Ecole Polytechnique Fédérale de Lausanne

Big Data
2015

Group Project Report

A Study of Linguistic Drift On *Le Temps* Newspaper Corpus

Authors:
Nicolas Bornand
Farah Bouassida
Malik Bougacha
Gil Brechbühler
Tao Lin
Cynthia Oeschger
Marc Schär
Jérémy Weber

Supervisors:
Immanuel Trummer
Christoph Koch

May 19, 2015

Contents

1 Introduction **3**
1.1 Problem Statement . . . 3

2 Statistics **3**
2.1 Language coverage . . . 3
2.2 Sentences and punctuation statistics . . . 4
2.3 Appearance of new words . . . 5
2.4 Synonyms . . . 7

3 TF-IDF **9**

4 OCR Correction **9**
4.1 Why do we need to clean the OCR? . . . 9
4.2 Clean Strategy . . . 10

5 Metrics **10**
5.1 Basic Distance . . . 11
5.1.1 The formula . . . 11
5.1.2 Computations on the corpus . . . 11
5.1.3 Analysis . . . 12
5.2 Chi-Square Distance . . . 12
5.2.1 The formula . . . 12
5.2.2 Computations on the corpus . . . 12
5.2.3 Analysis . . . 13
5.3 Cosine Similarity . . . 13
5.3.1 The formula . . . 13
5.3.2 Computations on the corpus . . . 13
5.3.3 Analysis . . . 14
5.4 Kullback Leibler Divergence . . . 14
5.4.1 The formula . . . 14
5.4.2 Computations on the corpus . . . 15
5.4.3 Analysis of results . . . 17
5.5 Out of Place Measurement . . . 18
5.5.1 Metric description . . . 18
5.5.2 Computations on the corpus . . . 18
5.5.3 Analysis of results . . . 19

6 Dating Articles **19**
6.1 Metrics . . . 20
6.1.1 Basic Distance . . . 20
6.1.2 Chi Square . . . 20
6.1.3 Cosine . . . 21
6.1.4 Cosine with TF-IDF . . . 21
6.1.5 Kullback-Leibler Divergence . . . 22
6.1.6 Out of Place . . . 23
6.1.7 Punctuation and sentences length . . . 25
6.2 Cross Validation . . . 25

7 Topics **29**
7.1 Metrics with Topics 30
7.1.1 Cosine Similarity 30
7.1.2 Kullback-Leibler Divergence 31

8 Conclusion **32**

1 Introduction

1.1 Problem Statement

Intuition helps us think that, through time, language is changing because of the variation of the vocabulary but also in the way words of this vocabulary are used. Along with discoveries and civilization evolution, some words disappear and others emerge. Moreover, the 21st century is characterized by secondary source documents that influence us in our way of expressing ourselves. Thus, our study tries to define the features that could prove that there is a linguistic drift and to quantify the changes with a metric. Our dataset is a french articles corpus from 1840 to 1998 from two Swiss newspapers. We are provided with xml files of OCRized articles. The fact that there is less articles during the first years is sometimes to be considered, as is the fact that for years 1917 to 1919 we just have one newspaper's articles and that in 1998, we just have the articles of a few months.

Within this project, we want to observe the evolution of the language : "Do some words appear and then disappear ?" "How much different words do we need to cover a certain percentage of the language ?" "Is the correction of the OCR useful to observe the linguistic drift ?" "Does giving more weight to uncommon words improve the drift ?" "Is it possible to date an article with precision using only its vocabulary ?" "Are topics helpful for observing the linguistic drift ?" These are the questions we want to answer through our project.

This report is structured as follows : first, in section 2 we present some statistics about the corpus using the number of words, the sentences, their length, the use of punctuation. Secondly, in section 5, we present and explain different metrics we used to try to observe a linguistic drift. Then, in section 6, we try to date articles using the metrics presented in section 5. After that, in section 7, we present some topics we found and we try to date articles according to topics. Finally, in section 8, we answer the questions stated in this introduction.

2 Statistics

First, we wanted to do some statistics about the given corpus. We have articles from two newspapers covering years 1840 to 1998. The first step we did was to extract the 1 to 5-grams. Then we could work on those grams. The grams are separated by their sizes in different folders (1-grams, 2-grams, ...) and there is one file by year in those folders. The files contain entries in the form : gram <tabulation> count, where count is the number of times the gram appears in that year.

In the following sections, we are going to talk some statistics we did based on the obtained grams that could be considered as features to compute the linguistic drift.

2.1 Language coverage

The first statistic consists of determining how much of the language is covered by the k most used words. The result is shown in figure 1. We could see that, with the same amount of words, we cover less and less of the language from 1840 to 1995.

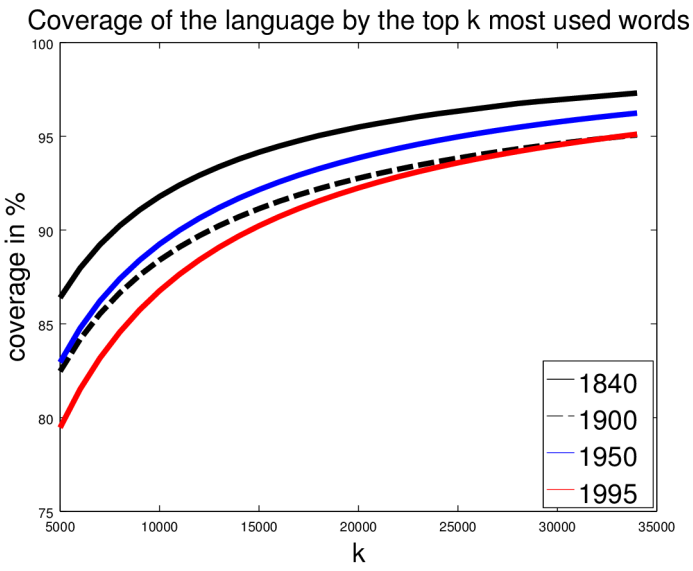

Figure 1: Percentage of text covered by the top k most used words

We can interpret this result in the way that there are more new words than disappearing words, and thus, against our intuition, our everyday vocabulary becomes wider. This can be due to all the vocabulary that appeared with the internet and new technologies.

2.2 Sentences and punctuation statistics

We then did some statistics on the sentences and the punctuation. We started by analysing the average length of the sentences (see figure 2).

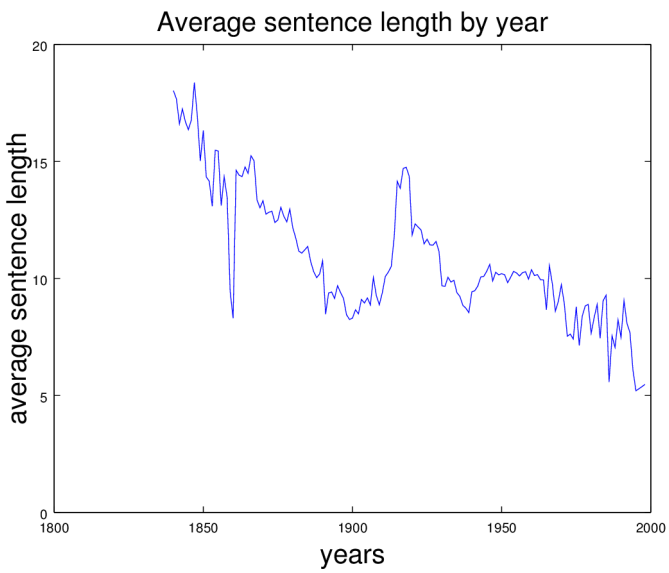

Figure 2: Average length of sentences by year

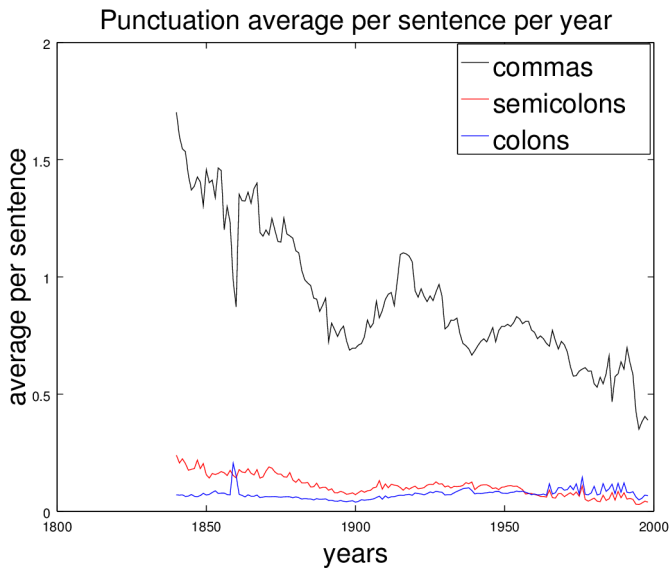

Figure 3: Average number of commas, semicolons and colons per sentence per year

If we ignore the down peak near 1850, coming from some unknown problem in two years (maybe less articles or big OCR problems) we see that the length of the sentences tends to decrease through time. In fact, the decrease is significant, going from 15-20 words by sentence in 1840 down to 8-10 words by sentence in 1990.

Note : another problem apart from the bad OCR is that from year 1917 to 1919 (included), we were provided with only one of the two newspapers merging into "Le Temps" articles, what affects the statistics for those years as there are less articles than in the other years.

We can correlate those results with the statistics on punctuation. In figure 3, the black curve represents the average of commas per sentences per year, the red curve the average of semicolons and the blue curve the average of colons. The two last curves are not really interesting in terms of linguistic drift as they do not vary much. In fact, the articles are using approximatly the same number of colons and semicolons from 1840 to 1990. The black curve however show that, there is a tendency to use less commas on articles, what is quite logical as sentences are shorter so we need to use less commas.

Again, those statistics show the linguistic drift, between 1840 and 1990 by showing the fact that, through time, shorter sentences are used in our corpus. This trivially affects the punctuation used.

As those observations show a real variation between years, we thought it could be a good idea to use them for a metric. Results are presented in section 6.

2.3 Appearance of new words

We consider the rate at which new 1-grams are incorporated in the vocabulary, and the nature of those words, as a crucial component of linguistic drifts. Thus, it was important to find with a reliable way to quantify the linguistic changes with this feature.

A simple approach to estimate the apparition of new words would be to sum up the

number of distinct 1-grams for each year and then monitor its evolution. But this suffers from several drawbacks : first the remaining OCR errors, that still account for a substantial portion of the distinct 1-grams, would be incorrectly considered as emerging words. Furthermore, it would not provide any clue about the actual amount of words entering/exiting the set of used words, but only snapshots of how many words are in fashion at any point in time. In our opinion, those two aspects would have strongly distorted the resulting figures, this is why we opted for a different procedure.

The idea was to define a pair of thresholds, one below which a word would be considered as not existing and a higher one meaning that the 1-gram is commonly used. The lower bound is not zero such as to take possible OCR correction into account. Once the higher threshold has been crossed, we require the frequencies to stay above the threshold R% of the time. This filter will only detects words that appear quite suddenly. After checking, we find that this method is accurate for determining the precise time where the words appeared (inferred from *zerosSeqAtStart*). Figure 4 is meant to illustrate the algorithm.

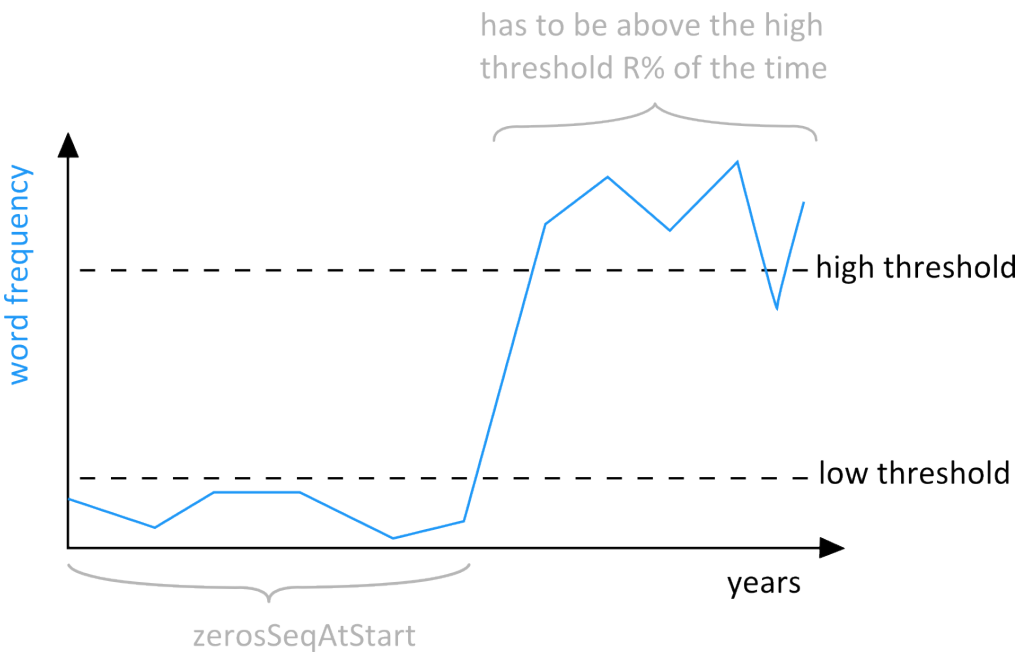

Figure 4: Visualization of the detection algorithm's behavior

Figures 5 and 6 show the amounts of words spotted as appearing, using frequencies averaged over range of 5 years. In our opinion, the results for the first 15 years are overly influenced by the fact that they contains very few articles compared to other years (factor of 6). We see a clear increase in term of word creation recently, no matter the set of parameters selected.

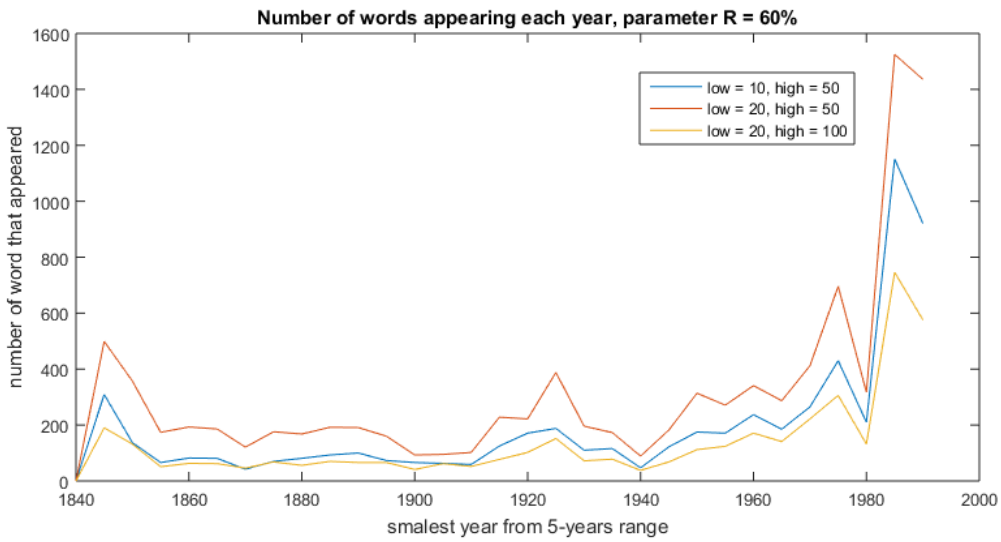

Figure 5: New words detection with tolerance rate R = 60%

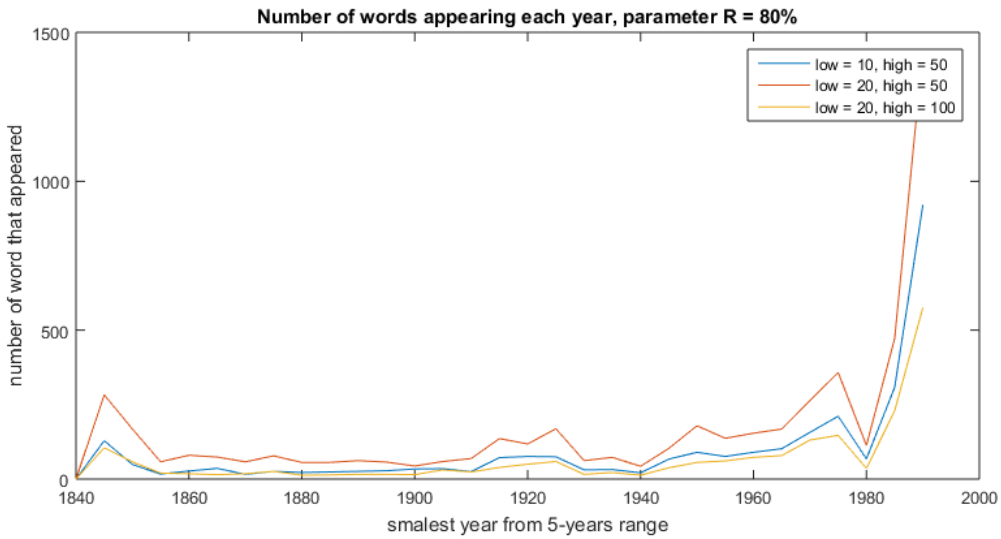

Figure 6: New words detection with tolerance rate R = 80%

2.4 Synonyms

We have to remember that beyond n-grams frequency changes, there is also a shift in term of topics that are covered by newspapers over the years. To try to isolate our analysis from that effect, we planned to compare frequencies of words that have similar meaning using a synonym dictionary. After trying a few of them, we have chosen to use the myThes v2.0 synonym dictionary from OpenOffice that seemed exhaustive and accurate enough.

Frequency vectors were generated for every entry in the synonym database (using mapReduce) before being inserted in an SQL database. We have chosen to average frequencies to ranges of 5 years such as we limit the data size. Then we build a little web application that allows to check for words in the thesaurus database, then select a subset of the proposed similar words before displaying a plot with the evolution of frequencies.

This interactive application enabled us to quickly explore hypothesis we had about expected frequencies changes while, before, we needed to run a mapReduce job on the cluster for any chosen word. But most of the time, the synonyms proposed happen to have various meanings and were unsuitable to perform an automated analysis on our data. The intervention of the user was crucial during the selection phase to choose the appropriate ones. For instance a synonym for ”voiture” (car) is ”caisse” (slang) that also means box (by far its most frequent usage). Thus the perceived frequency of ”caisse” will be strongly overstated and we cannot easily, at least for the scope of this project, tell apart the different meanings for a given 1-gram and split the occurrences accordingly.

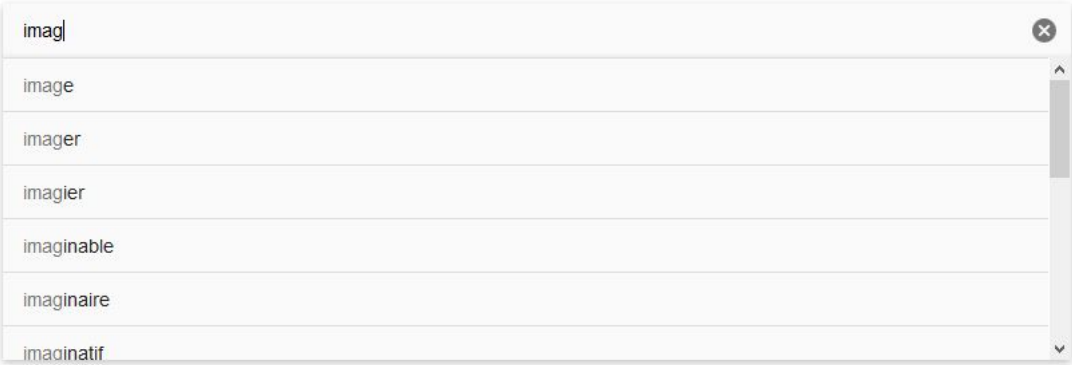

Figure 7: Seaching the thesaurus database with autocomplete

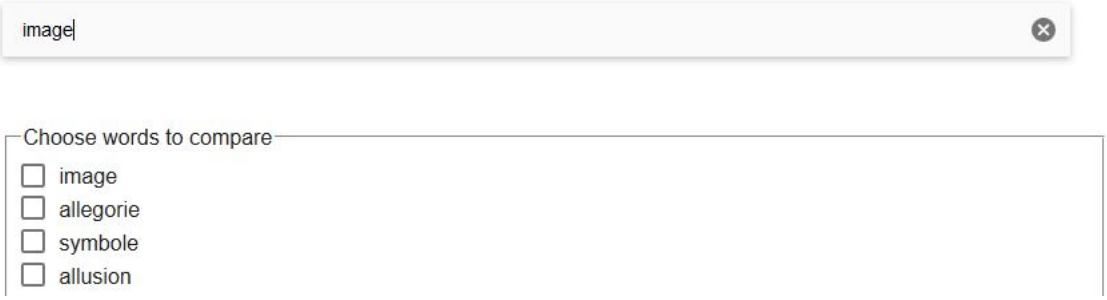

Figure 8: Check the proposed synonyms to display

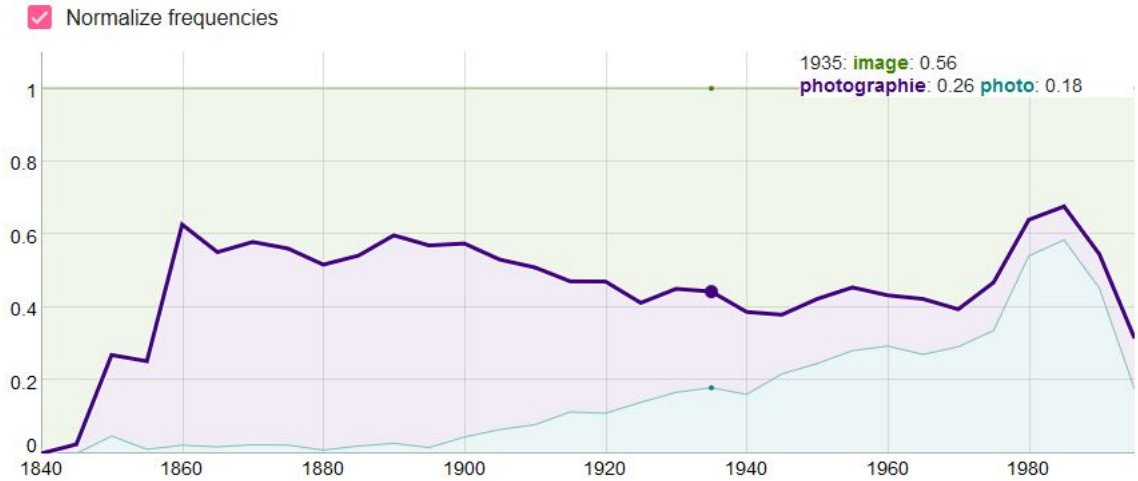

Figure 9: Resulting frequency graph (with normalization option enabled)

3 TF-IDF

To try to accentuate the linguistic drift over the corpus, we thought it was a good idea to implement the *Term Frequency* for each word in each year combined with the *Inverse Document Frequency* which divides the number of documents in the corpus (years) by the number of documents where the word appears.

$$TF = \frac{\#\ of\ occurrences\ of\ a\ word\ in\ a\ year}{\#\ of\ words\ in\ a\ year} \tag{1}$$

$$IDF(t, D) = log\frac{N}{|\{d \in D : t \in d\}|} \tag{2}$$

where N is the total number of documents (years) and $|\{d \in D : t \in d\}|$ is the number of documents where the word t appears. We will see in the next sections of this paper that the TF-IDF values can be useful or not depending on the metric used.

4 OCR Correction

4.1 Why do we need to clean the OCR?

Limited by the optical character recognition, OCR documents comprise of a significant amount of erroneous words. Since our project rely heavily on word matching between documents, we need to deal with OCR error before we apply advance similarity calculation methods.

Moreover, with all that said, the practical reality is that many OCR errors are quite predictable. So it is possible, and useful, to produce a pretty-good list of rules that simply translates individual tokens back into correctly spelled words. In this part, we utilized one simplest approach, i.e. to develop a set of rules to translate individual tokens produced by OCR errors back into correctly-spelled forms.

In order to found the impact of OCR correction, in the later section 5 about metrics, we also experimented different metrics on OCR corrected dataset and OCR non-corrected dataset to find out whether OCR will play a significant effect upon the accuracy of metrics.

4.2 Clean Strategy

Since we need to pay more attention to the linguistic shift, here, for the OCR correction, we chose simplest approach and identified a set of predictable character substitutions.

1. Basic Replacement Rule

   We apply some common replacement rules to make the OCR word to be the same form.

   1. Remove capital letter
   2. Remove alone letter
   3. Remove punctuation
   4. Remove accent
   5. Remove special letter (e.g., ;, /, -, ', &)

2. Apply a set of predictable character substitutions

   In this part, we identified a set of common OCR error and do basic word replacement.

   1. Replace complete word
      We replaced complete word if current n-grams exists a match with the word that in the dictionary, e.g., $so \rightarrow se$, $pir \rightarrow par$, $pavs \rightarrow pays$.
   2. Replace part of word
      For the matching of the partial word, we also use regular expression to find the corresponding word and its replacement word, e.g., $iii \rightarrow m$, $ii \rightarrow n$, $q.([a-z]) \rightarrow qu?$. Actually, most of wrong OCR words are local errors, and the substitution of local error plays a significant role in the OCR correction.

5 Metrics

We look to determine a metric (or distance) between the years of the corpus that can show a linguistic drift. Intuitively, the language nowadays and 10 years ago is relatively the same, but there is more differences with the language of 200 years ago. It should be a smooth evolution and we want to find a way to characterize it mathematically. Here are some properties that we defined for a distance d between two subsets of the corpus C_i and $C_{i'}$:

- Normalization: $0 \leqslant d(C_i, C_{i'}) \leqslant 1$.
- Distance between a subset and itself is zero: $C_i = C_{i'} \Rightarrow d(C_i, C_{i'}) = 0$.
- The distance should be almost similar for consecutive years while increasing as years become distant.
- The distance should not depend on the size of the corpus.

In the following, distances will be represented in matrices of size $N \times N$, with $N = 159$ years corresponding to the interval of years building our corpus:

$$
\begin{array}{cccccc}
 & year1 & year2 & ... & yearN \\
year1 & \mathbf{0} & \mathbf{d(year1, year2)} & ... & \mathbf{d(year1, yearN)} \\
year2 & \mathbf{d(year2, year1)} & \mathbf{0} & ... & \mathbf{d(year2, yearN)} \\
... & ... & ... & ... & ... \\
yearN & \mathbf{d(yearN, year1)} & \mathbf{d(yearN, year2)} & ... & \mathbf{0}
\end{array}
$$

We tried several metrics that are presented in the next subsections. We also put a set of criteria that we could use to judge that a metric is better than another one.
For the robustness, the size of the corpus shouldn't influence the distance. Thus, the distance between a subset of a set C_i and a subset of another set $C_{i'}$ should be the same as the distance between the whole sets C_i and $C_{i'}$.
Secondly, the distance should be robust considering OCR unresolved errors.
We also proceeded a datation of articles for each metric to compare the efficiency of each one. We would estimate that a datation is efficient if, when computing the distance of an article with each year or years intervals, then the smaller distance should assign the text to a year with 10 years margin.
Finally, in the matrix plot, the colour should fade out homogeneously from the diagonal to the opposite corners. There should be a kind of 'band' around the diagonal.

5.1 Basic Distance

In order to have a template and a point of comparison for our following work, we started with a simple and intuitive metric.

5.1.1 The formula

The distance calculation is based on the common words between two subsets C_i and $C_{i'}$ of the corpus. The distance $d(C_i, C_{i'})$ is defined as follows:

$$d(C_i, C_{i'}) = 1 - \frac{2\|C_i \cap C_{i'}\|}{\|C_i\| + \|C_{i'}\|} \tag{3}$$

where $\| \cdot \|$ denotes here the number of distinct words and $C_i \cap C_{i'}$ is the set containing the common words between C_i and $C_{i'}$.

5.1.2 Computations on the corpus

We applied the basic distance on the years contained in the corrected corpus, producing the following matrix:

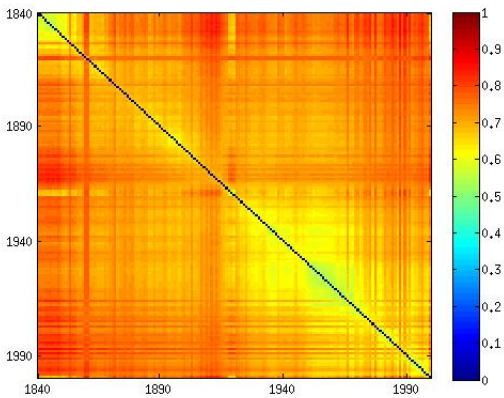

Figure 10: Basic distance for 1-gram with OCR correction

5.1.3 Analysis

As our goal is to find a metric that can show a linguistic drift, consecutive years should have a distance near zero and distant years should have a distance close to one. In figure 10, if we take for example the year 1940, the distance with itself is zero as wanted. But the distance with 1939 or 1941 is jumping over 0.5. Those kind of jumps around the zero diagonal are the behaviour we want to avoid and this distance is definitely not good to quantify the linguistic drift.

5.2 Chi-Square Distance

The *Chi-Square* distance is often used in textual data analysis to find similarities between documents. It is very close to the euclidean distance (sum of squared differences between the elements) weighted by a term associated to the frequency of each word in the whole corpus.

5.2.1 The formula

In each subset C_i of the corpus, we can denote $f_{i,j}$ the frequency of the word w_j. The distance is computed as follows:

$$d(C_i, C_{i'}) = \sum_j \frac{1}{f_{.j}} \left(\frac{f_{i,j}}{f_{i.}} - \frac{f_{i',j}}{f_{i'.}}\right)^2 \tag{4}$$

where $f_{i.} = \sum_j f_{i,j}$ is the total number of words contained in C_i and $f_{.j} = \sum_i f_{i,j}$ is the frequency of the word w_j over the entire corpus.

The term $f_{i,j}$ can also be replaced by the TF-IDF term.

5.2.2 Computations on the corpus

We applied the *Chi-Square* distance on the years contained in the corrected corpus, producing the following matrices:

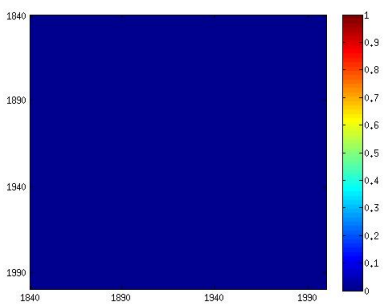

Figure 11: Chi-Square distance for 1-gram with OCR correction

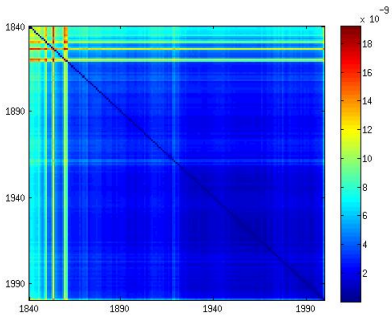

Figure 12: Chi-Square distance with reduced scale

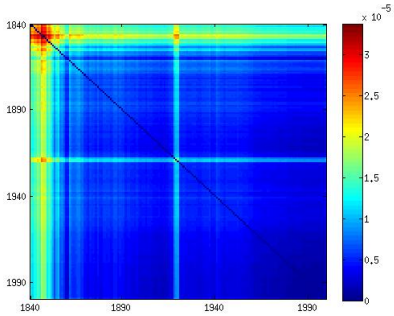

Figure 13: Chi-Square distance weighted by TF-IDF

5.2.3 Analysis

As we have a huge amount of articles for each year and that *Chi-Square* is used to compare documents (with a relatively small amount of words compared to our case), the maximal distance achieved is very low as shown in figure 11. In order to observe something we reduced the scale of the colours in figure 12. This distance is definitively smoother than the basic one but it doesn't show any evolution through the years.

Figure 13 shows the results when the terms of the distance are weighted by the TF-IDF coefficients. The values look similar than without the weighting.

5.3 Cosine Similarity

The cosine similarity is a measure between two vectors that is often used in documents research. It computes the cosine of the angle between both term vectors of the documents.

5.3.1 The formula

As the measure computes the similarity between two documents, we will use $d(C_i, C_{i'}) = 1 - cosine_similarity$ to have a distance. In each subset C_i of the corpus, we can denote $f_{i,j}$ the frequency of the word w_j. It is computed as follows:

$$d(C_i, C_{i'}) = 1 - \frac{\sum_j f_{i,j} \cdot f_{i',j}}{\sqrt{\sum_j f_{i,j}^2}\sqrt{\sum_j f_{i',j}^2}} \tag{5}$$

The term $f_{i,j}$ can also be replaced by the TF-IDF term.

5.3.2 Computations on the corpus

We applied the cosine distance on the years contained in the corpus, producing the following matrices:

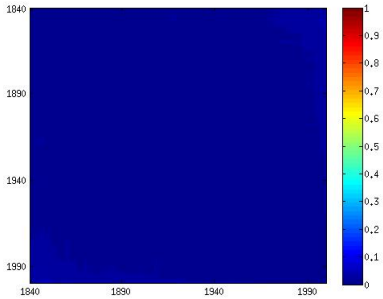

Figure 14: cosine distance for 1-gram with OCR correction

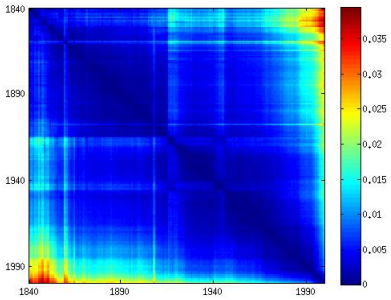

Figure 15: cosine distance with reduced scale on corrected corpus

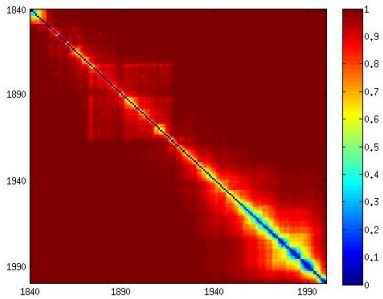

Figure 16: cosine distance weighted by TF-IDF

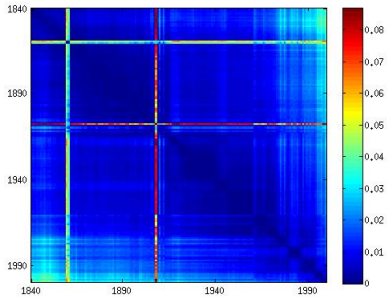

Figure 17: cosine distance on the corpus without correction

5.3.3 Analysis

In the same manner as with the *Chi-Square* distance, this measure is usually used to compare documents and not a huge set of articles containing a lot of common words. But in the rescaled matrix in figure 15, we begin to have a more suited behaviour for our purpose. Indeed, it is smooth in the way that consecutive years have almost the same colour while there is a kind of band around the diagonal and a higher distance in the upper right and top left corners.

With the TF-IDF terms in figure 16, it is like if the values where all brought closer to the diagonal. We can see that 10 consecutive years look similar with each other and that the years after are much farther. It is not as smooth as expected for a good distance but it can be very useful and precise for dating.

In figure 17, we can see the result of the cosine distance applied on the original OCR, without any correction. The red and yellow lines suggest that the OCR correction is useful and well done. It shows also that some particular years are affected a lot by OCR oddities.

5.4 Kullback Leibler Divergence

5.4.1 The formula

The *Kullback-Leibler* (KL) Divergence does not mesure of the distance between subsets of the corpus, but the divergence between them. Indeed, let us consider sets P and Q containing articles for one year each. The KL divergence is defined as follows :

$$D_{KL}(P||Q) = \sum_i P(i) ln \frac{P(i)}{Q(i)} \tag{6}$$

It computes the sum of multiplications between the probability of the word i in corpus P and the natural logarithm of the same probability over the probability of the word i in Q. Thus, P and Q need to be vectors containing the same words at the same indexes. We can also explain the *Kullback-Leibler* Divergence as the loss of information when Q is used to approximate P or, in other words, as the fact that Q tries to simulate P.

Obviously, this divergence is not symmetric. Indeed, let us take for example two years y_1 and y_2 from the corpus. Suppose now that y_1 contains less distinct words than y_2. We can conclude that y_2 will have a better approximation of y_1 than the other way around, because y_2 has a bigger subset of words of y_1, so it can simulate it easily. We can observe this phenomenon in the following plots. The years around 1840 have some difficulties to simulate 1990's years but in the other direction it is easy because there is more vocabulary in the 20th century than in the 19th.

To compute this formula, as we need to have vectors with the same size representing the same words in the same order but some words appear only in certain years, we need to add the missing words in each year with a number of occurrences equal to 0. Doing so, we have for each year a vector containing all words over the whole corpus. With KL Divergence, we have to be careful, because if $P(i)$ or $Q(i)$ is 0, the computation will be totally wrong. Thus, we have to add some smoothing in the KL computation. We made the choice to add a really small value (10^{25}) and to weight it with the probability of the word over the whole corpus. Thus, we keep a small difference between words that appear a lot in the corpus and the ones that are less frequent[1].

5.4.2 Computations on the corpus

We computed the *Kullback-Leibler* Divergence with the non-corrected corpus and the OCR-corrected corpus. We also used the probability that a word appears in a year and the TF-IDF value of each one. The explanations about the plots are located below.

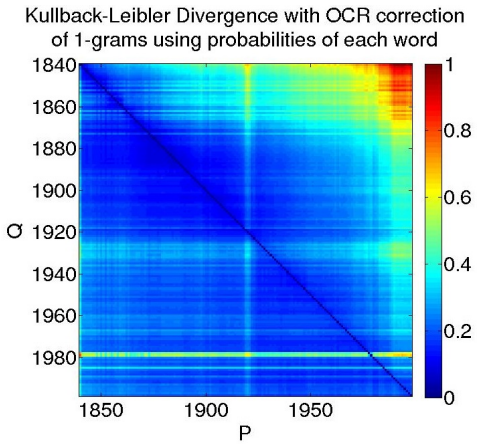

Figure 18: KL for 1-gram with OCR correction and probability of word

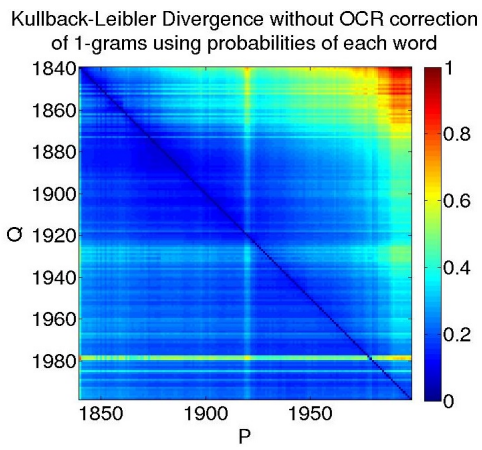

Figure 19: KL for 1-gram without OCR correction and probability of word

[1]It is a way to take less into account for words that appear not that much, like words that does not exist in french and that we were not be able to correct perfectly.

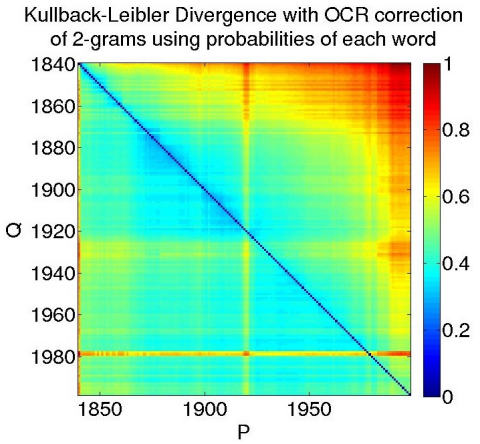

Figure 20: KL for 2-grams with OCR correction and probability of word

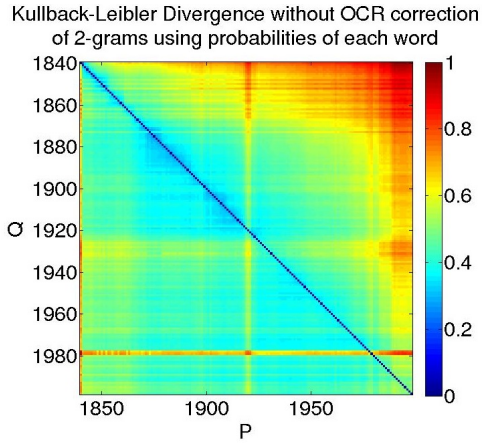

Figure 21: KL for 2-grams without OCR correction and probability of word

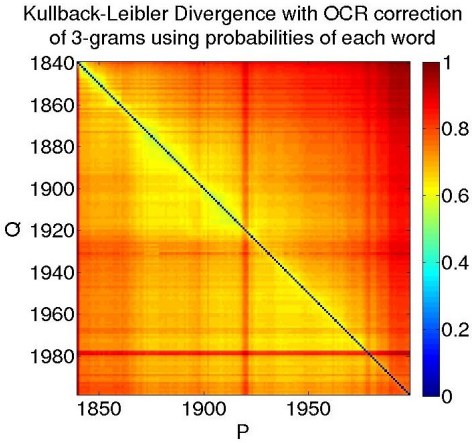

Figure 22: KL for 3-grams with OCR correction and probability of word

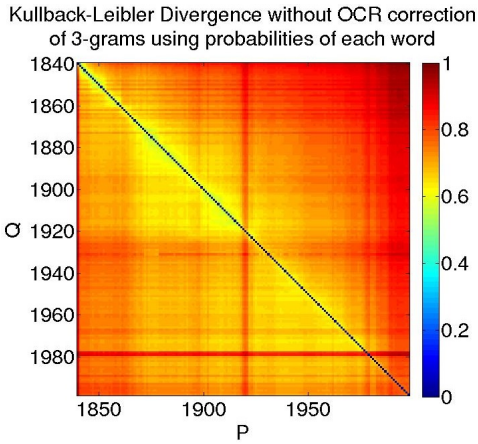

Figure 23: KL for 3-grams without OCR correction and probability of word

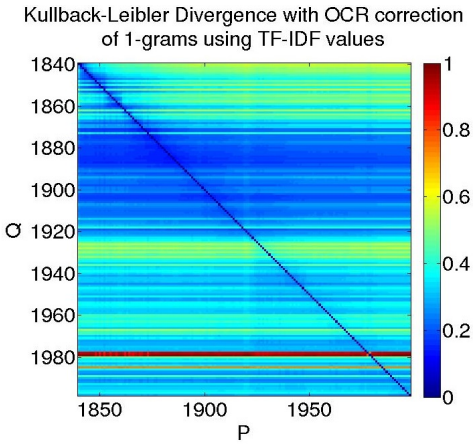

Figure 24: KL for 1-gram with OCR correction and TF-IDF

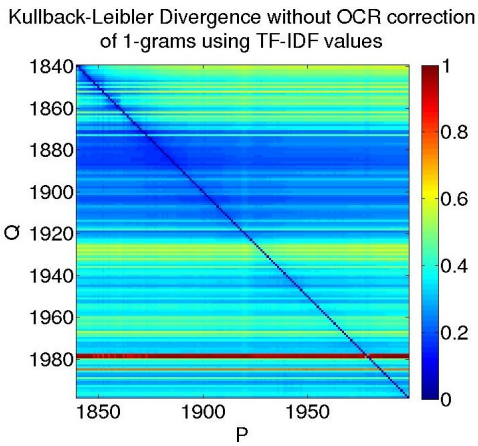

Figure 25: KL for 1-gram without OCR correction and TF-IDF

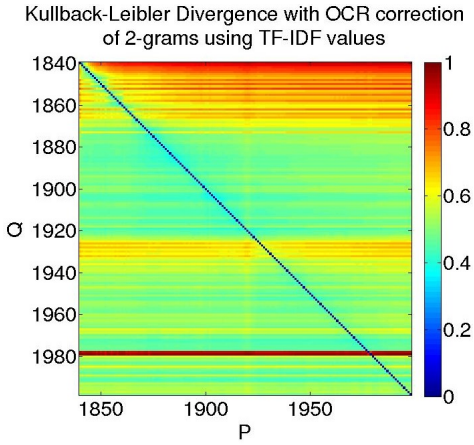

Figure 26: KL for 2-grams with OCR correction and TF-IDF

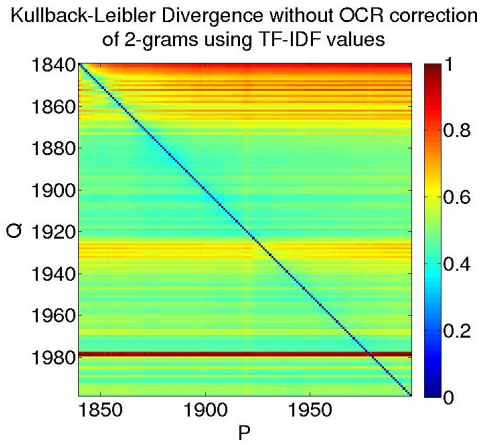

Figure 27: KL for 2-grams without OCR correction and TF-IDF

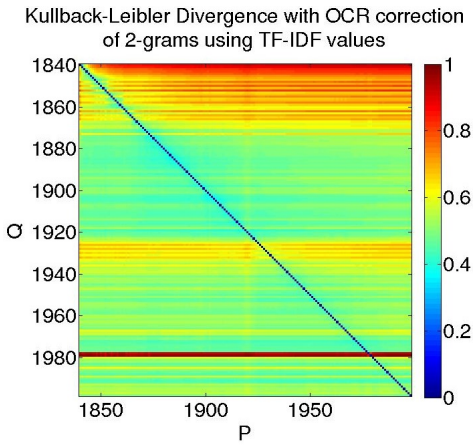

Figure 28: KL for 3-grams with OCR correction and TF-IDF

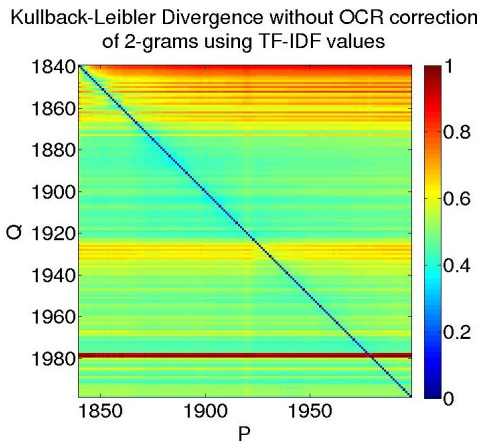

Figure 29: KL for 3-grams without OCR correction and TF-IDF

5.4.3 Analysis of results

As we can observe, there is no real difference between the results of the metric applied to the non-corrected corpus and the results of the metric applied to the corrected one. Thus, we can assume that the OCR correction is not really helping to compute the KL divergence between articles from each year. The interesting fact is that the upper right corner in each plot with the probability of a word (figures 18 to 23) has a higher divergence value than the rest of the plot. It is, as explained in the section of this report about synonyms, due to the fact that words tend to increase in number of words rather than disappear. Indeed, if we have less words in 1840 and much more in 1998, it is obvious that when we approximate 1840 with 1998, the divergence is low and when we approximate 1998 with 1840 it is more difficult (right upper corner).

Augmenting the degree of n-grams, we can observe that, generally, the plots become more red because the divergence is bigger, which is totally understandable (we have n-grams that occur less, so the approximation of a year with another one becomes more difficult). With the 2-grams, the "red part" is larger in the right upper corner and with the 3-grams, all the plot becomes red. In the 3-grams plot, we can also see that there is a

light tight band around the diagonal, which indicates us that the articles from close years share a bigger set of words.

The results of *Kullback-Leibler* divergence with TF-IDF (figures 24 to 29) are really strange and have always the same form (for 1,2 and 3-grams). It is difficult to explain why it has this form, but a possible explanation could be that the TF-IDF gives more importance to the words that appears less in the corpus. Knowing that we try to approximate a year with another one, it is obvious that it becomes difficult to approximate a year if the uncommon words have more weight. It can explain the lines in the plots, because a year could have the same difficulties to approximate each other year.

5.5 Out of Place Measurement

5.5.1 Metric description

This measure determines how far out of place an n-gram in one profile is from its place in the other profile. It allows us to consider the change in the importance of common word of the two corpuses. The rank of the n-gram is its order by its frequency. If a word appears or disappear it has a maximum value rank. Otherwise, we sum to the metric the change of rank of the n-gram. The process of computation can be concluded as follows:

$$D(i,j) = \sum_{x \in D_i \bigcap x \in D_j} |R_x^i - R_x^j| + \max_{x \in D_i \bigcap x \in D_j} (R_x^i - R_x^j)(\sum_{x \in (D_i - D_j)} 1 + \sum_{x \in (D_j - D_i)} 1) \quad (7)$$

5.5.2 Computations on the corpus

In the experiment, the *Out of Place* measurement is divided to two methods (i.e., whether to use the unmatched word's frequency. We also computed the *Out of Place* with the non-corrected corpus and the OCR-corrected corpus to see the impact of OCR correction. The figures illustrated below explain the performance of *Out of Place* under different conditions in the dataset of 1-gram. For the experiment on 2-gram and 3-gram, they have similar performance to 1-gram, and details are omitted.

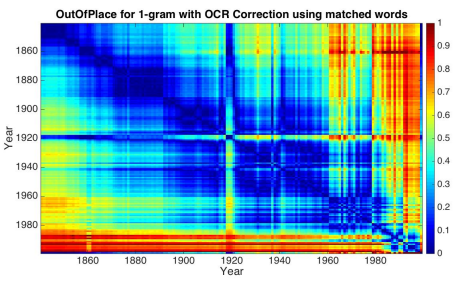

Figure 30: Outofplace for 1-gram with OCR correction using matched words

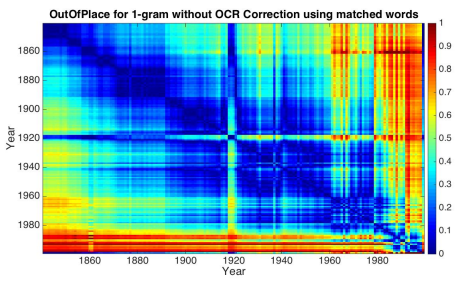

Figure 31: Outofplace for 1-gram without OCR correction using matched words

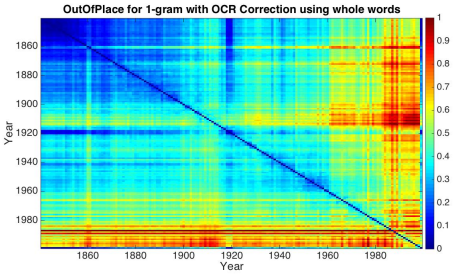

Figure 32: Outofplace for 1-gram with OCR correction using whole words

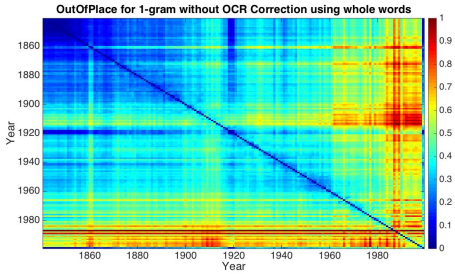

Figure 33: Outofplace for 1-gram without OCR correction using whole words

5.5.3 Analysis of results

We can witness that for this method, the effect of OCR correction is not clear: For the same metric strategy but using different dataset (i.e., OCR corrected dataset and OCR non-corrected dataset), it has few changes in ratio but more in count (see figures 30 and 31; figures 32 and 33). More accurately, the increased count obtained by *Out of Place* measurement does not affect the global picture, since most of data also increase by a similar ratio. But the feature shown by comparison is close to the description of original paper that *Out of Place* provides high robustness in the face of OCR error.

However, the decision that whether we should use those unmatched words greatly impact the final distance between different years. If we only use matched words, then its result seems acceptable (see figures 30) where we can witness that the distance is small when comparable year is close to current year, and the distance normally increased followed by the distance between different years. But there also exists a strange phenomenon around the year 1920 that the word appears in this year is close to the word in previous 40 to 80 years. Checking the performance of other metric in the year 1917 to 1919, we can also find a similar condition (e.g., Kullback Leibler Divergence, Chi-square) because one of the two newspapers is missing for those years.

When it comes to the result that utilized whole words including unmatched word (see 32), we can witness that before year 1960, the metric shows a good performance that closes to the rule that it ought to be. However, after year 1960, the distance between current year and comparable year is extremely high and badly decreases the quality of this metric. We will be using these two approaches in the part of dating article to find out a proper approach to fit ours dataset.

6 Dating Articles

Having some metrics to compare the articles between years, we thought it was a good idea to try to date a subset of articles from a year with these metrics. Thus, we wrote a *Scala* application that selects a subset of articles in a year.
Nevertheless, we had to adapt a bit our codes because it was unnecessary to compare years with years but only articles with years.
To simulate the fact that the selected subset of articles is new and cannot be taken into account in the 'already known' corpus, we had to remove the articles from the year they belong to.
Finally, some titles of articles in the newspaper are considered as a complete article. So we selected a subset of articles and not only a single one to avoid having the impossible

job of dating only 4 or 5 words.
The dating is done as follows: the distance is computed between the selection of articles and each year of the corpus. The prediction is then given by the year for which the distance is the smallest.

You can find below the results for the different metrics with the years 1880, 1920 and 1995 and with taking a subset of 15 articles in each year.

6.1 Metrics

6.1.1 Basic Distance

Here are some example where the distance is applied between the selected subset of articles and each year of the corpus:

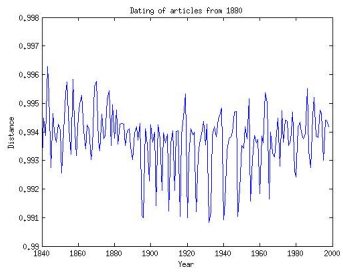

Figure 34: Dating articles from 1880 with the basic distance. Prediction is 1932.

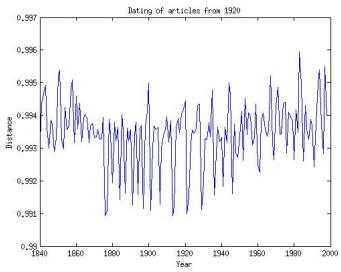

Figure 35: Dating articles from 1920 with the basic distance. Prediction is 1913.

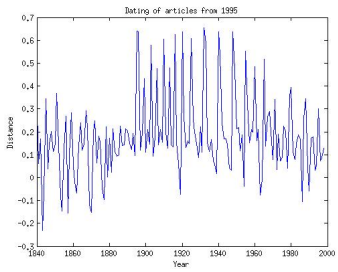

Figure 36: Dating articles from 1995 with the basic distance. Prediction is 1843.

With the basic metric, the distances look very random, and the minimum doesn't give a good prediction. In our opinion, this "sawteeth" behavior can be explained because the amount of words suddenly drop in certain years. Indeed, the basic distance is subject to the number of words in each year and knowing that in 1840's years there are less words than in the others and that between 1917 and 1919 and in 1998 there are less articles too, the predicted year will almost always be one of the previous ones.

6.1.2 Chi Square

In the same fashion as for the basic distance, here are the distances computed between the subset of articles and each year of the corpus:

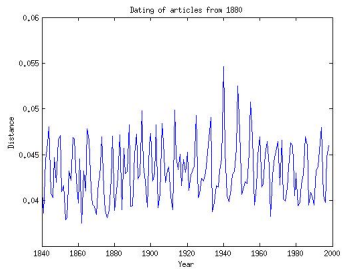

Figure 37: Dating articles from 1880 with the chi-square distance. Prediction is 1890

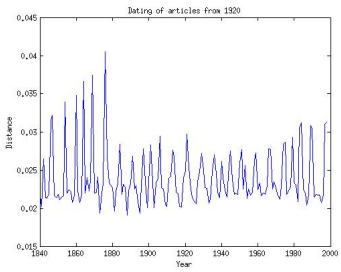

Figure 38: Dating articles from 1920 with the chi-square distance. Prediction is 1904.

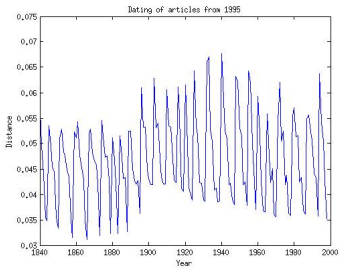

Figure 39: Dating articles from 1995 with the chi-square distance. Prediction is 1858.

With the *Chi-Square* metric, the distances also look relatively random and there is again no evident prediction in these examples.

6.1.3 Cosine

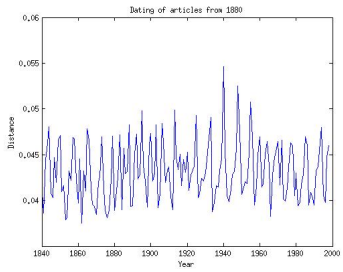

Figure 40: Dating articles from 1880 with the cosine distance. Prediction is 1862.

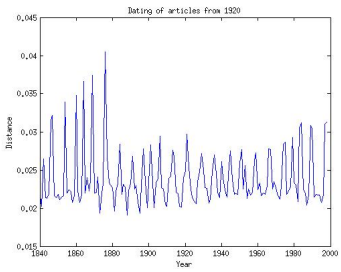

Figure 41: Dating articles from 1920 with the cosine distance. Prediction is 1889.

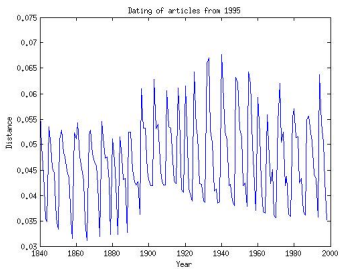

Figure 42: Dating articles from 1995 with the cosine distance. Prediction is 1866

The cosine distance seems to have the same behaviour as the basic and *Chi-Square* distances for dating. Its predictions are also not accurate.

6.1.4 Cosine with TF-IDF

Here are the distances computed by the cosine metric with the TF-IDF terms:

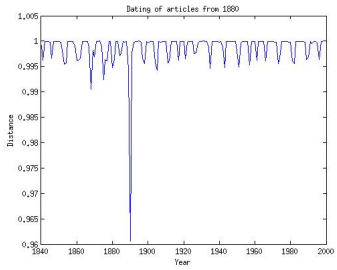

Figure 43: Dating articles from 1880 with the cosine distance with TF-IDF. Prediction is 1890.

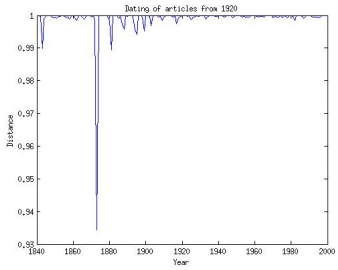

Figure 44: Dating articles from 1920 with the cosine distance with TF-IDF. Prediction is 1873.

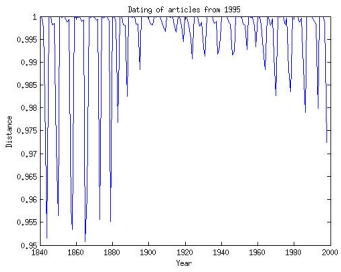

Figure 45: Dating articles from 1995 with the cosine distance with TF-IDF. Prediction is 1865.

The prediction is much more precise than with the previous metrics. As we saw in figure 16, low distances are very close to the diagonal which means that $d(year1, year2)$ is lower than 1 only if $year1$ and $year2$ are close to each other. With this property, the prediction of the cosine metric with TF-IDF terms will often be very close to the real year of the articles.

6.1.5 Kullback-Leibler Divergence

Figures 46 to 51 show the results for the 3 different years trying to represent the articles with each year :

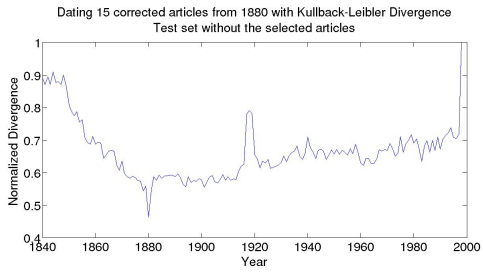

Figure 46: KL for 15 articles with OCR correction for year 1880

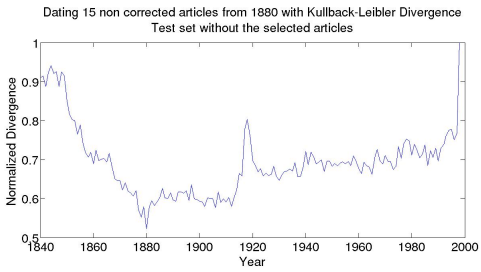

Figure 47: KL for 15 articles without OCR correction for year 1880

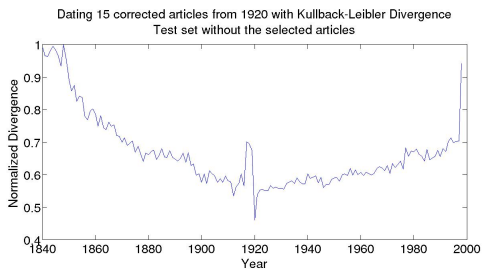

Figure 48: KL for 15 articles with OCR correction for year 1920

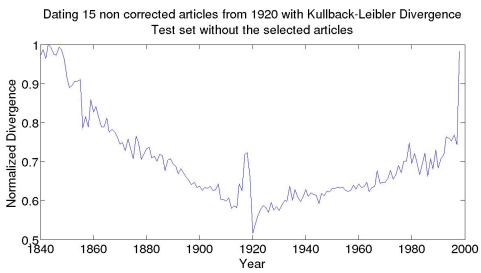

Figure 49: KL for 15 articles without OCR correction for year 1920

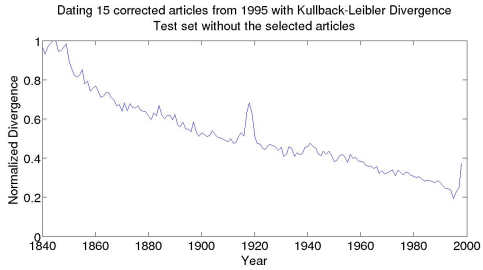

Figure 50: KL for 15 articles with OCR correction for year 1995

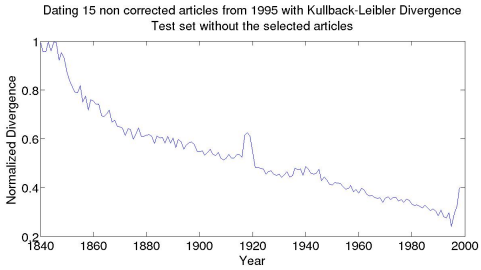

Figure 51: KL for 15 articles without OCR correction for year 1995

We took only the direction where the year approximates the articles because it has no sense to do the inverse. Indeed, to approximate a year that contains thousands of words with a subset of articles that contains a few hundred words will lead to a dating that will very often predict the years with less words (in the 19*th* century or 1998). We should note that for the years 1917 to 1919, there is only one newspaper in the corpus. It explains the peaks in the plots.

We can observe again in these figures that the OCR correction does not help so much for dating. But, we can observe that the *Kullback-Leibler* Divergence is really good to date the articles. Our explanation for these excellent results is that when we remove some articles from a year, the probability that the words are present in this year is higher than in other years. So, even if we remove the articles from the year, it stays hardly attached to the article and when the year approximates the articles, it has a really good matching.

6.1.6 Out of Place

In this part, we using *Out of Place* measurement to date the article by two different kind of approaches (i.e., whether use unmatched word or not). See figure 52, 53, 54, 55, 56 and 57 for more details.

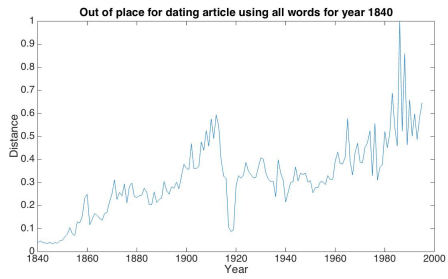

Figure 52: Out of place for dating article using all words for year 1840

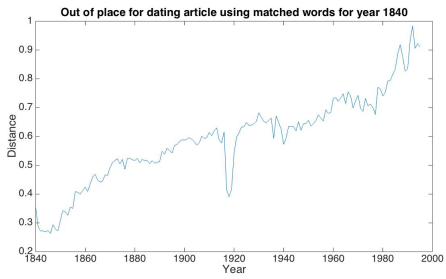

Figure 53: Out of place for dating article using matched words for year 1840

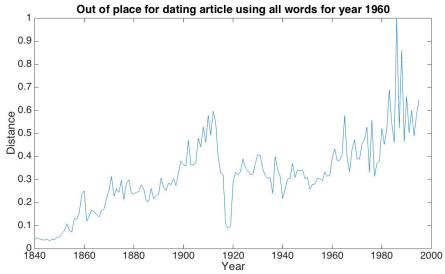

Figure 54: Out of place for dating article using all words for year 1960

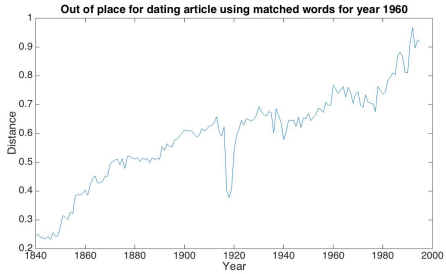

Figure 55: Out of place for dating article using matched words for year 1960

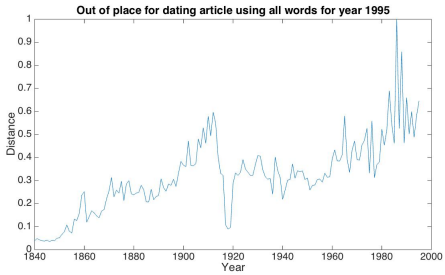

Figure 56: Out of place for dating article using all words for year 1995

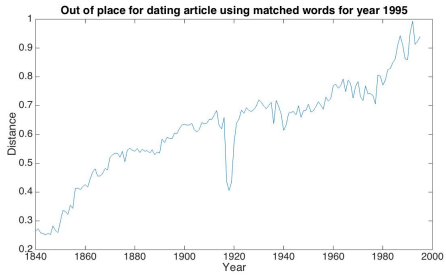

Figure 57: Out of place for dating article using matched words for year 1995

If we only look at the result of year 1840, we can say that it is a good metric for its high-accuracy of prediction. However, when tracking the condition on other years, the result is really odd since all of them give a same plot if we just classify them by eyes. Check its detailed information, there exists a slight difference between different pairs of article and years for the *Out of Place* measurement. However, we cannot deny that it works bad on dating article.

Returning to the formula of *Out of Place* measurement and also considering the size of article, this odd phenomenon then can be explained. For dating article, we only randomly select 15 articles whose size is too small to affect the rank of words that appears in the year data even though we have already removed word's frequency of article from the year it belongs to. As a result, noises are introduced to the final result or useful information are missing.

If we use all words, due to the big gap between the size of article and years, valuable information will be hidden by large volume noises which explains the unchangeable phenomenon on the distance of different article and years. If we only use matched words, for articles that from different years, the matched words are greatly limited by the year data, thus the computation process will always happen on the similar article dataset and year dataset if article data varies but year data remains the same. That is why the *Out of Place* measurement met some changes among part of relationships but still remain a similar trend.

Thus, as explained above, the *Out of Place*, at least for the function of dating article, is not a good metric.

6.1.7 Punctuation and sentences length

As the statistics on punctuation and sentences tend to change between years (especially for the sentences lengths and number of commas), we tried to implement a metric with those statistics.

Those statistics are extracted for each year and combined in a single entry by year. For the metric, we take a sample of articles from one year, we merge them, as it was one big article, we compute the statistics on this article and we use a simple euclidean distance to compute the distance with each year. The result for a sample of 15 random articles from 1925 is observable in figure 58.

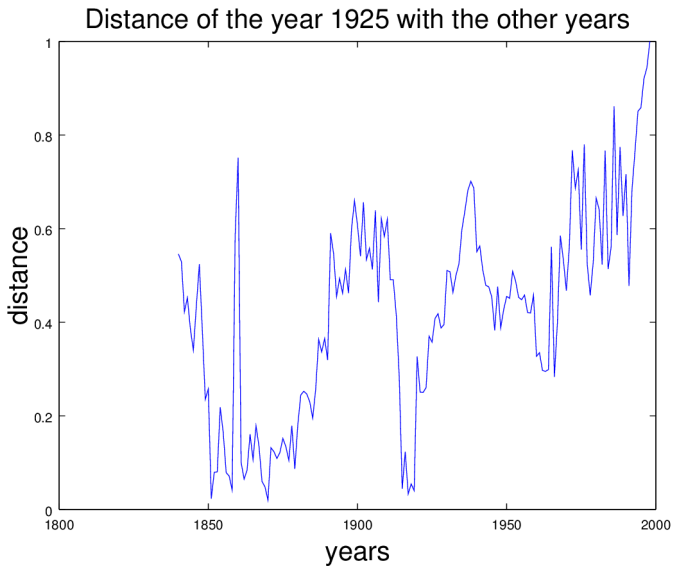

Figure 58: Distance between a sample of 15 random articles from 1925 and each year

As we can see, it does not work that well, but with more time, we could have used machine learning methods instead of a simple euclidean distance, we think it could be useful to combine the result obtained them with other distances to enhance the classification.

6.2 Cross Validation

To have stronger values when we date articles, we wrote a script that runs several times each metrics for a different subset of article of the same year y. The idea here is to have an average on the error of each metric to be able to compare them together as follows:

$$Mean_error = \frac{1}{N}\sum_{i=1}^{N}|p_i - y|$$

where p_i is the prediction at iteration i and N is the number of iterations. We obtained the following results for random years:

Mean error for different metrics for 15 articles in year 1843 with 15 iterations

Distance1 :	0
Cosine :	10.73333333333333333333
Cosine-TFIDF :	.40000000000000000000
Chi-Square :	112.40000000000000000000
Kullback-Leibler :	11.33333333333333333333
OutOfPlace :	3.00000000000000000000
Punctuation :	33.13333333333333333333

Mean error for different metrics for 15 articles in year 1854 with 15 iterations

Distance1 :	8.00000000000000000000
Cosine :	9.86666666666666666666
Cosine-TFIDF :	0
Chi-Square :	111.86666666666666666666
Kullback-Leibler :	0
OutOfPlace :	8.00000000000000000000
Punctuation :	15.80000000000000000000

Mean error for different metrics for 15 articles in year 1888 with 15 iterations

Distance1 :	42.00000000000000000000
Cosine :	15.13333333333333333333
Cosine-TFIDF :	0
Chi-Square :	62.40000000000000000000
Kullback-Leibler :	.26666666666666666666
OutOfPlace :	42.00000000000000000000
Punctuation :	32.80000000000000000000

Mean error for different metrics for 15 articles in year 1905 with 15 iterations

Distance1 :	65.80000000000000000000
Cosine :	17.13333333333333333333
Cosine-TFIDF :	.06666666666666666666
Chi-Square :	48.40000000000000000000
Kullback-Leibler :	3.06666666666666666666
OutOfPlace :	59.00000000000000000000
Punctuation :	21.06666666666666666666

Mean error for different metrics for 15 articles in year 1918 with 15 iterations

Distance1 :	78.93333333333333333333
Cosine :	4.73333333333333333333
Cosine-TFIDF :	0
Chi-Square :	36.33333333333333333333
Kullback-Leibler :	2.66666666666666666666
OutOfPlace :	72.00000000000000000000
Punctuation :	30.86666666666666666666

Mean error for different metrics for 15 articles in year 1934 with 15 iterations

Distance1 :	65.60000000000000000000
Cosine :	17.46666666666666666666
Cosine-TFIDF :	.80000000000000000000
Chi-Square :	128.93333333333333333333
Kullback-Leibler :	.33333333333333333333
OutOfPlace :	88.00000000000000000000
Punctuation :	37.66666666666666666666

Mean error for different metrics for 15 articles in year 1954 with 15 iterations

Distance1 :	44.00000000000000000000
Cosine :	24.60000000000000000000
Cosine-TFIDF :	0
Chi-Square :	0
Kullback-Leibler :	.66666666666666666666
OutOfPlace :	108.00000000000000000000
Punctuation :	48.06666666666666666666

Mean error for different metrics for 15 articles in year 1984 with 15 iterations

Distance1 :	14.00000000000000000000
Cosine :	7.20000000000000000000
Cosine-TFIDF :	1.66666666666666666666
Chi-Square :	3.73333333333333333333
Kullback-Leibler :	0
OutOfPlace :	138.00000000000000000000
Punctuation :	81.80000000000000000000

We observe that two metrics are in general better than the others. These two metrics are the *Kullback-Leibler* Divergence and the Cosine with TF-IDF. Indeed, as explained in section 5, these two metrics are good and we know why. The other metrics are not so bad, even the basic distance which has surprisingly an error of 0 when dating articles from 1843. We think this distance has a good result because the year 1843 is one of the years with the fewer words, so the matching is better than for other years. For the other metrics, we can observe that they vary depending on the year or certainly on the subset of articles.

To judge which metric is the best one, we can simply compute the mean error over all our samples. As it takes a lot of time to compute 15 times each metrics, we can't compute it for each year. Here is the overall mean over all the years we tested (1843, 1854, 1888, 1905, 1918, 1934, 1954 and 1984) :

Mean error for different metrics for 15 articles in the years selected above with 15 iterations per year

Distance1 :	39.7875
Cosine :	13.35875
Cosine-TFIDF :	0.36675
Chi-Square :	63.008625
Kullback-Leibler :	2.292
OutOfPlace :	64.75
Punctuation :	37.65025

In figure 59, we can observe all these results put together in a plot :

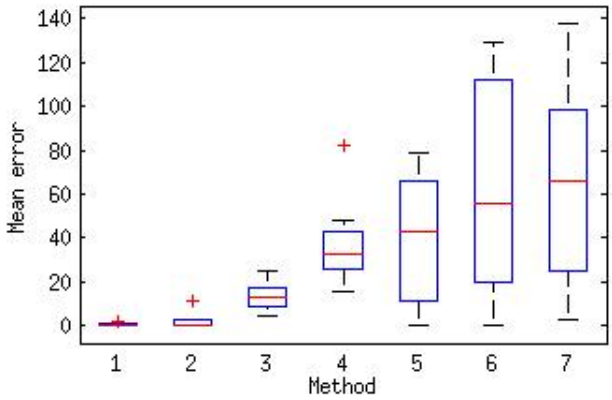

Figure 59: Mean error for 15 articles selection and 15 iterations. 1.Cosine TF-IDF, 2.Kullback-Leibler, 3.Cosine, 4.Punctuation, 5.Basic, 6.Chi-Square, 7.Out-of-place

We can observe that, as expected, Cosine with TF-IDF distance and *Kullback-Leibler* Divergence are the best ones. Not only their error means are the lowest but also their error variance are the tightest. In the remaining metrics, the simple Cosine and our punctuation metrics also have a relatively tight variance and a mean error below 40 years. We can observe that the *Chi-Square* and *Out of Place* metrics can have up to 159 years of error, which is the maximum gap in our corpus

Finally, we made some tests to compare, for a given year, if the number of selected articles to date changes a lot the results of the metrics. We chose to run for a given year (1934) the dating of articles with subsets of articles of sizes 5, 15, 30, 45, 60, 75 and 90. The results are shown in the following table (the smallest distance is colored in green, the biggest in red in table 1).

We observe that in general, more we have articles, better is the dating. First, we have to note that more we take articles, more we remove words in the article's year which can make the number of words in this year consequently smaller than other years. It could be an explanation why taking a lot of articles to date them is not necessarily a good thing (for cosine distance for example). Nevertheless, we observe that the punctuation distance prefers a bigger set of articles, because it can have a better approximation of the sentences length, number of commas and so on for the article's year. It is also the case for all the metrics. Indeed, for *Kullback-Leibler*, Cosine, Cosine with TF-IDF and Out of

# of articles	5	15	30	45	60	75	90
Basic Distance	64	64	64	64	64	64	64
Cosine	35	18.7334	7.7334	14.1334	6.4667	4.7334	10.7334
Cosine-TFIDF	2.5334	0.6667	0.1334	0	0	0	0
Chi-Square	2.1334	0	0	0	0	0	0
Kullback-Leibler	0.0667	0.0667	0	0	0	0	0
Out of Place	89.2	88.2667	88	88	88	88	88
Punctuation	46.6	52.4667	38.0667	29.6667	38.4667	29.0667	32.8667

Table 1: Mean error distance for the year 1934 with different numbers of articles for 15 iterations

Place distance, once the minimum distance is obtained, it seems that it does not change. We can explain this phenomenon by the fact that adding more information for dating is useless once we have reached the minimum possible distance. For the basic distance, the gap of 64 years is probably because the predicted year is 1998. This distance predict 1998 because in this year the number of words is much smaller than in other years (the data cover not all the year because of a merging of newspaper). Due the size of 1998, the basic distance should always have a better connection with it.

7 Topics

We wanted to see how is the evolution of the language inside a specific topic of the corpus (e.g. swiss politics, sport, art,...). Our idea was that a topic clustering would help us capture the linguistic drift more precisely when comparing two articles that talk about the same subject. This new approach would give us different results.
For topic clustering, we chose to use *Latent Dirichlet Allocation* (LDA) implemented in spark since spark 1.3.0.

The algorithms take as input a matrix of article and 1-gram occurrences and cluster the articles into a specified number n of topics. We could also set the parameters for topic distribution over terms β and articles distribution over topics α. Empirical results led us to set the β to 17, a relatively high value because articles contain lot of stop words that could have too much weight for lower values. In the same way, we set α to 1.1, the minimum value, as articles in a newspaper would rather belong to a single topic. For a number of topics equal to n equal to 15.
The LDA algorithm gave us as output the following sample of topic(the words are giving in the decreasing order of weight to describe a topic):

- **Sport** : match, equipe, club, championnat
- **Religion** : oeuvre, histoire, auteur, oeuvres, religion
- **Undefined** : bien, sans, homme, fait
- **Politics** : politique, gouvernement, conseil, parti
- **Banque**, Stock : banque, action, bourse, paris
- **Government** : ministre, etats, unis, gouvernement
- **Swiss Politic** : conseil, president, etat, canton

- **Swiss Market** : dutch, nestle, shell, steel, pacific
- **Theatre and Music** : theatre, dimanche, concert, musique, samedi

We can see that some topics are well defined and we can identify them easily. Others contains words that does not allow us to find a topic by inspection. But in general the result we have is good and can perform well.

7.1 Metrics with Topics

In order to measure the linguistic drift over the topics we constrained ourselves with two metrics: Cosine distance and *Kullback-Leibler* divergence. Simply because these two gave us the better results beforehand.

7.1.1 Cosine Similarity

The topics clustering helps the cosine based distance. It seems that splitting articles into topics makes a better contrast between years. This result can be well interpreted. Indeed, because the years are represented by vectors containing the frequencies of words. The dimension of these vectors are reduced which is a consequence of that a word belongs to a specific topic. Therefore the angle between them is big and due to the language itself over time and not over a topic.

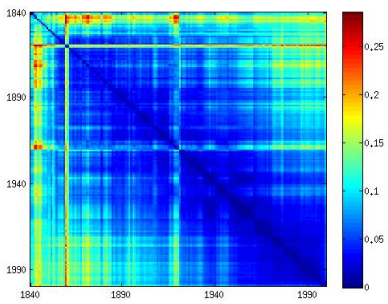

Figure 60: Cosine distance on "Religion" topic

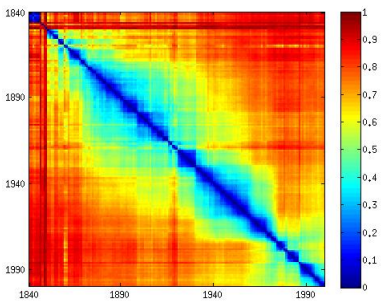

Figure 61: Cosine distance on "Undefined" topic

There is some topics where we cannot distinguish any drift where an example is shown in figure 60. The concerned topic is about history and religion, so it seems that it doesn't change over years and the language drift is hard to enhance for that kind of subject. On the contrary, there is some topics where a drift over years is well marked such as in figure 61. For those cases, the fact that the articles are separated in topics gives an improvement for the cosine metric.

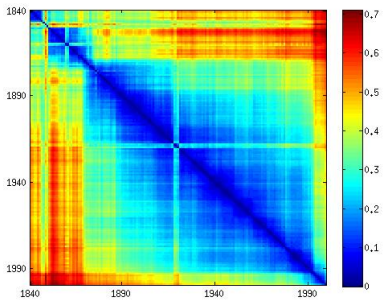

Figure 62: cosine distance on "Swiss politics" topic

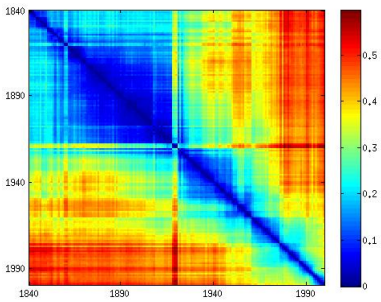

Figure 63: cosine distance on "Theatre and Music" topic

We have also interesting topics such as Swiss politics or music in figures 62 and 63 where a drift is also enhanced compared to the distance on the whole corpus but the changes are not proportional along the time. This can suggest that there is a kind of acceleration or deceleration over time in some specific topics.

7.1.2 Kullback-Leibler Divergence

The computations between years for a given topic give some waited results. Indeed, as you can see in figures 64 and 65, the divergence between years is really small. Especially for the topic 'swiss politic' because the words that are present in this topic ('confédération', 'suisse', 'fédéral', ...) are words that are used every year. Thus, when a year approximate another one, this is almost perfect. There is just an exception for the years 1840 to 1848 that have difficulties to approximate other years. It is easily explicable because the swiss federal state was created in 1848 so the vocabulary for the 'swiss politic' was not exactly the same between 1840 and 1848.

For a more general topic like 'culture', we can observe the same behaviour that have the *Kullback-Leibler* Divergence without using the topics (see section 5). Indeed, the vocabulary for 'culture' should increase over years, so years of the $19th$ century have more difficulties to approximate the years of the $20th$ century and it is not the case in the other direction.

We can also note that use *Kullback-Leibler* divergence to date articles is not a good idea, due to the general plots with the topics.

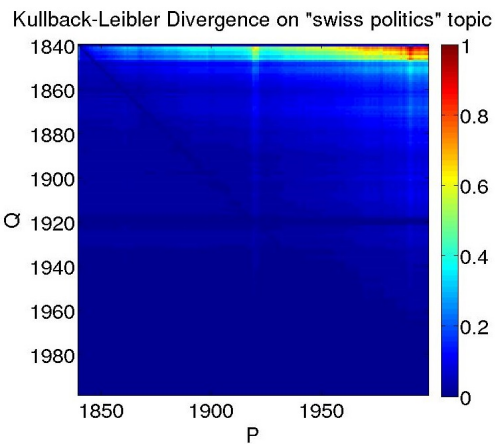

Figure 64: KL for the topic 'swiss politic'

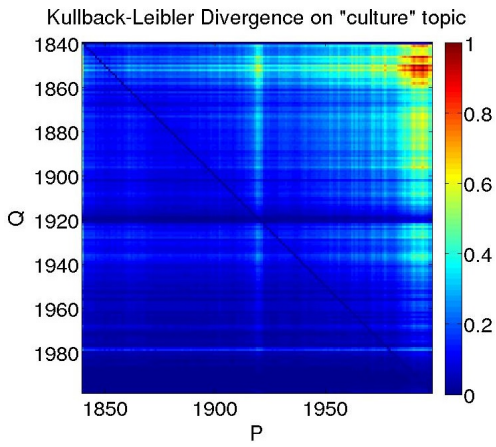

Figure 65: KL for the topic 'culture'

Finally, for all the topics that were extracted from the corpus, the divergence between years are well represented by the 2 examples in figures 64 and 65. For topics that have vocabulary that not really changes over time, the divergence will be of the same kind as for the 'swiss politic' topic and and for the other it will be like the 'culture' topic.

8 Conclusion

To conclude this report, we can say that there exists a linguistic drift in our newspaper's corpus. Indeed, a straightforward statement is that we use more vocabulary to express the same ideas and we do shorter sentences. We also observed that the gain on words through is more important than words disappearances.

About the OCR correction, we saw that for some metrics the correction doesn't have any influence but for other ones (in particular for cosine distance), it helps to improve the observation of the linguistic drift. Then, with the TF-IDF values, it is again the cosine metric that benefits from the importance of uncommon words. For the other metrics, it is less useful.

Then, when we tried to date articles, we obtained some really good results using the *Kullback-Leibler* Divergence and the cosine distance with TF-IDF. Indeed, with those two metrics, we were able to date a subset of articles with less than 10 years error in general what could be considered as an indicator of the goodness of these metrics. With the other metrics, the error is a bit bigger, but they can have sometimes some good results due to some specific reasons (like the number of words in the year).

The synonym approach has revealed interesting points such that words that have similar meaning evolve in a different manner through time and also by their popularity. But we need to use this with precaution because of the several meaning of words (e.g. "caisse").

Finally, with the topics, we found some interesting ones for which we were able to found some properties using the *Kullback-Leibler* Divergence and the cosine distance. Indeed, with the KL divergence, we saw that years in general are not divergent to each other and with cosine we saw that years that are close have a small distance between them.

References

[1] Vincent Buntinx, Human Language Technology: A study of linguistic drift on a corpus of 4 000 000 of digitized newspaper, DHLAB / CDH / EPFL

[2] M. Rajman and L. Lebart. Similarités pour données textuelles. Proc. of 4th International Conference on Statistical Analysis of Textual Data (JADT'98), pp. 545-555, Nice (France), 1998

[3] R. Besançon, M. Rajman and J.-C. Chappelier. Textual Similarities based on a Distributional Approach. Proceedings of the Tenth International Workshop on Database and Expert Systems Applications (DEXA99), pp. 180-184, Firenze (Italy), 1999

[4] R. Besançon and M. Rajman. Evaluation of a Vector Space similarity measure in a multilingual framework. Proceedings of the Third International Conference on Language Resource and Evaluation (LREC'2002), May, 2002

[5] St-Jacques and C. Barrière. Similarity judgments: philosophical, psychological and mathematical investigations. In Proceedings of the Workshop on Linguistic Distances (LD '06). Association for Computational Linguistics, Stroudsburg, PA, USA, 8-15, 2006

[6] J.-B. Michel et al. Quantitative Analysis of Culture Using Millions of Digitized Books. Science 331, 17, 2011. DOI: 10.1126/science.1199644

[7] F. Kaplan. Linguistic Capitalism and Algorithmic Mediation. In Representations, vol. 127, num. 1, p. 57-63, 2014

[8] V. Niculae, M. Zampieri, L. P. Dinu and A. M. Ciobanu. Temporal Text Ranking and Automatic Dating of Texts. Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Gothenburg (Sweden), 2014

[9] Zampieri, Marcos. Using bag-of-words to distinguish similar languages: How efficient are they?. In : Computational Intelligence and Informatics (CINTI), 2013 IEEE 14th International Symposium on. IEEE, 2013. p. 37-41.

[10] C Buckley, G. Salton. Term weigthing Approaches in automatic Text retrieval. Information Processing and management, 24:513-523, 1988

[11] https://databricks.com/blog/2015/03/25/topic-modeling-with-lda-mllib-meets-graphx.html

[12] Hanna M. Wallach David Mimno Andrew McCallum Department of Computer Science University of Massachusetts Amherst Amherst, MA 01003 Rethinking LDA: Why Priors Matter