', 'imagen' and 'forma'), and the right side covered in blue is the neological sense. What we see in this figure is an original meaning that has been surpassed by its two neological meanings. This high frequency of the new neological sense of 'icono' in contrast with the difference with the other meaning is what resulted in the high KL score of 0.9258.

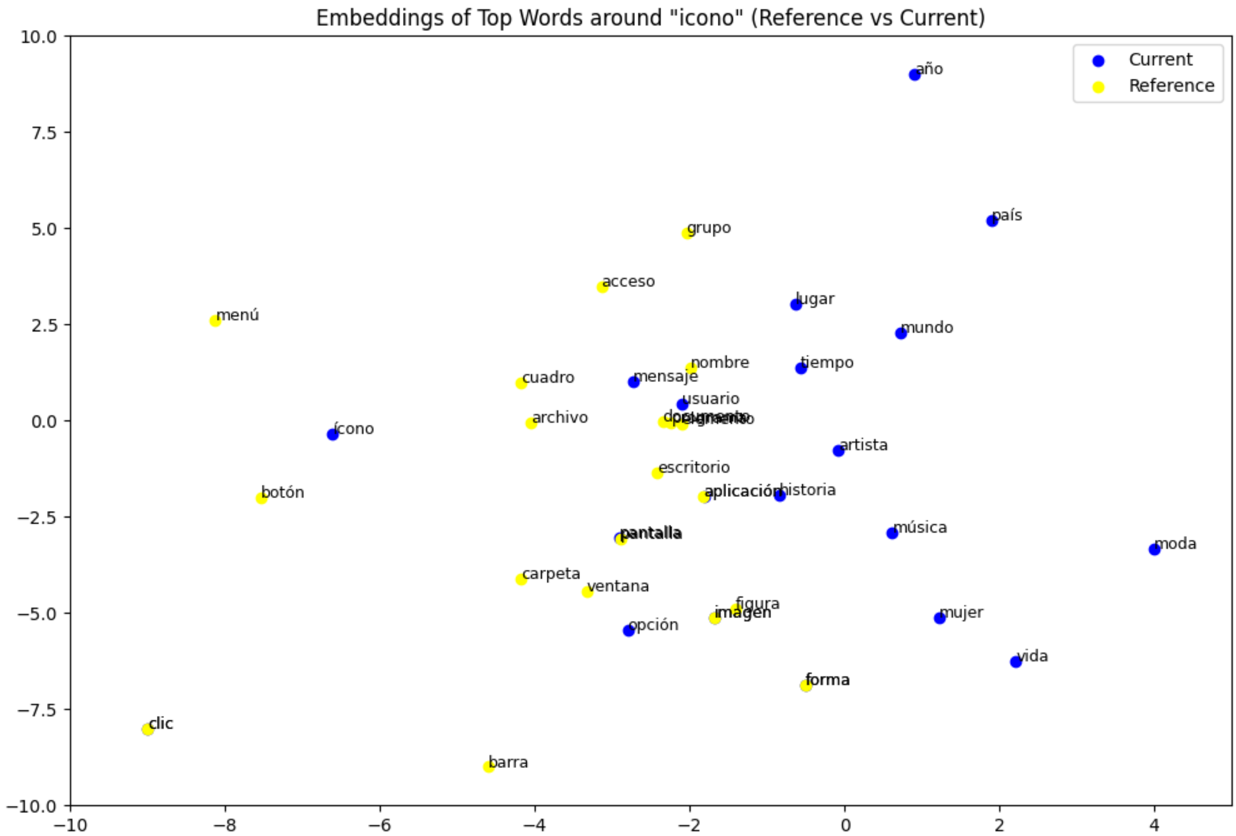

Figure 6: 2D plot of embeddings of Top Words around the term 'icono'

Further proof that frequency of use was not the only important factor is the term 'viral' which had the same amount of concordances in both files but still got a KL score of 0.9542. The situation with this word was that, even if the code was checking the same amount of concordances, for the Current file most of those concordances were using the newer meaning, therefore, the code still detected that there had been a big contextual change. This goes in line with Lamaczová (2018) that noticed that the newer sense usage could eventually surpass that of the original sense. This difference is so stark because for this test we only had concordances from 2020-2023 and the 1970s, so the code was not looking at a gradual change. In Figure 7, we can see the scatter plot with the top words of ‘viral’. Using Adelstein (2022) classification, ‘viral’ could fit the *complete polysemous senses* class, as the new meaning is very different from the original but has some connection to the idea that it is something that propagates quickly. This can be seen in the visualization as the two meanings occupy different spaces. Drawing a diagonal line from top left to bottom right would almost perfectly separate words related to the original sense from the neologic one. Terms in yellow from the Reference files would all stay on the left bottom side, while the blue terms from the Current one are split between both sides showcasing this new polysemy of the word.

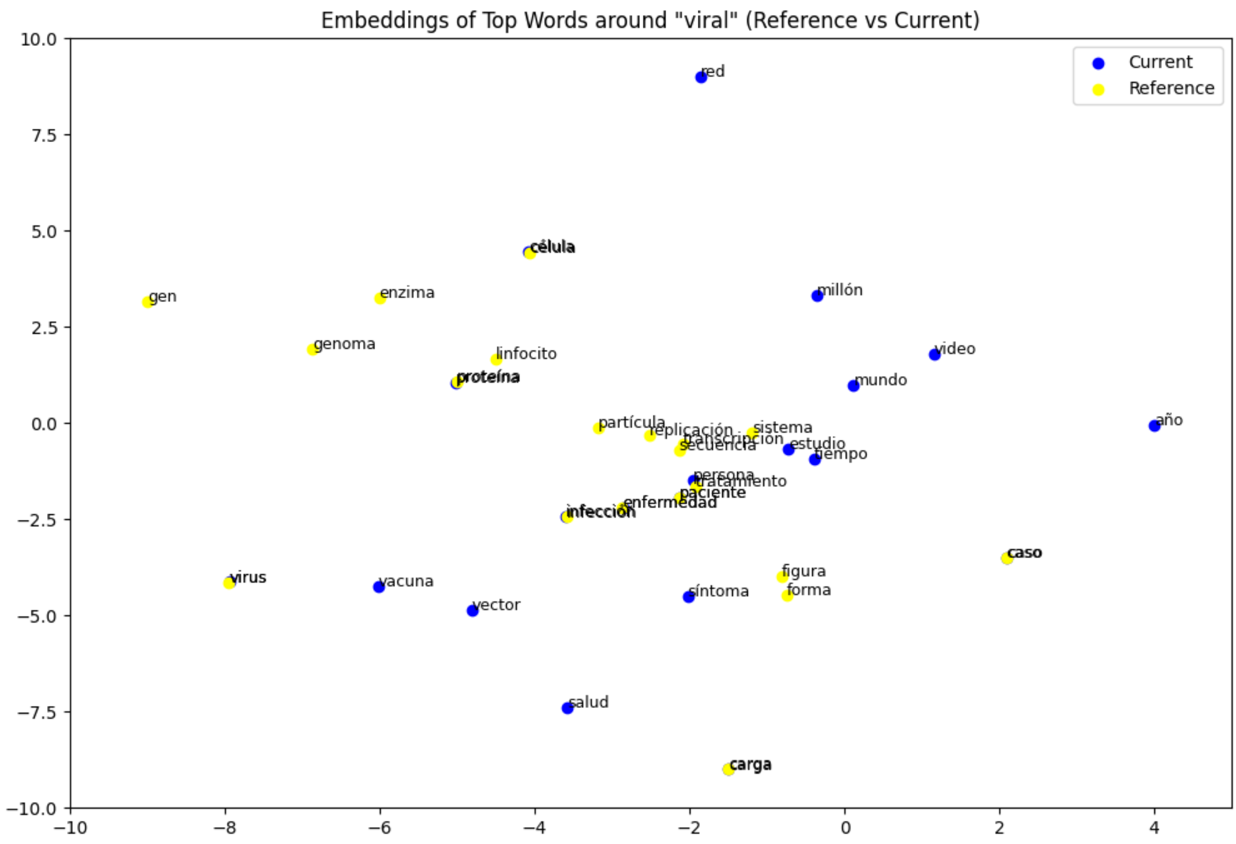

Figure 7: 2D plot of embeddings of Top Words around the term ‘viral’

A more extreme example is ‘enlace’, also classified as a *complete polysemous senses* SN. Figure 8 is the visualization of the term’s top words and while it can also be divided into two meanings by a diagonal line from bottom left to top right, the difference with Figure 7 is that in this case the Current file’s top words mostly occupy only the space of the new neological meaning. In this case, the plot shows how the neologic sense has reached a point where it has more frequency of use than the original one, as most of the top words in the recent years' context belong to it.

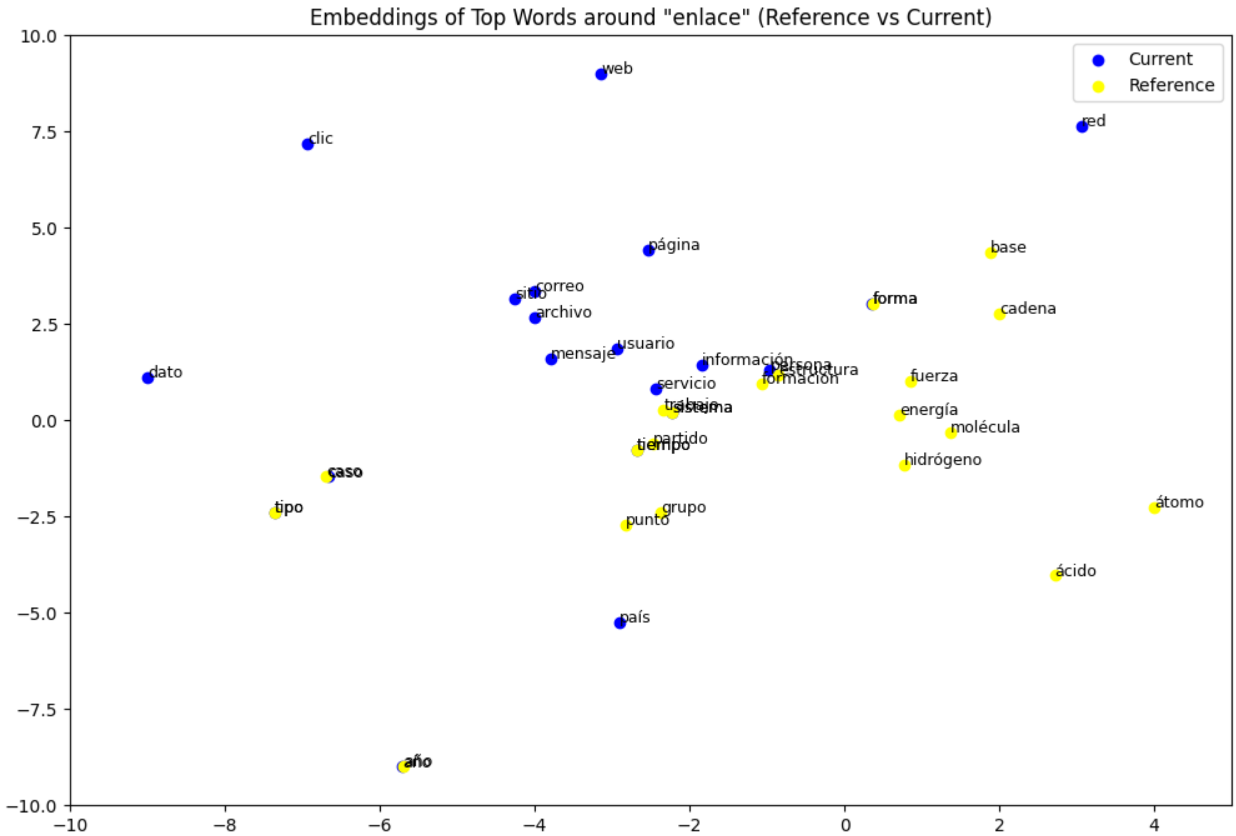

Figure 8: 2D plot of embeddings of Top Words around the term 'enlace'

To summarize the Spanish test results, three determinant factors for the code's success were found: the Reference files should not include a lot of concordances with the neological sense, a big change in the general frequency of the word, and the newer sense having a high frequency of use relative to the original meaning. Because the code is not based on the semantic meaning of words, the semantic distance of the senses does not matter. This means that the code does not differentiate or act differently depending on the classifications of the SN. For example, it detected easily both 'unicornio' and 'hibernar' even though the former has gone through a metaphoric process and the second is a generalisation. The visualizations presented were much more useful at showing these differences in the SNs and representing the changes the words had gone through.

5 Conclusions

In this present study, a total of 76 Spanish and 26 Catalan words were analysed, 65 of which were confirmed neologisms by BOBNEO and 37 were words that had not suffered a change in meaning recently. For the detection, a Python code was created that used the Kullback-Leibler Divergence as its basis by using context variation indexes to classify target words into potential SNs or not. The first step after choosing the words was extracting their concordances from periods 1975-2023 on three different corpora and using them to calculate the KL Divergence between diachronic concordance files to observe the difference between SN and non-SN which would allow us to establish a threshold for detection.

This initial test was conducted for both Spanish and Catalan. The results showed a big difference between the average KL scores of both languages, a difference which was attributed to the divergence in the data obtained from the different corpora. The CREA/-CORPES corpora only allowed for a download of a maximum of 1000 concordances while CTILC had no limit. Because of this, only the oldest and newest concordances could be downloaded for the Spanish corpus making the term's change much more stark in comparison with the gradual change that is experienced in the Catalan corpus. Moreover, the average words per concordance between the corpora was also significant, 64 words for the Spanish one against the 25 of the Catalan. These were some of the hypotheses behind the big difference in the average KL score between languages and it means that the corpora used for this detection method will greatly influence the KL scores obtained.

Essentially, before detection, multiple tests should be conducted to determine the most appropriate threshold for a specific corpus. This is seen in the big difference in KL scores and ultimately in the chosen optimal threshold for each language, with 0.35 for Spanish but a much higher 0.6 for Catalan. However, the evaluation metrics showed that despite this variation, the method could achieve similar levels of performance, with Spanish obtaining an F1 score of 0.7878 in the 26-term test and Catalan getting a 0.7058. The second evaluation with 50 new Spanish terms further consolidated these results by obtaining an F1 score of 0.7857. This indicated that the method is consistent and with a relatively high accuracy.

When observing the results more closely, some key factors were identified for a more successful detection. For starters, the Reference file should include none or very few instances of the term used with its neological sense, because this would stop the code from detecting a contextual change as it happened with 'ventana'. Then, it was detected that most of the words with the highest KL had a common element: they had experienced a frequency increase in the Current concordance files. As observed, several authors had mentioned word frequency as a potential parameter for SN automatic detection and it is not surprising that the KL method being an approach based on context variation indexes is highly influenced by it. A word with very low frequency suddenly experiencing an increase of it suggests that it has undergone some type of change. However, it was observed that a general frequency change alone was not always enough to guarantee detection, as some terms that had experienced a big increase still obtained a score close to the threshold. It was then concluded that it was the combination of a frequency change and a big shift in context that resulted in these terms achieving such high KL scores.

For words that had a similar number of concordances in both concordance files, the main factor that was attributed to successful SN detection was the frequency of usage of the neological meaning in comparison with the original sense. For example, some SNs that the code failed to detect were those that were very common in the language but whose new neological sense only occurred very rarely. This meant that the relative frequency of the words in the new context was very low and not enough to impact the KL divergence rate. Other false negatives were dialectal SNs, which were already unfavoured by only being used in particular regions and also suffered from those regions being unequally represented in the corpora where the data was extracted. In contrast, the SNs with the same number of occurrences that were correctly identified where those were the neological sense seemed to be used at an analogous frequency as the original sense, or even had substituted the original meaning and are used at a much higher rate.

The method managed to detect most of the SNs that had high frequency in BOBNEO, as well as, some SNs used as examples in other previous studies. We have demonstrated that contextual change and divergence between diachronic concordance files is a good method for SN identification. However, the code also proved that not in all cases can one equal contextual change to being a SN. For example, world events or technological advancements also participated in greatly altering a term's context without the term undergoing any change in meaning. At the same time, verbs and adjectives that are less defined and more abstract and can therefore fit any context also tricked the code into being classified as SNs.

In comparison with other state-of-the-art methods, the F1 score of 0.78 was similar to Torres Rivera and Torres-Moreno (2019)'s 0.84 —their best weighted average F1 score using Sense2Vec on nouns—. However, the basic software needed for this method of basic detection is much less. This would be very advantageous, as software maintenance is currently a big concern of semantic neologism automatic detection.

This approach is also not limited by semantic fields or type of SN. As the formation process or semantic distance between senses did not influence the KL score. This was advantageous because it meant that SNs that were formed by reduction of meaning could also be identified the same way that words with a full change of meaning would. Nevertheless, it denies the possibility of shedding some light on the continuous debate on SN classification. Still, some interesting conclusions about the topic could be drawn from the top words embedding visualizations. As the semantic distance Adelstein (2022) mentions could be identified in different separations from the words in a 2D space.

5.1 Limitations

One limitation of this study was the corpora used. Because of resource and time constrictions, already existing corpora were employed as tools to obtain the data necessary. This meant that the author did not have control over factors such as making sure that the data belonging to each genre, domain, and region was equally distributed. For example, CREA and CORPES have a lot more text from Iberian Spanish than any other variety of the language. Plus, the percentage of text that is *Fiction* and *Literature* is much higher than any other genre. As commented in the analysis, these factors affected the detection of some SNs.

5.2 Future Work

There have been various topics that have surfaced from this study that would be interesting to look at more thoroughly. On the one hand, further tests using the KL method for detection would be recommended. It would be intriguing to look at how other languages apart from Spanish and Catalan perform. On the other hand, dwelling deeper into these found differences between the two languages would be interesting, establishing whether the big divergence in scores comes entirely from the differences in corpora or if reasons are emerging from each language's stance on neologisms. Finally, there could be some research into how word embeddings, their cosine similarity, and their representation into 2D plots could work alongside creating classifications for SN.

5 References

Adelstein, A. (2022). Neología y semántica: grados de neologicidad en el ámbito nominal. *Neología y semántica: grados de neologicidad en el ámbito nominal*, 327-345.

Bastuji, J. (1974). Aspects de la néologie sémantique. *Langages*(36), 6-19.

Cabré, M. T., Domènech-Bagaria, O., & Solivellas, I. (2021). La classification des néologismes. révision critque et proposition d'une typologie multivariée et fonctionnelle. *Neologica*, *2021*(15), 43–62.

Cabré, M. T. (2006). La clasificación de neologismos: una tarea compleja. *Alfa*, *50*(2), 229-250.

Cook, P., Lau, J. H., Rundell, M., McCarthy, D., & Baldwin, T. (2013). A lexicographic appraisal of an automatic approach for detecting new word senses. *Proceedings of eLex*, *2013*.

Díaz Hormigo, M. T. (2007). Aproximación lingüística a la neología léxica. In *Morfología: investigación, docencia, aplicaciones: Actas del ii encuentro de morfología: investigación, docencia* (p. 33-54).

Firth, J. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 10–32.

Institut d'Estudis Catalans. (n.d.). *Corpus textual informatitzat de la llengua catalana.* Retrieved 2024-04-27, from `https://ctilc.iec.cat/scripts/index.asp`

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, *22*(1), 79–86.

Lamaczová, K. (2018). Análisis de los neologismos semánticos viral y bizarro: un estudio de corpus. *Language design: journal of theoretical and experimental linguistics*, *20*, 0093-120.

Martínez Linares, M. (2015). Cuestiones en torno al tratamiento de los neologismos semánticos: entre la intuición, el análisis lingüístico y el diccionario. In A. Cuniţă & C. Lupu (Eds.), *Neologie, neologism. conepte, analize* (p. 69-100). Bucarest: Universidad de Bucarest.

Merriam-Webster. (n.d.). *Unplug.* Retrieved 2024-05-05, from `https://www.merriam-webster.com/dictionary/unplug`

Moset Estruch, L. (2024). *Neología semántica y contexto: Un estudio de la detección de los neologismos semánticos tóxico, unicornio y hibernar usando dos corpus diacrónicos.* (Unpublished work)

Observatori de Neologia. (1989–). *Banc de dades dels observatoris de neologia (bobneo).* `https://bobneo.upf.edu/`. (Data de consulta: 15/03/2024)

Real Academia Española. (n.d.-a). *Banco de datos (CORPES XXI).* Retrieved 2024-03-19, from `http://www.rae.es`

Real Academia Española. (n.d.-b). *Banco de datos (CREA Anotado).* Retrieved 2024-03-19, from `http://www.rae.es`

Real Academia Española, & Asociación de Academias de la Lengua Española. (s.f.). *Diccionario panhispánico de dudas (dpd)* (2nd ed.). Retrieved 2024-04-26, from `https://www.rae.es/dpd`

Renouf, A. (2014). Semantic neology: The challenges for automatic identification. *Neologica*, *2014*(8), 185–220.

Sánchez Manzanares, M. d. C. (2009). Procedimientos trópicos en la neología semántica: sistematicidad y creatividad. *Revista de investigación lingüística*, *12*, 123-146.

Torres Rivera, A. (2019). *Detección y extracción de neologismos semánticos especializados: un acercamiento mediante clasificación automática de documentos y estrategias de aprendizaje profundo* (Unpublished doctoral dissertation). Universitat Pompeu Fabra.

Torres Rivera, A., & Torres-Moreno, J.-M. (2019). Detecting new word meanings: a comparison of word embedding models in spanish. In *Coria*.

University of Illinois Urbana-Champaign. (n.d.). *Kullback-leibler divergence.* Retrieved 2024-03-03, from `https://hanj.cs.illinois.edu/cs412/bk3/KL-divergence.pdf`

Vellón, J. (2009). El uso del inglés como estrategia discursiva en el texto publicitario. *Cultura, lenguaje y representación*, *7*, 157–182.

Wijaya, D. T., & Yeniterzi, R. (2011). Understanding semantic change of words over centuries. In *Proceedings of the 2011 international workshop on detecting and exploiting cultural diversity on the social web* (pp. 35–40).