Loanwords and their sentiments

Daniel Mol, Tim Döring, Victor Gordan

June 2022

1 Introduction

Language is an inherently dynamic construct. We constantly shift the meaning of words, add new words and remove words according to our needs. After all, the point of language is to streamline communication, and with new events new words are necessary to effectively describe them. One consequence of the dynamicity of language are loanwords: words used in one language (the source) are 'borrowed', translated or not, by speakers of a different language. The motivation for the use of loanwords is debated, but one theory suggest that it is driven by the social identity attributed to the users of the language [14], which is to say that the reason for using a loanword has to do with the sentiment associated with the source language. A prime example of this theory is the use of French loanwords in English: after the Norman conquest of England in 1066, French became the language of the English nobility as William II of Normandy declared himself king of England [4]. The reasoning for this fits the aforementioned theory; simply put, the English nobility considered English to be a language of low prestige, and as such would rather speak the more 'noble' French. Consequently, French words tend to have a more positive sentiment attributed to them. We posit that similar loanword sentiments for other languages can be ascribed to historical relations, and as such we will research the extent of the effects that historical relations have had on the sentiments of loanwords in different language. To this extent, we will perform a sentiment analysis on a multitude of datasets, including tweets and etymological origins, and then predict the sentiment of loanwords by their appropriate historical relation.

2 Research context

Research on loanwords and historical relations is all but sparse. As such, we provide some research on loanwords in general. Loanwords are sometimes not positively looked upon due to preservation of cultural identity. For example, recently in France, words such as 'streamer' and 'e-sports' were banned by the Académie Française [3], in order to preserve its language purity. Similar efforts are made in Asian languages due to differences in alphabets. For example, the word 'Amsterdam' is written as 阿姆斯特丹 in Mandarin Chinese, which is phonetically pronounced similar to the original (Āmǔsītèdān). This method is also often employed in Korean, though South-Korea is a much more anglicized country due to its affiliation with the U.S. during the Cold War. Although Korean also has its language purists, English loanwords call for more mixed feelings as found by the research done by Rüdiger [11]. Generally, the responses were either mixed or negative; almost all interviewed participants of the study thought overusing English makes you look arrogant and/or pretentious. Academic contexts and a clear unavailability of a Korean alternative were considered acceptable reasons for the use of English loanwords. Similar results were found for the use of English loanwords in Japanese [6]. However, Japanese loanwords in Korean are regarded in a vastly different manner, and this is due to their relations. Firstly, Japanese business relations are not considered as important as English ones, and few people learn Japanese compared to English, since: "Japanese is considered to be unimportant but easy, English is considered to be important but difficult" [9]. This all supports our theory of identity-driven loanword usage.

Another analysis of Romance language origins can be found in [5]. In it, the author gives reasoning for the meaning of loanwords in Romanian from Turkish versus Slavonic. For example, 'a iubi' comes from the Old Slavonic 'Ijubiti', which means 'to love', and 'duşman' comes from the Ottoman-Turkish 'düşmen', which means 'enemy'. This can be explained due to the historical relations between Romania and the Ottomans as they were enemies, and conversely the Romanians and Slavic countries were allied.

3 Data collection

To acquire our data, we employed a variety of techniques. Firstly, for our etymological dataset we made use of a Python module called BeautifulSoup to be able to do Webscraping. We then scraped the lists on EZGlot (for example, the Spanish words of Latin origin in [2]) and Wiktionary (similar example, in [12]), which gave us our dataset for words and their origins. From EZGlot we collected both the word of the 'receiving' language and that of the 'origin' language, but for Wiktionary that would have made the webscraping a lot more complicated, so we only collected the word in the 'receiving' language which is actually the only part that we use. We don't care what the original word was, just from which language it came. The sizes of our origin data sets vary in size from around 200 to 25000 words. Languages that are studied more and languages that were more related tended to have more words, for example the data set for English words of Latin origin was 23000 words, while something like Italian words of Old French origin was 200.

For our tweet dataset, we used a dataset from Kaggle, where each tweet had been labeled automatically to have a positive ('4'), neutral ('2') or a negative ('0') sentiment. The dataset can be found here [7]. Due to the very small amount of neutral tweets, we decided to disregard them and instead label each tweet as 1 or -1 if they were positive or negative, respectively. This dataset consists of 1.6 million tweets, with a 50% equal split of positive and negative tweets.

Finally, the sentiment dataset for English words was a combination of three different datasets acquired from the following sources: SentiWordNet [1], MPQA [13] and Opinion Lexicon [8]. These were referred to us by the Sentiment Symposium website of C. Potts [10]. The sentiments in Sentinet were determined using Machine Learning, the other two datasets were manually constructed by the authors. These three datasets combined make up about 27.000 words. In order to then get sentiments for words in other languages, we would simply use Google translate to translate a word from English to a different language and then attribute the sentiment from the English word to the translated word. We understand that this might be wrong in some cases, but in general, words without context will be translated to words with a similar meaning. A small, manual check of the Romanian translation was done and the sentiments were determined to be correct in most cases.

4 Methods

In order for us to obtain an answer to our research question we shall now discuss the methods used to process our datasets into something useful. For the tweet dataset, we were interested in the context appearences of loanword. We used the tweet dataset for these 5 languages: Dutch, French, German, Latin and Portuguese. For every word in a tweet, which we acquired by tokenizing the tweets, we checked if that word was in our etymology list (the ones we webscraped), and if it was, we added 1 to the appropriate language for the appropriate sentiment. For example, the word 'calamity' can be expected to be found in a negative tweet and when searching through our etymology list we will find it comes from French ('calamité', similar meaning). As such, our negative French loanwords will be incremented by 1. In other words, we counted for each language the number of occurrences in negatively labeled tweets as well as the number of occurrences in positively labeled tweets. Even if appearing in a negative context does not necessarily imply that the word is mainly used in such a context, however, wee deemed the 1.6 million labeled Tweet dataset to be large enough to yield statistically relevant results nonetheless. We also recorded the amount of times that a source language

was dominant, meaning the majority of the loanwords appearing in a given Tweet stemmed from this source language increased the corresponding counter by one. We dealt with ties in three different ways and recorded all of our results: Random, meaning a random language is chosen among those that have the same amount of tweet tokens; fraction, meaning each language in the list of tied source languages gets 1 / L points added to their counter instead of 1, where L is the number of languages that have the same number of matching tokens in a tweet; and ignore, where no language counter changes in case of a tie.

For our second dataset, we did the same except the difference is the words in our Sentiment dataset are labeled individually, which we expected to yield better results. So, for each word in our etymology dataset, we search the sentiment dataset to find its sentiment and save the result. For most languages, we looked at the loanwords they borrowed from other languages. The languages we looked at were only Indo-European languages, namely Dutch, English, Finnish, French, German, Italian, Polish, Portuguese, Romanian and Spanish. We also included ancient, medieval, and early-modern languages for the origin of words such as Latin, Old French, Old Norse, etc. The results of these methods are discussed in the next section.

In order to determine if there was a link between the historical relations and our predictions we obviously had to come up with a set of predictions (discrete, not continuous, unlike our results). We thought of ways to make this objective by webscraping the Wikipedia pages of different countries that used certain languages and maybe counting the number of wars they had with the other countries which spoke the languages we are interested in, but we thought that there are too many things that could go wrong with that. So we instead decided to manually set the relations between countries/languages and argument our decision. The predictions and the arguments can be seen in the 'functions.py' file. For example we said that French has a positive sentiment of English loanwords since the two countries have been allies ever since the world wars, but that's not the case for Middle French loanwords in English due to the 100 years war.

5 Results

For the part where the sentiment of loanwords was analyzed directly from a sentiment lexicon we found a few things. First of all, we saw that not all languages are equal when it comes to the sentiment of the loanwords they have. Results varied quite a bit from language to language. For example in Figure 1 we can see that words of a Dutch origin are generally negative while words of a Portuguese origin are positive. Our sentiment data set has 55% negative words in it and we noticed that if we artificially changed that, our results would tend towards that ratio. So rather than looking at absolute ratios, we instead decided to look at the difference from the base ratio which is 0.5561. So that means, for example, that 37% $[(0.55-0.18)\cdot 100]$ of English words that originate from Portuguese have a negative sentiment to them. Ironically enough, the idea for this project started from hearing that words of an Ottoman-Turkish origin are generally negative in Romanian, but our data shows they are actually more positive than average. All graphs can be seen in the Appendix or the repository of the project.

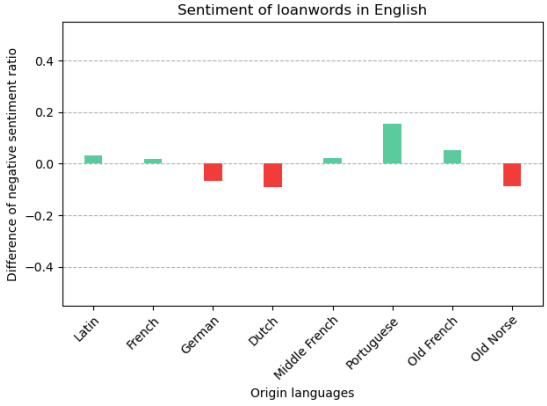


Figure 1: Sentiment of English loanwords

In the context analysis of the 1.6 Million Tweet data set, we found (rather unsurprisingly) that Latin loanwords appeared most often in the data (excluding names and tweet descriptions). However, this analysis also revealed that Dutch and French loan-

words clearly appear more often in positive contexts than in negative contexts, while German and Latin loanwords show the opposite trend. The findings about Portuguese loanwords are rather inconclusive despite showing a tendency towards positive appearances, as there were much fewer matches compared to the other languages (only about 1800). As the random method of dealing with ties in the 'random' dominant language counting method mathematically converges to the same result as the fractions method, its graphs are not displayed here. The absolute number of matches, the dominant-fraction matches and the dominant-ignore matches are displayed in graphs that can be found in the Appendix or the repository. There are two versions for each, one not scaled and one scaled. The scaled versions show the positive-negative ratios more clearly, especially for those source languages that have fewer matches overall, while the unscaled figures pronounce exactly these unbalanced appearances between source languages. Our results from the tweet dataset versus the other sentiment sets do not quite match up: Latin was more negative in the tweet dataset, for example. This could be attributed to the style of writing tweets represent, which may consist of more negative Latin words. Or, it could be due to our assumption that a positive tweet contains mostly positive words.

Originally (before the presentation) we had 15 'receiving' languages and when we compared our results with the historical predictions we got 63% of them right, which has a 2% chance of happening randomly and so is statistically relevant with a 95% confidence interval. We then enlarged our word origin data sets using Wiktionary, but reduced the number of 'receiving' languages to 10 by cutting out languages where we were less sure about the historical relationship as well are reducing the number of 'origin languages'. This time we got 57% of them right, which has a 11% chance of happening (a lot higher since there were less languages). Our second result is only relevant with a confidence interval of 85% which is not great.

6 Conclusion

The purpose of this paper was twofold: firstly, we wanted to find the sentiment of loanwords for different languages, and secondly we wanted to research if the determined sentiment could be attributed to historical relations. Having set off with background research that seemed to support our hypothesis that there is a positive relation between the two, we analyzed both etymology lexicons and tweet datasets to find the sentiment of loanwords. The acquisition of these datasets and the subsequent processing required techniques such as web scraping, and tokenization and lemmatization. After having the data find what sentiment loanwords from a source language in a recipient language had, we proceeded to manually and qualitatively (we didn't look at how negative or positive the sentiment was, just its direction) predict the sentiment of the loanwords based on the historical relations between the recipient and source country.

This resulted in a, albeit statistically dubious, conclusion that our hypothesis was true. However, some flaws of our research can explain this answer. First and foremost, our basis for historical relations were sometimes superficial and subjective. It is not so simple to determine a very strong negative or positive relation between countries, due to hundreds of years of history with many different events. Categorizing these events for a proper understanding of relations is a whole paper on its own. Even then, we do not know *when* a word was borrowed: was it during a period of war, or maybe it was during a small period of good relations in an otherwise troubled history? This would most probably influence the sentiment of a loanword. Secondly, and this is more of a caveat, the quality of the dataset can be doubted in some specific cases. For example, 'fall out of grace' was determined to have a positive sentiment, together with many other words containing 'fall'. Therefore, we must conclude that our results are inconclusive, and require further optimization, larger and better quality datasets, and a more objective and encompassing way of determining historical relation to achieve a proper final result.

Division of tasks: Victor - searched for and processed the sentiment lexicons, Wiktionary webscrap-

ing, wrote initial reports, made graphs. Daniel - EZGlot webscraping, writing up most of the report, cleaning up code and comments. Tim - wrote sentiment results function, did twitter part of project

References

[1] Andrea Esuli and Francis Bond. Sentiwordnet 3.0.1. `https://github.com/aesuli/SentiWordNet`.

[2] EZGlot. List of spanish words of latin origin. `https://www.ezglot.com/etymologies.php?l=spa&l2=lat&submit=Compare/robots.txt`.

[3] Agence France-Presse. France bans english gaming tech jargon in push to preserve language purity. `https://www.theguardian.com/world/2022/may/31/france-bans-english-gaming-tech-jargon-in-push-to-preserve-language-purity`, May 2022.

[4] Elly Gelderen. A history of the english language. *A History of the English Language*, pages 1–358, 2014.

[5] Robert A. (Robert Anderson) Hall. *Comparative Romance grammar / [Vol. 1], External history of the Romance languages.* Foundations of linguistics series. American Elsevier Pub. Co, 1974.

[6] Mark Irwin. *Loanwords in Japanese*. John Benjamins Publishing, 06 2011.

[7] Kazanova. Sentiment140 dataset with 1.6 million tweets. `https://www.kaggle.com/datasets/kazanova/sentiment140`.

[8] Bing Liu. Opinion lexicon. `https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon`.

[9] Joseph Sung-Yul Park. The local construction of a global language: Ideologies of english in south korea. *The Electronic Journal for English as a Second Language*, 14(4), 2011.

[10] Christopher Potts. Sentiment symposium tutorial: Lexicons. `http://sentiment.christopherpotts.net/lexicons.html`.

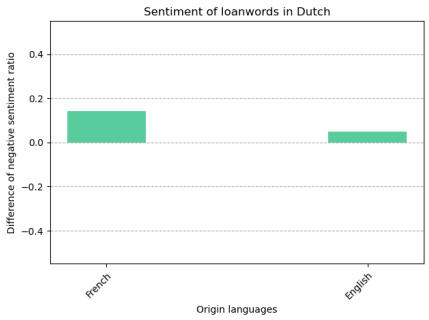
[11] Sofia Rüdiger. Mixed feelings: Attitudes towards english loanwords and their use in south korea. *Open Linguistics*, 4(1):184–198, 2018.

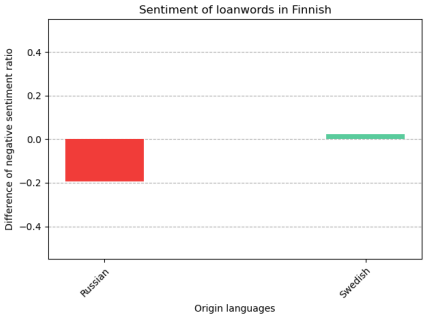
[12] Wiktionary. Category:spanish terms derived from latin. `https://en.wiktionary.org/wiki/Category:Spanish_terms_derived_from_Latin`.

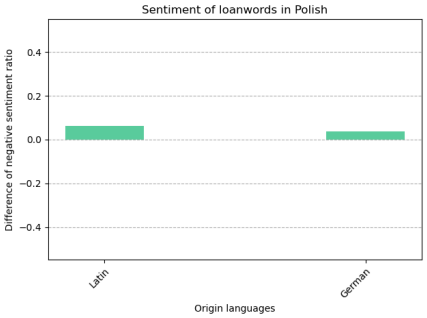
[13] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Subjectivity lexicon. `https://mpqa.cs.pitt.edu/lexicons/subj_lexicon/`, 2005.

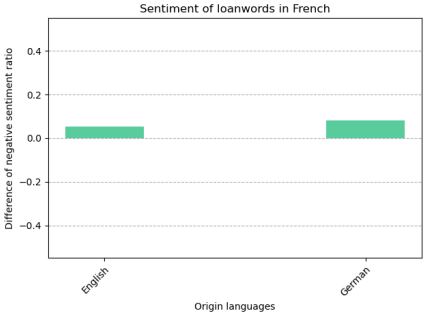
[14] Eline Zenner, Laura Rosseel, and Andreea Simona Calude. The social meaning potential of loanwords: Empirical explorations of lexical borrowing as expression of (social) identity. *Ampersand*, 6:100055, 2019.

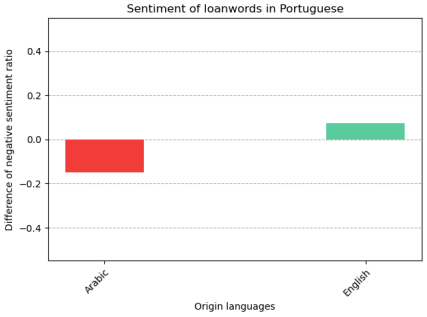
Appendix

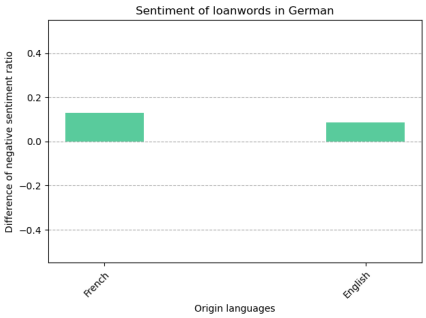


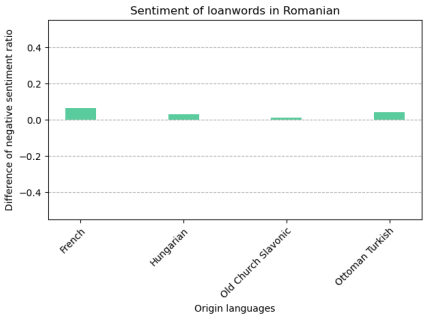

Sentiment of loanwords in Finnish
Difference of negative sentiment ratio
0.4
0.2
0.0
−0.2
−0.4
Russian
Swedish
Origin languages

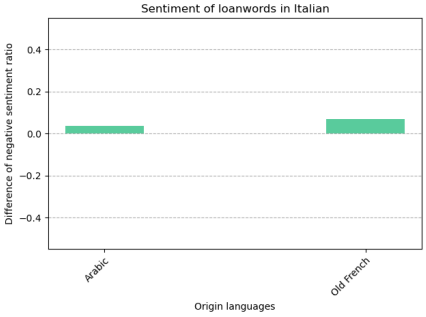

Sentiment of loanwords in Polish
Difference of negative sentiment ratio
0.4
0.2
0.0
−0.2
−0.4
Latin
German
Origin languages

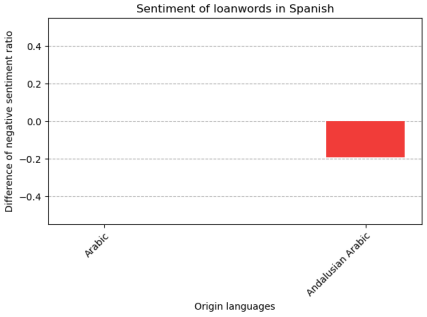

Sentiment of loanwords in French
Difference of negative sentiment ratio
0.4
0.2
0.0
−0.2
−0.4
English
German
Origin languages

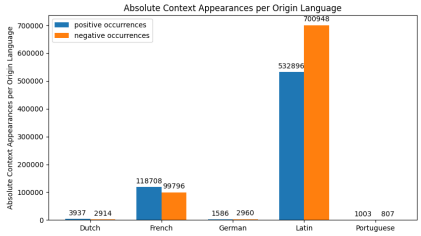

Sentiment of loanwords in Portuguese
Difference of negative sentiment ratio
0.4
0.2
0.0
−0.2
−0.4
Arabic
English
Origin languages

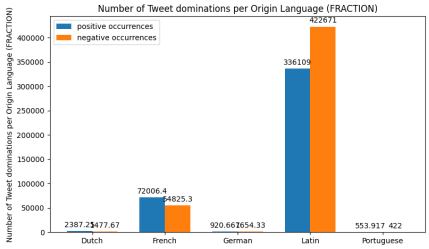

Sentiment of loanwords in German
Difference of negative sentiment ratio
0.4
0.2
0.0
−0.2
−0.4
French
English
Origin languages

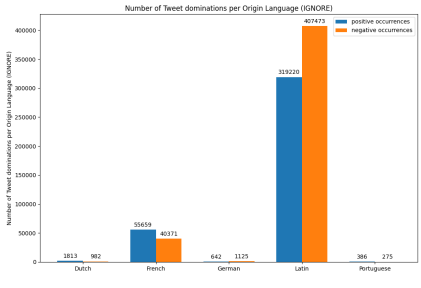

Sentiment of loanwords in Romanian
Difference of negative sentiment ratio
0.4
0.2
0.0
−0.2
−0.4
French
Hungarian
Old Church Slavonic
Ottoman Turkish
Origin languages

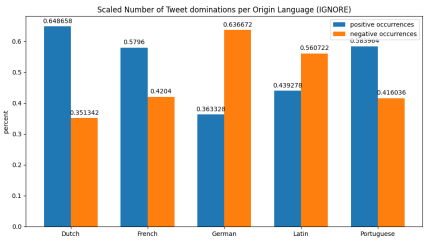

Sentiment of loanwords in Italian
Difference of negative sentiment ratio
0.4
0.2
0.0
−0.2
−0.4
Arabic
Old French
Origin languages

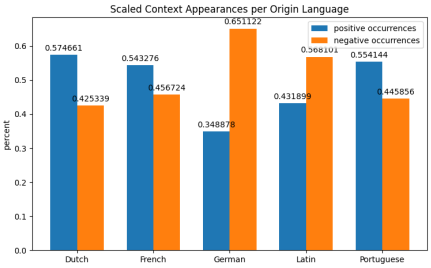

Sentiment of loanwords in Spanish
Difference of negative sentiment ratio
0.4
0.2
0.0
−0.2
−0.4
Arabic
Andalusian Arabic
Origin languages

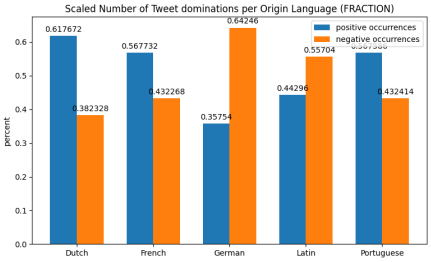

Absolute Context Appearances per Origin Language
positive occurrences
negative occurrences
Absolute Context Appearances per Origin Language
3937
2914
118708
99796
1586
2960
532896
700948
1003
807
Dutch
French
German
Latin
Portuguese


Number of Tweet dominations per Origin Language (FRACTION)
positive occurrences
negative occurrences
Number of Tweet dominations per Origin Language (FRACTION)
2387.25
477.67
72006.4
54825.3
920.66
654.33
336109
422671
553.917
422
Dutch
French
German
Latin
Portuguese


Number of Tweet dominations per Origin Language (IGNORE)
positive occurrences
negative occurrences
Number of Tweet dominations per Origin Language (IGNORE)
1813
982
55659
40371
642
1125
319220
407473
386
275
Dutch
French
German
Latin
Portuguese


Scaled Number of Tweet dominations per Origin Language (IGNORE)
positive occurrences
negative occurrences
percent
0.648658
0.351342
0.5796
0.4204
0.636672
0.560722
0.439278
0.583584
0.416036
Dutch
French
German
Latin
Portuguese


Scaled Context Appearances per Origin Language
positive occurrences
negative occurrences
percent
0.574661
0.425339
0.543276
0.456724
0.348878
0.651122
0.431899
0.554144
0.445856
Dutch
French
German
Latin
Portuguese


Scaled Number of Tweet dominations per Origin Language (FRACTION)
positive occurrences
negative occurrences
percent
0.617672
0.382328
0.567732
0.432268
0.35754
0.64246
0.44296
0.55704
0.567586
0.432414
Dutch
French
German
Latin
Portuguese