Statistical analysis of 1-grams flow among Indo-European languages

Name1 Surname [1,2]❶, Name2 Surname [2]❶, Name3 Surname [2,3]¤, Name4 Surname [2], Name5 Surname [2]‡, Name6 Surname [2]‡, Name7 Surname [1,2,3]*, with the Lorem Ipsum Consortium¶

1 Affiliation Dept/Program/Center, Institution Name, City, State, Country
2 Affiliation Dept/Program/Center, Institution Name, City, State, Country
3 Affiliation Dept/Program/Center, Institution Name, City, State, Country

❶These authors contributed equally to this work.
‡These authors also contributed equally to this work.
¤Current Address: Dept/Program/Center, Institution Name, City, State, Country
†Deceased
¶Membership list can be found in the Acknowledgments section.
* correspondingauthor@institute.edu

Abstract

Author summary

Introduction

In recent years, the field of linguistics has been benefited from the development of more sophisticated computational tools, helping to process a greater amount of data in less time and allowing the study of linguistics from a statistical perspective. This statistical study began with the works of George Zipf [?], in them Zipf argues that if the words used in a text are ranked by their frequency of appearance, where the lower ranks belong to the most frequent words, then the frequency f of any word and its rank k are related by a power law of the form f 1/k. The previous expression is known as Zipf's law, and it has not only been tested on language datasets, Morales et al. [?] have also proved on sports and games data, and Cristelli et al. [?] on the gross domestic product of several countries, wealth of American citizens, and population of cities.

Although Zip's law has opened several statistical studies in linguistics, nowadays few studies have been done about how within the vocabulary of a language, words from the language itself and from other languages are mixed. Currently in the spanish language, there are words from the english language that do not have a translation or that sometimes displace those that already exist in spanish. For example, for native spanish speakers it is common to hear the word marketing instead of mercadeo when dealing with economic or business issues; also the word online has replaced en línea, when referring to issues related to the internet, a word without translation in spanish.

This trending is not only affecting spanish, but also other languages that are being influenced by topics where english is the main and common language for communication. However, in different periods of time, the flow of words came from other languages. D'Amore [?] discusses with linguistic rigor the flow of words between english and spanish, showing historical and cultural causes that allowed such flow; in addition to mentioning the influence of arabic in spanish and french in english.

In this work, we use the Google Books N-gram [**?**] dataset of the most frequent words in books published in english, french, german, italian and spanish languages. With that dataset, we develop an algorithm that identifies the words of one language and that are being used by others. Once these words have been classified, we construct two models to quantify the influence that one language has had on another during the 20th century. The first model, we count the number of new words that a language received from another, while in the second method, we develop the concept of the use of one language in another, from quantifying the relative frequency of the words of a language that are being used in another language. In both we identify historical, social and cultural causes that are responsible for the flow of words and the constantly use of them.

Finally, we use the concept of rank diversity, that shows the number of words occupying a certain rank across the time. This study shows that regardless of who is the language the flow comes from or who language receives them, the lower ranks are always occupied by fewer words, and as the rank increases the diversity curve also increases as a logarithmic.

Metodology

For the development of this work, we used the Google Books Ngram data set. This dataset is made up of listing for each year and for each language of publication the most used "n-grams". The "n-grams" are the words or set of words that make up the text of a book, where the number n indicates the number of words that make up the gram, being a 1-gram an individual word, a 2-gram a phrase composed of two 1-gram, 3-gram the set of three 1-gram and so on.

From this data set, the lists of the five thousand most used 1-grams each year between 1740 and 2009 were extracted for the English, French, German, Italian and Spanish languages. In each list, the words are ranked according to their frequency of appearance, where the most frequent words have the lowest ranks.

To determine the presence of one language in another, an algorithm was developed to find the words that are common between at least two languages, these must be the same in writing, letter by letter. These words were defined as **migrant words**.

A migrant word is associated with a **source language** and a **receiving language**, where the source language is the one where the word appeared for the first time within the five thousand most used words, while the receiving language is the one where the word is also present, being a different set from the source language. To determine the source language, we established that this will be the language where the word appeared for the first time within the five thousand most used words; in the case of a migrant word has appeared in the same year in two or more languages, the source is the one where the word has the lowest rank.

The previous criteria of looking for words with the same writing and later associating them with a source language, is not perfect. One of the most common errors was finding words with the same writing, but with different meanings, for example mayor in English refers to the representative of the government in a locality, while in Spanish, mayor is an adjective to indicate that something is bigger or older. Another recurring error was to not distinguish words with the same meaning but with different endings, for example, the word imagine is written imaginer in French and imaginar in Spanish. Finally, in some cases, the authentic source language is some other language for which there is no information in the data set, for example the word natural comes from Greek, but there is no data from Greek language in the google books n-grams data set, consequently this word was associated with English as its source language.

The above errors were detected by individually analyzing each of the migrant words and their corresponding source and receiving languages. One way to have cleaner data

is by consulting an expert in each language, who review the words and decide which ones were classified properly, however, this is not practical since if there were more languages in the database, it would be necessary to consult an expert for each language.

New words

The purpose of our work is to find a method to quantify the influence that one language has had on another, however there is no a generalized process that shows how much influence the things have. We built a first method to quantify the influence that one language has on another, this consisted of analyzing the times when a migrant word with a certain source language appears for the first time in the receiving language, finding relationships between those times and the words involved on the move. Migrant words with the same source language and that appear for the first time in the receiving language were referred to as **new migrant words**.

The previous process was carried out in three different ways, the first by taking a source language A and counting the new migrant words in the different receiving languages for each decade, this shown in which languages A has been influential; the second way was by taking language A as the receiving language, and counting how many new migrant words of different source languages are present in each decade, this shown the languages that have been influential on A; finally, by taking a language A and a language B, and counting how many new migrant words from A are in B, and how many from B are in A, this shown the influence between A and B.

A complementary part of interpreting the influence between languages is to recognize the causes that originate the migrations of words. These were inferred after analyzing the meaning of the new migrant words in a given year of migration. According to [?] a semantic field is a set of associated words that share part of their meaning. Then, from classifying the migrant words within a semantic field, the cause or event that originated the migrations might be found, since the migrations occur in the same year (or in the years around) where the event occurred.

The fact of looking for relationships between the migrant words and their meaning, we decided to omit the functional words from the results, these words are those that help to structure a message according to the grammar of the language, such as articles, prepositions and conjunctions. In this way, all new migrant words are content words, that is, those that carry the information and meaning of the message.

Analysis of new migrant words

In order to quantify how many new words have migrated between the English, French, German, Italian and Spanish languages, Fig 1 shows, for each decade of the 20th century, the behavior of each language, being the source language or the receiving language of new words, representing how one language has influenced others and also how it has been influenced. Throughout the 20th century, the English language has contributed almost three times as many words as it has received, where its greatest apogee occured in the 1940s, during which World War II occurred; and in the first decade of 2000, where access to technology was feasible for larger sectors of the population, after the development of globalization.

It is also notable that the French and German languages were mainly languages influenced by others, since they received more words than they contributed, this trend being more appreciable since 1940 after the end of World War II. In addition, since 1980, German has obtained 30 new words every decade compared to the previous decade, consequently it has been the largest receiver in the last 30 years.
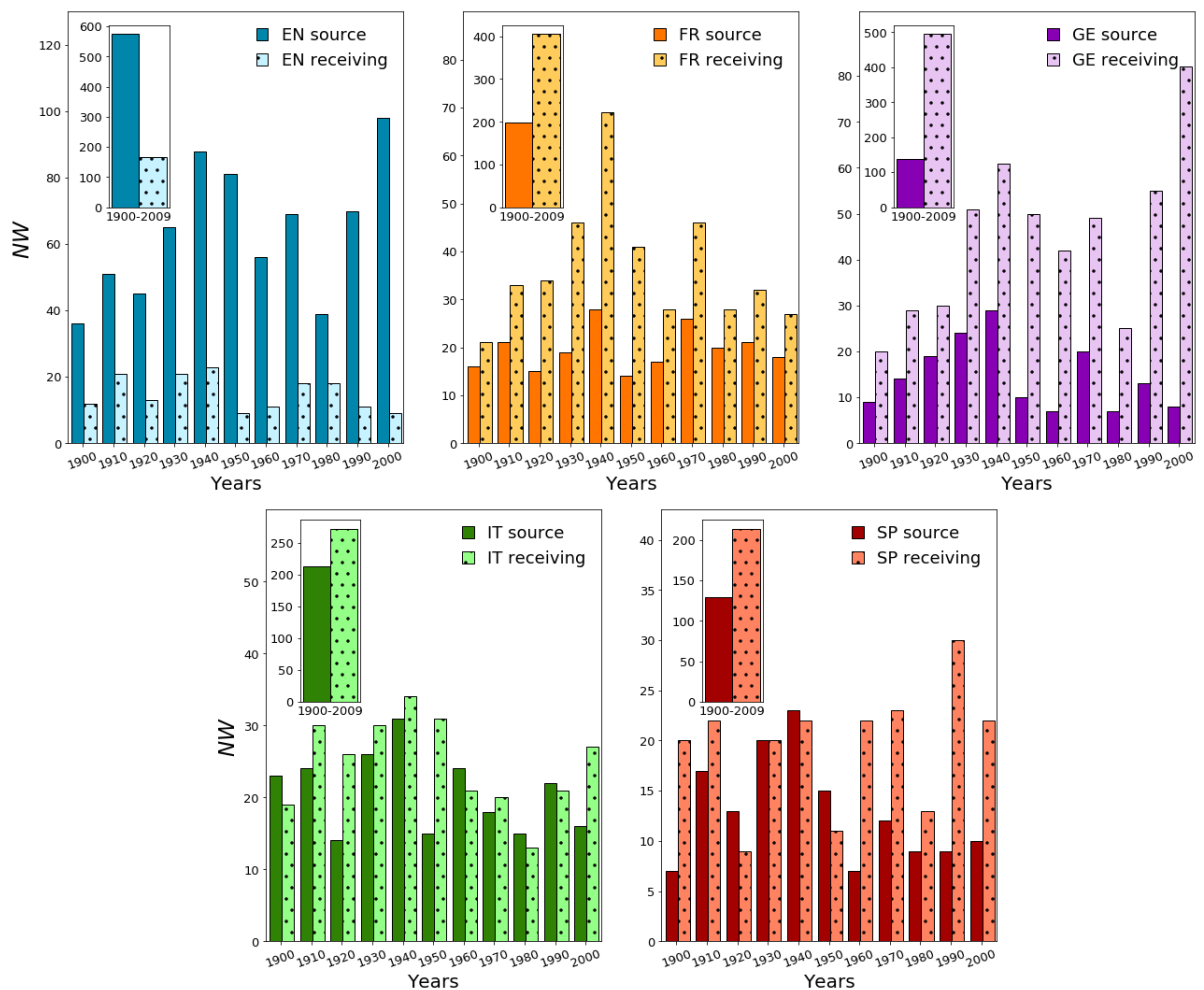
Fig 1. Languages as sources and receivers of new words. During the 20th century, only English language has been the language that has migrated more words than it has received. French, German, Italian and Spanish have received more words during the second half of the century, after the end of World War II.

The Italian language decade by decade, contributes almost the same number of words as those it receives, the only exceptions occurred in the decades of the 1920s and 1950s after the end of the First and Second World War, decades where it was mainly a receiving language. Finally, the Spanish language in 1930 and 1940, also contributed the same number of words as those it received; in the other decades it was primarily a receiving language.

The previous results only manage to show in which decades the languages contributed or received more words, it being evident that half a century after the end of World War II, the English language began to be primarily the source language that contributed the most words to others.

Another way to interpret the previous results is shown in Fig 2. This shows how the new migrant words that each source language contributed are distributed in the different receivers for each decade. To exhibit that migrant words are related by their meaning, some new migrant words that were related to certain semantic fields are display in table 1; specifying the language they come from, to which languages they
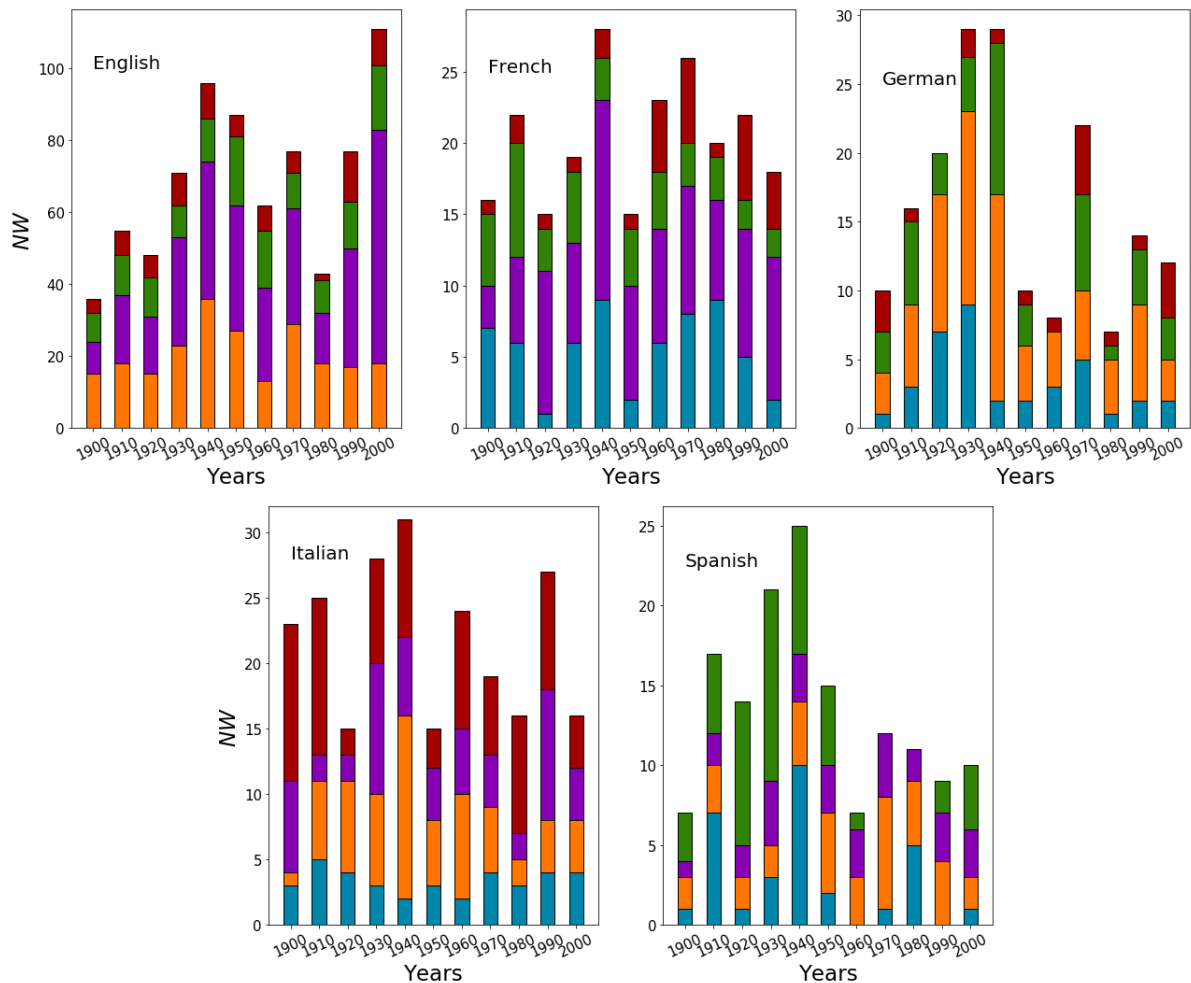
Fig 2. New migrant words that migrate from one source language to the different receiving languages (English ■, French ■, German ■, Italian ■, and Spanish ■).

Table 1. Table

Heading1				Heading2			
*cell1row*1	cell2 row 1	cell3 row 1	cell4 row 1	cell5 row 1	cell6 row 1	cell7 row 1	cell8 row 1
*cell1row*2	cell2 row 2	cell3 row 2	cell4 row 2	cell5 row 2	cell6 row 2	cell7 row 2	cell8 row 2
*cell1row*3	cell2 row 3	cell3 row 3	cell4 row 3	cell5 row 3	cell6 row 3	cell7 row 3	cell8 row 3

Comentarios

The English language as the source language, contributed more words to the French, German and Italian languages, between the 1930s and 1950s. In this period the words found are related to the semantic field of World War II; where countries such as the United States of America, the United Kingdom, France, Italy and Germany were involved. In the last twenty years (1980-2000), the French, German, Italian and Spanish languages have received from English words with content on economics, technology and globalization; we associate this increase with the fact that the English language has been positioned as the common language for communication and for spreading information. 139 140 141 142 143 144 145 146

Supporting information

Acknowledgments

References

1. Conant GC, Wolfe KH. Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet. 2008 Dec;9(12):938–950.

2. Ohno S. Evolution by gene duplication. London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag.; 1970.

3. Magwire MM, Bayer F, Webster CL, Cao C, Jiggins FM. Successive increases in the resistance of Drosophila to viral infection through a transposon insertion followed by a Duplication. PLoS Genet. 2011 Oct;7(10):e1002337.