Statistical analysis of word flow among five Indo-European languages

Josué Ely Molina[1,2†], Carlos Gershenson[3,4,5,6†], Tomás Basile[2,7†],
Carlos Pineda[2*†]

[1]Facultad de Ciencias, Universidad Nacional Autónoma de México, Coyoacán, Mexico City, 01000, Mexico .
[2]Instituto de Física, Universidad Nacional Autónoma de México, Coyoacán, Mexico City, 01000, Mexico .
[3]Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas, Universidad Nacional Autónoma de México, Coyoacán, Mexico City, 01000, Mexico .
[4]Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Coyoacán, Mexico City, 01000, Mexico .
[5]Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA .
[6]School of Systems Science and Industrial Engineering, Binghamton University 4400 Vestal Parkway East, Binghamton, NY 13902, USA .
[7]Faculty of Science, University of Copenhagen Bulowsvej 17 Frederiksberg, Copenhagen 1870, Denmark .

[†]These authors contributed equally to this work.
* carlospgmat03@gmail.com

Abstract

A recent increase in computational processing has allowed the possibility to perform different linguistic studies with large datasets. Here we use the Google Books Ngram dataset to analyze word flow among English, French, German, Italian, and Spanish. We study what we define as "migrant words", a type of loanwords that do not change their spelling. We quantify migrant words from one language to another for different decades, and notice that most migrant words can be aggregated in semantic fields and associated to historic events. We also study some properties of accumulated migrant words and their rank dynamics. We propose a measure of *use* of migrant words that could be employed as a proxy of cultural influence.

Keywords: Languages, ranking, word flow

1 Introduction

In recent years, the increase of data availability [1] and the development of computational tools [2] has benefited various statistical studies to understand certain characteristics of the human population. For example, we are able to predict with a high confidence the growth rate of a city [3, 4], the number of people who have watched a movie [5], the user traffic on a web page [6], and even the way we use words in written language [7, 8]. The previous examples are cases of Zipf's law, formulated by George Zipf in the 1930s [9–14] upon discovering that if the words used in a text are ranked by their frequencies 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 $f \sim 1/k$.

Zipf's law has been mostly used to study the structures of language. Nonetheless, not enough studies have been made to understand the historical and cultural features that language provides. One way to begin such a study is by noting that the languages themselves are mixed, since within the vocabulary of a language, words from other languages are continuously added.

Currently in the Spanish language, there are loanwords from English that do not have a translation or that sometimes displace those that already exist in Spanish. For example, for native Spanish speakers in Mexico, it is common to hear the word *marketing* instead of its translation *mercadeo* when dealing with economic or business issues; also the word *online* has replaced *en linea*, when referring to issues related to the *Internet*, a word officially adopted in Spanish.

This trend has not only affected 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 [15] discusses with linguistic rigor the flow of words between English and Spanish, showing historical and cultural causes that allowed such a flow; in addition to mentioning the previous influence of Arabic in Spanish and French in English [16–18].

In this work, we use the Google Books Ngram dataset [19] of the most frequent words in books published in English, French, German, Italian, and Spanish. With this dataset, we develop an algorithm that identifies the words of one language and that are being used with exactly the same spelling by others (see section 2). Once these words have been classified, we construct two approaches to quantify the influence that one language has had on another during the 20th century. In the first approach (section 3), we count the number of new words that a language received from another. In the second approach, we develop the concept of the *use* of one language in another, by quantifying the relative frequency of the words of a language that are being used in another language (section 4). In both approaches, we identify historical, social, and cultural causes that are related to such words. In section 5, we measure rank diversity [20], that quantifies the variety of words occupying a certain rank across time. This study shows that regardless of the original or receiving language, the lower ranks are always occupied by fewer words, and as the rank increases, the diversity curve also increases following a universal sigmoid curve. Finally, we measure the robustness of our results by removing migrant words and comparing the resulting sets with the original ones in section 6, before closing with a discussion (section 7).

Studies like the one we present can be complementary to detailed linguistic studies of loanwords. Certainly, we do not attempt to replace such studies, but to add insights and suggest further avenues of research. We do not need to sacrifice precision or large amounts of data [21] when we can have both.

2 Methodology

We used the Google Books Ngram dataset [19]. This dataset contains the usage frequency, for each year and language, of the most used “N-grams” in Google Books. 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 contiguous words that make up the gram, being a 1-gram an individual word, a 2-gram a pair of words, a 3-gram a sequence of three words, and so on.

We polished our final dataset by removing function words, that is, words that do not carry significant information and only give structure to the text. This includes words such as articles, pronouns, prepositions, conjunctions and determiners. To select the words to be removed, we used lists of function words from the natural language processing Python library spaCy [22]. After removing the functional words from the dataset, the lists of the five thousand most used 1-grams each year between 1800 and 2009 were extracted for the English, French, German, Italian, and Spanish languages. We are performing this cut as all the lists of the five languages (between 1800 and 2009), have at least this amount of 1-grams. It is important that all languages have the same number of words, so that there is no bias towards one of them. For each language and each year, 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 have exactly the same spelling. These words were defined as *migrant words*, which are a particular case of loanwords.

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, but appeared in the top five thousand most used words at a later time. If a migrant word appeared in the same year in two or more languages, the source is the one where the word has the lowest rank.

The previous criterion for searching words with the same spelling and later associating them with a source language is not perfect. There are some cases that our method did not detect and were established as mistakes. One of the most common errors was finding words with the same writing, but with different meanings (polysemy). 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 greater, bigger, or older. Another recurring error was not distinguishing words with the same meaning but with slightly different spellings. 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 dataset, for example the word *natural* comes from Greek, but there is no data from Greek in the Google Books N-grams dataset, nor from the years that the migration occurred. Consequently, our algorithm sets Spanish as source language for this word. To reduce this type of mistakes, in our algorithm we only consider migrant words that appeared in the receiving language after 1850. That way, words that simply are common to different languages and do not really represent an influence of one over another will be less likely to be included as migrant words, since it is more probable that they will appear in their languages already between 1800 and 1850.

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 reviews the words and decides 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. Notwithstanding of this requirement to regulate errors, we established a method to determine the importance (weight) of these errors in the results, that will be presented in section 6.

3 New migrant words

The purpose of this work is to establish the influence that one language has on another. A first method to quantify such influence is by counting the *new migrant words* (NMW). These are words that appear for the first time in a receiving language and that come from a source language.

We study the flow of NMW in two ways. First, we count the number of NMW_{out} that a fixed language exports as a source language. Second, counting, for a fixed language, the number of new migrant words (NMW_{in}). In this second way, we can study from which language are the NMW_{in} coming. The results are presented in Fig. 1 for each decade of the 20th century.

From this figure, we can see that the English language has migrated on average two times more words than it has received, where the greatest influence of English occurred in the 1940s and 2000s. Consequently, the largest proportion of migrant words in the other languages come from English. It is worth noticing that French, German, Italian, and Spanish exported more words during the 1940s, but their export rate has remained roughly stable, with minimums for English in the 1900s and 1950s, French in the 1980s, German in the 1920s and 2000s, Italian in the 1920s, and 1950s, and Spanish in the 1960s and 1970s.

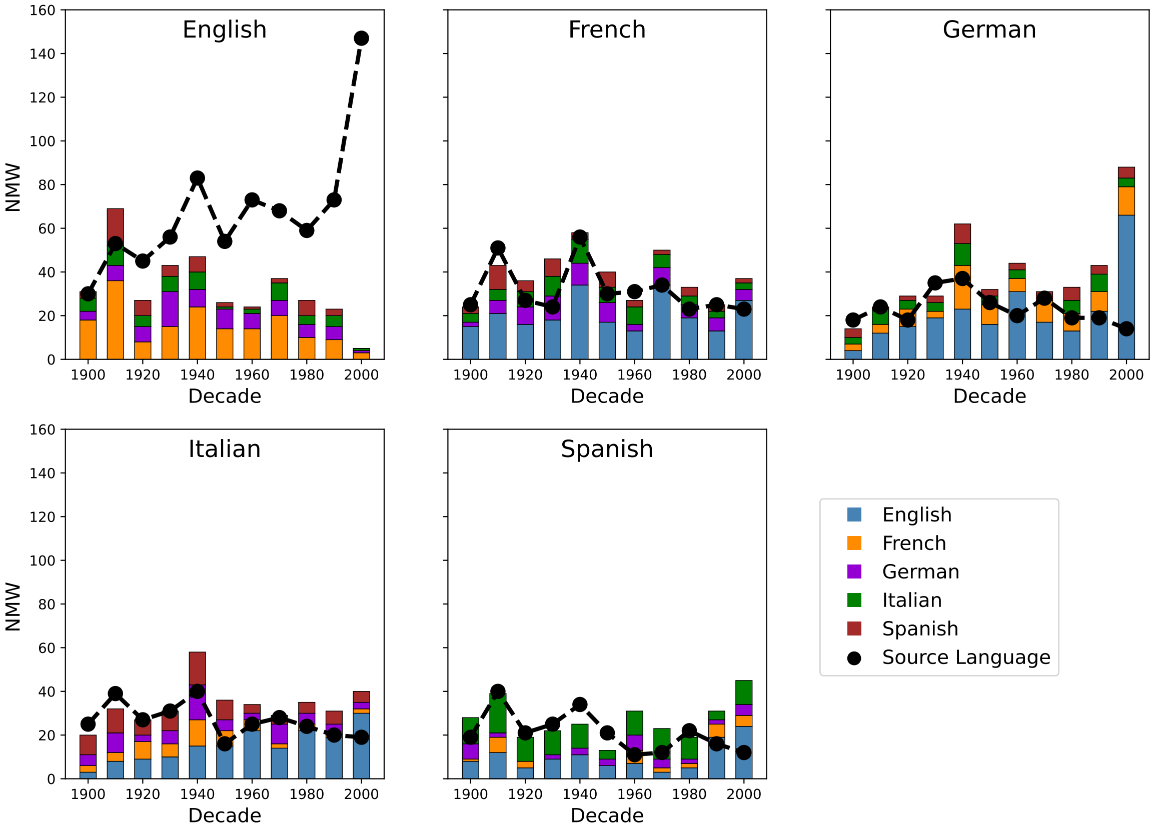

Fig. 1: **New migrant words, per language and per decade.** NMW are considered for each language. Each panel contains one language. The dotted line displays the number of $\mathrm{NMW_{out}}$ that originate in the corresponding language, and the bars the $\mathrm{NMW_{in}}$ coming to that language, separated by the origin of the different NMW.

The major influencer of English has been mostly French. Apart from English, French has received more influence from German and Italian, German from French and Italian, Italian from Spanish, and Spanish from Italian.

Analyzing the lists of migrant words, we realized that these can be grouped into semantic fields. According to [23], a semantic field is a set of words that are related based on their meaning. Table ?? shows the words grouped by semantic fields, as well as the pairs of source language and receiving language involved. We note that some of the migrant words are related to historical or cultural events. For example, between the 1930s and 1940s, words historically related to the Second World War migrated between all languages; while since the 1990s, words referring to technology and globalization migrated from English.

These kind of groupings allow us to understand which languages are most influential and when. The English language has migrated words to others because of

New Migrant Words by Semantic Field

Semantic Field	New migrant words	Source language
World War I	austro, russie, prusse, versailles.	FR
	kaiser, reich.	GE
World War II	roosevelt, churchill, nazis, stalin.	EN
	berchtold, hitler kaiser, lenin, gestapo.	GE
	duce, mussolini, regime.	IT
Aftermath of WWII	onu, urss, vietnam	FR
Historic figures in arts, science and philosophy	bernoulli, laplace. bach, beethoven, engels, freud, hegel, heidegger, marx, mozart, nietzsche.	GE
Ideologies and political terms	burgueoise, diplomatie, politique.	FR
	capitalista, comunista, fascismo, marxismo, socialista, terrorismo.	IT
Economy	depression, dollar, economic, economy, financial, investment, market, marketing, value.	EN
Technology	digital, internet, mail, online, software.	
Globalization	business, customer, management, market, marketing.	
Presidents of the United States of America	roosevelt, kennedy, johnson, nixon, reagan, bush, clinton.	
Medicine	anestesia, lepra, metabolismo, virus, aorta.	SP
Latinoamerican countries and cities	argentina, aires, colombia, chile, panama.	

technological development and globalization in the last thirty years. French, Italian and German were influential after the war events of the 20th century, in addition to the academic influence of Germany seen through surnames of historic figures. Finally, Spanish was influential after economic crises in Latin American countries [24]. The fact that locations from one country become frequent words in another language suggests that some people speaking the latter are interested in the former. Similarly, influence can be seen e.g. as USA presidents become commonly used words (in the top 5000) in other languages. This suggests that migrant words could be used as a proxy measure of cultural influence (see next Section).

Another interesting feature that we observed is that migrant words also fulfill Zipf's law [9]. In Fig 2 we present all language pairs, grouped by receiving language and we observe (within fluctuations), an asymptotic power-law decay with an exponent close to one.

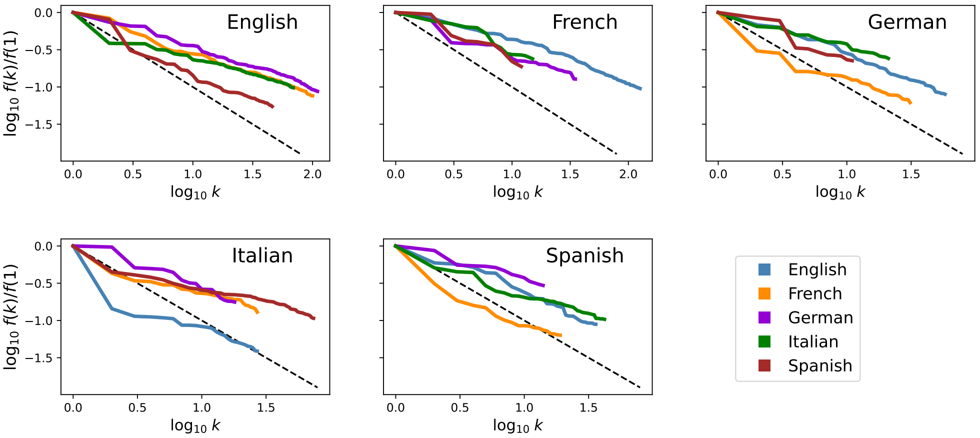

Fig. 2: **Zipf's law of the accumulated migrant words, grouped by receiving language.** We display a frequency-rank plot for all language pairs, for the migrant words during the year 2000. Indeed, after a transient, Zipf's law is observed (a dashed line with slope −1 is provided for comparison).

4 Usage of migrant words

The previous results show that words travel from one language to another in groups belonging to a common semantic field. Nevertheless, we still cannot associate them with a number that quantifies how much influence one language has on another. To obtain such a number, we will focus on migrant words in the years after the first year they migrated, observing how their frequencies vary over time. For example, a migrant word will begin to be influential if its frequency increases over time.

Consider all words that up to a given year t have migrated from language A to B. We call these words the accumulated migrant words from A to B up to year t. At said year t, some of these words will be in the top five thousand most used words in the receiving language B and each of them will have a frequency $f(j)$, where j is the ranking of the word in said year. We now add the frequencies of the accumulated migrant words of A to B at year t and normalize this quantity by dividing it by the sum of frequencies of the first five thousand words that make up the list of the receiving language at year t:

$$\underset{\text{A}\to\text{B}}{U}(t) = \frac{\sum_j f(j)}{\sum_{k=1}^{5000} f(k)}. \tag{1}$$

We define this new value as the *use* of A in B at year t, and interpret its value as a measure of influence. It will then be said that the influence of A has increased on B, if in an interval of time Δt the use of A on B, $\underset{\text{A}\rightarrow\text{B}}{U}$, increases.

We obtained the accumulated migrant words for all possible combinations of source and receiving languages from 1800 to 2009, but only kept those that reach the receiving language after 1850, so as to get rid of as many words with a common ancestry as possible, as mentioned in the methodology. Afterwards, we calculated the use, Eq. (1), for each pair of languages between 1900 and 2009, so as to have a time period (1850-1899) to build a large enough dataset to have meaningful migrant words. The results are presented in Fig. 3, grouped by source language.

It should be noted that our method is not perfect, as there are some homonym words that are not migrant between these languages, but are considered as such. One case is that of words with common origins. For example, *social* is a word in English, Spanish, and French (*sociale* in Italian and French, *sozial* in German). But they all have the same root in Latin *socialis*. Still, *social* (with this spelling) migrated into Italian from English in 1951. On the other hand, *similar* is native to both English and Spanish, but is classified as having migrated to Spanish from English in 1940, probably simply because it had not made it to the top 5000 until then. To avoid these errors, data from Latin and other influencing languages would be necessary, unless languages with few common origins are being compared.

Another limitation comes from homonym words with different meanings. For example, *kino* was a popular word in Spanish in the 1700s because of Jesuit missionary Eusebio Francisco Kino, who was active in the then northwest of New Spain (now northwest Mexico and southwest USA). In preliminary results, then the word *kino* was classified as an influence from Spanish into German (which means cinema). We updated the criteria to consider only words from 1800 (data is sparse before that year, which made the word *kino* more relevant than it really was) and this error was no longer present. Still, few are other similar errors, such as *miles* which appears as migrating from English into Spanish in 1895, but it simply has a different meaning in Spanish (thousands). Conversely, *sales* appears as migrating from Spanish (where it means both "salts" and "you go out") into English in 1903.

English

English has influenced French the most, with a slow but steady increase through the century. It has also greatly influenced German, with a slight increase before and during the Second World War, then a slight slowdown, and then a rapid increase since 1990. Some influential words from English have been *university*, *film*, *computer*, *internet*, *web*, *software*, *dna*, *marketing*, *management*, *international*, *london*, *cambridge*, *oxford*, *york*, *chicago*, *william*, *george*, *james*, *charles*, *american*, *time*, *design*, and *life*.

English has been influenced most by French (which is well known), then by German, and then by Spanish (which has been influenced by English roughly in the same degree). Still, these influences have decreased roughly since 1980. The most relevant words that are considered migrants into English (remember that they need to reach the top 5000 in the year of their migration) are names of people (*francisco* from Spanish in 1861, *jean* from French in 1871) and places (*florida* from Spanish in 1866,

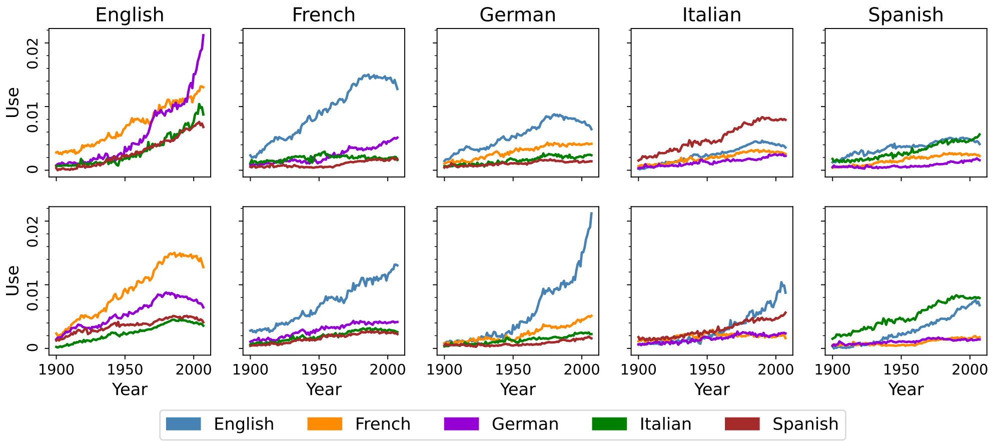

Fig. 3: **The use U among languages.** We plot the use, as defined in Eq. (1) for all language pairs between 1900 and 2009: on the top row, how column languages influence plotted languages, while the bottom row shows how column languages are influenced by plotted languages. Results are discussed in the main text.

vietnam from French in 1964), and other specific words (e.g., *piano* from Italian in 1908, *plasma* from German in 1951 (which has a Latin origin, but it seems it was popular in German physics earlier)).

French

From the languages studied, French has influenced most English, as already mentioned. It has also influenced Italian, and German, especially after 1950. Apart from words with Latin origin that were more common in French first, names such as *jean*, *marie*, *pierre*, and *foucault* have also migrated from French into other languages.

English has been the most influential language for French, while Italian and Spanish have influenced it only slightly (is spite of all three being Romance (Neo-Latin) languages). Other names that have migrated into French (apart from those mentioned from English) include *gabriel*, *jose*, *aires* (1899, 1981, 1914 from Spanish), *freud*, *marx*, *heidegger*, *nietzsche*, *hitler* (1956, 1923, 1983, 1905, 1932 from German), and *franco* (1886, from Italian)

German

Some historic popular words (top 5000) in the early twentieth century included *bismarck* (1886), *kaiser* (1915), and *balkan* (1915). Before and during the Second World War, *lenin* (1931), *marx* (1934), *proletariat* (1934), *beethoven* (1927), *wagner* (1911), *reich* (1939), and *hitler* (1934) were popular. Some science-related words also appear as migrating from German to other languages, because many scientific advances were first published in German (before the 1930s).

German has been influenced mostly by English, specially after 1950, with a great increase since 1990. This might be related to the fall of the Berlin Wall. Apart from words already mentioned in previous sections, examples of migrant words into German include *germany* (1972), *law* (1950), *microsoft* (2004), *party* (1962), *copyright* (2007), and *xml* (2008).

Italian and Spanish

Our analysis shows that Italian has influenced Spanish the most. The majority of these words were first popular in Italian, and then became popular in Spanish. But they did not migrate directly from Italian, as they are homonyms with the same Latin root. This is also the case, but less frequent, for French.

The reciprocal is also observed, with words becoming popular first in Spanish and then in Italian. However, Spanish has influenced English more than French.

5 Rank diversity

In the previous sections, we quantified the influence of a language on another. However, one can wonder about how the migrant words change in time. Are the most important words the same, or do they change? In fact, since the accumulated migrant words are organized by year, and at the same time in each year the words are ordered in ascending order in rank, then over time, the same rank can be occupied by different words. One way to quantify this change is through rank diversity $d(k)$ [20]. This quantity is defined as the number of different elements that occupied rank k within the same dataset, divided by the number of time slots considered. Rank diversity has been used in datasets of the most used words in six Indo-European languages [20, 25, 26], in sports and game classifications [27], and in many other datasets [28]. Although in previous studies of rank diversity of languages and the current one the criteria for establishing rankings are different, in both there is a common result: the lowest ranks are always occupied by fewer elements, thereby as the rank increases, the number of different elements that occupied it also does.

After calculating the rank diversity (considering all years) for each source and receiving language pair, the diversity values resemble a sigmoid curve, as can be seen in Fig 4. This can be fitted with a curve that is cumulative of a Gaussian centered at μ and with deviation σ, i.e.

$$\Phi_{\mu,\sigma}(\log_{10} k) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\log_{10} k} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \mathrm{d}y. \tag{2}$$

The parameters μ and σ are obtained with a linear regression. It is observed that the behavior of diversity increases as the rank also increases, regardless of whether the corpus has few or many ranks (14 in German-Spanish, 290 in Spanish-Italian, etc). With this, it can be concluded that, the migrant accumulated words in the middle and high ranks are the ones that tend to change their position the most within a ranking over time.

These observations suggest that only (relatively) few migrant words are used frequently, during long periods of time, while most migrant words are used not so (relatively) frequently, and their usage varies (relatively) more with time.

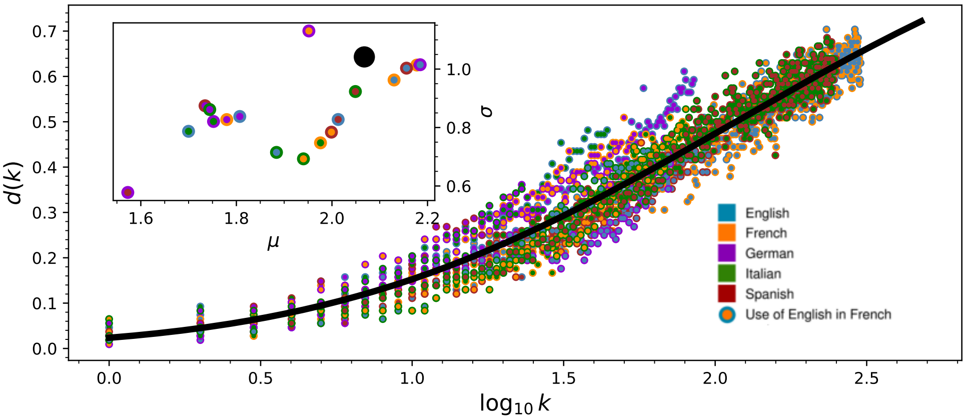

Fig. 4: **Rank diversity of accumulated migrant words among languages.** Diversity for all pairs of languages. Each pair is well fitted by the sigmoid proposed in Eq. (2), with fitting parameters μ and σ reported in the inset. As a reference, we show a global sigmoid (in black) obtained by fitting all data points. Its fitting parameters are also shown in the inset as a black dot.

6 Robustness

The method we used to build the set of migrant words relied on words having exactly the same spelling when going from one language to another. We know that that is not always an accurate description. Some words change their spelling. For example the word *parquear* in Spanish comes from to the verb *to park* in English.

To check the stability of the results presented and the importance of omitting certain words, we proceeded as follows. Take the original set (the one used in the previous section) of the accumulated words of a pair of source language and receiving language. From this set, eliminate a certain group of words, in order to obtain a reduced set; in both, equation 1 is used to obtain the modified use between the years 1900 and 2009. The next thing is to determine how similar the use of both sets are. We normalized the values of both sets, after dividing them by the average value of each one; then for each year t we obtain the distance between each value of original use u_t and its corresponding value in reduced use $\tilde{u}_t$. The average of them gets the *average distance* $\langle D \rangle$, which will be the one that quantifies the similarity of the results, indicating a greater similarity if it is close to zero. This distance is defined as

$$\langle D \rangle = \frac{1}{N} \sum_{t=1}^{N} |u_t - \tilde{u}_t| \,. \tag{3}$$

Recalling that migrant words have frequency inversely proportional to the rank, it is clear that some words are more important than others (see Fig. 2). Thus, care must be taken when one removes a fixed proportion of words, or a fixed frequency, as it can cause a big difference. One way to explore such aspect is to remove words from higher ranks or lower ranks. With these ideas, in each source language and receiving language pair, we carry out two types of elimination, in the first we begin to eliminate the words with the lowest ranks gradually increasing the proportion of words removed R_p (from 1% to a 99%); in the opposite way, for the second case, we begin by eliminating those words with highest ranks. In both cases, each time the eliminated portion was increased, the average difference was calculated to observe the similarity.

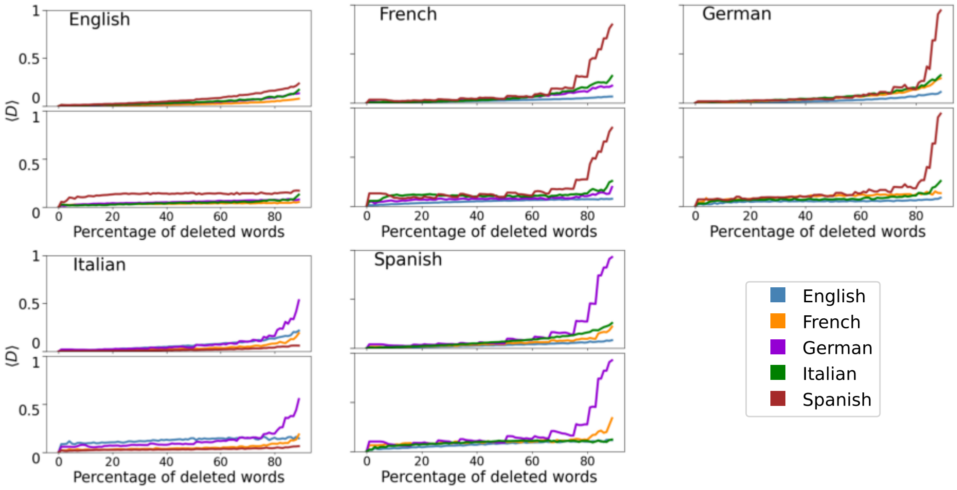

Fig. 5: **Similarity of use when some words are eliminated.** We study the similarity of the shape of usage Eq. (1), as measured by $\langle D \rangle$ [see Eq. (3)], when some words are eliminated. The upper plots correspond to the case in which the lower ranked words (more frequent) are being eliminated first, whereas in the lower plots we start eliminating the higher ranked words (less frequent).

In Fig. 6 we can observe how much the shape of the curves for usage changes, with an increasing proportion of words eliminated. Clearly, when removing the lower ranking words, the deformation is greater. However, we see that in most cases, removing the 60% of the higher ranked words produces a deformation with $\langle D \rangle < 0.1$ (exceptions being German influencing English, Italian influencing German and Spanish influencing German and Italian). This result implies that care must be taken when doing this analysis with respect to words that have low ranks. Thus, using this automated approach to yield quantitative statements should pay special attention to the most frequently used words.

7 Discussion

We analyzed how migrant words (loanwords with the same spelling) have spread across five Indoeuropean languages using a “blind big data” approach, that was based on the migration of words from one language into another. This methodology allowed to analyze 5000 words per year along 110 years in five different languages (English, French, German, Italian and Spanish). From the number of words that have migrated, we can draw some conclusions. English has been the language that has contributed more words to other languages, while at the same time, the number of words coming from other languages has diminished. Consistently, the other languages analyzed have English as its most important contributor for new words. Interestingly, we observed that often such word migration is associated with historical events, such as the First and Second World Wars, economical crises, or a burst of technological development. We have also analyzed the usage of migrant words, which quantifies how often those words are actually used in the receiving language. This has actually been consistent with the results for usage of migrant words, where it is also clear that English has the biggest influence on other languages. Moreover, this influence keeps on growing, while the influence coming from other languages has declined or stagnated. However, the effect of historical events on the usage is not as clear as it is on the number of migrant words, except for the sharp increase of influence of English since the 1990s, related to technological development and globalization.

Other aspects, such as Zipf's law and rank diversity were also studied, both of which were consistent with previous studies, namely that the frequency of migrant words also follows a power law (as languages [9], and many other phenomena [10–14]) and that the rank change across time follow a universal pattern seen across a wide range of systems [28]. Additionally from the linguistic aspects of these studies, our findings can be useful to study cultural influence as well. The fact that a name of a place or person from one place is used frequently in another implies relevance. Thus, migrant words can be used also as proxies of cultural influence.

However, our study does have limitations. Languages often alter the spelling of words as they migrate, and our method does not account for these variations. A more sophisticated algorithm would be necessary to include such cases. Furthermore, our methodology is limited to languages that use the same alphabet, though automatic transliteration could potentially broaden our analysis to include languages like Russian. To explore the impact of some of these limitations, we tested the robustness of our results by artificially removing some words and found that the usage patterns remained largely unchanged unless a substantial percentage of words were excluded. This gives us confidence that a more precise analysis would yield similar conclusions.

Looking ahead, advances in computational processing power and the availability of diverse linguistic data are making more sophisticated statistical studies possible. While these methods have limitations, they offer valuable insights that complement traditional approaches in linguistics and culturomics [29, 30]. Future research could explore more specific data sources, such as particular journals, newspapers, and social media. Although our focus was on migrant words due to the nature of our dataset, similar methods could be applied to study the origin, spread, and adoption of cultural information, such as memes in Dawkins's sense [31]. On the one hand, statistical approaches can be used to explore and find potential patterns or insights that should be interpreted by linguists. On the other hand, linguists can exploit novel data availability to test and contrast hypothesis about language usage and change. Ultimately, progress in this field will require collaboration between disciplines traditionally seen as separate in academia.

Abbreviations

NMW, New Migrant Words; En, English; Fr, French; Ge, German; It, Italian; Sp, Spanish.

8 Declarations

Availability of data and materials

The datasets generated and analysed during the current study are available in the Statistical-analysis-of-word-flow-among-five-Indo-European-languages repository, https://github.com/tbasile/Statistical-analysis-of-word-flow-among-five-Indo-European-languages.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by UNAM-PAPIIT IG101324 and a SUNY Empire Innovation Professor startup grant.

Authors' contributions

All authors contributed equally to this manuscript. All authors read and approved the final manuscript.

Acknowledgements

We gratefully acknowledge the contributions of Jorge Flores, whose work on this project during its early stages was invaluable. We are deeply saddened by his passing.

References

[1] Hilbert, M., López, P.: The world's technological capacity to store, communicate, and compute information. Science **332**(6025), 60–65 (2011) https://doi.org/10.1126/science.1200970 http://science.sciencemag.org/content/332/6025/60.full.pdf

[2] Shalf, J.: The future of computing beyond moore's law. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **378**(2166), 20190061 (2020) https://doi.org/10.1098/rsta.2019.0061 https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.2019.0061

[3] Batty, M.: Building a science of cities. Cities **29**(S1), 9–16 (2012) https://doi.org/10.1016/j.cities.2011.11.008

[4] Murcio, R., Morphet, R., Gershenson, C., Batty, M.: Urban transfer entropy across scales. PLoS ONE **10**(7), 0133780 (2015) https://doi.org/10.1371/journal.pone.0133780

[5] Sinha, S., Pan, R.K.: Blockbusters, bombs and sleepers: The income distribution of movies. In: Econophysics of Wealth Distributions, pp. 43–47. Springer, ??? (2005)

[6] Barabási, A.-L., Albert, R., Jeong, H.: Scale-free characteristics of random networks: the topology of the world-wide web. Physica A: Statistical Mechanics and its Applications **281**(1), 69–77 (2000) https://doi.org/10.1016/S0378-4371(00)00018-2

[7] Montemurro, M.A.: Beyond the Zipf–Mandelbrot law in quantitative linguistics. Physica A: Statistical Mechanics and its Applications **300**(3–4), 567–578 (2001) https://doi.org/10.1016/S0378-4371(01)00355-7

[8] Cancho, R., Solé, R.V.: Zipf's law and random texts. Advances in Complex Systems **5**(1), 1–6 (2002) https://doi.org/10.1142/S0219525902000468 http://www.worldscientific.com/doi/pdf/10.1142/S0219525902000468

[9] Zipf, G.K.: Selected Studies of the Principle of Relative Frequency in Language. Harvard University Press, Cambridge, Massachusetts (1932)

[10] Petruszewycz, M.: L'histoire de la loi d'Estoup-Zipf: documents. Mathématiques et Sciences Humaines **44**, 41–56 (1973)

[11] Newman, M.E.: Power laws, Pareto distributions and Zipf's law. Contemporary Physics **46**(5), 323–351 (2005)

[12] Baek, S.K., Bernhardsson, S., Minnhagen, P.: Zipf's law unzipped. New Journal of Physics **13**(4), 043004 (2011)

[13] Perc, M.: Evolution of the most common English words and phrases over the centuries. Journal of The Royal Society Interface **9**(77), 3323–3328 (2012) https://doi.org/10.1098/rsif.2012.0491 http://rsif.royalsocietypublishing.org/content/9/77/3323.full.pdf+html

[14] Font-Clos, F., Boleda, G., Corral, A.: A scaling law beyond Zipf's law and its relation to Heaps' law. New Journal of Physics **15**(9), 093033 (2013)

[15] D'Amore, A.: La influencia mutua entre lenguas: anglicismos, hispanismos y otros préstamos. Revista Digital Universitaria **10** (2009)

[16] Gorlach, M.: A Dictionary of European Anglicisms: A Usage Dictionary of Anglicisms in Sixteen European Languages. OUP Oxford, USA (2005). https://books.google.com.mx/books?id=qhFREAAAQBAJ

[17] Haspelmath, M., Tadmor, U.: Loanwords in the World's Languages: A Comparative Handbook. Walter De Gruyter, Berlin, Germany (2009). https://books.google.com.mx/books?id=HnKeVbwTwyYC

[18] Durkin, P.: Borrowed Words: A History of Loanwords in English. OUP Oxford, Oxford, UK (2014). https://books.google.com.mx/books?id=4W6JAgAAQBAJ

[19] Michel, J.-B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Team, T.G.B., Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A., Aiden, E.L.: Quantitative analysis of culture using millions of digitized books. Science **331**(6014), 176–182 (2011) https://doi.org/10.1126/science.1199644 https://science.sciencemag.org/content/331/6014/176.full.pdf

[20] Cocho, G., Flores, J., Gershenson, C., Pineda, C., Sánchez, S.: Rank diversity of languages: Generic behavior in computational linguistics. PLoS ONE 10(4): e0121898 (2015)

[21] Harford, T.: Big data: A big mistake? Significance **11**(5), 14–19 (2014) https://doi.org/10.1111/j.1740-9713.2014.00778.x

[22] Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017)

[23] Matthews, P.H.: The Concise Oxford Dictionary of Linguistics. Oxford University Press, Oxford, UK (2014). https://doi.org/10.1093/acref/9780199675128.001.0001 . https://www.oxfordreference.com/view/10.1093/acref/9780199675128.001.0001/acref-9780199675128

[24] Ortega, M.L.: La crisis de 1914-1924 y el sector fabril en chile. Historia (Santiago), 45(2), 433-454 (2012)

[25] Morales, J.A., Colman, E., Sánchez, S., Sánchez-Puig, F., Pineda, C., Iñiguez, G., Cocho, G., Flores, J., Gershenson, C.: Rank dynamics of word usage at multiple scales. Frontiers in Physics **6**, 45 (2018) https://doi.org/10.3389/fphy.2018.00045

[26] Cocho, G., Rodríguez, R.F., Sánchez, S., Flores, J., Pineda, C., Gershenson, C.: Rank-frequency distribution of natural languages: A difference of probabilities approach. Physica A: Statistical Mechanics and its Applications **532**, 121795 (2019) https://doi.org/10.1016/j.physa.2019.121795

[27] Morales, J.A., Sánchez, S., Flores, J., Pineda, C., Gershenson, C., Cocho, G., Zizumbo, J., Rodríguez, R.F., Iñiguez, G.: Generic temporal features of performance rankings in sports and games. EPJ Data Science **5**(1), 33 (2016) https://doi.org/10.1140/epjds/s13688-016-0096-y

[28] Iñiguez, G., Pineda, C., Gershenson, C., Barabási, A.-L.: Dynamics of ranking. Nat. Comm. **13**(1), 1646 (2022) https://doi.org/10.1038/s41467-022-29256-x

[29] Michel, J.-B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Team, T.G.B., Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A., Aiden, E.L.: Quantitative analysis of culture using millions of digitized books. Science **331**(6014), 176–182 (2011) https://doi.org/10.1126/science.1199644 http://www.sciencemag.org/content/331/6014/176.full.pdf

[30] Bollen, J., Thij, M., Breithaupt, F., Barron, A.T.J., Rutter, L.A., Lorenzo-Luaces, L., Scheffer, M.: Historical language records reveal a surge of cognitive distortions in recent decades. Proceedings of the National Academy of Sciences **118**(30), 2102061118 (2021) https://doi.org/10.1073/pnas.2102061118 https://www.pnas.org/doi/pdf/10.1073/pnas.2102061118

[31] Dawkins, R.: The Selfish Gene. Oxford University Press, Oxford, UK (2016)

Figure legends

- **Figure 1: New migrant words, per language and per decade.** NMW are considered for each language. Each panel contains one language. The dotted line displays the number of $\mathrm{NMW}_{\mathrm{out}}$ that originate in the corresponding language, and the bars the $\mathrm{NMW}_{\mathrm{in}}$ coming to that language, separated by the origin of the different NMW.

- **Figure 2: Zipf's law of the accumulated migrant words, grouped by receiving language.** We display a frequency-rank plot for all language pairs, for the migrant words during the year 2000. Indeed, after a transient, Zipf's law is observed (a dashed line with slope -1 is provided for comparison).
- **Figure 3: The use U among languages.** We plot the use, as defined in Eq. (1) for all language pairs between 1900 and 2009: on the top row, how column languages influence plotted languages, while the bottom row shows how column languages are influenced by plotted languages. Results are discussed in the main text.
- **Figure 4: Rank diversity of accumulated migrant words among languages.** Diversity for all pairs of languages. Each pair is well fitted by the sigmoid proposed in Eq. (2), with fitting parameters μ and σ reported in the inset. As a reference, we show a global sigmoid (in black) obtained by fitting all data points. Its fitting parameters are also shown in the inset as a black dot.
- **Figure 5: Similarity of use when some words are eliminated.** We study the similarity of the shape of usage Eq. (1), as measured by $\langle D\rangle$ [see Eq. (3)], when some words are eliminated. The upper plots correspond to the case in which the lower ranked words (more frequent) are being eliminated first, whereas in the lower plots we start eliminating the higher ranked words (less frequent).

Table legends

- **Table 1:** Examples of new migrant words for all pairs of languages, grouped together by semantic field. A notorious influence of historic events on word migration is observed. We use the following abbreviations: EN for English, FR for French, GE for German, IT for Italian and SP for Spanish.