ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

SEMESTER PROJECT

Chronocloud, Kernel and word resilience on Google N-grams

Author:
Ciprian I. TOMOIAGĂ

Supervisor:
Vincent BUNTINX

Professor:
Prof. Frédéric KAPLAN

Digital Humanities Laboratory
EPFL

June 9, 2017

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Abstract

In the following, we present an exploratory analysis of the Google books corpus over English, French, Spanish and German. The aim of this work is to model and quantify linguistic drift using word resilience and kernel distance. We also look at word birth and death rates and resilience spectra. Our results indicate that the studied languages have slowed down in more recent years and, with the exception of important events like the world wars, they follow the same evolution model.

Contents

List of Figures

1 Introduction

Language is, without doubt, among the most complex systems the humans developed and powers all other cultural and technological advances. It is fundamentally dynamic and understanding how it changes can help shed light on a number of important questions in cognition, anthropology and sociology.

As new concepts emerge, they are either assigned to existing words or new words get invented for them. This makes, in a very broad sense, two different types of linguistic change: lexico-morphological and semantic. The former induces changes to the language vocabulary through additions, removals and simplifications while the latter represents shifts in word meanings. In this project we will focus on the lexical part.

1.1 Data

Our data comes from a massive digitisation effort by Google in 2012, after having scanned approximately 4% of the world's books. It is made available as n-gram tokens, which are series of n words along with their counts in the corpus. Each n-gram's data is split over the years in which it appears, with the oldest dating from the 1500s. This allows us to track through time the evolution of different tokens, from single words up to five-word expressions (5-grams). For the rest of this work we will use *token* and *word* interchangeably to refer to the same concept, namely an n-gram.

Since this database is suitable for processing in Google's data centres, we are forced to apply filters to bring it down to more manageable sizes. As such, we concentrate our efforts on data after the year 1800 and only on the following four languages: English, French, German and Spanish. We note there is a general frequency filter, as Google removed n-grams which appear less than 40 times across the corpus. Then we impose a more aggressive filter based on the cumulative count across the considered time range (1800–2009): we only consider those which exceed a threshold $T = 35 \cdot (2009 - 1800)$. This has implications especially for more recent words, which are ignored unless they are very important. Figure 1.1 shows the theoretical sizes of the considered corpora, while figure A.1 shows the number of distinct entries in each sub-corpus.

1.2 Visualisation – Chronocloud

The Chronocloud is a visualisation tool based on temporal profiles of words (e.g. figure 1.2) and it highlights linguistic shifts in different epochs over a time range. We generated them for each of our sub-corpora and noted interesting changes associated

with different events. Since they are quite dense in information, they are best viewed online; we only include a few interesting ones here (figures 1.3 to 1.6).

Observing the one for English 4-grams, we see the kernel captures some useful expressions (`at the same time`, `on the other hand`, `from time to time`) but also quite a few related to USA. It is therefore probable the corpus is based on American documents. We also note that in the sector from 1860 to 1870 there is a significant fraction of chemistry related terms, but these are not special in any way to conclude that they would be related to an important concern of the time, as is the case of `carbon dioxide` in the 2-grams (not shown). Some of these trends are carried over in the 5-grams with simple prepositional additions, such as the political tendency in the 1850s. In general, however, there are new meaningful expressions which show the interests and main events of different epochs.

The French 4-grams contain more varied expressions in the kernel but still have a chemistry focus in 1860s. We also notice unfamiliar spellings of words before 1880, which were "corrected" by the orthographic reform of 1878.

The Chronocloud is a great tool for getting a feeling of external drivers of linguistic change, such as cultural or political events and technological breakthroughs. Unfortunately it also has some drawbacks: it is difficult to compare importance of words among different groups and it relies on a stochastic positioning within the same group. This makes it difficult to asses how much different words contribute to the "total" drift and how the drift is influenced by adding new resources to the corpus.

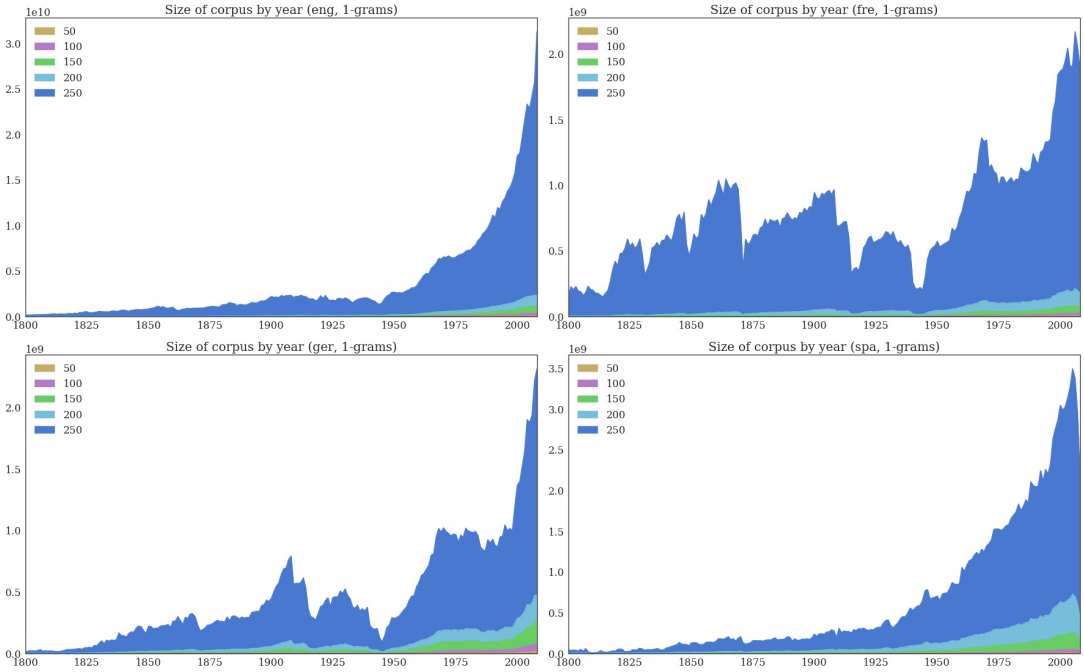


FIGURE 1.1: Corpora sizes

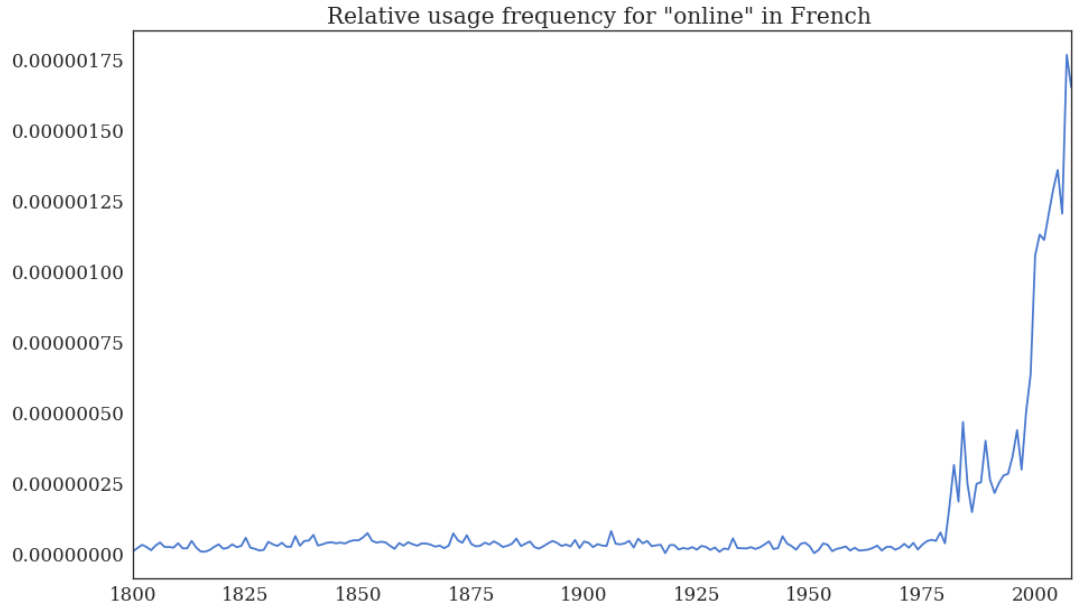


FIGURE 1.2: Example word profile

The temporal profile of the word `online` in the French corpus is given by its relative frequency in each year.

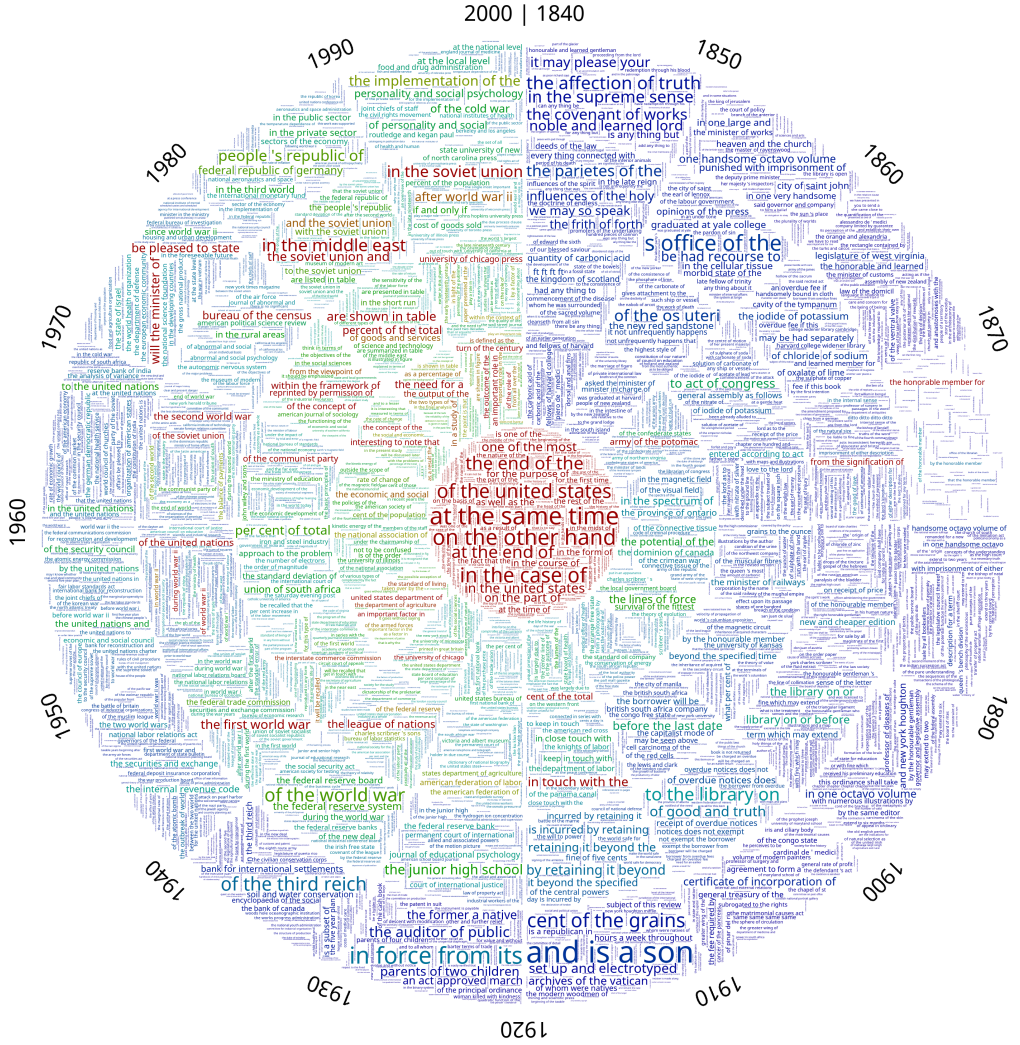


FIGURE 1.3: Chronocloud for English 4-grams

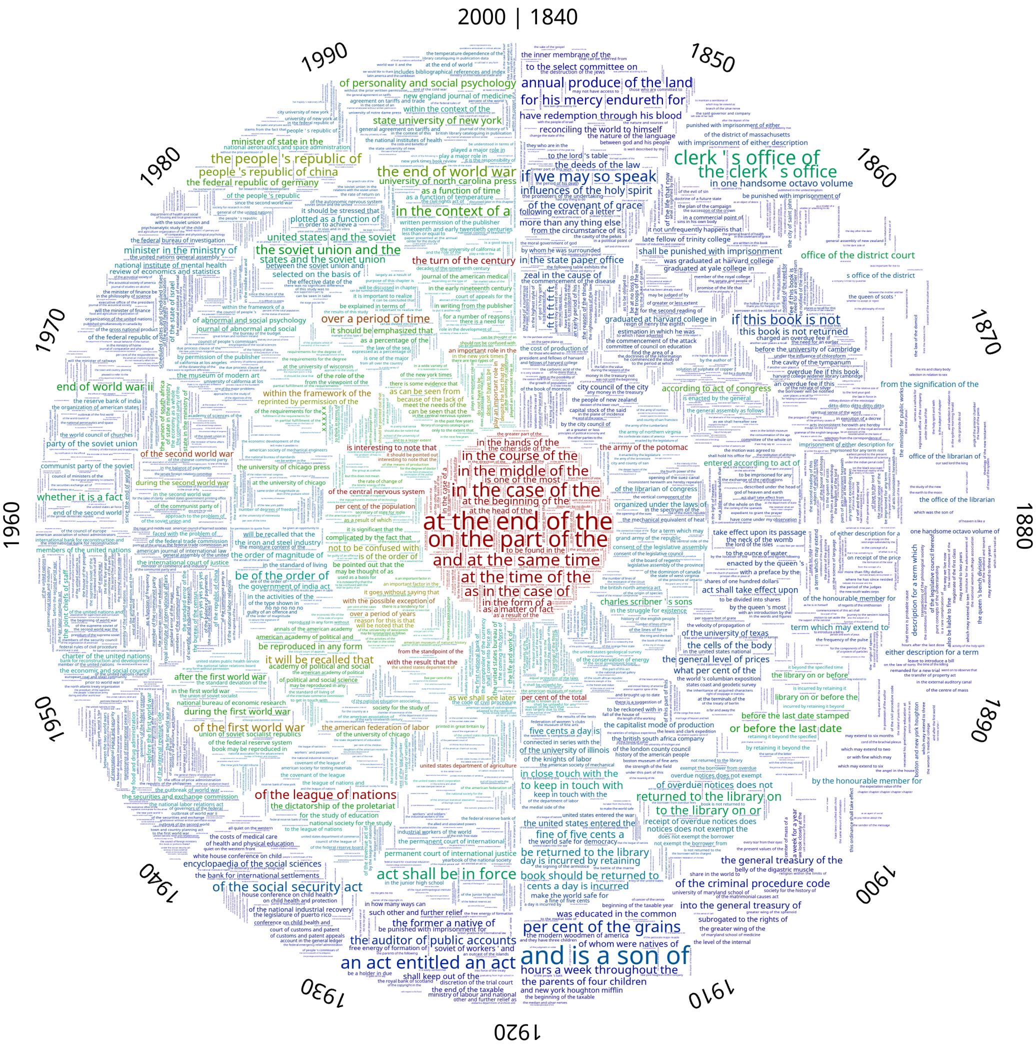


FIGURE 1.4: Chronocloud for English 5-grams

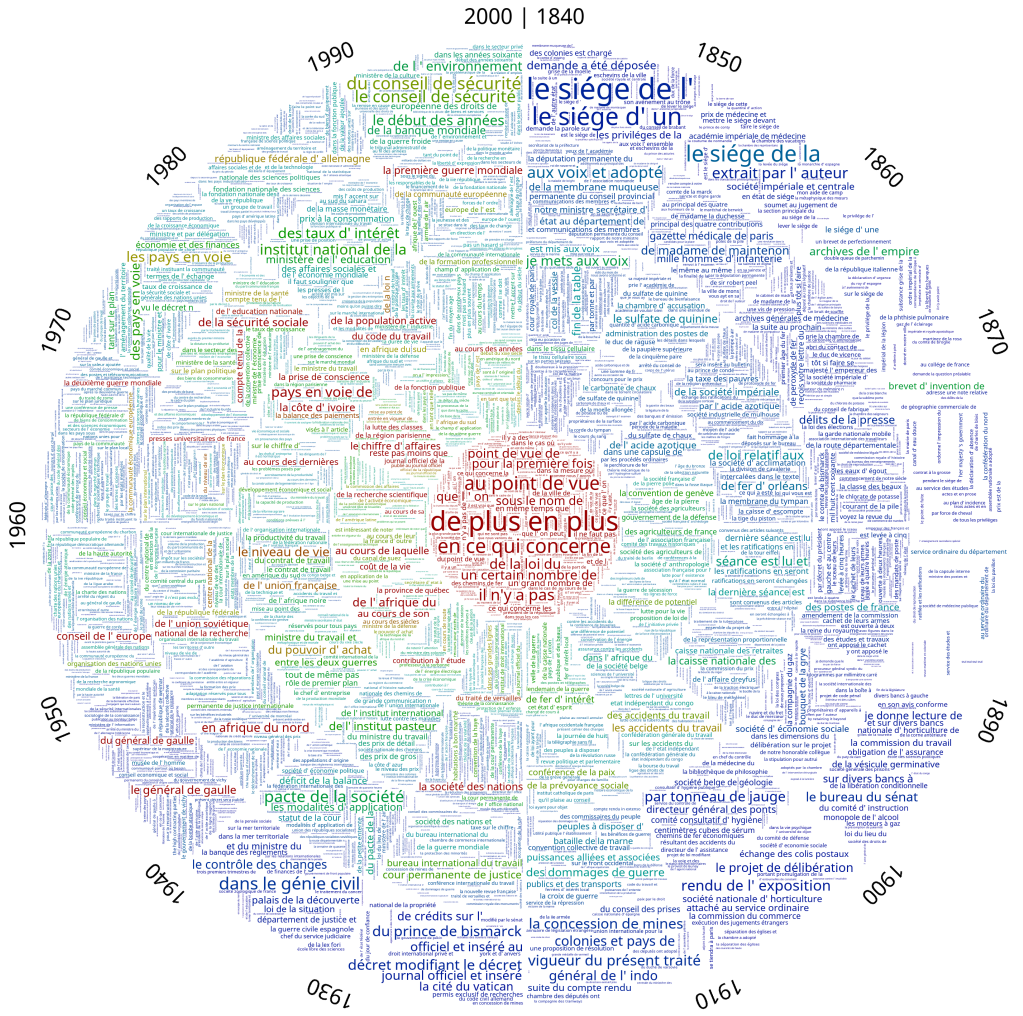


FIGURE 1.5: Chronocloud for French 4-grams

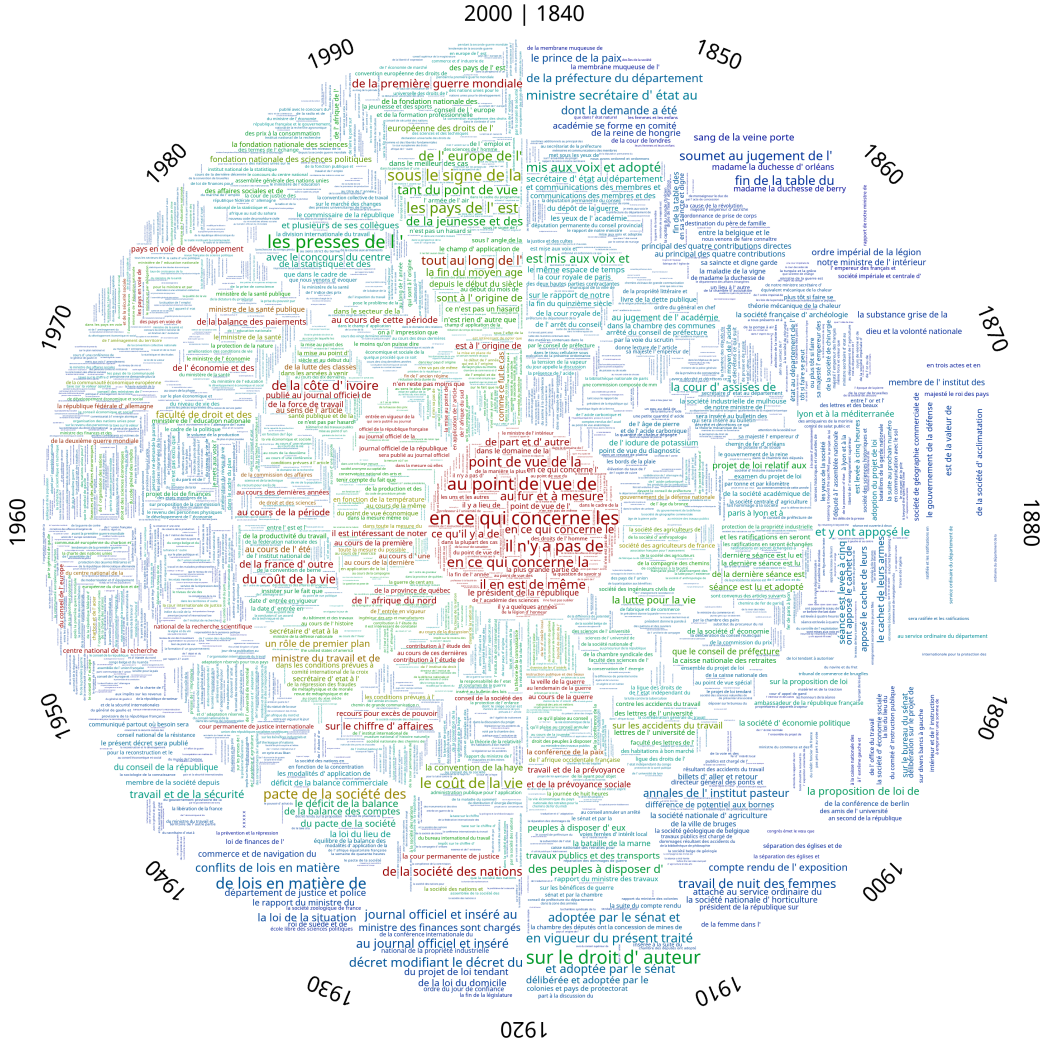


FIGURE 1.6: Chronocloud for French 5-grams

2 Measures of linguistic drift

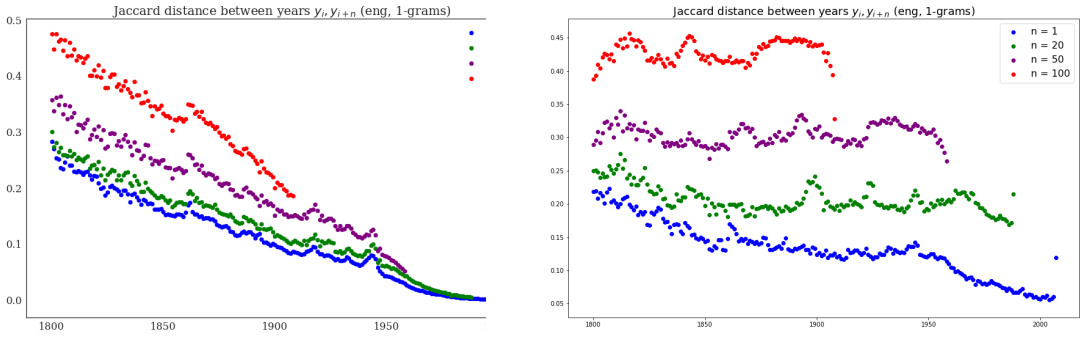
2.1 Jaccard distance

Given the corpora of two different years C_i, C_j, we can use the Jaccard distance [1] on their lexica to assert their similarity. This is defined as

$$d(L_i, L_j) = 1 - \frac{|L_i \cap L_j|}{|L_i \cup L_j|}.$$

Figure 2.1 (a) shows the speed of language drift for the English corpus as captured by this metric. It appears that the language ceases to evolve in more recent years, and this phenomenon is replicated among the other languages as well. However, we note that Jaccard distance is based exclusively on the presence or absence of words in the lexicon and completely ignores their usage counts. Therefore, following Heap's law (also known as Herdan's law [2]), we can argue that this measure only shows the increasing size of the corpus (figure 2.2) and gives little information about meaningful or unexpected additions.

One alternative would be to consider a noise cutoff threshold f_c which would set to zero all frequencies below it. This would mark the word as absent in years where it is very rare, while still capturing the overall trend; for example, the word `online` would be seen as absent in all corpora before 1817 (figure 1.2). This gives better behaviour of the measure as can we can observe in figure 2.1 (b). However, the threshold f_c differs from one corpus to another and is, in general, difficult to get right since it applies corpus-wide, on all words.



(a) Without noise cutoff (b) With noise cutoff at $1/10.000.000$

FIGURE 2.1: Drift speed measured with Jaccard distance

2.2 Language kernel

Every language has a base vocabulary, a set of words indispensable for communication. We call this the language kernel and it comprises words for day-to-day actions and objects, along with linkage words and placeholders (prepositions, conjunctions and pronouns). The extended vocabulary then adds words for advanced or new concepts, which are less frequently encountered and often exist temporarily. To capture these properties we use the resilience, an indicator of how long a word or a token lasts in the language.

2.2.1 Word resilience

Given a token (n-gram) w, its resilience is given by the longest sequence of years where the word is *alive*. Buntinx et al. [3] consider a word *alive* based on presence in the corpus, i.e. when its usage frequency is non-zero. We believe this poses similar problems as in the Jaccard case so we offer an extension to the resilience measure by applying a noise threshold. This can be either global or local (per word).

For the global case, a word is considered *alive* in year y when $f_w(y) \geq f_c$. The cutoff f_c is still chosen arbitrarily and differs among corpora. To counteract this issue we use a word-local significance threshold f_s:

$$alive(w, y) \equiv f_w(y) \geq f_s \cdot Median[f_w(y_i)], \qquad i \in [1800, 2008]\,.$$

This follows the human intuition when looking at individual n-gram plots and has the advantage of applying uniformly for all words regardless of the actual values of their frequencies, therefore regardless of corpus. We believe this definition is better adapted for our analysis than the global cutoff since it gives a more fair split of the corpus in resilience groups (figure 2.2). In our case, we used $f_s = 0.1$, but different values we tried ($f_s = 0.05$, $f_s = 0.15$ did not show a significant qualitative difference.

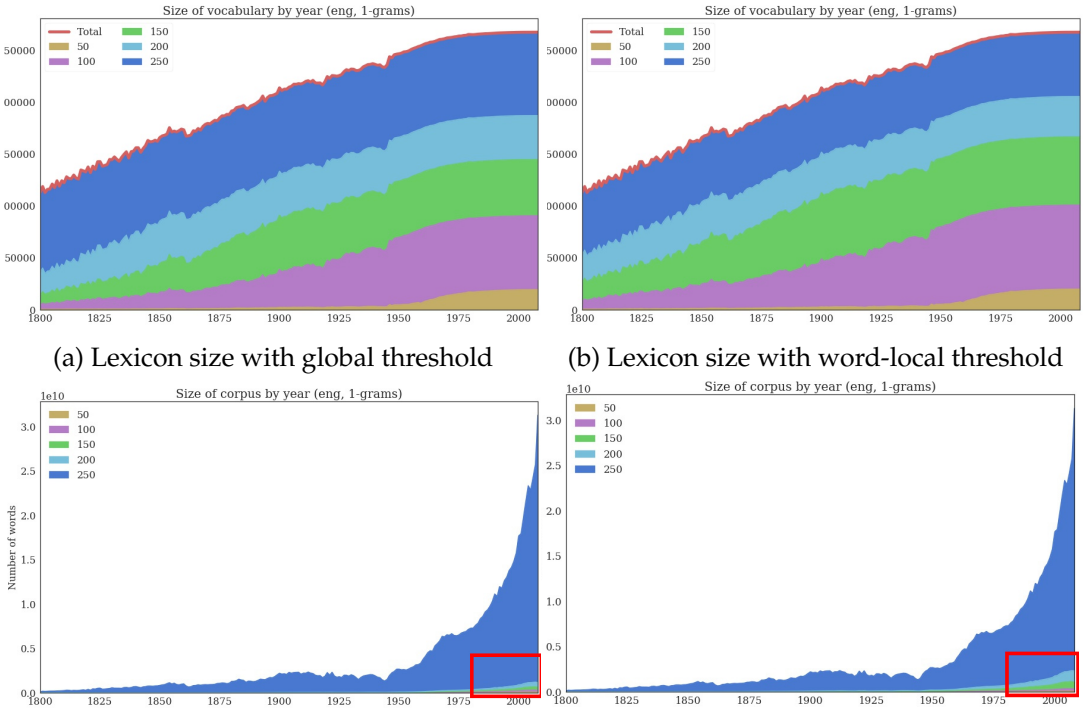
2.2.2 Kernel distance

Having equipped the words with the resilience property, we can define the language kernel K as the set of all words with maximum resilience: $K = \{w | w \in L_y, \forall y\}$. We can now use it to define a drift measure that is robust to corpus size and, therefore, tracks shifts in words' priorities. The kernel distance between two corpora C_i, C_j with lexica L_i, L_j is defined as:

$$d_K(L_i, L_j) = \frac{1}{\|K\|} \sum_{w_i \in K} |rank_{L_i}(w_i) - rank_{L_j}(w_i)|,$$

$$\|K\| = \frac{1}{2} \sum_{i=1}^{N} |N - 2i|, \qquad N = |K|.$$

One interesting question is whether we should compute the words' ranks among the whole lexicon L_i of a year or only on the subset that is included in the kernel. We chose the latter, since the former is heavily influenced by rank jumps of words



(a) Lexicon size with global threshold

(b) Lexicon size with word-local threshold

(c) Corpus size with global threshold

(d) Corpus size with word-local threshold

FIGURE 2.2: Kernel classes in corpus and lexicon

The top row shows the evolution of lexicon size and the share taken by each resilience group. Note the groups are exclusive, so beige depicts words with resilience up to 50 years, violet words with resilience between 50 and 100 years and so on. As explained in section 1.1, the smaller group (beige) is underrepresented.

The bottom row depicts the overall size of the corpus, clearly showing that more common words make up most of the counts.

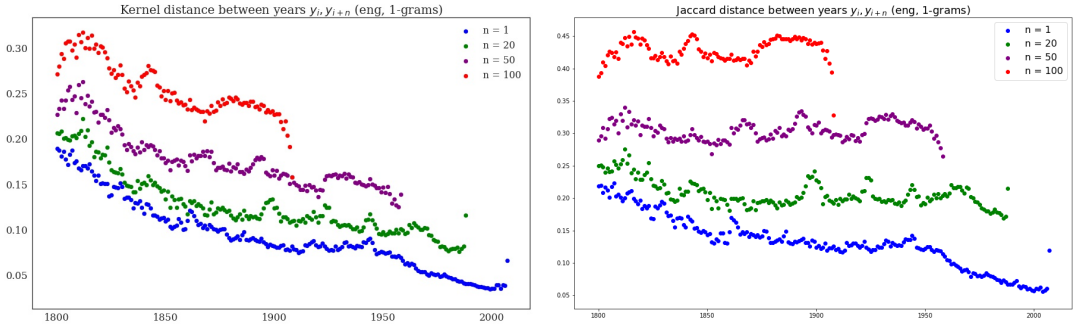
In both cases, the choice of resilience definition influences the share each group takes. We see that using a global threshold (left column) generally assigns longer resilience for words, which means the threshold is too low.

which we do not consider. As such, figure 2.3 (a) shows that this measures a potential slowdown of linguistic drift in more recent years. The trend seems consistent with deltas higher than 1, but the peaks around the years of world wars, which are associated with smaller corpus sizes, are still evident. The results in chapter 3 justify that they are relevant and not just an effect of corpus size.

2.3 Comparison

Taking the English 1-gram corpus as an example (figure 2.3), we observe that for data that is 1 year apart (blue), both Jaccard and Kernel distance indicate a slow down of linguistic drift in more recent years with the exception of the world wars, where a strong increase is detected. However, when considering sub-corpora separated by more than 20 years (green, violet, red), the Jaccard distance levels off whereas the kernel indicates the same downward trend as for the 1 year case. This could mean that the kernel better captures variability of language despite the deliberate omission of new constructs.

We applied the above measures on all available corpora and different n-grams and found consistent results (figure 2.4). Overall, we also noted a very strong similarity among all the plots for different languages and n-grams, which could mean we are measuring an intrinsic model of language evolution, difficult to further uncover or interpret. This premise motivates the next chapter of our work. However, an interesting case is given by the Spanish corpora which seems almost unaffected by the world wars. This further strengthens the hypothesis that the two events brought an important contribution to the languages of the involved parties, similar to a "globalisation" effect.



(a) Kernel distance

(b) Jaccard distance (with noise cutoff)

FIGURE 2.3: Measures comparison for linguistic drift

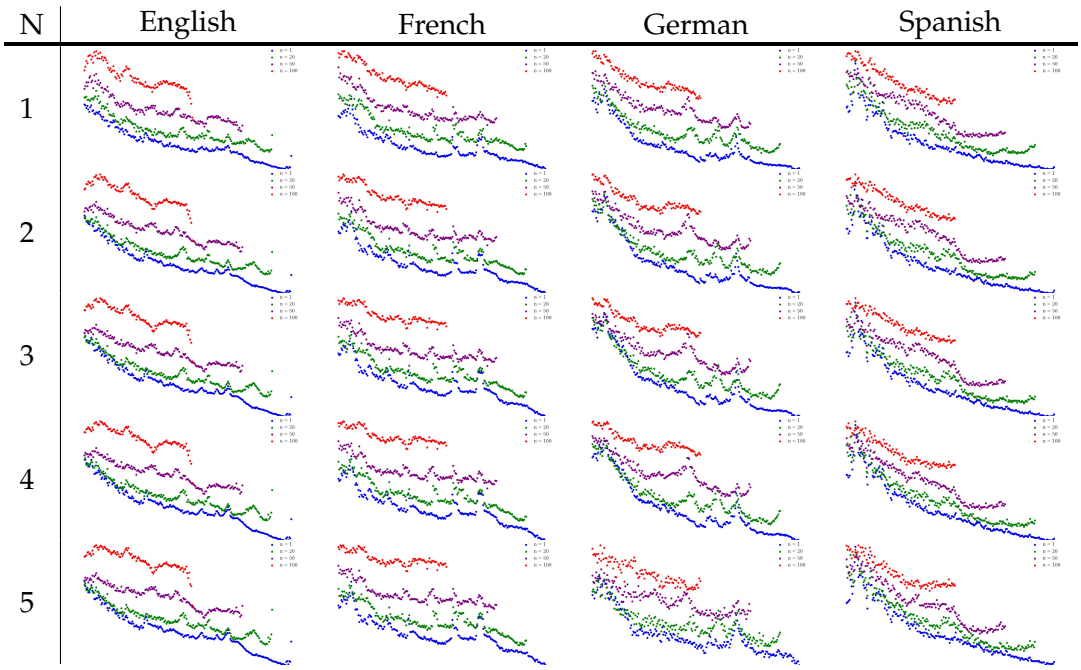


FIGURE 2.4: Overview of kernel distance on all corpora

Apart from the differences around the world wars, there is a striking similarity of language evolution both when measured over different representations (n-grams, vertically) and also, more importantly, when measured across different languages. This could mean that languages, in general, follow a constant evolution model

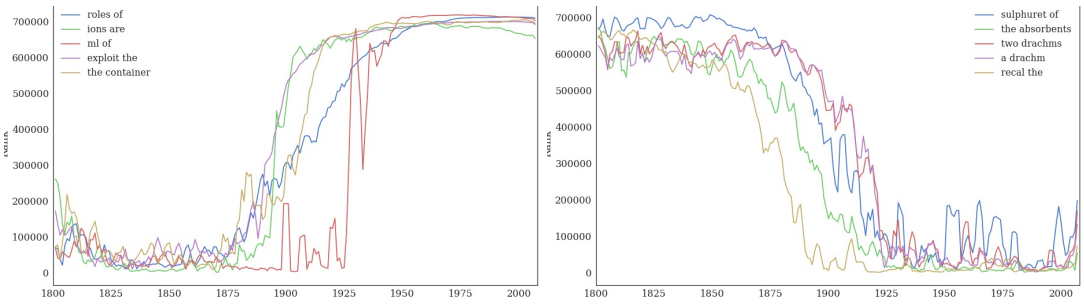
3 Sources of linguistic drift

In chapter 2 we have seen that linguistic drift is present but is decelerating so the languages become more stable. As it was difficult to get a solid confidence in the measures, we will try to further dig in the corpora to better understand if the drift is real and where it comes from.

3.1 Ranks variation

Building directly on the concept of word resilience, we note that some words are more stable than others. For example, the top 6 words of the English corpus never change order, as a direct effect of Zipf's law. This implies that when computing a distance, most of it is given by the words in the tail of the distribution. Another effect of the same law is that the tail comprises most words of the lexicon, which makes it difficult to visualise the actual source of drift.

We can, however, identify the main "culprits" by considering the tokens with the highest variation of rank change. While this could be due to intense jiggling around a value, we could be interested in the tokens that "pierce" through the lexicon or are phased out quickly. Practically, we fit a line to their ranks (among all years) and get the ones with largest positive and negative values. English 2-gram examples are shown in figure 3.1 and complete list is in **??**. This example indicates that there was a big addition of chemistry related corpus in the years 1875 to 1925; it is impossible to say, without background knowledge, whether this is due to selection bias or is actually correlated with inventions of the time.

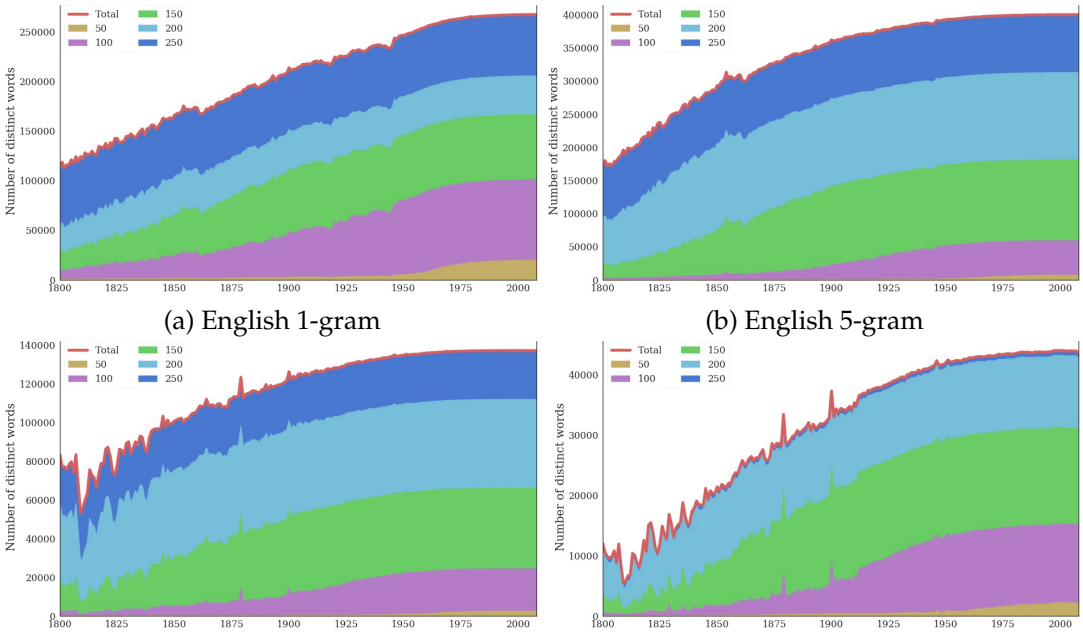


(a) Ascending trends (b) Descending trends

FIGURE 3.1: Top 5 fastest word births and deaths in English 2-grams

3.2 Resilience analysis

We use the words' resilience to classify them in disjunctive groups spanning increasing numbers of years. Replicating this for n-grams beyond one shows us that some



(a) English 1-gram

(b) English 5-gram

(c) Spanish 1-gram

(d) Spanish 5-gram

FIGURE 3.2: Resilience grouping

languages such as English have more stable expressions than, for example, Spanish (figure 3.2).

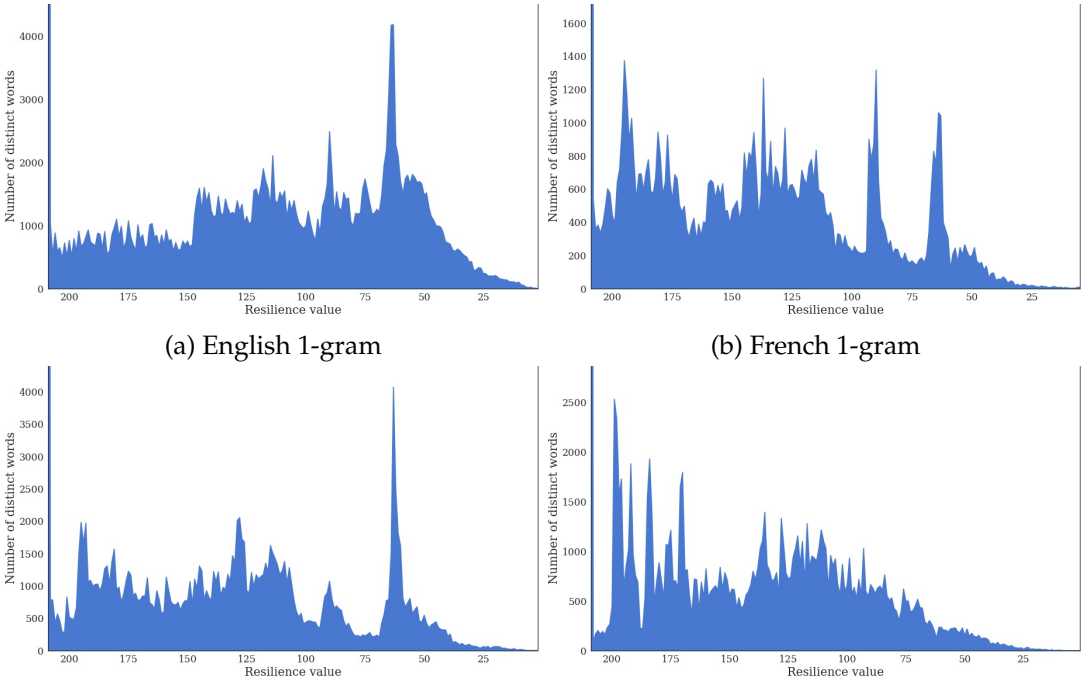
Going into more details, we show the variation of lexicon size in terms of resilience only (figure 3.3). It is interesting to see spikes at relatively well defined intervals. We believe these represent percolation duration, i.e. the average amount of time necessary for words to become stable; it is probably related to the minimum usage necessary for dictionaries to include the word.

3.3 Birth and death rates

As detailed in chapter 1, we are only dealing with lexical language changes. While it is impractical to predict individual trajectories and particular histories given the stochastic nature of language change [4], we can use aggregates of certain trajectories in order to see the general trend.

Given the utility of a word as its relative frequency in a year, we can consider that all words compete to be part of the shared corpus and vocabulary. For example, in the French corpus the words `géognosie` and `stratigraphie` both refer to the study of Earth's crust. They were in competition in 1860, as the usage changed from one to the other, with `stratigraphie` finally winning. Similarly, word pairs such as `<cortége,cortège>`, `<siége,siège>`, `<collége,collège>` were equally used before 1878, when the `-ège` form was imposed by the linguistic reform (figure 3.4).

To quantify the evolution of language in terms of lexical change, we can use the resilience of words to define their birth and death times (see section 2.2.1). Then we can aggregate these per year and normalise by the corpus size, to obtain the birth



(a) English 1-gram

(b) French 1-gram

(c) German 1-gram

(d) Spanish 1-gram

FIGURE 3.3: Vocabulary size per resilience value
The plots are truncated to show most interesting values, since the amount words in the kernel (max resilience) is better depicted in figure 3.2.

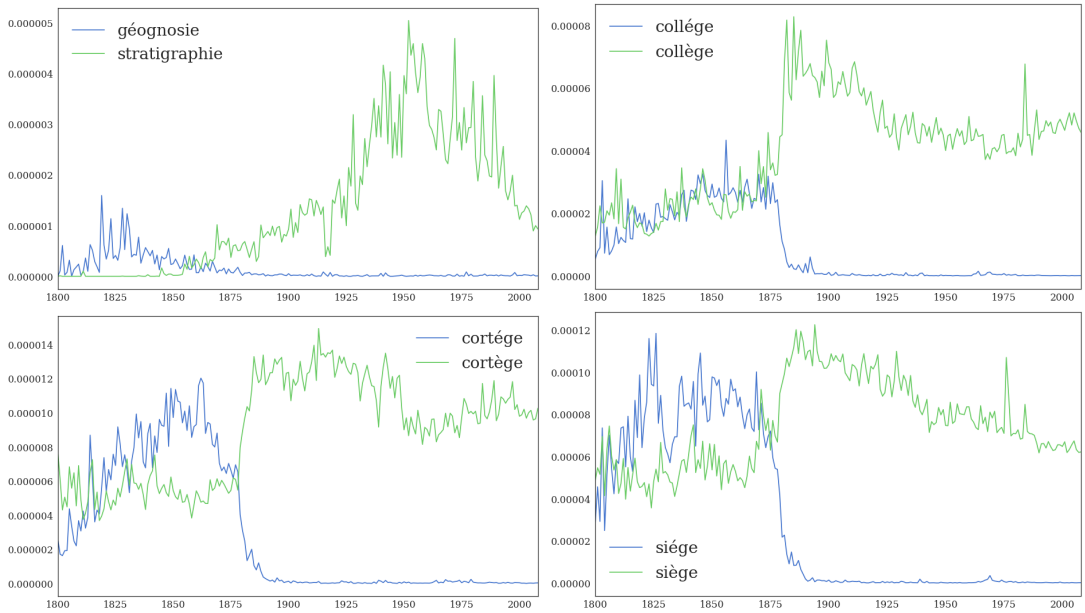


FIGURE 3.4: Competition of related words in French 1-grams

and death rates:

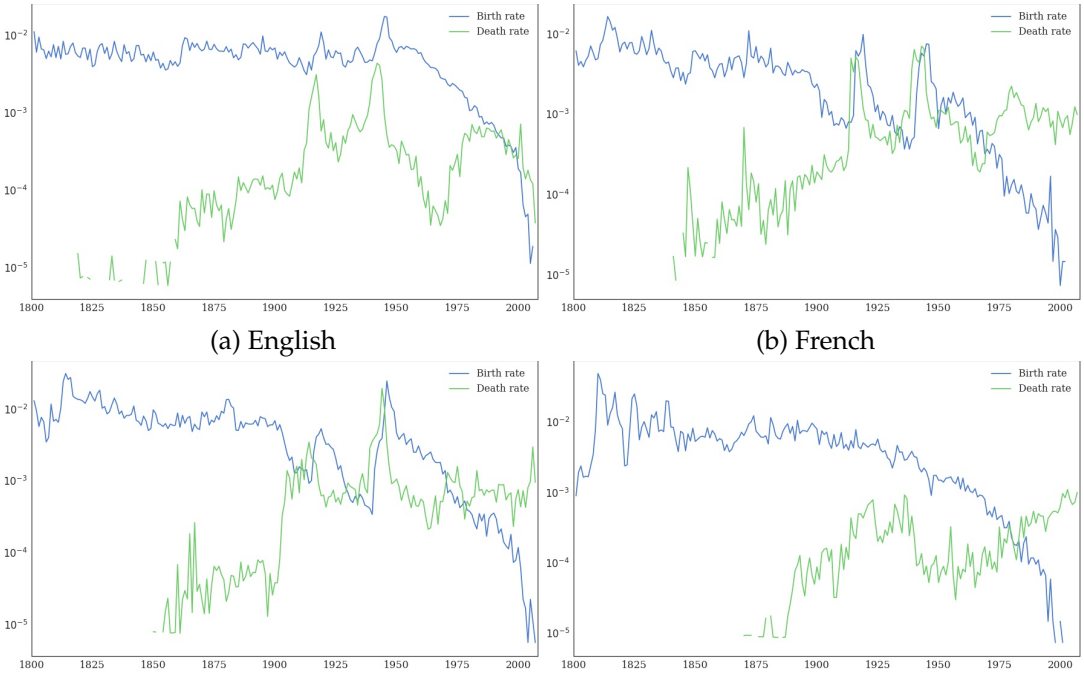
$$r_b(t) \equiv \delta_b(t)/N_w(t)$$
$$r_d(t) \equiv \delta_d(t)/N_w(t),$$

where δ is the number of words emerging or disappearing in year t.

Studying these variations across languages (figure 3.5) indicates there is a significant slow down of words added in more recent years (note the logarithmic scale), which is in accordance to our previous measures. We keep in mind that the birth rate is biased by the limitations in the preprocessing step, which practically allow only recent words with a significant impact in the corpus. However, this does not affect the death rate, which is generally increasing among all corpora.

We also note that, as opposed to the kernel distance, this measure of linguistic change shows clear differences between corpora of different languages (and cultures) and each of them is impacted differently my major events.

Finally, we try to identify whether these rates are genuine and what is the composition of the language at these points. We investigated sets of words that were born or died in 1850 and compared to those in 1950. For births, it seems most words are genuine and resisted until the most recent years. For deaths, however, we identified a non-negligible amount of noise or unexpected tokens. This consisted mostly of words in other languages (e.g. Italian was present among French dead words) and misspellings. In the first study over this corpus, Michel et al. [5] mention that 51% of English lexicon in 1900 consists of non-English words, and 32% in 2000. We can ascribe this difference to the improvement in printing technology as well as the apparition of automatic spell-checkers.



(a) English (b) French

(c) German (d) Spanish

FIGURE 3.5: Birth and death rates for all 1-grams

4 Conclusions

In this project we performed an exploratory analysis of linguistic drift in order to identify its sources and some possible evolution models. We generated Chronocloud instances over a variety of corpora and noted external factors that influence the set of words and expressions in use at a given time. These are available for further exploration and to guide other research over the large dataset of Google books.

Then we adopted two different metrics (Jaccard and Kernel distance) and used them to quantify and model the linguistic drift. We observed that the kernel distance is more robust to corpus size and seems to capture a slow-down of language evolution. However, this is heavily replicated among all corpora, which could confirm the hypothesis of constant evolution of languages. We looked at particular examples of words that contribute the most to the kernel distance and identified two potentially useful types: words that quickly become popular and words that disappear suddenly. Based on this, we further analysed the language in terms of birth and death rates and reached the same conclusion indicated by the distance: less words are born in recent years while at the same time more and more disappear, phenomenon which leads to a stabilisation of the language.

Finally, we discovered that many words compete for a place in the lexicon. Therefore we explored the spectrum of resilience as an indicator of words' adoption. The four languages present different distributions of words' resilience, but surprisingly none follows a smooth curve. On the opposite, languages seem to share a certain periodicity, indicated by equally spaced peaks. We suppose this could indicate that words have a "maturation age" when they become accepted and stable in the language. Further investigation is needed to confirm this.

4.1 Future work

There are still some ideas we started to test over the corpus, but did not get to finish:

- test and explain the general shift in colour of the Chronocloud, as it could be related to the "piercing" power of a word
- test and incorporate logarithmic growth measure over the corpus; these incorporate Zipf's and Heap's law into their model
- automatically find pairs (even triplets) of competing words based on their temporal profiles
- compare languages between them by the amount of n-grams generating (n+1)-grams; this comes from the observation that the 5-gram English Chronocloud has many expressions also found in the 4-gram plus a preposition, whereas the French 5-gram generally presents more unique expressions than the 4-gram

- given more processing power, determine more accurate birth and death rates over the original corpus, without frequency filters

A Extra figures

A.1 Vocabulary sizes and resilience

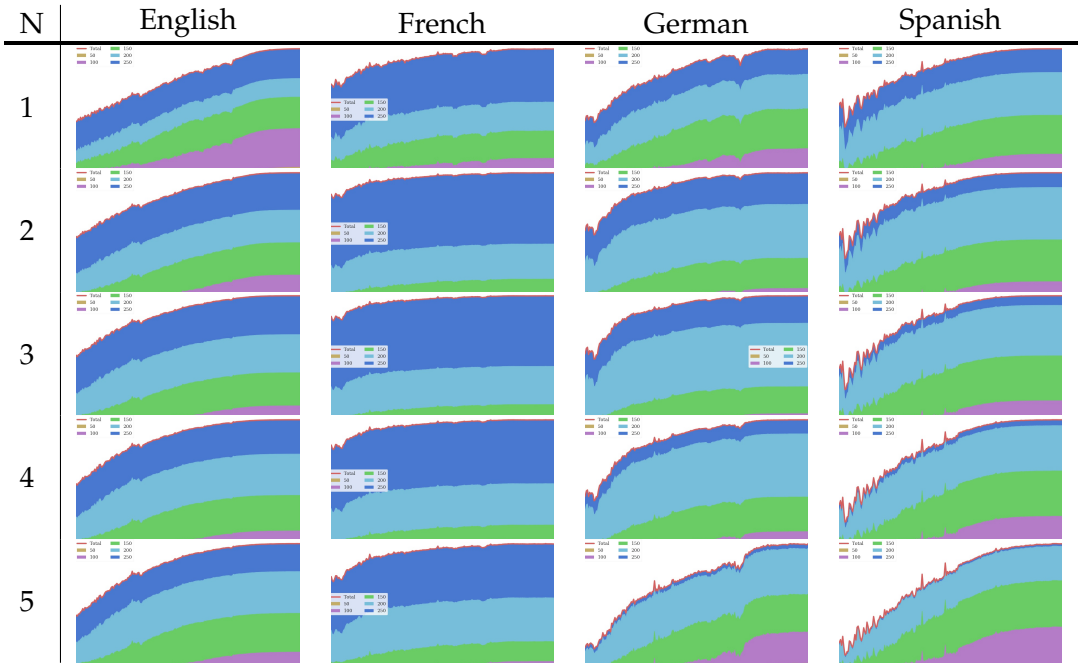


FIGURE A.1: Overview of lexicon sizes

A.2 Resilience spectra

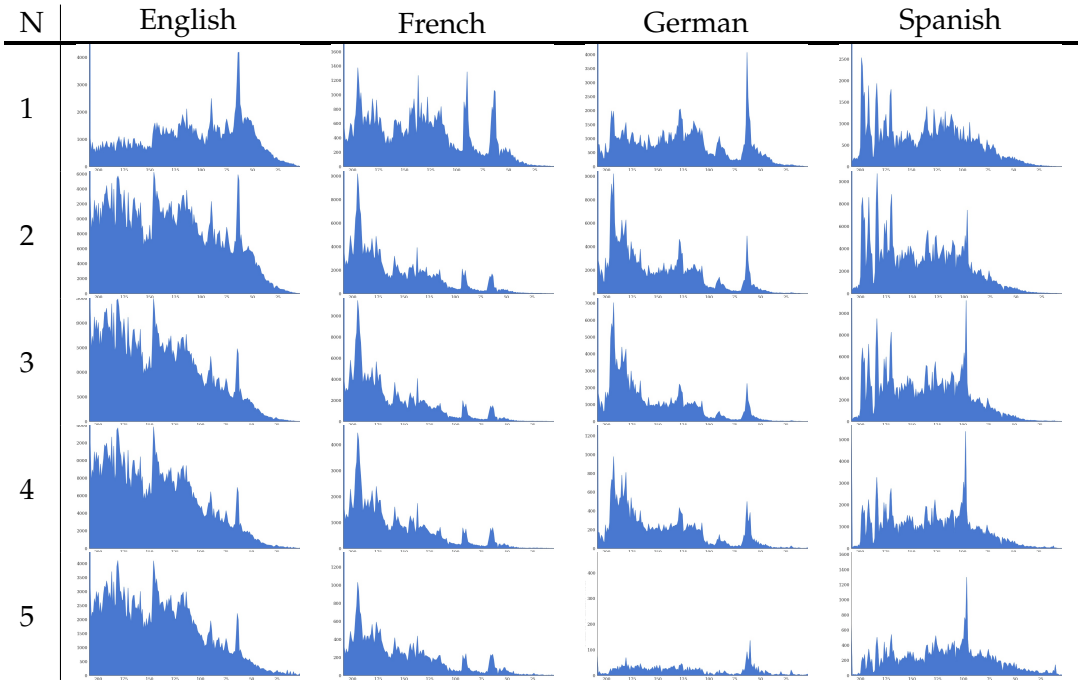


FIGURE A.2: Overview of resilience spectrums
The number of unique words for each resilience value. Note that the vertical axes are not aligned.

B Examples

B.1 Word deaths

B.1.1 French 1850

augmentoit, livroit,

B.1.2 French 1950

ablancourt, accipiat, accroistre, adan, adynamique, aidede, alque, amez, asseline, assés, auctorité, avés, bautain, bayles, bennigsen, blasme, bukarest, bysance, cerna, ceulx, châtel, cognoist, colonage, combatre, comitats, commenca, compaignie, congnoissance, conseillier, consiliis, conventu, copiam, corum, coseigneurs, cousté, couzon, daignant, debvoir, deffense, demeurans, demourant, desuper, dictes, dinaires, discordia, dolgorouki, eliam, enffans, esclavons, escript, eslevé, estouteville, estoyent, faictz, fayel, flaccidité, fouilloux, foulahs, fétidité, gaietés, galabert, grailly, gresse, gruaux, henrys, hospodars, hématocèle, héry, illec, impost, innocens, insultées, intercessions, jeg, jouyr, jubet, landwehr, larda, léopol, malléoles, manans, marcigny, mesmement, moinet, moncade, moys, multitudinem, papuleuses, parcequ'elle, pardes, parentis, parolles, parquoy, passement, pauthier, peyne, philomèle, pyrogallique, quartiergénéral, recepte, recut, remonstrances, remontra, rossillon, ruffi, rupt, sandomir, sassafras, saturnines, scel, seel, sensorium, shrapnels, simarre, sirmium, soubz, soyecourt, statutum, steenstrup, subditis, ségalas, sénarmont, tarvis, temporalibus, tengo, testât, touron, toz, traicte, troys, tumulaires, unsern, venans, vertot, veues, viart, vidin, villarceau, vimes, vindrent, vitia, vivandiers, volentes, voluerint, volunté, voulans, voullu, yorck, évéque, évéques,

B.1.3 English 1950

Note English 1850 has almost no deaths due to the initial filter

achillis, algonkins, anchylosis, aneurism, apothem, aretin, bagration, barbo, betther, bnt, bohannan, buol, calcaneo, chalier, cherusci, contagium, coquerel, corlear, cuv, cuyuni, dermod, duras, eailroad, ejusmodi, eneugh, equino, foresaid, genga, guelderland, happe, homoeopath, jasher, kabylia, keepit, kilom, lacedemonian, lancisi, loaner, m'donald, maartens, machault, martinico, massimiliano, miggs, moncey, mylo, nalle, ormulum, parietes, phthisical, picou, plassy, preceptorship, pringles, puisieux, pultusk, purifieth, pylephlebitis, qat, rarotongan, refluent, rosevear, schwalbach, seignior, seigniors, simoda, soemmering, sordes, suzon, tirailleurs, toula, tympanitis, undauntedly, urethrotomy, vanhomrigh, vatel, velpeau, venerabili, viiith, villeroy,

Bibliography

[1] Paul Jaccard. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bull. Soc. Vaud. Sci. Nat.*, 37:241–272, 1901.

[2] Gustav Herdan. *Type-token mathematics*, volume 4. Mouton, 1960.

[3] Vincent Buntinx, Cyril Bornet, and Frederic Kaplan. Studying linguistic changes over 200 years of newspapers trough resilient words analysis. *Frontiers in Digital Humanities*, 4:2, 2017.

[4] Daniel J Hruschka, Morten H Christiansen, Richard A Blythe, William Croft, Paul Heggarty, Salikoko S Mufwene, Janet B Pierrehumbert, and Shana Poplack. Building social cognitive models of language change. *Trends in cognitive sciences*, 13(11):464–469, 2009.

[5] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182, 2011.