doi.org/10.26398/IJAS.0029-015

ABOUT FRAGMENTED ANALYSIS OF TEXTS. SOME INFERENTIAL ISSUES IN TEXT MINING (VARIATIONS ON THE "INAUGURAL ADDRESSES CORPUS")

Ludovic Lebart[1]

Télécom-ParisTech, Paris, France

Abstract. *After a brief reminder about the geometrical aspects of data analysis, we contrast the supervised approach (leading to straightforward external validation) and the unsupervised approaches (leading to several methods of internal validation based on re-sampling techniques). In the case of a corpus of texts comprising several parts, a fragmentation of the text provides an unsupervised variant of the analysis of the global lexical table (parts x words). We present then in the unsupervised case some validation procedures allowing for a critical use of the methods and thus providing an assessment of the results. These procedures could be described as variants of bootstrap techniques adapted to the complex nature of textual data. The application example concerns the corpus of Inaugural Addresses of US presidents.*

Keywords : *Statistical inference, Validation, Bootstrap, Textual data analysis.*

1. INTRODUCTION

There are many varieties of statisticians and probably as many varieties of linguists. This gives an idea of the number of all possible combinations of skills involved in textual data analysis. These combinations, crossed in turn with application fields that grow and multiply, give rise to numerous studies. Conceptual enrichment for texts was immediate. The fertility of the paradigm of exploratory analysis of textual data has been a source of enthusiasm and passion, not all extinct today, even if we should rename it *text mining* to communicate more easily.

The passage from the concept of scalar to that of vector, i.e., from words to lexical profiles, was decisive: We can calculate lexical profiles from sentences, paragraphs, chapters, books, responses, articles, speeches, and we can compute distances between these lexical profiles. Several tools allow for representing and classifying these distances. Counts and frequencies of the pioneers of statistical analysis of texts that might seem dull or dry are now complemented with forms, structures, typologies, hierarchies, trees. Data are becoming increasingly extensive and complex, and the produced results alike. Importantly, these results are not

[1] Corresponding author: Ludovic Lebart, email: ludovic@lebart.fr

binary, as at the end of a hypothesis test. A great ambition: these results are supposed to be closer to the mind than the raw data. But there will be a price to pay for these technical developments: investment and training for the researcher, and / or division of labor, never desirable in a process of knowledge acquisition.

1.1 STORIES AND HISTORY OF RELUCTANCES…

The approach of the data analyst (dealing with numerical or textual data) is often misunderstood, for various reasons, sometimes opposed. Let's take a historical example, which concerns the early stages of multiple correspondence analysis (MCA). In 1941, Louis Guttman (physicist by training) came up with a technique devised to build a scale (Guttman, 1941) which is no other than the MCA in all its analytical details and under a quite modern presentation. In this seminal article that went quite unnoticed during the war, he recommends using only the first extracted dimension (the first principal axis) to build a scale. Almost ten years later, the psychologist Cyril Burt rediscovered the method (Burt, 1950), to whom it lends exploratory virtues, advocating to both keep and interpret several dimensions. A controversy ensued concerning the priority of the method and its potential (Guttman, 1953; Burt, 1953). Burt actually concedes the paternity to its predecessor, but finds is very difficult to convince Guttman to look at and interpret several dimensions. Many criticisms have been made for other reasons to Burt, much later, but we must recognize that the psychologist, accustomed to a complex reality and the concept of multidimensional space, was able to perceive things that the mathematician, interested with quantification, visibly badly conceived. This example is typical of a first type of misunderstanding: pure exploration appears either unnecessary or unworthy of a scientific status, or simply incongruous, because many users do not feel that there is a complex space to explore before devising models.

1.2 RICHNESS AND DIFFICULTIES

The analysis of textual data faces the same difficulties and ostracism. It is true that to analyze texts is not necessarily a scientific activity, even with the help of scientific tools. But to visit texts with some powerful visualization tools is on the one hand an enjoyable activity for those who love texts, on the other an indispensable stage for any scientific approach of texts. This is the phase of *systematics* that precedes the making of all sciences, as was the case, for example, of botany or geology. Simply put, we must look at the data and texts before modeling. But this is not easy to put into practice because if we use what we have learned from a data set within the framework of a model, we cannot legitimately test the model on the same data. Such embarrassing situation has been analyzed a long time ago by Cox (1977). We

will show in Section 2 (non-probabilistic aspects of data analysis) that the tools we use are not merely statistical, and that the status of the results still remains to be elucidated. Then we will remind some basic concepts of the theory of learning that can help us work on the texts (Section 3: supervised and unsupervised models). Section 4 (The tests of validity adapted to texts) deals with the validation tools that can be applied to multi-dimensional data, and therefore to texts. Finally, the last section will be devoted to an application that will try to illustrate the previous phases, and help us answer the question: *How to move from contemplation to an exploration, and then to conclusions?*

2. NON-PROBABILISTIC ASPECTS OF DATA ANALYSIS

Can there be a statistic without frequency or repetition? To what extent statistical schemes applied to texts are valid? Realistic? Useful?

Etienne Brunet (1984), in a deep and pleasant article entitled "The violated urn", responds with patience and pedagogy to a mathematician who vehemently questioned the urn scheme used by lexical statistics. Brunet recalls with several arguments and examples that "the urn scheme is an ideal figure, constantly belied by the reality of discourse." We can rephrase his argument by saying that a model can develop more useful tools than the model itself. This was the case of classical factor analysis discovery from a psychological model considered simplistic (Spearman, 1904): The model has been criticized and invalidated over the years, but the method has both survived and been diversified.

The reference to the urn scheme is enriched with the use of data analysis because there is a geometrical component (not probabilistic) in exploratory tools that goes beyond the model of independence (or of conditional independence) at the basis of most statistical models (cf. Le Roux and Rouanet, 2004). We will outline two examples: the analysis of a graph, and an image compression.

2.1 DESCRIPTION OF GRAPHS

Suppose we ask the inhabitants of each of the 32 Irish counties to answer the open question: "What are your neighboring counties?". The first two responses (laconic!) could be:

1) for the county "Galway": *Mayo, Roscommon, Offaly, Clare, Tipperary*
2) for the county "Leitrim": *Sligo, Roscommon, Longford, Fermanagh, Cavan, Donegan*
3) … etc.

Figure 1: Sketch of a map of Ireland blindly reconstructed from a textual description of the 32 counties

Evidently, we are dealing here with special texts. However a correspondence analysis of the lexical table "counties x words", gives us the map in Figure 1, where counties are positioned without any inversion (the edges of the graph represent the contiguity relationship between counties, as defined by R.C. Geary [1954] in his seminal paper about contiguity).

The initial structure is reconstructed from an adjacency relationship. Therefore, some structures can be detected (that of planar, or approximately planar, graph is a rather favorable case) but no statistical tools allow for validating such representation. The recognition of the geographical map depends on an external validation. The use of "supplementary variables" projected *a posteriori* on the principal plane may also play the role of an external validation, as we shall see in the examples of application of Section 5.

2.2 IMAGE COMPRESSION

This second example of a data set without random components is a photograph, i.e. an array that contains 145 lines and 294 columns (98 pixels x 3 colors) representing the former US President Bill Clinton. Each cell contains a number between 0 and 255 (levels of Red, Green and Blue). The example illustrates the compression properties of correspondence analysis.

Figure 2: Color Image : 3 numbers (< 256) per pixel.
Table 294×145 [$294 = 98 \times 3$]

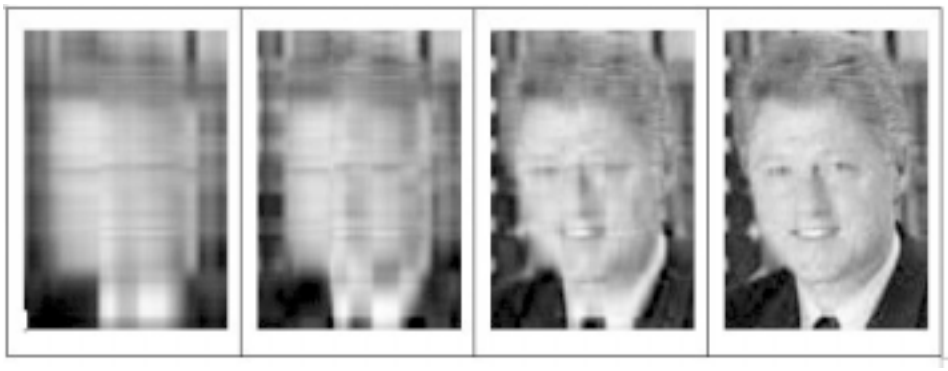

Figure 3: A portrait of former US President Bill Clinton. Correspondence analysis: Images reconstructed successively with 4, 10, 20 and 100 principal axes

With four axes (left of figure 3), a simple spot color is obtained. With 20 axes (third in the left), which corresponds to only 14% of the volume of the original table (and 95 % of the trace – sum of all eigenvalues), the person is already recognizable. Therefore, the tool we use has the power to produce summaries (more numerical than statistical) from data tables that describe non-statistical realities. Most notable in this example is that the compression (like any SVD algorithm) does not depend on the order of rows and columns (this is not the case of the usual compression algorithms involving *Fast Fourier Transforms* currently used in image processing). It could be reassuring for those who are worried to consider texts as bags of words, ignoring the orders of these words: Redundancy is everywhere, in the images as well as in the texts.

2.3 POWER AND LIMITS OF THE TOOL

So we have mathematical tools able to detect some forms or structures and to perform some syntheses. Their purely heuristic interest is undeniable. We are dealing with *instruments of observation* and not with modeling. In this respect, as

in microscopy, for instance, there may be independence between the laws or rules governing the observation tools and those governing the observed reality. Such independence familiar to *data miners* is almost the opposite of the methodology of the pioneers of factor analysis at the beginning of the last century.

3. SUPERVISED AND UNSUPERVISED MODELS

In the practice of statistical learning (cf., e.g., Vapnik, 1998; Hastie et al., 2001), it is customary to distinguish between "unsupervised approach" (meaning approximately "exploratory approach or descriptive approach") and "supervised approach" (closely related to the "confirmatory or explanatory approach"). Mostly, principal axes techniques and clustering techniques are unsupervised, while discriminant analysis (assigning elements to existing classes) and multiple regression are supervised. External validation is a standard procedure in the case of supervised learning models. Once the model parameters are estimated (learning phase), external validation is used to assess the model (generalization phase), usually with cross validation methods.

3.1 EXTERNAL VALIDATION IN THE CONTEXT OF CORRESPONDENCE ANALYSIS (CA)

External validation can be used in the unsupervised case in the context of CA in the following two practical circumstances:

a) When the data set can be divided into two or more parts, one part being used to estimate the model and the other part(s) used to verify the adequacy of the model.

b) When certain metadata (external information) are available to supplement the description of the elements to be analyzed. We will assume that the external information is provided by supplementary elements (additional rows or columns of the data table). In practice, the supplementary elements are projected onto the main viewing planes subsequently. Their positions can be evaluated using conventional statistical tools (e.g. Student *t*) or from Bootstrap validation (see section 4). The technique of additional or supplementary variables can be viewed as a visualized regression. In this sense, it is a supervised technique. It can give an answer to the question: are these additional variables independent of the structure revealed by the active variables?

3.2 ABOUT A CORPUS FRAGMENTATION OPTION

It is possible to create new "artificial observations" in a text corpus. We deliberately use the oxymoron "artificial observations" to highlight the originality of the

approach proposed by Reinert (1983, 1986) on the basis of a procedure known as ALCESTE methodology, and more recently IRAMUTEQ (Ratinaud, 2016).

The text is then considered a "potential supplier of observations." Such text is somewhat arbitrarily divided into units called elementary context units (ECU) having equal or similar lengths (for example 20 consecutive words, or a sentence, or a line, a block of lines). The underlying assumption is that such units deserve to be taken into consideration because they contain valuable information on *local co-occurrences of words* (types or lemmas). Note that the creation of these artificial observations is possible only because the corpus of texts has a sequential or chronological structure. If, for example, we process a set of 50 political discourses through correspondence analysis (CA) of the lexical contingency table (50×1000) cross-tabulating the 50 speeches with the 1000 most frequent words, we are in fact in the case of a supervised approach. We use our knowledge of the partition of the corpus to aggregate the words, and in doing so we limit the calculation of distances between words to their overall frequency in each speech. Otherwise, if we fragment the text into 2000 ECUs, for example, and if we analyze the obtained partition of the corpus in 2000 ECUs, *ignoring the partition into discourses,* we are in the case of an unsupervised analysis. If we afterwards project the 50 centers of gravity (averages) of the ECUs pertaining to each speech (as 50 additional categories), we perform an external validation of the unsupervised analysis. Note that the answers to open questions in a sample survey can be considered as natural ECUs while respondent categories could be used to define (artificial) speeches. In fact, the two approaches complement each other: on the one hand, the analysis of the supervised contingency table (50×1000) [speech $\times$ words], on the other the unsupervised analysis of the table (2000×1000) [ECUs $\times$ words], with its subsequent confrontation with discourse partition. Section 5 will exemplify this methodology.

To summarize the advantages of the fragmentation of the corpus into elementary context units:

– The structure of the text within each speech is taken into account, a piece of information neglected in the traditional approach to the aggregated table.
– A deeper understanding of the internal structure of each text, a finer granularity.
– A convincing external validation can be performed using the partition of the original corpus

However, in the framework of this external validation, the quality of the visualizations remains to be assessed. The resampling tools presented in the next section will be an indispensable complement to the method discussed here.

4. THE TESTS OF VALIDITY ADAPTED TO TEXTS

4.1 THE PARTIAL BOOTSTRAP

The bootstrap technique that will be called partial bootstrap (without recalculation of the eigenvalues for the duplicates samples) proposed in particular by Greenacre (1984) in the context of correspondence analysis, addresses several of the concerns of users. A replication is a resampling with replacements of *n* individuals (vectors observations), followed by the plot of the new *p* variables obtained, these variables having the status of "supplementary variables" in the first *q* axes of the basic analysis. Instead of each variable-point in the visualization plane, we have a cloud of *s* replicates of that point. We obtain, as a byproduct, the variance on each axis which is distinct from what would be replicates of eigenvalues. The *s* replications being projected onto the same system of axes (axes from the original analysis) we can graphically characterize the dispersion of replications of a given variable either by the convex hull of all of its replicates or by an ellipsoid fitted to the cloud of these replicates (computed through a principal component analysis for each of the *p* clouds of replicates). The convex hull has the advantage of completeness (all replications are wrapped up), the ellipsoid has the advantage of describing the density of the cloud of replication, and to be less sensitive to possible outliers. The following applications (Section 5) include examples of these ellipses.

4.2 TOTAL BOOTSTRAP

The total bootstrap consists of carrying out as much principal axes analyses as there are replications. However, the system of axes is no longer the same from one replicated table to another (Milan *et al.*, 1995; Chateau *et al.*, 1996). There may be changes in signs (the principal axes have arbitrary directions), axes permutations, axes rotations. It is therefore necessary to perform a series of transformations to find homologous axes during the successive diagonalizations of the *s* matrices from replicated samples C_k (C_k is the *k-th* replication). The three types of possible transformations, leading to three types of stability tests are:

4.2.1 TOTAL BOOTSTRAP TYPE 1

Total bootstrap type 1 (conservative test, very pessimistic): simple change (if necessary) of the directions of axes for the replications. A simple scalar product between original axes and replicated axes suffices to unify the directions of original and duplicated axes.

4.2.2 TOTAL BOOTSTRAP TYPE 2

Total bootstrap type 2 (fairly conservative): bootstrap type 1 is now complemented with a correction of possible change in the ranks of the axes. Replicated axes are assigned (sequentially) the rank of the original axes with which they are most correlated. Then we proceed to a possible change of sign of the axes, as in type 1 bootstrap.

4.2.3 TOTAL BOOTSTRAP TYPE 3

Total bootstrap type 3 (test rather lax if one is interested in the stability of the axes, but able to describe the stability of dimensions greater than 1 sub-space): a Procrustean rotation (see Gower and Dijksterhuis, 2004) allows for closer coincidence between replicated axes and original axes.

4.2.4 SUMMARY OF USES

Total bootstrap type 1 ignores the possible inversions of axes and rotation of axes. It validates stable and robust structures. Each replication must produce the initial axes with their ranks (order of eigenvalues).

Total bootstrap type 2 is ideal if we want to validate axes, latent dimensions, without giving particular importance to their ranks.

Finally total bootstrap type 3 can globally validate a subspace spanned by the principal axes corresponding to the first eigenvalues. For example, if the subspace of the first four replicated axes coincides with that of the first four initial axes, we can find a rotation in four-dimensional space that will align the axes (which approximately brings us back to the case of partial bootstrap). Like partial bootstrap, total bootstrap type 3 can be considered as lax by users who are interested in the individuality of axes, instead of subspaces generated by consecutive axes (Lebart, 2003, 2007).

4.3 THE SPECIFIC (OR HIERARCHICAL) BOOTSTRAP

The specific bootstrap occurs when there are several levels of statistical units, or levels of hierarchy. In the case of responses to open-ended questions, there is a population of respondents, and a "population" of occurrences (tokens) of words (types). It is usual to deal with the lexical table words × [categories of respondents] (occupation, region, gender, age, etc.). Bootstrap methods described above stipulate drawings with replacement of the words in a contingency table.

But if one wishes to make statistical inferences to the general population from which the sample of respondents is extracted, it is then advisable to proceed to a drawing with replacement of the respondents themselves. In such a case, each

respondent is for us a "bag of words" (Tuzzi et al., 2000).

It is conceivable that the perturbation of the contingency table data is stronger then, especially if these "bags" are of different sizes (some words could appear several times within the same response, etc.). Naturally, this kind of bootstrap can also be partial or total, which does not facilitate the task of the user.

5. APPLICATIONS: VARIATIONS AROUND EIGHT PRESIDENTS

5.1 THE CORPUS OF TEXTS

We will illustrate the previous considerations with the analysis of a medium sized corpus: the *State of the Union* speeches of the last eight American presidents, excerpt from the "Inaugural address" corpus (that can be extracted from the *nltk.book* corpuses: see e.g. Bird et al. 2009) [see also, for example, the website: http://www.usa-presidents.info/union/ that contains all the texts back from the speeches of George Washington in 1789]. This corpus clearly cannot represent all the typical situations that may be encountered in the analysis of texts (long time series, surveys with open questions and closed questions, interviews, document databases).

The whole corpus of the 44 presidents from Washington (1789) to Obama (2009) contains 1,738,048 words (tokens) with 25,246 distinct words (types). The differences in both languages and events lead to a marked and predictable chronological structure. Before focusing on the sub-corpus of eight consecutive presidents (from R. Nixon to B. Obama), we find it useful to present a visualization of the trajectories of the last thirteen presidents during the period (1940– 2009; section 5.2.1).

5.2 PARTIAL OVERVIEWS OF THE CORPUS OF TEXTS

5.2.1 FROM FRANKLIN DELANEY ROOSEVELT (1940) TO BARACK OBAMA (2009)

The whole sub-corpus contains 296,905 words (tokens) with 11,030 distinct words (types). In Figure 4, the pattern of the trajectory of presidents in the CA first plane (axes 1 and 2) is not obviously chronological. However, the two convex hulls that could be drawn around the series (Roosevelt – Johnson) on the one hand, and the series (Nixon – Obama) on the other do not overlap.

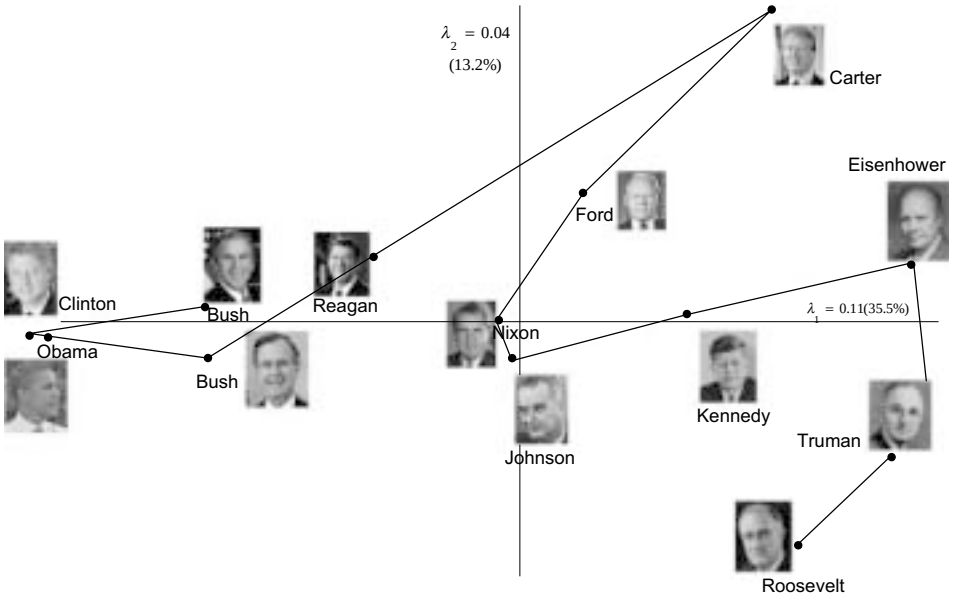

Figure 4: Sketch of the trajectories of the 13 last presidents «1940 - 2009» Plane (spanned by axes 1 and 2 of the Correspondence Analysis of the table cross-tabulating 13 presidents and 836 words appearing at least 50 times (with snapshots of the presidents)

5.2.2 FROM RICHARD NIXON (1969) TO BARACK OBAMA (2009)

This is the sub-corpus which will be fragmented into lines and blocks of various sizes in Section 5.3 up to Section 5.5. As part of this purely illustrative example, the corpus was lemmatized using the software *TreeTagger* (Schmid, 1994), with elimination of function words and prepositions.

After that pre-processing, the corpus of the last eight presidents has a length of 139,899 words and contains 8306 distinct words (in the following, we will talk either of words or lemmas). We actually restrict the text to the 117,099 words (tokens) generated by the 583 words (types) that appear at least 50 times. This corpus contains 12,854 lines of 120 characters, detail that will matter to us because we will successively consider as *elementary context units* each pair of consecutive lines, and then blocks of 20 consecutive lines, before considering the fully agglomerated lexical contingency table (8×583) (presidents $\times$ words) again.

Figure 5 produces a kind of zoom on the upper left part of Figure 4. We note incidentally that chronology is no more a noticeable trend within such a relatively short time span.

We have chosen to complement this display with a small subset of words,

many of them being located beyond the frame of the display (see the arrows on the graphical display). This gives a hint of the richness of the whole map containing the projections of the 583 active words. Such working documents are unfortunately unpublishable in a standard format journal. The (small) confidence ellipses of the points-president will be dealt with in Section 5.5.

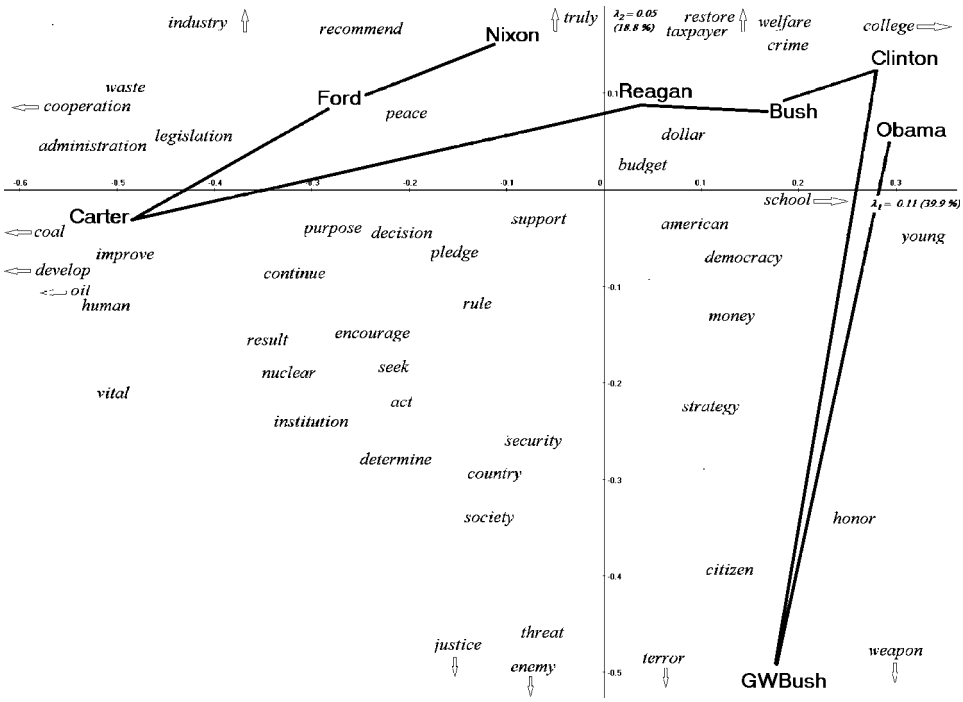

Figure 5: Correspondence Analysis of the lexical table (583 × 8): 583 words appearing at least 50 times cross-tabulated with 8 presidents, with a small sample of 50 words

We now start the fragmentation of the text into context units of increasing sizes.

5.3 ANALYSIS (STILL UNSUPERVISED) OF 6430 PAIRS OF LINES

Figure 6 represents neither the lines nor the words. It merely shows the 8 locations of the presidents (an indicator variable with eight categories considered as supplementary elements). Each pair of consecutive lines is assigned to one of the 8 presidents (i.e.,: we deal with an *a posteriori* projection of the dummy variable "President", using the transition formulas).

Since the partition of the corpus into 8 presidents was not used to build the factorial axes, the projections of the 8 cluster centers (the 8 presidents) are an evidence of the specificity of the context units (ECUs) of some presidents. The bootstrapped 95 % confidence ellipses are built here by drawing with replacement 30 times the 6430 pairs of lines, and thereby allowing to estimate the variability of the locations of presidents-points. The four points (*Nixon*, *Carter*, *GW Bush* and *Reagan*) have a typical location on the axes, however, the points (*Clinton*, *Obama*) are not significantly different in this plane; likewise for the two points (*Reagan*, *Bush_Sr*). We could see that this change goes in the direction of a gradual stabilization of the pattern vis-à-vis the aggregation structure of the lines.

Clinton and *Obama* points – although indistinguishable – occupy a common typical position on the vertical axis. We have observed that they remain indistinguishable at almost all levels of aggregation.

A triangle whose three vertices are *Carter*, *G.W. Bush*, and the pair (*Clinton*, *Obama*) will actually remain stable for blocks of 5 lines (not presented here) and blocks of 20 lines (Section 5.4).

Carter point is particularly isolated whatever the size of the blocks is (farmer before being president, and later Nobel Peace Prize, President Jimmy Carter was considered an outlier, sometimes described as "a UFO" by political commentators). His vocabulary is actually specific: the following words are overused: *administration, development, policy, international and underemployment America, child,* and verbs: *to do, to say, to let, to know*).

The particularity of the analysis applied on blocks of two lines is partially due to very conventional sentences at the start and the end of speeches (typical lemmas: *God, bless, you, America, honor, members, thank, Fellow, congress,* etc.). In various forms, these terms are common to every president except Jimmy Carter (in the present corpus). These salient features which, however, isolated Carter, would gradually dissolve with the increase of the block sizes.

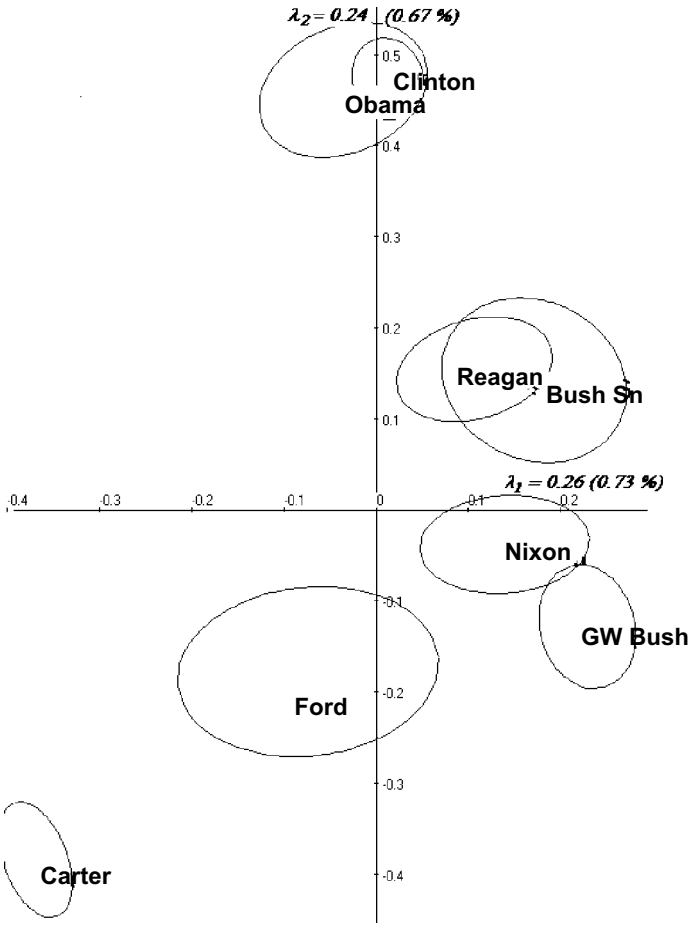

Figure 6: Projection of the supplementary variable "President" on the first factorial plan of the CA of the sparse table 583 x 6430 (6430 pairs of lines considered as context units), with *specific* bootstrap ellipses (lines [*instead of words*] drawn with replacement). As usual with such large sparse tables, the percentages of inertia of the first eigenvalues (out of 582 non-zero eigenvalues) are small, and not to be interpreted in terms of information

5.4 ANALYSIS (UNSUPERVISED) OF 646 BLOCKS OF 20 CONSECUTIVE LINES

As announced, aggregation in blocks of *20* lines reproduces a similar pattern of points in the principal plane. One might be surprised to find ellipses with similar sizes, while the number of blocks decreases significantly. The implemented specific bootstrap consists in drawing with replacement the *646* blocks, while the

ellipses in Figure 5 were obtained from 6430 drawings with replacement. Note that a drawing with replacement induces a perturbation which does not depend much on the size *n* of the sample: The probability that an observation is missing from the drawing tends rapidly to *1 / e (e = 2.71828 ...)*. While the act of removing blocks seems to have a great impact on the results, the structure calculated from these blocks is also better established, and there is a kind of compensation between the severity of the bootstrap and the stability of the structure.

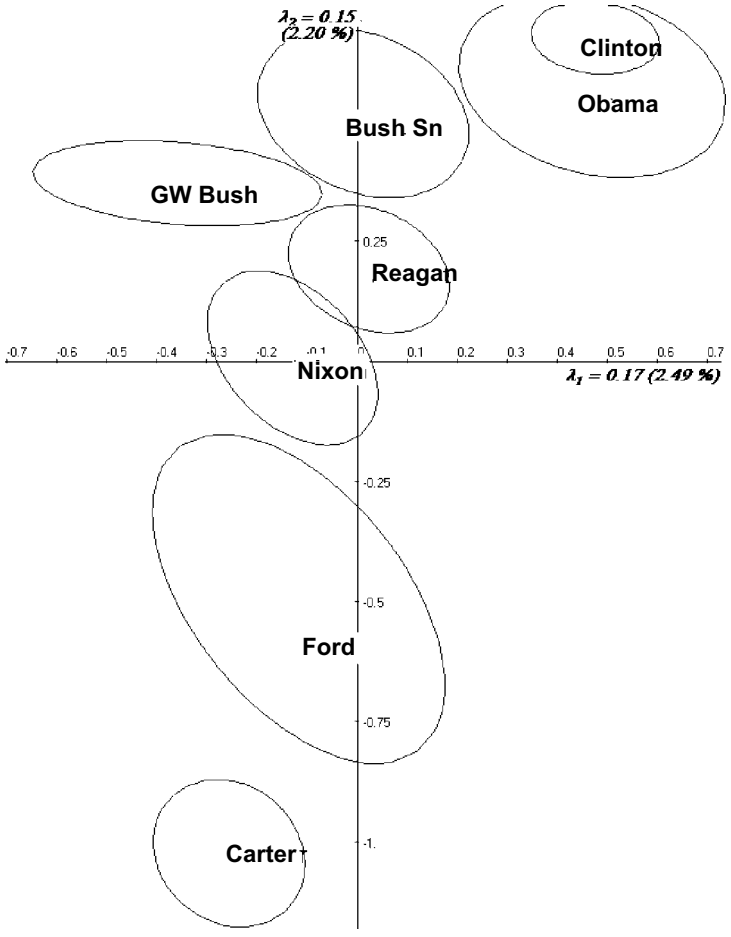

Figure 7: Projection of the additional variable "President" on the first principal plane of the CA of the table 583 × 646 (646 blocks of 20 lines considered as context units), with bootstrap

Figure 7 comprises both a *geometric component* (locations of the president-points) and a *statistical component* (sizes and shapes of the bootstrapped 95 % confidence ellipses). Figure 7 taught us that the differences between presidents are (probably) not due to chance alone (except for Obama and Clinton, whose confidence ellipses largely overlap). Now we could take advantage of the fact that the number of blocks becomes printable to see how these blocks are distributed in the first principal plane and how they overlap the blocks of different presidents. This fragmentation process allows us to observe the dispersion of blocks of 20 lines *within each speech.* It allows us to identify *typical blocks* and incites us to scrutinize them more carefully. Finally, we should not forget to look at the following principal axes (3, 4, …), which can also receive both their confidence ellipses and their scattering diagram for blocks. The third axis allows us, in most of these analyses, to separate the Democrats (Carter, Clinton, Obama) from the Republicans.

5.5 SUPERVISED ANALYSIS OF 8 FULL SPEECHES OF 8 PRESIDENTS

The analysis of lexical contingency table *(583 x 8)* crossing the *583* lemmas and the eight presidents, already sketched in Section 5.2 and Figure 5, constitutes the classical approach. This phase of analysis could be said to be **supervised** because the partition into eight presidents is used to build the principal axes, which was not the case in Sections 5.3 to 5.5, for which the partition was involved *a posteriori* as a supplementary variable characterizing the blocks of lines. Which could be amazing in Figure 9 are the small sizes of confidence ellipses. Note that the total bootstrap type 1 is a bootstrap involving as basic statistical units the words, not the ECUs (being lines or blocks of lines). The *117,099* occurrences of words are drawn with replacement within the contingency table (words x presidents) to create a replication of this table. This kind of bootstrap still shows that Clinton and Obama are now distinguishable dots on this map. The underlying statistical model takes into account the **interdependence** between words and observed presidents (the urn scheme assumes $583 \times 8 = 4664$ different colors for the *117,099* balls in the urn), but on such important numbers of occurrences, it reminds us that most individuals (or scribes/ghost writers) are different.

All these figures should be completed by the underlying spaces of words, as sketched in Figure 5. However, the corresponding graphical displays would not be compatible with the size of the publishable figures in the format of a scientific paper.

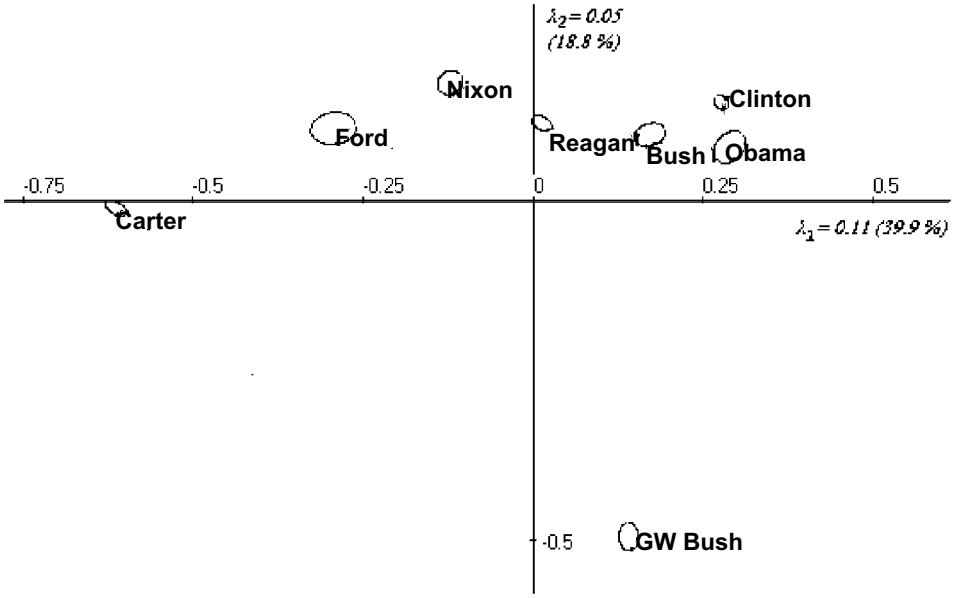

Figure 8: Same principal plane as Figure 5. Confidence ellipses derived from the "Total bootstrap type 1" (the most conservative bootstrap). 8 active variables "President" onto the first principal plane of the CA of the contingency table (583 × 8)

6. CONCLUSION

Data Analysts could say, according to Plato: "Let no one ignorant of geometry enter here." Geometrical representations allowed by visualization tools are indeed indispensable when dealing with the complexity of the relationships between texts, words, words and texts.

The same data analysts should add: "Let no one ignorant of statistics get out of here shouting *Eureka*". We have at hand several tools to explore, discover, and learn, and others, no less important, to conclude, prove, assess. The first tools are perhaps the most attractive one, the latter are sometimes experienced by non-statisticians as a necessary evil. Much remains to be done to define something that looks like a processing strategy. In fact, we have chosen here to preprocess the data instead of devising a new method. The basic idea being that a few versatile and robust techniques mastered by the user (here Correspondence Analysis, but it could be Principal Components Analysis as well in some other contexts), together with a deep knowledge of the data (in collaboration with the scientist) are more productive than a weak grasp of many seemingly more adapted methods.

In this presentation, we simply tried to highlight the contribution of both fragmentation in blocks and re-sampling techniques to pinpoint the intricate links between exploration and inference in textual data analysis. (Note that Data and software (DtmVic) can be freely downloaded from: www.dtmvic.com).

REFERENCES

Bird, S., Klein, E. and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Sebastopol, CA, USA.

Brunet, E. (1984). Le viol de l'urne. In: *La recherche française par ordinateur en langue et littérature*. Slatkine-Champion, Genève-Paris.

Burt, C. (1950). The factorial analysis of qualitative data. In *British Journal of Statistical Psychology*. 3(3): 166-185.

Burt, C. (1953). Scale analysis and factor analysis. Comments on Dr Guttman paper. In *British Journal of Statistical Psycholo*gy. 6: 5-20.

Chateau, F. and Lebart, L. (1996). Assessing sample variability in visualization techniques related to principal component analysis: *bootstrap* and alternative simulation methods. In A. Prats, editor, *COMPSTAT96*, Physica Verlag, Heidelberg: 205-210.

Cox, D.R. (1977). The role of significance tests. In *Scandinavian Journal of Statist*ics. 4: 49-70.

Geary, R.C. (1954). The contiguity ratio and statistical mapping. In *The Incorporated Statistician*. 5(3): 115-145.

Gower, J.C. and Dijksterhuis, G.B. (2004). *Procrustes Problems*, Oxford University Press, Oxford.

Greenacre, M. (1984). *Theory and Applications of Correspondence Analysis*, Academic Press, London.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of a scale construction. In P. Horst, editor, *The prediction of personal adjustment*. SSCR New York: 321 -348.

Guttman, L. (1953). A note on Sir Cyril Burt's factorial analysis of qualitative data. In *British Journal of Statistical Psychology*. 6: 1-4.

Hastie, T., Tibshirani R. and Friedman, J. (2001). *The Elements of Statistical Learning*, Springer, New York.

Le Roux, B. and Rouanet, H. (2004). *Geometric Data Analysis.* Kluwer, Dordrecht.

Lebart, L. (2003). Validation techniques in text mining. In S. Sirmakessis, editor, *Text Mining and its Applications*. Springer: 169-178.

Lebart, L. (2007). Which *bootstrap* for principal axes methods? In P. Brito, P. Bertrand, G. Cucumel and F. De Carvalho, editors, *Selected Contributions in Data Analysis and Classification*. Springer Heidelberg: 581 – 588.

Milan, L. and Whittaker, J. (1995). Application of the parametric bootstrap to models that incorporate a singular value decomposition. In *Applied Statistics*. 44(1): 31-49.

Ratinaud, P. (2016). [http://www.iramuteq.org/Members/pierre.ratinaud].

Reinert, M. (1983). Une méthode de classification descendante hiérarchique: Application à l'analyse lexicale par contexte. In *Cahiers de l'Analyse des Données.* 3 : 187-198.

Reinert, M. (1986). Un logiciel d'analyse lexicale: [ALCESTE]. In *Cahiers de l'Analyse des Données*. 4 : 471–484.

Schmid H. (1994). Probabilistic part of speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Spearman, C. (1904). General intelligence, objectively determined and measured. In *American Journal of Psychology*. 15: 201-293.

Tuzzi, A. and Tweedie, F.J. (2000). The best of both worlds: Comparing Mocar and Mcdisp. In M. Rajman and J-C. Chappelier, editors, *JADT2000* (*Cinquièmes Journées Internationales sur l'Analyse des Données Textuelles*). EPFL, Lausanne : 271-276.

Vapnik, W. (1998). *Statistical Learning Theory*. Wiley, New York.