frontiers in Digital Humanities

ORIGINAL RESEARCH
published: 03 February 2017
doi: 10.3389/fdigh.2017.00002

Studying Linguistic Changes over 200 Years of Newspapers through Resilient Words Analysis

Vincent Buntinx, Cyril Bornet and Frédéric Kaplan*

Digital Humanities Laboratory (DHLAB), Swiss Federal Institute of Technology, Lausanne, Switzerland

This paper presents a methodology to analyze linguistic changes in a given textual corpus allowing to overcome two common problems related to corpus linguistics studies. One of these issues is the monotonic increase of the corpus size with time, and the other one is the presence of noise in the textual data. In addition, our method allows to better target the linguistic evolution of the corpus, instead of other aspects like noise fluctuation or topics evolution. A corpus formed by two newspapers "La Gazette de Lausanne" and "Le Journal de Genève" is used, providing 4 million articles from 200 years of archives. We first perform some classical measurements on this corpus in order to provide indicators and visualizations of linguistic evolution. We then define the concept of a lexical kernel and word resilience, to face the two challenges of noises and corpus size fluctuations. This paper ends with a discussion based on the comparison of results from linguistic change analysis and concludes with possible future works continuing in that direction.

Keywords: linguistic change, corpus studies, newspapers archives, textual distance, corpora kernel, word resilience

OPEN ACCESS

Edited by:
Taha Yasseri,
University of Oxford, UK

Reviewed by:
Martin Gerlach,
Northwestern University, USA
Tom Nicholls,
University of Oxford, UK

****Correspondence:***
Vincent Buntinx
vincent.buntinx@epfl.ch

Specialty section:
This article was submitted to Big Data, a section of the journal Frontiers in Digital Humanities

Received: *10 November 2016*
Accepted: *12 January 2017*
Published: *03 February 2017*

Citation:
Buntinx V, Bornet C and Kaplan F (2017) Studying Linguistic Changes over 200 Years of Newspapers through Resilient Words Analysis. Front. Digit. Humanit. 4:2. doi: 10.3389/fdigh.2017.00002

1. INTRODUCTION

This research investigates methods to study linguistic evolution using a corpus of scanned newspapers, continuing the work presented in conference paper (Buntinx et al., 2016). Language changes quantification in large corpora is a problem widely addressed since the recent availability of large textual databases. One commonly used method is to compute a distance measure between subsets of the corpora and analyze the temporal evolution of such measure. In Bochkarev et al. (2014), authors used Kullback–Leibler divergence in the form of symmetrized relative entropy between two sets of word frequencies. They applied this measure on the Google Books N-Gram Corpus (Michel et al., 2011) in order to compute lexical evolution for multiple languages. Others studies (Pechenick et al., 2015a,b) have used the Google Books Corpus computing Kullback–Leibler and Jensen–Shannon divergence. They analyzed the specific contributions to the distance of most frequent words in order to combine quantitative and qualitative analysis. Another work (Cocho et al., 2015) used the frequency rank evolution of words and addresses the linguistic change analysis through the concept of rank diversity of languages. In a recent work, physicists and mathematicians used the generalized entropy on symbolic sequences with heavy-tailed frequency distribution (Gerlach et al., 2016). Their method is particularly suited for textual corpora words distribution, which follow the well-known Zipf law (Zipf, 1935; Piantadosi, 2014). The corpus we used is composed of 4 million press articles, indirectly documenting the evolution of written language, covering about 200 years of archives. The corpus is made out of digitized facsimiles of Le Journal de Genève (1826–1997) and La Gazette de Lausanne (1804–1997). For each newspaper, the daily scanned issues were algorithmically transcribed using an optical character recognition (OCR) system. The whole archive represents more than 20 TB of scanned data (including text, metadata, pdf, and images) and contains about two billion words, placing their study beyond the capabilities of most usual analysis techniques used by regular desktop computers. This corpus has already been the subject of several studies (Buntinx and Kaplan, 2015; Buntinx et al., 2016; Rochat et al., 2016). The corpus can easily be divided into subsets corresponding to the year of publication. However, the number of pages and their content fluctuates greatly depending on the year, ranging from 280,000 words per year in the early 19th century to about 18 million in the later years of the 20th century. **Figure 1** shows the relative size of each subset in terms of number of words per year for Le Journal de Genève (JDG) and La Gazette de Lausanne (GDL). The textual data contain some OCR errors and present other potential perturbations due to the nature of some of the content (noise). For example, bus schedules, stock market, or cinema tables contain repeated words that serve the purpose of their informative content but do not reflect the linguistic evolution. This corpus must therefore be considered as potentially noisy. Some periods, like the one from 1900 to 1915 for JDG and the one from 1965 to 1998 for the two newspapers, present higher noise levels than others. It is usual to apply a frequency filter in order to manage this problem. The main contribution of this work is to design a robust method allowing to measure linguistic changes avoiding possible misinterpretations due to noise fluctuations and corpus size variations.

Considering the lack of data for Le Journal de Genève for the years 1837, 1917, 1918, and 1919, we left these years out in all further graphs and analyses. In addition, some years had to be removed because the scanning quality was too poor (1834, 1835, 1859, and 1860 for JDG and 1808 for GDL).

2. USING CLASSICAL DISTANCES TO STUDY LINGUISTIC DRIFT

A straightforward approach to the problem consists in computing a textual distance between subsets of the corpora. One could, for instance, easily compute the so-called Jaccard distance (Jaccard, 1901, 1912) between two lexical sets. Considering two different corpora C_1 and C_2, and their lexica, i.e., the list of unique (non-lemmatized) words, $L(C_1) \equiv L_1$ and $L(C_2) \equiv L_2$, the Jaccard distance $d(L_1, L_2)$ is defined as follows:

$$d(L_1, L_2) = 1 - \frac{|L_1 \cap L_2|}{|L_1 \cup L_2|} = 1 - \frac{|L_1 \cap L_2|}{|L_1| + |L_2| - |L_1 \cap L_2|}$$

In the same way, other distances could also be explored, such as those given by Kullback and Leibler (Kullback and Leibler, 1951; Kullback, 1987), Chi-squared distance (Sakoda, 1981), or Cosine similarity (Singhal, 2001).

The Jaccard distance is an intuitive measure that determines the similarity of two texts using the relative size of their common lexicon. This distance, which is complementary to the notion of lexical connexion (Muller, 1980), is exclusively based on

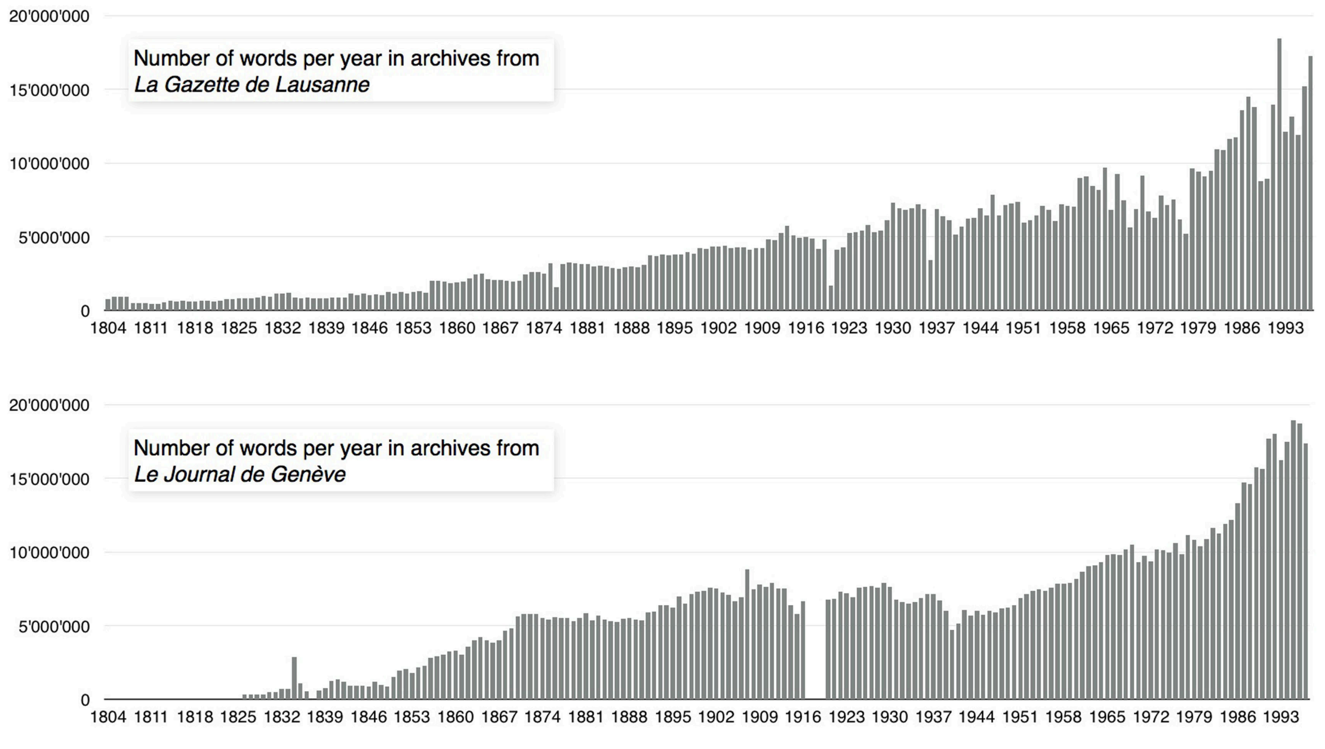

FIGURE 1 | **Corpus size versus years for GDL (top) and JDG (bottom)**.

the presence/absence of words in the lexicon and ignores their frequency.

The Jaccard distance is a metric (Levandowsky and Winter, 1971) satisfying the following classical distance properties:

- Separation: $d(L_1, L_2) = 0 \equiv L_1 = L_2$;
- Symmetry: $d(L_1, L_2) = d(L_2, L_1)$;
- Triangular inequality: $d(L_1, L_3) \leq d(L_1, L_2) + d(L_2, L_3)$.

Since the Jaccard distance measure is based only on the presence/absence of word in the corpus subsets, noise can affect the measure of linguistic evolution. In order to reduce this effect, $L(C_1)$ and $L(C_2)$ are filtered to keep only the words whose frequency is greater than 1/100,000. However, the frequency threshold is quite arbitrary, and filtered data still present OCR errors and noises. The computation of the Jaccard distance between all subsets yields a symmetric matrix $M \times M$ where M is the number of distinct years for a given newspaper. This matrix contain all distances between each pair of years $L(C_i)$, $L(C_j)$ normalized in the interval [0, 1]. The heatmaps of the Jaccard distance matrix of Le Journal de Genève (JDG) and of La Gazette de Lausanne (GDL) are given in **Figure 2**.

The values on the matrix's diagonal are equal to zero by definition (property of separation). We observe the expected behavior of the values outside the diagonal, which should be highly correlated with the difference between the compared years. In addition, level lines of the heatmap suggest the hypothesis that the linguistic evolution is not linear but evolves period of time by period of time. Indeed, in the case of a linear evolution, the level line would be parallel to the diagonal of the matrix. The same data are presented in a more convenient

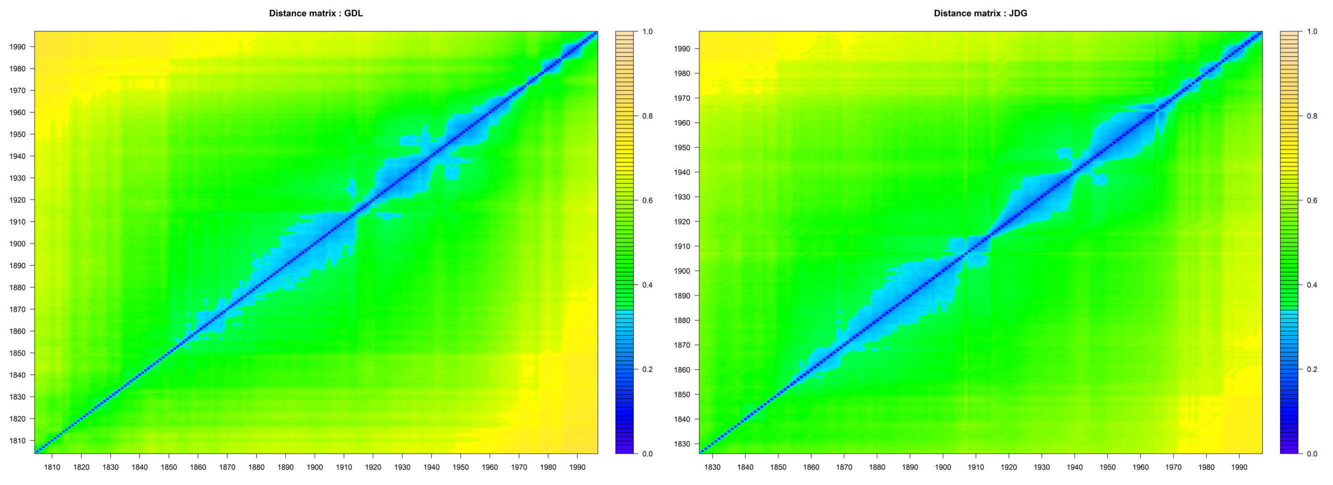

FIGURE 2 | **Heatmap of the Jaccard distance matrix of GDL (left) and JDG (right)**.

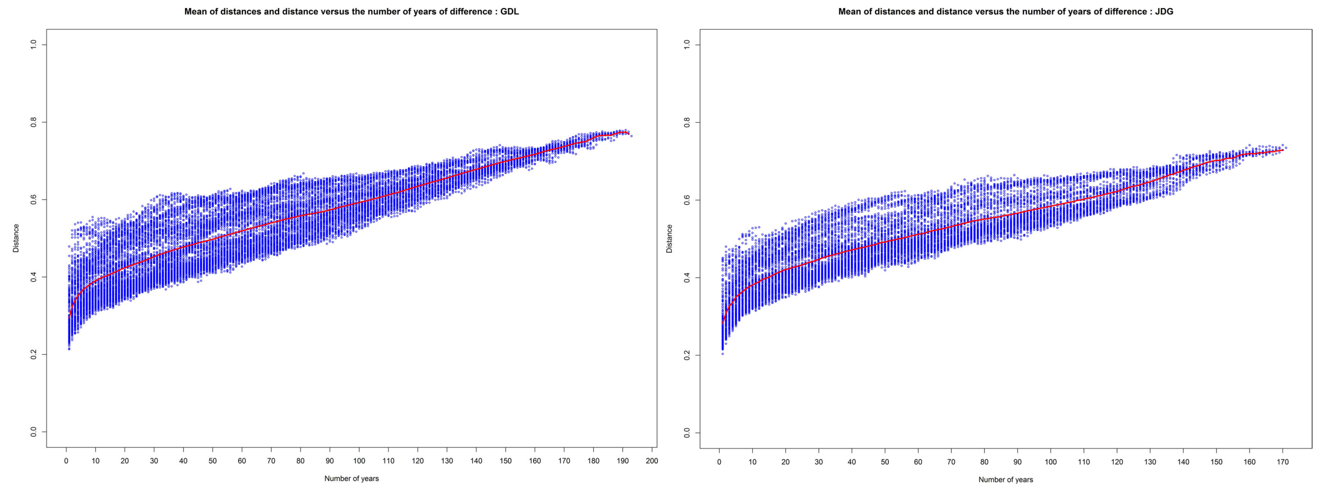

FIGURE 3 | **Jaccard distance (blue) and mean of distances (red) versus the time difference (in number of years) between the compared subset from the GDL (left) and JDG (right)**.

form in **Figure 3**. We have plotted the matrix's values in a two-dimensional graph showing the distance values versus the time differences between subsets (blue) with the mean value over time (red). In this representation, we observe that the distances seem to be overall proportional to the number of years separating the two subsets. This observation immediately suggests that the linguistic drift exists and can be quantified by the Jaccard distance. The more time separates the textual corpus, the more the subsets are indeed considered to be distant. However, it is showed in **Figure 1** that the corpus size is correlated with time and can have the role of a hidden variable affecting the distance value more than just the amount of time separating subcorpora. Two windows of time are particularly sensible in term of size fluctuation, which are the period before 1870 (with very low data representativity) and the period after 1965 (showing a sudden increase in the corpus size).

If we restrict the Jaccard distance matrix to using only the data from the most stable years in terms of size and recompute the **Figure 3** visualization, we observe the same Jaccard distance evolution. As showed in **Figure 3**, the behavior of the mean of distances (red curve) is more sensible to the first years of separation for the two newspapers. In order to measure the evolution of linguistic changes and to clarify if these changes are accelerating, decelerating, or remain stable, we show a final visualization of the distance matrix by plotting only distances between years y_i and y_{i+n} with n equal to 1, 20, 50, and 100 in **Figure 4**.

On the distance $d(y_i, y_{i+1})$ showed in **Figure 4**, we observe that the distance over 1 year decreases slowly before stabilizing from year 1920. This suggests the hypothesis that the language is more stable after 1920. We observe a brutal instability in years after 1965, matching the noisy periods of time. On the distance $d(y_i, y_{i+n})$ with $n = 20$, 50, and 100, we observe that n-years distance evolution for dates separated by more than 20 years slowly decrease before increasing significantly in more recent years because of the "contamination" of this distance matrix by the data of the perturbed years. Same graphs without that noisy period do not show any increase of the distance value, and this can therefore not be interpreted by an acceleration of the linguistic evolution. The Jaccard distance matrix indicates an overall effect that could possibly be caused by a linguistic drift, including the appearance of new words and disappearance of some old ones. However, the Jaccard distance is known to be affected by big size differences (Muller, 1980; Brunet, 2003), and other distance definitions and characterizations have been designed in order to correct this unwanted property. An improved Jaccard distance is given in a study of text similarities (Brunet, 2003) with the purpose to remove size difference sensibility from the Jaccard distance. We computed the improved Jaccard distance, and it appears that this distance has the same behavior, but with a different normalization, as the classical Jaccard distance. In addition, OCR errors and noise can affect the Jaccard distance because of its binary nature and because of its lack of frequency consideration. Frequency filters can be used to decrease noise influence, but the applied threshold is quite arbitrary.

3. LEXICAL KERNELS AND WORD RESILIENCE

3.1. Definition and Basic Measures

The uneven distribution of the size of corpus subsets (**Figure 1**) causes methodological difficulties for interpreting the distances defined in the previous section. Fluctuations in the lexicon size and noise cause an indirect increase of the linguistic drift as measured by the Jaccard formula. Under such conditions, it is difficult to untangle the effects of the unevenness of the distribution of corpus subsets from the actual appearance and disappearance of words.

These difficulties of interpretation motivate the exploration of another, possibly sounder approach to the same problem. We define the notion of the lexical kernel.

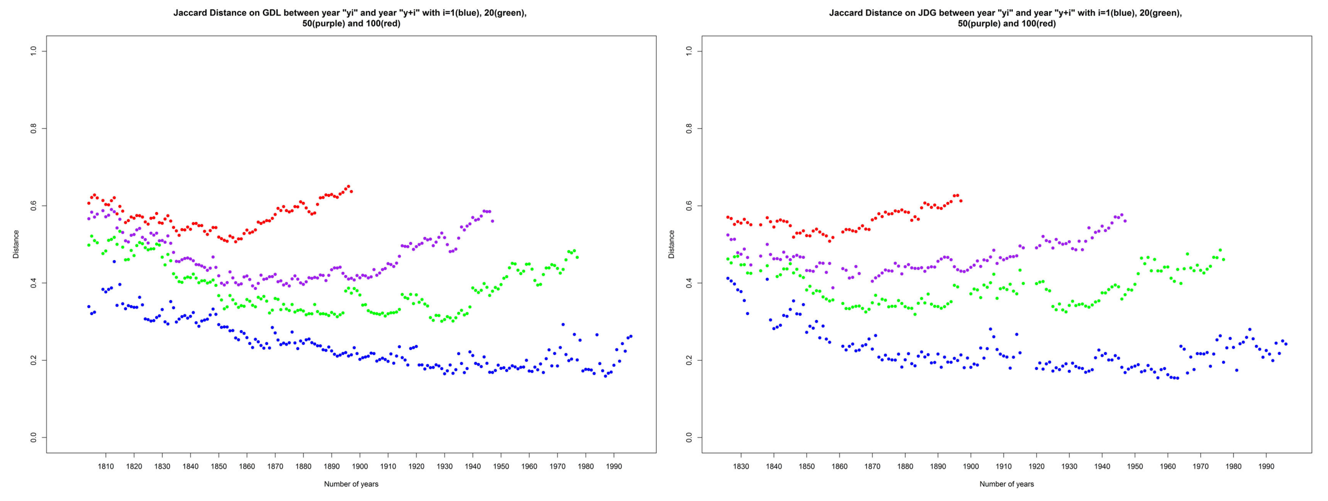

FIGURE 4 | **Jaccard distance between the years y_i and y_{i+n} with $n = 1$ (blue), $n = 20$ (green), $n = 50$ (purple), and $n = 100$ (red).**

Definition 1. *The lexical kernel $K_{x,y,C}$ is the sequential subset of unique words common to a given period of time starting in year x and finishing in year y of a corpus C.*

$K_{1804,1997,GDL}$ is, for instance, the subset of all words present in the yearly corpus of La Gazette de Lausanne. It contains 5,242 unique words that have been used for about 200 years. The kernel $K_{1826,1997,JDG}$ contains 7,485 unique words, covering a period of about 170 years. As the covered period is smaller, the time constraint is smaller, and the kernel is naturally larger.

It is interesting to note that 4,464 words are common to the two kernels. **Figure 5** shows the statistical distribution of word typologies for both kernels.

Extending the notion of a kernel, it is rather easy to study the resilience of a given word.

Definition 2. *The resilience set $R_{d,C}$ is the union of all kernels $K_{x,y,C}$ corresponding to a duration of $y - x \leq d$ years.*

The definition of word resilience is naturally derived from the resilience set notion.

Definition 3. *The resilience r of a given word w in the corpus C is given by the following formula: $r(w,C) = max\{d \mid w \subset R_{d,C}\}$.*

For instance, $R_{100,GDL}$ contains all the words that are maintained in the corpus *GDL* for at least 100 years. R subsets are organized as concentric sets: $R_{1,C} \subset R_{2,C} \subset \ldots \subset R_{i,C} \subset R_{i+1,C}$. The relative proportion of each subset sheds light on both the stability and dynamics of language change. **Figure 6** shows the distribution of word resilience for both newspapers.

The GDL resilience curve in **Figure 6** is normalized (to the same time scale as JDG) in order to make the two curves comparable. This representation of R_d shows a similar overall word resilience trend for both JDG and GDL. However, we notice that the two curves intersect when considering the longest durations.

These definitions pave the way for a formulation of the study of linguistic change in terms of algebra of sets. Instead of analyzing what is rapidly changing in the language, we study the most stable elements of language through the notions of kernel and word resilience. We can then apply a new definition of distance to the set of resilient words, which is the maximum duration kernel. Indeed, reducing the analyzed set of words to the more resilient ones allows us to exclude noise efficiently. In addition, the issue of distance sensibility to the corpus size is reduced, and the method targets linguistic evolution more precisely because the lower use of resilient words can be the result of semantic evolution, punctual journalistic events, or linguistic diversity induced by the newspaper layout evolution. The number of words is the same for each year, but the corpus size influences the frequency of kernel words when the size is small. Indeed, the smaller the corpus size, the higher the frequency fluctuation. In order to reduce these effects, we defined a distance based on word ranks ordered by their frequencies.

3.2. Distances Analysis Applied to Kernels

In order to compare the same kernel from two different years, let us consider their ordering according to the frequency of those words in those years. We may then define their distance as the computational cost to reorder one into the other. Again, we require a metric that can satisfy the mathematical properties of a distance. One way to do so is to consider a distance equal to the sum of each differences of position for each word in two given lists.

Definition 4. *Let be $I_j(w_i)$ the index of the word w_i in the list L_j. The kernel distance is given by* $d^{K}_{L_1,L_2} = \sum_{w_i \subset K} |I_1(w_i) - I_2(w_i)|$.

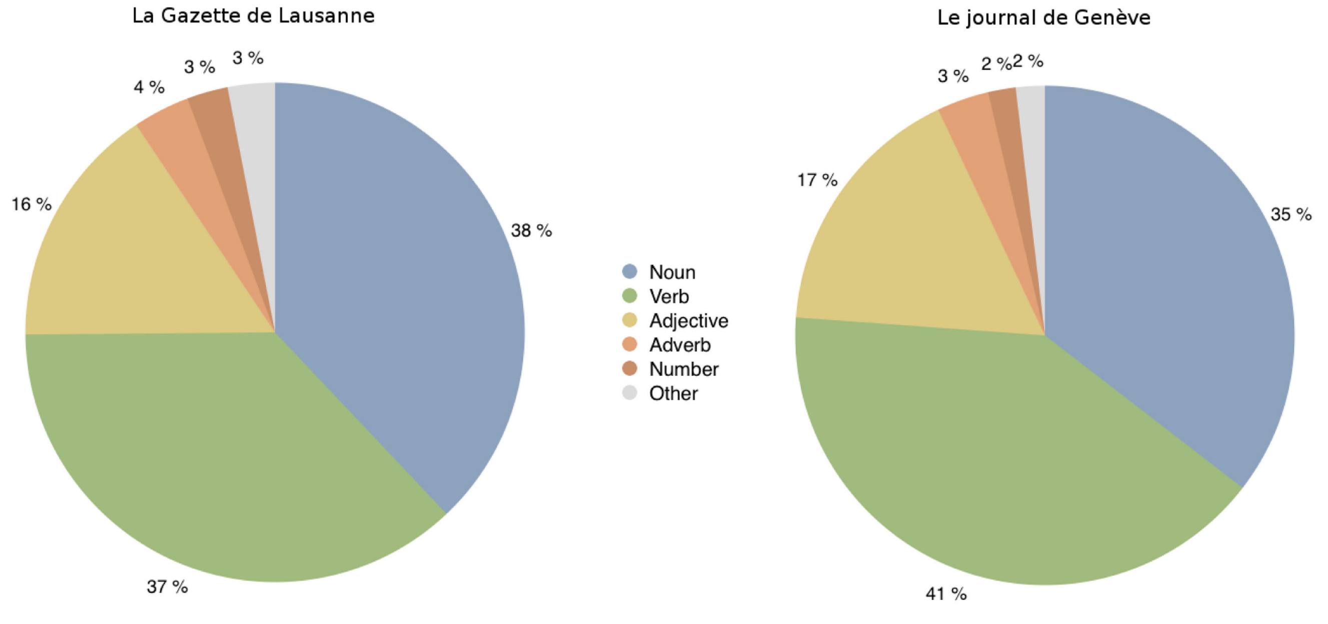

FIGURE 5 | **Distribution in terms of typologies of words contained in the kernel of $K_{1804,1997,GDL}$ (left) and $K_{1826,1997,JDG}$ (right).**

We have applied this new distance definition to the list of words from the kernel set, ordered by frequency for each years, and we have plotted the same analysis than for the Jaccard distance in the **Figure 7**, showing a representation of the distance matrix across years and the **Figure 8**, showing the distance between years y_i and y_{i+n} with n equal to 1, 20, 50, and 100.

The Jaccard distance represented in **Figure 3** and the kernel distance represented in **Figure 7** are normalized on the interval [0, 1]. They are based on different elements by definition, one on the presence/absence of words in the lexica and the other on the frequency order of kernel words. The two distances increase with increasing time difference, supporting the hypothesis that the linguistic drift exists. The Jaccard distance on the whole lexica is distributed from a mean of 0.25, when the two lexica are separated by 1 year, to 0.8, when these lexica are separated by the maximum number of years. The kernel distance, applied by definition to only a very reduced set of resilient words, is distributed from a mean of 0.1 to 0.4. The two distances share a common behavior on the two corpora of JDG and GDL. However, the kernel distance can be viewed as a lower bound of the linguistic drift showing the evolution of the most stable words. It is remarkable that the plotted evolutions share the same behavior even though the distances are based on different information types. Indeed, the Jaccard distance applied to the kernel would be equal to zero, and

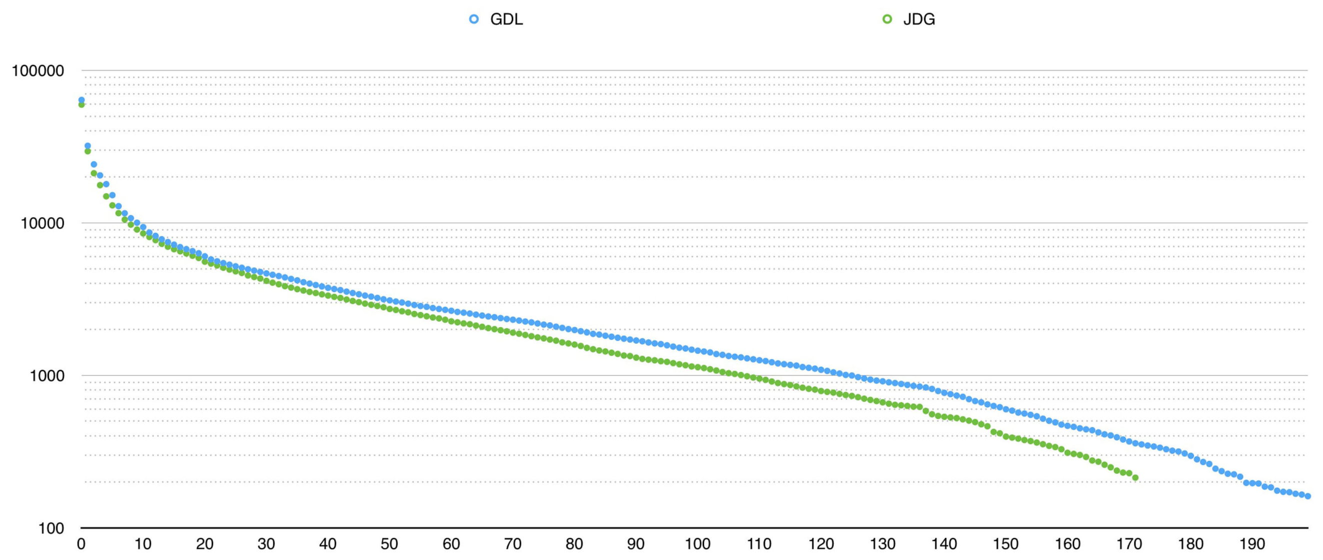

FIGURE 6 | **Size of R_d versus the number of maintained years d (logarithmic scale) showing the word resilience distribution for JDG (green) and GDL (blue)**.

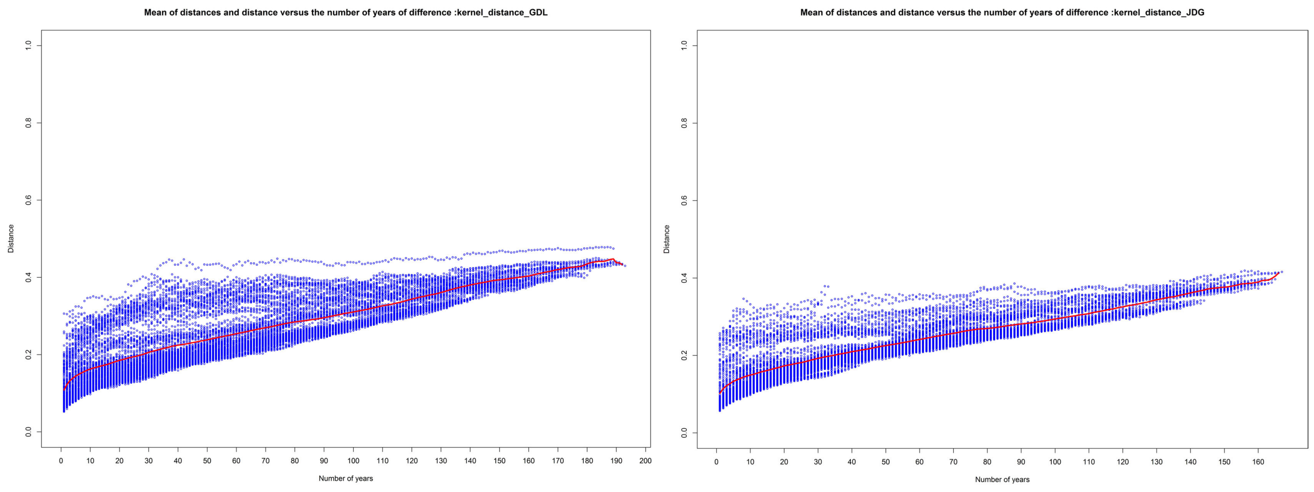

FIGURE 7 | **Kernel distance between the years y_i and y_{i+n} with n = 1 (blue), n = 20 (green), n = 50 (purple), and n = 100 (red)**.

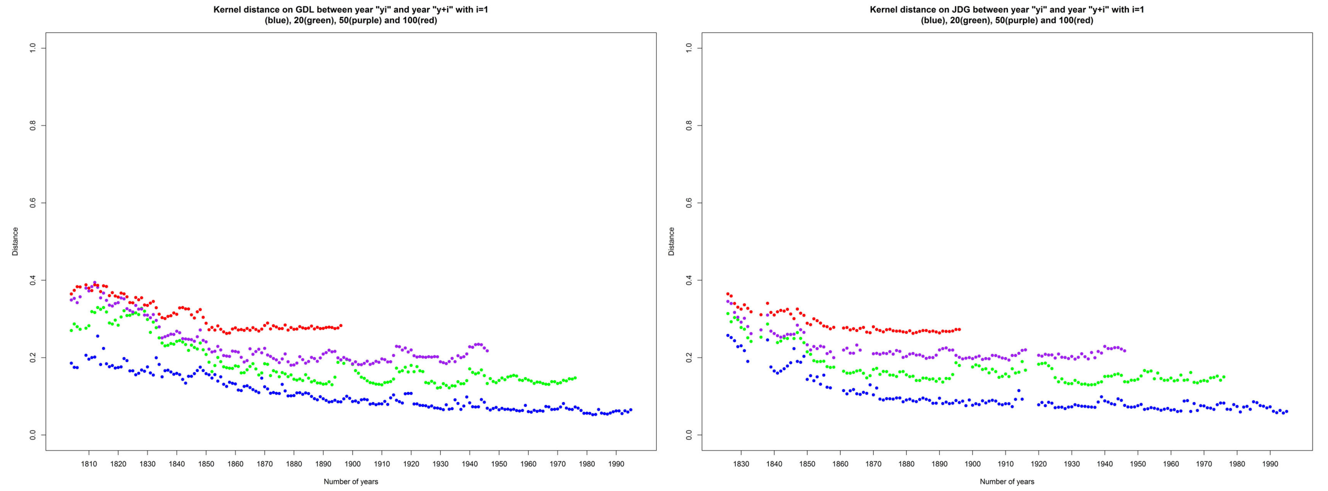

FIGURE 8 | **Kernel distance between the years y_i and y_{i+n} with $n = 1$ (blue), $n = 20$ (green), $n = 50$ (purple), and $n = 100$ (red).**

the kernel distance use information about the frequency of a very reduced set of words.

When comparing the Jaccard distance and the kernel distance in **Figures 4** and **8**, we observe that the distances between subcorpora for the oldest years share the same fluctuation as with the Jaccard distance and decrease continuously. However, this effect can be due to the low language representativity of data before 1850 (low corpus size). In general, the kernel distance decreases slowly and continuously. We observe that there is no increase but rather a very stable phase when considering two subcorpora separated by more than 20 years. The kernel distance is also more stable in more recent years. In order to attest the robustness of this measure even with noise fluctuations, we have done a linear regression on our data and on the specifically unstable period with noise (1965–1998) for the two newspapers. We hypothesized that nature of linguistic evolution excludes brutal variations and randomness around a given trend even with a simple linear model, we so expect the coefficient of regression is higher for the more robust method of evolution measurement. The two regressions for GDL and JDG on the whole data are represented in **Figure 9**. We observe that kernel distances has better regression coefficients (0.8218 for GDL and 0.6196 for JDG) than Jaccard distance (0.6294 for GDL and 0.3339 for JDG). The regressions for GDL and JDG on the noisy period are represented in **Figure 10**. We also observe that kernel distances has better regression coefficients on this short unstable period (0.2790 for GDL and 0.1635 for JDG) than Jaccard distance (0.0174 for GDL and 0.00002 for JDG). These results suggest that even if still impacting it, the kernel distance is more robust to noise than Jaccard distance.

4. DISCUSSION

Several distance definitions have been applied to the corpus of GDL and JDG in order to quantify linguistic changes. We first used the Jaccard distance on the whole corpus with a filter frequency. Our observations from the **Figures 2–4** support the hypothesis of the existence and quantifiability of the linguistic changes, even though we observe that the Jaccard distance is potentially sensible to noise. In addition, the Jaccard distance is known to be sensible to corpus size fluctuations (Muller, 1980; Brunet, 2003), so we defined the concept of kernel and word resilience in order to study the most stable part of the language.

We defined a kernel distance based on frequency rank comparison between 2 years on kernel words. Surprisingly, **Figures 7** and **8** show the same behavior than the Jaccard distance on the whole corpus. This supports the hypothesis that the linguistic distances' information extracted by word presence/absence on the whole corpus can be retrieved using a reduced set of resilient words from the kernel with the kernel distance. In addition, the kernel distance clearly overcomes the noise problems, canceling the effect of the contamination of years with higher noises like the period of 1900–1915 for JDG and the period of 1965 and more for the two newspapers. Observation for this distance, like the one for the Jaccard distance, shows a decrease before the period of time prior to 1870, which is a period with low probability of being representative of language because of the small corpus size. After this period of time, distances from a year to the next year seem to decrease slowly but with more stability. Additionally, the distance from a year to 20, 50, or 100 year later remains stable.

From our experiments on the two corpora of GDL and JDG, we have made a series of observations that support the existence of a continuous and relatively constant linguistic drift. We tried several methods to quantify this linguistic change with success in overcoming problems of noise, corpus size fluctuation, and precise targeting of linguistic change instead of other cumulated effects on the corpora's textual data like topics, OCR quality, or noise evolution. If these measures show a quantization way of the linguistic drift, we do not have any serious indicators or proof of a potential acceleration or deceleration of the language change evolution on the periods of 1804–1997 (GDL) and 1826–1997 (JDG).

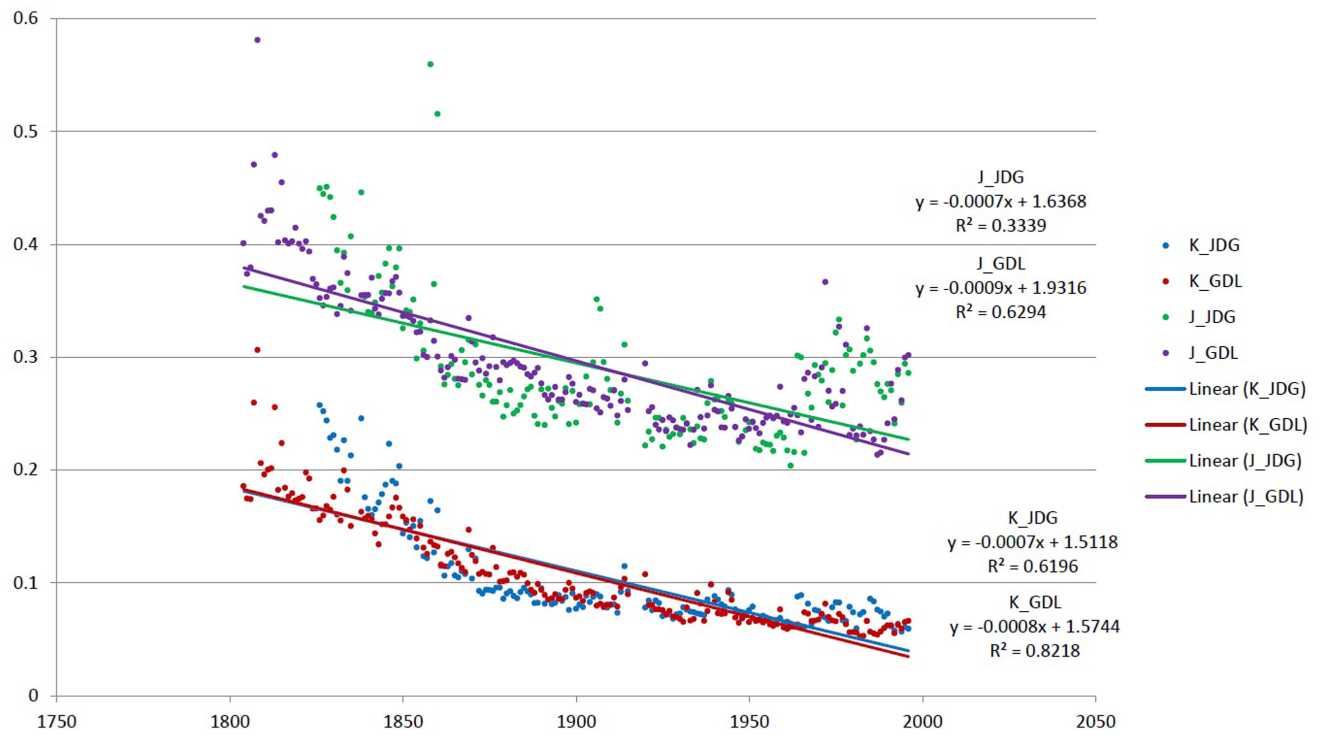

FIGURE 9 | **Jaccard distance (purple for GDL and green for JDG) and kernel distance (red for GDL and blue for JDG) versus years with their linear regressions and regression coefficients**.

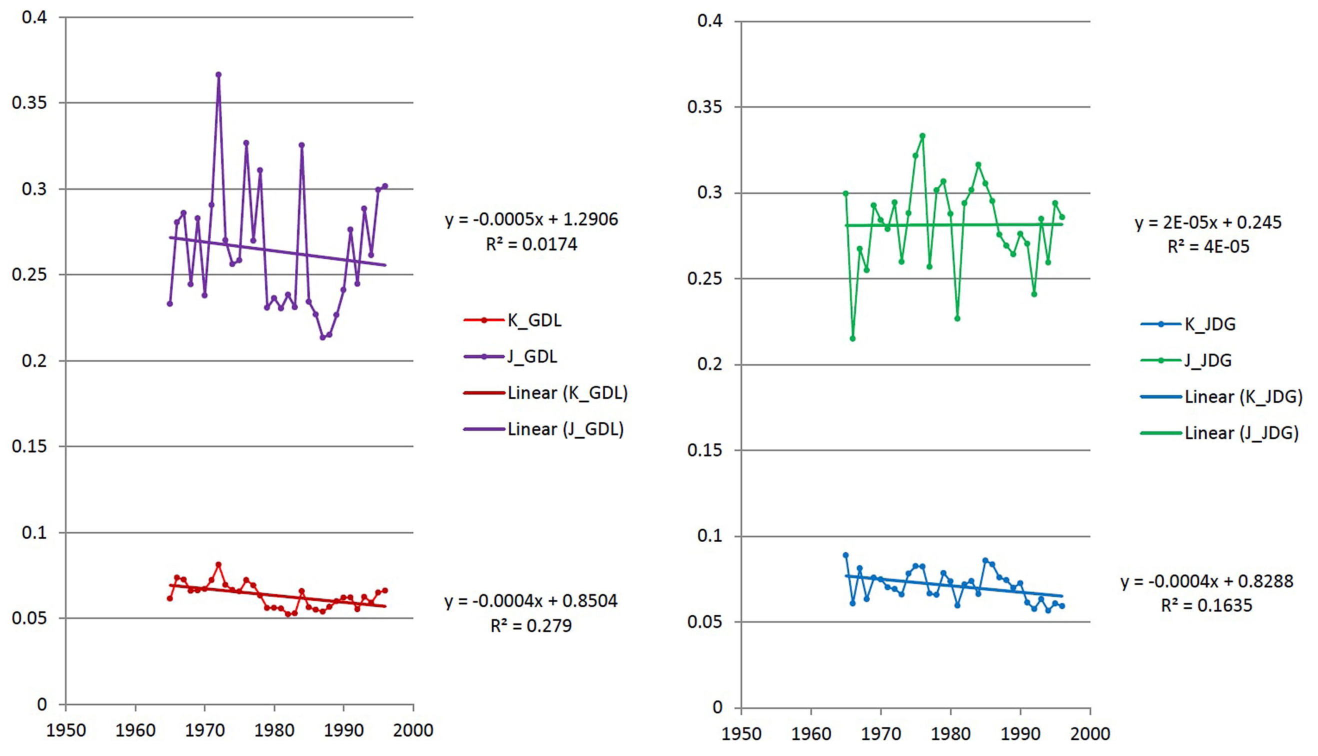

FIGURE 10 | **Jaccard distance (purple for GDL and green for JDG) and kernel distance (red for GDL and blue for JDG) versus years with their linear regressions and regression coefficients on the period of 1965 and more**.

However, these methods should be applied on a corpus where data are available after 1997 in order to verify if this observed stability is maintained during the period of 1998–2016 where a lot of technologies mediating our language have potentially accelerated linguistic evolution (Kaplan, 2014).

5. CONCLUSION AND FUTURE WORK

Large databases of scanned newspapers open new avenues for studying linguistic evolution (Westin and Geisler, 2002; Fries and Lehmann, 2006; Bamford et al., 2013). However, these studies should be conducted with sound methodologies in order to avoid misinterpretation of artifacts. Common pitfalls include misinterpreting results linked to the size variation of the subsets or overgeneralizing results obtained from one particular newspaper corpus to general linguistic evolution.

In this paper, we introduced the notion of a kernel as a possible approach to studying linguistic changes under the lens of linguistic stability. Focusing on stable words and their relative distribution is likely to make interpretations more robust. Results were computed from two independent corpora. It is striking to see that most of the results obtained from each of them are extremely similar. The kernels compositions in terms of grammatical word typologies are very similar.

The kernel distance, applied to the kernels words in order to measure the linguistic changes, has showed to be robust when it comes to OCR errors and noise. In addition, we observed that the study of kernel words allows the extraction of the same linguistic distance's information as the Jaccard distance applied to the whole corpus. This suggests that our methods are indeed measuring general linguistic phenomena beyond the specificity of the corpora chosen for this study. Future works and analysis should include the case where corpus kernel size is too small and implement a distance measuring the linguistic change between subset of resilient words that are not necessarily part of the kernel. In addition, our results still need to be confirmed with subsequent studies involving other corpora, such as non-journalistic texts and texts written in other languages.

AUTHOR CONTRIBUTIONS

The three authors have contributed equally to the conception and design of this work through ideas and results discussion. VB has performed data acquisition, computation and analysis, visualizations of computed results, and has written the article. CB has provided visualizations of computed results and suggested to reduce the analyzed set of words to those shared by each subcorpora. FK has provided deeper formalization of the developed concepts of kernels and words resilience. The three authors have participated in the process of reviewing the articles' final version, ensuring its accuracy and integrity.

ACKNOWLEDGMENTS

We thank the team of Le Temps newspaper and the BNS (Bibliothque Nationale Suisse) for giving us the opportunity to work on those 200 years of archives.

FUNDING

This study is funded by FNS and is part of the project "How algorithms shape language," number 149758.

REFERENCES

Bamford, J., Cavalieri, S., and Diani, G. (2013). Variation and Change in Spoken and Written Discourse: Perspectives from Corpus Linguistics. *Dialogue Studies* 21.

Bochkarev, V., Solovyev, V., and Wichmann, S. (2014). Universals versus historical contingencies in lexical evolution. *Journal of the Royal Society Interface* 11: 20140841. doi:10.1098/rsif.2014.0841

Brunet, E. (2003). Peut-on mesurer la distance entre deux textes? *Corpus*. Available at: http://corpus.revues.org/index30.html

Buntinx, V., Bornet, C., and Kaplan, F. (2016). Studying linguistic changes on 200 years of newspapers. In *Digital Humanities 2016*.

Buntinx, V., and Kaplan, F. (2015). Inversed N-gram viewer: searching the space of word temporal profiles. In *Digital Humanities 2015*.

Cocho, G., Flores, J., Gershenson, C., Pineda, C., and Snchez, S. (2015). Rank diversity of languages: generic behavior in computational linguistics. *PLoS ONE* 10:e0121898. doi:10.1371/journal.pone.0121898

Fries, U., and Lehmann, H.M. (2006). The style of 18th century English newspapers: lexical diversity. In *News Discourse in Early Modern Britain*, 91–104.

Gerlach, M., Font-Clos, F., and Altmann, E.G. (2016). Similarity of symbol frequency distributions with heavy tails. *Physical Review X* 6: 021009. doi:10.1103/PhysRevX.6.021009

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Société Vaudoise des Sciences Naturelles* 37: 547–79.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist* 11: 37–50. doi:10.1111/j.1469-8137.1912.tb05611.x

Kaplan, F. (2014). Linguistic capitalism and algorithmic mediation. *Representations* 127: 57–63. doi:10.1525/rep.2014.127.1.57

Kullback, S. (1987). Letters to the editor. *The American Statistician* 41: 338–41. doi:10.1080/00031305.1987.10475510

Kullback, S., and Leibler, R.A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics* 22: 79–86. doi:10.1214/aoms/1177729694

Levandowsky, M., and Winter, D. (1971). Distance between sets. *Nature* 234: 34–5. doi:10.1038/234034a0

Michel, J.-B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K.; Google Books Team, et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science* 331: 176–82. doi:10.1126/science.1199644

Muller, C. (1980). *Principes et méthodes de statistique lexicale*. Vol. 2. Bulletin des bibliothèques de France (BBF), 80.

Pechenick, E.A., Danforth, C.M., and Dodds, P.S. (2015a). Characterizing the google books corpus: strong limits to inferences of socio-cultural and linguistic evolution. *PLoS ONE* 10:e0137041. doi:10.1371/journal.pone.0137041

Pechenick, E.A., Danforth, C.M., and Dodds, P.S. (2015b). Is language evolution grinding to a halt: exploring the life and death of words in English fiction. In *CoRR*. arXiv: 1503.03512v1, 1–12.

Piantadosi, S.T. (2014). Zipf's word frequency law in natural language: a critical review and future directions. *Psychonomic Bulletin & Review* 21: 1112–30. doi:10.3758/s13423-014-0585-6

Rochat, Y., Ehrmann, M., Buntinx, V., Bornet, C., and Kaplan, F. (2016). Navigating through 200 years of historical newspapers. In *Proceedings of iPRES 2016*, 186–195.

Sakoda, J.M. (1981). A generalized index of dissimilarity. *Demography* 18: 245–50. doi:10.2307/2061096

Singhal, A. (2001). Modern information retrieval: a brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 24: 35–43.

Westin, I., and Geisler, C. (2002). A multi-dimensional study of diachronic variation in British newspaper editorials. *International Computer Archive of Modern and Medieval English* 26: 133–152.

Zipf, G. (1935). *The Psychobiology of Language: An Introduction to Dynamic Philology*. Cambridge, MA: M.I.T. Press.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer TN and handling Editor declared their shared affiliation, and the handling Editor states that the process nevertheless met the standards of a fair and objective review.