Beaudouin, V., 2016. Statistical Analysis of Textual Data: Benzécri and the French School of Data Analysis. Glottometrics 33, 56–72.

Statistical analysis of textual data: Benzécri and the French School of data analysis

Valérie Beaudouin
Télécom ParisTech, I3 (UMR 9217)
valerie.beaudouin@telecom-paristech.fr

0. Introduction

While the dream of artificial intelligence (AI), of a machine capable of dialoguing in a natural language, of understanding texts and so of generating them, or even of translating them, has run up against a wall, inductive approaches for the exploration of texts have been developed, with lower theoretical ambitions but greater efficacy. The purpose of such approaches is to identify phenomena and regularities in a corpus of texts and to infer laws from them.

A discourse, or text, being the raw material of numerous human and social sciences, this current has not been restricted to a particular discipline, such as linguistics. These methods have been, and still are, widely used in many different disciplines.

From the 1960s to the 1990s, long before "text mining" became fashionable, France witnessed an exceptionally active period in the field of automated text analysis, exploiting the new affordances provided by IT: digital corpora, statistical algorithms and computing power.

A research field in this territory has grown up, with its laboratories, academic journals, reference books, symposiums, internal controversies, and currents… It brings together researchers coming from different disciplines (literature, linguistics, politics, sociology…). Its multidisciplinary aspect, and the diversity of the objects of research that its methods have been used on, comes from the very ubiquity of human language as a tool. Beyond their different goals and disciplines, the actors of this field are motivated by the common need to mine the text that is the material of their research.

The diffusion of these methods within the social sciences has been associated with the commitment of researchers who have devoted a large part of their activities to developing and diffusing the tools and software that put these methods into practice. The French school of Data Analysis was a major actor in this development, and at its core were Jean-Paul Benzécri and his colleagues; the

influence of these founders is still vivid in the practice of text mining, because the algorithms and software carry their philosophy, as we will show below.

In this article, we have attempted to trace the history of the statistical analysis of textual data, focusing on the influence of Benzécri's work and school, and to make explicit their theoretical positions, clearly opposed to AI and to Chomskyan linguistics. After a presentation of the intellectual project, as an inductive approach to language based on the exploration of corpora, we present the principles of correspondence analysis, which is the main method developed in the Data Analysis School, used for corpus analysis but also for many other types of datasets. Then, we will focus on textual data analysis, a set of methods to analyse a corpus of texts (answers to open-ended questions, set of newspapers articles, corpus of literary works…). Based on the fact that software programmes have played a major role in the use of these statistical techniques, we shall examine a selection of these, display their specificities and their underlying theoretical bases.

In the process, we had to face the question of how to name this field, which has evolved considerably. For purposes of clarity, we shall use as the generic term 'textual data analysis', as used during the emblematic colloquium of this community, the JADT (*Journées Internationales d'Analyse des Données Textuelles – Textual Data Statistical Analysis*), even if the most currently used term today is text mining. This JADT conference was founded in 1990 (in Barcelona), with a scientific committee head by Ludovic Lebart. Since then, this international conference takes place every second year in a different European country.

1 The origins of textual data analysis

From the middle of the 1960's, Jean-Paul Benzécri, his colleagues and students introduced and developed a series of methods, which is commonly designated as "Analyse des Données" (Data Analysis) and that we can consider as the precursor of data mining and "big data". The methods could be applied to all kinds of data, textual data being a particular kind. .

Jean-Paul Benzécri, born in 1932, alumnus of the Ecole Normale Supérieure, obtained his Ph.D. in mathematics (topology) in 1955 at Princeton University under the direction of mathematician Henri Cartan. He started his career at the University of Rennes as an assistant professor in 1960. In 1965, he was promoted as a professor at ISUP, the Statistical Institute of the University of Paris, where he spent the rest of his career (Armatte, 2008). He is a mathematician, mainly interested in linguistics. When he was in Rennes, he introduced a mathematical linguistics course that revealed his turn to linguistics and the beginning of data analysis.

Benzécri is unanimously considered the father of the French School of Data Analysis.

In a nutshell, the principle of correspondence analysis consists in setting the data in rectangular "tables", in the form of matrices, in order to be able to apply data analysis methods to these tables. The tables were initially contingency tables (or cross tables that represent the frequency distribution of two qualitative variables). Correspondence analysis, initially adapted to contingency or cross tables, was extended to other kinds of tables, as disjunctive tables (Multiple Correspondence Analysis) and can be used on all kinds of tables with positive numbers. The idea is to identify the pattern of the relation between two sets of elements put into the table. In the case of a text corpus, the tables contain texts in their rows and words in their columns; at the intersection of a row and a column, there is an indicator of the presence or frequency of the word in the text.

Data analysis algorithms allow the information contained in the matrices to be synthesised. Factor analysis attempts to reorganise the matrices so that the first dimensions contain the maximum amount of information; classification methods allow for the identification of homogenous subgroups of texts and words. The School of Data Analysis often combines factor analysis and classification.

1.1 The origin of data analysis

In *A History and prehistory of data analysis* written in 1975 and published in 1982, Benzécri traces the origins of data analysis, explains correspondence analysis and put it in relation to current related works (Benzécri, 1982). As he explains in his introduction, after a chapter on "chance science" ("science du hasard"), he distinguishes three steps for the improvement of multidimensional statistics (or multivariate data analysis): biometry from Quetelet to Pearson, the works of Sir Ronald Fisher and psychometrics (from Spearman to Guttman). By these means, he draws a personal history of the origins of correspondence analysis (Armatte, 2008) to which he dedicates the last part of the book. Although he underlines the originality and homogeneity introduced by his method, he also presents related works.

The origins of data analysis go back to the beginning of the century. Psychologists were the pioneers in the exploration of multidimensional data and factorial analysis, as analysed by Olivier Martin (Martin, 1997). Spearman, the British psychologist, by analysing the links between students' academic results and their mental aptitudes (Spearman, 1904), believed that he had shown the existence of a general aptitude or intelligence *factor*, which was later given the letter G. Subsequently, not just one, but several factors were sought from increasingly numerous data. Here lie the origins of *factor* analysis.

Correspondence analysis, a branch of factor analysis, started with Fisher, during the 1940s (Fisher, 1940). For Benzécri, by exploring discriminant analysis, Fisher developed the basic equation of correspondence analysis. Then, in 1961, Kendall and Stuart elaborated the canonical methods for the analysis of contingency tables (Kendall and Stuart, 1961). This allowed them to calculate the parameters used to test the hypothesis of independence between rows and columns.

Benzécri explains that he used the name of correspondence analysis for the first time in 1962 and presented the method in 1963 at the College de France (Benzécri, 1982, p. 101). Correspondence analysis is a generic term used as an umbrella.

He was aware of the work by psychometrists and was in contact with Shepard at Bell Labs who had introduced "multidimensional scaling" (Rouannet, 2008). His mathematical linguistics course at the University of Rennes lays the foundation of data analysis as it will be developed by the school.

1.2 The main contribution of Benzécri

Correspondence Analysis is often presented as an adaptation to categorical (or discrete) data of Principal Components Analysis (Greenacre and Blasius, 2006; Hill, 1974; Murtagh, 2005) or very close to muldimensional scaling (Hill, 1974). How can we specify the originality of the Benzécri's contribution to multidimensional analysis?

His main contribution was to show the full algebraic properties of the method and to display its interest: the testing of the independence of rows and columns, but above all the description of how data diverge from this hypothesis, by representing "proximities", the associations that exist between rows and columns, on factorial maps (Diday and Lebart, 1977). The map, a data visualisation of the proximities between individuals and between variables, is the central output for the interpretation. The accent on visualization methods is a key to understanding the success of the Data Analysis School. What was a complex set of data was organized as a "space" for the benefit of the analyst, and suddenly the cloud of data became accessible to interpretation as a whole, with a structure that could be explored, discovered, commented on and displayed. This approach differs from the more classic (and widespread in English literature at the time) approach of testing hypotheses on data sets.

Benzécri was not only interested in algorithms: data analysis constitutes for him a *global framework*, and this is his second main contribution. It first includes data preparation: how to transform any kind of data into a rectangular table with positive numbers that can be analysed. Correspondence analysis can be applied to almost all kinds of tables after suitable data transformation. It also includes a global set of aids to interpretation: the computation of contributions allows for measuring the quality of the representation on the map and the projection of supplementary variables gives to the practitioner complementary elements for interpretation. The association of correspondence analysis with clustering methods (in particular with ascending hierarchical classification) allows a deeper understanding of data, and a simpler interpretation.

Finally, the framework gives a unique method (correspondence analysis and classification) instead of a profusion of algorithms, hard to understand for non-statisticians.

The framework is clearly oriented for users and practitioners by offering a methodological frame, with a particular attention to the display of results.

Benzécri devised and authorised the diffusion of a global framework for analysing "large tables", but he was above all guided by a theoretical and philosophical ambition, which directly interests us here.

1.3 The philosophy of Benzécri

As a mathematician turning towards linguistics, Benzécri became interested in data analysis methods not as psychological tools (a discipline which has been at the origin of a very large number of developments), but instead as a research tool for linguistics: “Correspondence analysis was initially proposed as an inductive method for linguistic data analysis” (Benzécri, 1982, p.102), “It was mainly with a view to studying languages that we became involved in the factorial analysis of correspondences” (Benzécri, 1981, p. X). His theoretical ambition was to open the doors to a new linguistics, in an era that was dominated by generative linguistics. He was opposed to the idealistic thesis of Chomsky who, in the 1960s, considered that only an abstract modelling could reveal linguistic structures. Against this thesis, Benzécri proposed an inductive method of linguistic data analysis "with, on the horizon, an ambitious tiering of successive researches, leaving nothing about form, meaning or style in darkness" (Benzécri, 1981, p. X). In this sense, he was quite close to the objectives of Bloomfield and Harris, who aimed at constructing the laws of grammar from a corpus of statements, with a distributionalist approach. The methods Benzécri developed were from his point of view more efficient for an in-depth understanding of language than the works on statistical linguistics carried out by Guiraud or Muller (Guiraud, 1954; Muller, 1977) which he found interesting but too exclusively focused on vocabulary (Benzécri, 1981, p. 3).

> We propose a method aimed at the fundamental problems that interest linguists. And this method (…) will consist in a quantitative abstraction, in the sense of starting from tables of the most varied data, it will construct, through calculation, quantities that could measure new entities, situated at a higher level of abstraction than that of the facts that were initially collected. (Benzécri, 1981, p. 4)

By identifying factors, there can be doubt that an operation of *abstraction* has indeed been carried out. The computer gives neither any names nor meanings to the entities that it has extracted; it is up to specialists to provide their interpretations.

Benzécri’s philosophical ambition was to reassign value to the inductive approach, and thus to oppose idealism:

> For we condemn the idea that, from principles lightly received, idealism can through a dialectic, even if it is suborned to mathematics, derive certain conclusions; then, to such a priori deductions, we oppose induction which, a posteriori, from the basis of observed facts attempts to rise up to what orders them. (Benzécri, 1968, p. 11)

He criticised idealistic theories that suppose the existence of a model and check its relevance approximately through observation. He doubted that it was possible to reduce a complex object into a combination of elementary objects, "for the order of the composite is worth more than the elementary properties of its components" (Benzécri, 1968, p. 16).

The objective that he thought to be attainable through data analysis was being able to be extract "from the mush of data the pure diamond of true nature". The passage from data to abstract entities, from darkness to light, was made possible in his eyes thanks to data analysis and the "novius organum" of the computer: “The new means of calculation allow us to confront complex descriptions of a large number of individuals, and so place them on flat or spatial maps, in reliable images that are accessible to intuitions from the nebular of initial data” (Benzécri, 1968, p. 21). As an auxiliary for synthesis, the computer is a mental tool: after Aristotole’s *organum* and the *Novum Organum* conceived by Bacon, is not this *Novius Organum* "the newest tool"? (Benzécri, 1968, p. 24).

> After all, it can be seen just how much analysis is free from a priori ideas. From data to results, a computer, insensitive both to expectations and to the researcher’s prejudices, proceeds on the large and solid basis of facts that have previously been defined and accepted as a whole, then counted and ordered according to a programme which, given that it is incapable of understanding, is also incapable of lying. (Benzécri, 1968, p. 24)
>
> Finally, among all the, often contradictory, a priori ideas that each problem inspires in profusion, a fitting choice is made: even more, some ideas which, a posteriori, and after a statistical examination of the data, seem to have been quite natural a priori, would not always have occurred to the mind. (Benzécri, 1968, p. 24)

1.4 Influence

The contribution of Benzécri (a unified frame for data analysis oriented to users) greatly contributed to the diffusion of correspondence analyses in France in all the physical, social, human, and biological sciences: they were, and still are, extremely successful as a display of results. Pierre Bourdieu played an important role in the diffusion of the method as his influence in social sciences increased. Bourdieu’ theory was profoundly inspired by correspondence analysis when he analysed the social space as a field of tensions for example in *Distinction* (Bourdieu, 1984). Rouanet explains that “For Bourdieu, MCA provides a representation of the two complementary faces of social space, namely the space of categories - in Bourdieu’s words, the space of properties - and the space of individuals. Representing the two spaces has become a tradition in Bourdieu’s sociology” (Greenacre and Blasius, 2006, p. 167).

The Data Analysis School has been, and still is, widely present in the field of social sciences, and its approach continues to be used very regularly. Publication of such research, however, runs up against the fact that English-speaking publications favour hypothetic-deductive approaches. The purely exploratory

dimension, aimed at bringing out forms and models from data, does not have the same legitimacy as other approaches; they are too descriptive, instead of being explicative. Yet, it is well known that hypothetic-deductive methods are fragile, because of the order of causality which is pre-established at the moment when a hypothesis is determined. Consequently, the data analysis school had a wider diffusion in France than in other countries.

In Paris, Benzécri put together a large team of data analysis researchers, as can be seen in their numerous collective publications under his direction. The main publications of Benzécri consist of treaties, handbooks and a history.

The treaty on Data Analysis is constituted of two volumes: the first (Benzécri, 1973a) is dedicated to taxonomy and reviews all the classification and clustering methods, the second (Benzécri, 1973b) to correspondence analysis.

A *History and prehistory of data analysis*, redacted by Benzécri in 1975 and published in 1982 (Benzécri, 1982), constitute a state of the art of correspondence analysis and situates the originality of his approach.

For Benzécri this book is an introduction to the series of handbooks *Pratiques de l'analyse des données* published at the beginning of the 1980's: the first volume is dedicated to correspondence analysis (Benzécri, 1980), but in the 1984 edition, an added chapter concerns classification. The second is more theoretical and the third is dedicated to linguistics: *Pratique de l'analyse des données. 3 Linguistique et lexicologie* (Benzécri, 1981).

Each of his volumes involved a large number of contributors, 30 for example for *Linguistique et lexicologie*.

The Journal of data analysis (*Cahiers d'Analyse des Données*) based on an idea of Michel Jambu (Armatte, 2008) stands as the main outlet for articles in the field of data analysis, extended to textual data analysis. This journal was published from 1976 to 1997.

An element that distinguishes Benzécri's work is the organisation of his collective books that all propose: theory, examples of applications from very large fields (natural and human sciences) and programs to be reused in different computers. This structure is an element that explains the important diffusion of methods. The statistical procedures were explicit and shared (an open source approach before its time). At the end of the 1980', several correspondence analysis procedures were included in the leading statistical software packages of the time, notably SPSS, BMDP, and SAS (Greenacre and Blasius, 2006). Nowdays they are implemented in "R", the open source package for statistical computing (Husson et al., 2009).

At ISUP, Benzécri along his co-workers had an important flow of students, estimated at 180 master students per year and 40 Ph.D. (Armatte, 2008) who contributed to the diffusion of methods.

Although cluster analysis is also an important part of Data Analysis School, we will focus on Correspondence Analysis, which can be considered as the core of Benzécri's innovation.

2 Correspondence Analysis

The presentation of correspondence analysis in this section is based on the chapter dedicated to this topic in *Histoire et préhistoire de l'analyse des données* (Benzécri, 1982, p. 101-131), on the introduction in the volume dedicated to linguistics and lexicology (Benzécri, 1981, p. 73-135) and on the *Handbook* (Benzécri, 1992).

Correspondence analysis is a method that gives a geometrical representation of the associations between two sets of elements in correspondence as they appear in a table. It is applied to a specific kind of data: a table of correspondence between the two sets of elements (correspondence or concordance table). Statistical tests are usually used to reject the idea of independence of variables or attributes. The Benzécri's approach is exploratory and descriptive. The main originality of correspondence analysis is to represent, in a geometrical way, the extent to which the independence of observations and attributes is *not verified*. For Benzécri, independence between rows and columns lacks scientific interest; what is interesting is precisely the detail of *how* they interact.

2.1 From a correspondence table to profiles

Correspondence analysis firstly requires one to transform raw data, for example a corpus, into a contingency table, that crosses two sets of elements, a set I (individuals or observations) and a set J (variables or attributes). At the crossing point of a row and a column, we get the number of occurrences of the attribute j in the observation i, k(i,j). Two examples will clarify.

Suppose we are interested in analysing theatre plays. We can build a table, I representing the set of plays, and J the vocabulary that we can find in the plays. In this case, k(i,j) will represent the number of occurrences of the word j in the play i. In the table, there are as many rows as elements in the set I (plays), m, and as many columns as there are in the set J (words), n. Rows are individuals and columns are properties. Let's take another example from (Benzécri, 1982, p. 103). In order to analyse the distribution of nouns and verbs in a corpus, we can build a table where rows are nouns and columns are verbs and at the intersection of a row and a column, we have the number of sentences where the noun is the subject of the verb.

In order to compare the distribution of the two sets of elements, row and column profiles are calculated: f^i_j is *k(i,j)/ki.* (where $ki. = \sum_{j=1}^{n} k(i,j)$, ie the sum of frequencies on the line i). The profile of i will be f^i_J, a vector made of the sequence of f^i_j ($f^i_J = \{f^i_j \mid j \in J\}$)

Symmetrically, the profile of an element j will be $f^j_I = \{f^j_i \mid i \in I\}$.

2.2 Representing the distance between profiles

How do we compare the profiles of different elements (rows or columns of the table)? We need a space and a distance. Correspondence analysis uses a Euclidean space and a distributional distance, or the chi-square distance, which is a distinctive feature of correspondence analysis. The distance between i and i' will be defined as follows:

$$d^2(i, i') = \sum \{(f^i_j - f^{i'}_j)^2 / f_j \mid j \in J\}$$

Each element i (resp j) of set I is represented by its profile and is assigned a mass proportional to the total of the row. The set of the profiles fiJ constitutes a cloud N(I) in a multidimensional space. Respectively, a cloud N(J) is defined for the profiles fjI.

The main idea is to reduce the complexity of the cloud and to find a way to represent most of the information in a lower dimension space. For this, the center of gravity of the cloud is calculated and the dispersion of the cloud around its center of gravity is measured (inertia). Then the factor axes, or principal axes of dispersion, are constructed. Points are projected on those axes, and their coordinates on these axes are called factors. In the plan defined by the first two axes we can have the best projection of the cloud (which minimizes the loss of information).

A distinguishing feature of correspondence analysis is the perfect symmetry of the roles assigned to the two sets I and J in correspondence. This permits the simultaneous representation of the two clouds on the same axes.

The main objective is to visualize the distance between observations or attributes, i.e. the distance from a random distribution. The algorithm produces a set of 'aids to interpretation' that allows the researcher to interpret the results properly.

Often correspondence analysis is combined with hierarchical clustering: the classification is based on the coordinates of the elements on the factor axes.

3 Instruments at the service of the humanities and social sciences

Innovations rarely come from isolated individuals. They emerge and are diffused through networks, collectives and institutions, in which individuals meet and exchange, in which innovations circulate, are discussed, improved and criticised. The diffusion of textual data analysis is no exception to this rule.

Laboratories, journals and lectures have progressively contributed, thus stimulating exchanges and debates. But in this specific field of research, IT tools have become the major players in the diffusion of methods and the organisation of this network. On the one hand, they crystallised the theoretical debates within

the community and, on the other, raised the question of economic, or more modestly commercial, factors linked to these methods.

For the diffusion of these methods has been supported for economic reasons: in the sector of surveys and marketing, the possibility of conducting quantitative research on qualitative data, in other words to introduce measurement into the analysis of discourse, provides an interesting opportunity.

After quickly examining the institutions that have contributed to bring to life this scientific speciality of textual data analysis, we will then focus on a few emblematic textual statistics programmes, while showing how each tool bears the marks of the environment in which it was developed (the discipline, type of corpus and the questions raised by researchers) and how this milieu interacts with the researchers' own objectives.

3.1 Places

After Rennes, ISUP, in Paris, became the centre of elaboration and diffusion of data analysis. Benzécri's seminar at ISUP was attended by most prominent statisticians and researchers in this area. This field was far broader than just textual data analysis as we have seen, but the audience included key figures such as Ludovic Lebart, who also paid particular attention to texts.

Crédoc (*Centre de recherche pour l'observation des conditions de vie*) was for a long time a powerhouse in the field of textual statistics. Ludovic Lebart worked there for many years (1971-1988), setting up and directing the survey *Aspiration et Conditions de vie des Français*. With André Morineau, he was behind the development of Spad (*Système portable d'analyse de données*) (Lebart and Morineau, 1982) and its extension devoted to texts 'Spad.T' (Lebart et al., 1989) which was also based on the work and findings of Eric Brian (Brian, 1986). The Lebart & Morineau's programmes were, up to the year 1987, distributed by a non-profit organization, Cesia in a freeware context and served many researchers or data analysts in the pioneer era of what was to become text mining. Spad had been designed to analyse quantitative surveys, and Spad T for the analysis of answers to open-ended questions. The implementation of the algorithms was guided by the framework of surveys with open-ended questions. A data centre in the basement of Crédoc, shared with the Cepremap, another research centre on economics, and connected to Circé (a regional computing centre in Orsay, *Centre Inter Régional de Calcul Électronique*) provided the possibility to develop and test these tools on data and was the meeting point of a community also involving statisticians such as Jean-Pierre Fénelon (Fénelon, 1981) or Nicole Tabard (pioneer of geographic information systems) (Lebart et al., 1977). A few years later, in the "Prospective de la Consommation" department, Saadi Lahlou developed a research axis based on the applications of lexical analysis in the social sciences (Yvon, 1990; Beaudouin and Lahlou, 1993; Lahlou, 1992;). He contributed to the diffusion of these methods in the field of social psychology.

At Crédoc, Spad was used, but also Alceste, which had been developed by Max Reinert (Reinert, 1990, 1987), and could analyse sets of texts other than open-ended questions. Lexical statistics became a tool for the study of social representations (Lahlou, 1998) and led to a reflexion about the interpretation processes (Lahlou, 1995). Lahlou started a collaboration with M. Reinert to develop tools on the Unix platform and to process greater volumes of text. The large number of *Cahiers de recherche* from Credoc published on these subjects, and the contracts using these methods, bear witness to the dynamism of this centre at the time.

Portability on Mac, Unix and Windows ensured an enduring success of Alceste software in the social sciences in France, and as the software's dictionaries extended to other languages, to further countries.

The laboratory "*Lexicologie et textes politiques*" was set up in 1967 at the Ecole Normale Supérieure in St-Cloud. It has been attached to various different bodies over time, and some of its activities are now located in the Icare laboratory of the ENS in Lyon, while others are at Paris III. The analysis of political discourses stands as the backbone of the unit, with a methodological reflexion branch that explores the place occupied by machines in lexicometry, for the analysis of texts. Pierre Lafon (Lafon, 1984) and André Salem (Salem, 1987) undertook more specifically the setting-up of statistical analysis tools: "these two linguist-mathematicians […] were advised in their methods by the masters of 'data analysis' (Jean-Paul Benzécri) and of probability theory (Georges-Théodule Guilbaud)" (Tournier, 2010). It was in this laboratory that reflexions about corpus linguistics started in France (Habert et al., 1997) and more exactly reflexions regarding annotation systems and the enrichment of texts. André Salem's Lexico programme is one of the tools created in this context. It includes correspondence analysis. It can be distinguished from other software on two points: the identification and processing of repeated segments (sequences of words allowing for the introduction of a notion of syntax) (Salem, 1987) and a detailed processing that measures the chronological evolution in the corpus (Salem, 1995). Correspondence analysis allows to show the distances between sub-parts of a text corpus and to visualise, if relevant, the chronological evolution of texts. An attachment to political and trade-union discourses were specialties of this laboratory.

In the South of France, at the University of Nice, another laboratory was founded in 1980, which accorded a significant role to machines. Etienne Brunet, a literary scholar who had been a computer amateur since the end of the 1960s, set up an active research pole at the university, based in the laboratory *Bases, Corpus, Langage*. Brunet designed a tool, Hyperbase, which was particularly suited to the analysis of very large volumes of literary texts (Brunet, 1988), but also political texts (Mayaffre, 2000), which opened up bridges with the laboratory in St Cloud. The software includes a correspondence factor analysis from the programs developed by J-P Fénelon and his colleagues. It gives a visualisation of distances between words and sub-parts of texts projected on the map.

For example, figure 1 represents the result of the correspondence analysis applied to a table containing in rows the different works of Rabelais (capital letters, PANT for Pantagruel) and in columns the personal pronouns.

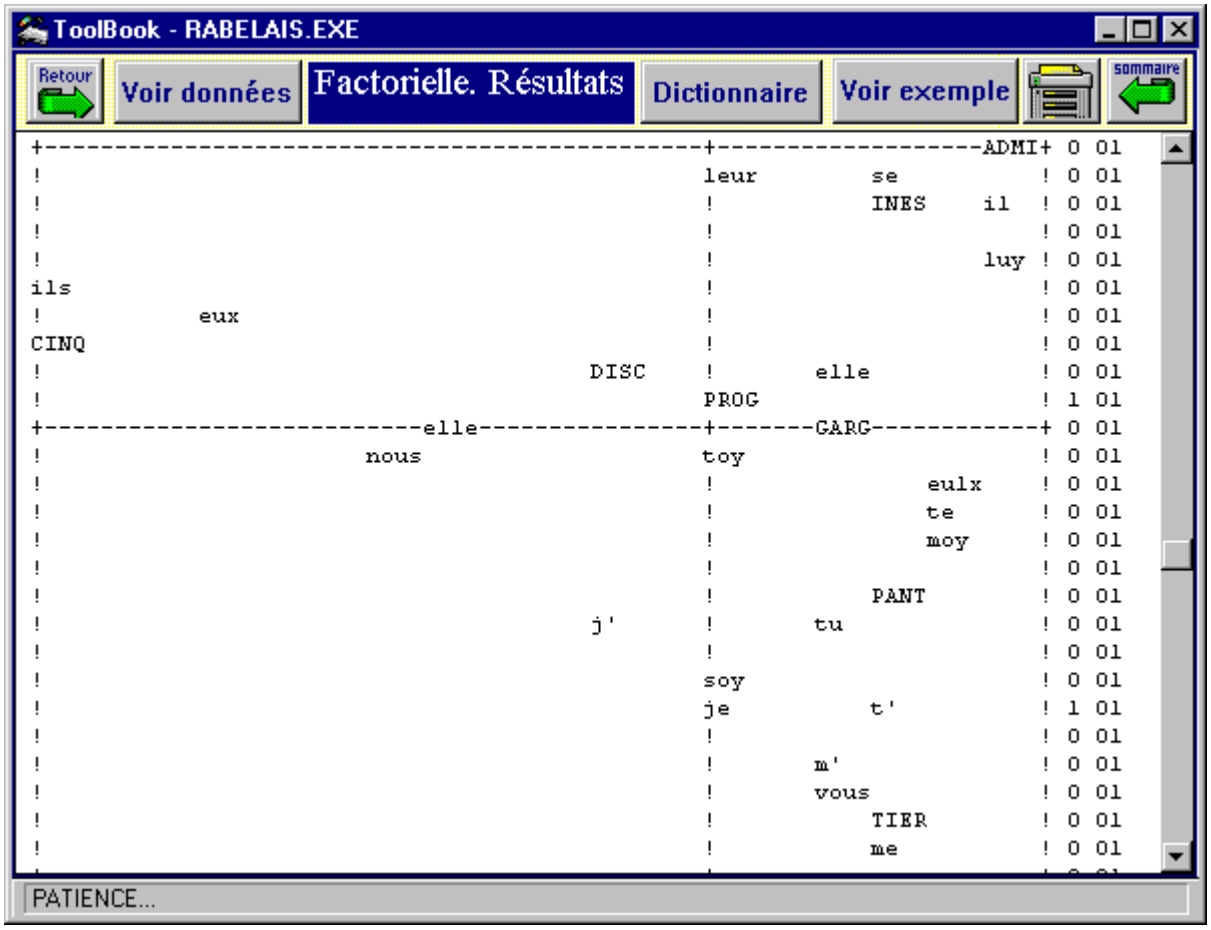

Figure 1. Hyperbase Factorial Analysis

(http://ancilla.unice.fr/~brunet/PUB/hyperwin/analyse.html)

This tool was distributed in the community of humanities researchers. This laboratory explored large corpora from the Frantext database, an exceptional collection of digitized literary works. Since 2001, it has had its own journal, *Corpus*, whose current editor-in-chief is Sylvie Mellet. Two volumes (Brunet, 2009, 2011) collected the main papers published by Etienne Brunet .

Other sites have also played an important role: the IBM scientific centre led by François Marcotorchino, the team headed by Dominique Labbé in Grenoble and other sites abroad, such as Sergio Bolasco's team at the Sapienza in Rome…

The *Journées internationales d'Analyse des Données Textuelles*, which have been organised every second year since 1991, stand as a point for rallying, but also enlarging, the community of researchers in this field. Mostly French-

speaking, it also welcomes Italian and Spanish researchers from the same field. The systematic publication of the papers and the availability online from André Salem and Serge Fleury's journal *Lexicometrica* (http://lexicometrica.univ-paris3.fr/jadt/) thanks to Paris III, constitute a corpus of experiences.
Lebart and Salem's book, *Analyse statistique des données textuelles*, published by Dunod in 1988 (Lebart and Salem, 1988) and republished in 1994 (Lebart and Salem, 1994), then translated into English as *Exploring Textual Data* (Lebart et al., 1998), has become the reference manual in this field .

3.2 Programmes

Publications played a decisive role in the diffusion of methods of textual analysis, explaining the algorithms, displaying possible usages on corpora, and multiplying examples of application. But the diffusion of usages has mainly taken place through the tools themselves, which have been major vectors in the appropriation of methods that are sometimes viewed with mistrust by the world of the social sciences and the humanities. In each case, we shall underline the particularities of the programme: preparation of corpora (selection of texts and variables), processing algorithms and interpretation. We will focus on two software programmes that where the most innovative for text analysis in Benzécri's tradition: Spad T and Alceste.

3.2.1 Spad T

As we have seen, Spad T is an extension of Spad (*Système portable pour l'analyse des données)* which allows for the analysis of answers to open questions in surveys. Spad and Spad T were both designed and coded by Ludovic Lebart and André Morineau at the data centre of Crédoc and Cepremap (see above).

The unit of analysis (each row of the table) is the individual in the survey, characterised by their answers to open and closed questions. But it can also correspond to a group of individuals, according to variables such as age, or level of education, with all the individuals having the same variable value constituting *one* text (a row in the table). For example, figure 2 is the result of the correspondence analysis of a cross tabulation between words (from answers to an open question[1]) and individuals grouped by educational level.

[1] The question was "What are the reasons that might cause a couple or a woman to hesitate having children?" (Lebart et al., 1998).

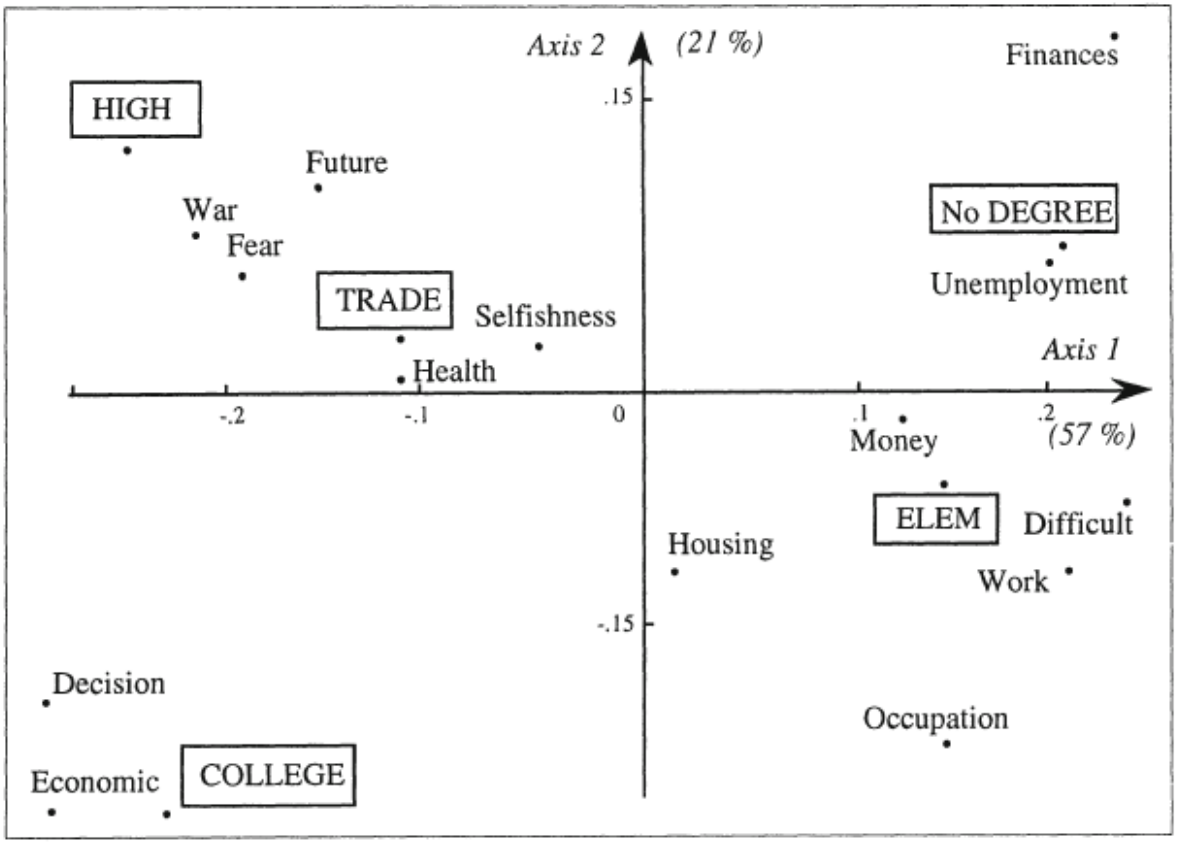

Figure 2. Proximities among words and among educational level
(Lebart et al., 1998, p. 52)

For the words entered in the tables (i.e. making up the columns of the table), Spad T proceeds as follows: it keeps the graphic forms and the words, as they appear in the text, and uses no form of lemmatisation (that is taking graphic forms back to their roots, or dictionary entries); with a frequency threshold, it eliminates rare and very short words (under 3 letters, for example), which is a way to exclude grammatical words (articles, pronouns…). As the answers are reduced throughout the chain leading from the survey to the processing (investigators tend to keep only the main points when noting down answers, the entry clerks often also simplify anyway), and the corpus in question is full of redundancies, this rather brutal "cleansing" has in practice little impact on the results.

Spad T offers a full palette of data analysis procedures. The most classic approach is to carry out a correspondence analysis in a table crossing the answers in the rows with the words used in the columns. Then, based on factor coordinates, an ascending hierarchical classification (clustering) is carried out. The principle consists of bringing together in pairs the answers that are most alike in terms of the vocabulary used, and to advance progressively so as to arrive at a predefined number of classes.

To assist interpretation, it is possible to obtain for each class its specific vocabulary (the words that are significantly more present in this class than in the others), and the most characteristic answers. As Spad T is consistent with Spad, it is possible also to add the values of other variables to the survey, which are over- or under-represented in the class. Spad T includes a most useful "Tamis" (sieve) procedure which systematically tests the interaction of a given modality with every other modality of every other variable in the survey, and orders them by decreasing degree of significance. This enables profiling a class and orienting interpretation and testing without any preconception, in the very explorative spirit of the Data Analysis School.

To sum up, Spad T[2] is particularly well suited to a specific usage context (quantitative surveys) and well-defined types of corpora (answers to open questions). The data analysis and interpretation assistance algorithms are extremely robust, and the usage context means that the simplistic vocabulary reduction creates no problems. The originality of the approach is the possibility to incorporate metadata (*i.e.* information on individuals who produced the text), and then to situate the texts regarding the characteristics of the speaker or writer.

It should be noted here that one of the flaming debates that animated the community was precisely on this issue of lemmatisation; some defended the idea of working on "raw" graphic forms (Lafon, 1984), while others considered that lemmatisation (the reduction of forms to their lemma) was an indispensable prerequisite to any processing, as can be seen in the defence mounted by Muller in his introduction to Lafon's book. The pros considered it was a necessary step to avoid ambiguity of forms (homonymy) while the cons thought it leads to a loss of information: plural/singular, masculine/feminine, person, time being meaningful. This debate provoked heated discussions at almost every JADT conference until the possibility of keeping at the time the raw and the lemmatized form was provided.

3.2.2 Alceste

The methodology of ALCESTE (*Analyse des Lexèmes Cooccurrents dans les Énoncés Simples d'un Texte*) was designed by Max Reinert (1993, 1983); it was inspired by the field of data analysis, Reinert being also a participant of Benzécri's seminar. However, Reinert's preoccupations took a particular orientation. He considered a corpus as a sequence of statements produced by a subject-utterer. Thus, the text is modelled in a table containing statements in rows, bearing the mark of the subject-utterer, and words or lexemes in columns, referring to objects in the world (without any preconceptions about the "reality" of these objects). The objective is then to bring out "lexical worlds".

[2] Ludovic Lebart has made available to the public a software programme, DTM-VIC (http://www.dtmvic.com/), which shares the same properties as Spad for analyzing both numerical and textual data.

> A lexical world is thus at once the trace of a referential site, and the index of a form of coherency linked to the specific activity of the subject-utterer, which we shall call a local logic. (Reinert, 1993, p. 9)

Thanks to statistical procedures, which associate statements using the same type of vocabulary, the method is able to identify different lexical worlds, which could be interpreted as "visions of the world". For example, in his study of *Aurélia* by Nerval, Reinert (Reinert, 1990) identified three types of world by classifying the statements: the imaginary world, the real world and the symbolic world, each of which bears the mark of a certain relationship with the narrator.

Let's describe Alceste in a nutshell. The input is a text or a set of texts, described by some extra textual variables, which describe the communication situation. The output is a typology of the statements that constitute the corpus. A statement is defined as a point of view from a subject about the world. The clustering process is based on the similarity / dissimilarity of words inside the statements. Each cluster of statements is interpreted as a lexical world, which reflects a world view.

This theoretical orientation has consequences on the way analysis is carried out. Let us start with textual units. Reinert attempted to identify the notion of a statement: a point of view about the world that bears the trace of a subject. But how to define automatically the notion of an statement given that it does not necessarily coincide with the notion of a sentence, and no punctuation marks allow it to be identified clearly? As there is no satisfactory solution to this problem, Reinert offered a heuristic: make two possible segmentations of the corpus into textual units while varying the length of the units. Thus, one table would contain in its lines the textual units from the first segmentation, and a second those from the alternative one.

What vocabulary elements are kept in the table's columns? As with Spad T, a frequency threshold allows rare words to be eliminated (this has virtually no impact on the final result since calculation is done on co-occurrences). A lemmatisation process reduces the words to their roots and above all provides an identification of the elements of speech (nouns, verbs, pronouns…). Given the perspective adopted by Reinert, only "full" words, with reference points, are kept for the analysis, and not grammatical words (articles, etc.), which form the text's cement.

On these matrices, which cross textual segments and lemmatised words, Alceste carries out a descending hierarchical classification, using an original algorithm devised in 1983 (Reinert, 1983) which is particularly suited to sparse matrices (with over 90% "0's"). The idea is to take all of the textual segments and to divide them into two groups, in such a way as the groups will be as homogenous as possible in terms of the vocabulary used, while also being as distant as possible from each other. The procedure is then reiterated on the larger remaining group until the requested number of classes has been obtained. This classification process is iterative and leads to a typology. Technically, the descending hierar-

chical classification uses factor analysis. Once the first axis is calculated, a hyperplane is slid along the axis to split the cloud into two sub-clouds until it maximises the inertia between both while minimizing the intra-class inertia. This defines the first two groups, and the process is reiterated (Reinert, 1983).

This is where the heuristic proposed by Reinert comes into play again: on each of the tables that have been made, a descending hierarchical classification is carried out, then the two analyses are compared, so that only the most stable typological classes in both analyses will be conserved. What is more, this provides a procedure which can optimise the number of end classes. For example, the figure 3 shows the result of the double classification on Aurélia (Reinert, 1990). At the end, three classes will be kept : 8 <->9, 10 <->11 and 11<->10.

Figure 3. Descending hierarchical classification with Alceste (Reinert, 1990)

In this process, the new main axis is calculated separately for each successive sub-cloud and the result is amazingly robust, compared to other classification techniques which are based on a single factor analysis.

Each class of the typology is characterised by a list of words that make up the specific vocabulary of the class, in comparison with the entirety of the corpus, using the most characteristic textual segments of the class, and the most representative values of the illustrative variables. The whole can be visualised on a factor analysis plane. These interpretation aids allow for a characterisation of the lexical-semantic field appertaining to each class and give a picture of which external production factors best explain its particularities. (Schonhardt-Bailey et al., 2012) provide what is so far the sole detailed and illustrated description of the Alceste algorithm in English. Alceste has been used for analysing corpora of answers to open questions, literary works, newspaper articles, semi-directed interviews, forum interactions, film reviews, dictionary articles…

4 Conclusion and perspectives

Jean-Paul Benzécri and his colleagues developed a global framework for data analysis (correspondence analysis and clustering methods). Those inductive methods were defined for linguistic purposes, but were widely used in other disciplines, for text analysis but also for quantitative data. The efficiency of those approaches for exploring data and for building hypotheses of research has been widely proven by thousands of publications.

In linguistics, textual data analysis opened the path to a systematic study of language based on corpora, corpus linguistics, with the assumption that field-collected texts, in natural contexts, are the best way to infer sets of rules.

Although the research in statistics and computing sciences has much evolved, in particular with machine learning techniques, it is interesting to note that those "old" techniques are still used by researchers in the social sciences. To do so, the textual data analysis tools have been adapted to larger corpora. While a corpus containing 2,000 answers was considered to be a large one during the 1980s, we now process ones with tens of thousands, or even millions of texts. The textual statistic tools were developed with programming languages which have sometimes since become obsolete, such as Fortran, and were often limited in their size when it came to processing. Updating them to make them appropriate to current volumes sometimes requires codes to be written anew. For example, Max Reinert's Alceste software was entirely reprogrammed by Pierre Ratinaud, and renamed Iramuteq (http://www.iramuteq.org/), with a more modern interface and the capacity to process far larger volumes. Such re-writing can raise problems of intellectual property rights, in that the approaches and the classification algorithms are virtually identical. In the same way, TXM developed for the Textométrie project (http://textometrie.ens-lyon.fr/), reuses and modernises old algorithms, while opening up an enrichment of the lexical data with morpho-syntactic, phonetic or other traits. In such cases, there have been no fundamental changes made to the algorithms of data analysis themselves which is a proof of their efficiency for social scientists.

The methods discussed above are based mainly on the analysis of the distribution of frequencies and co-occurrences of words in texts. The main unit of analysis is the word in its textual context. But, before long, the reduction of a text to a "bag of words" seemed too reductive and the introduction of finer descriptive traits of texts became necessary. Benzécri and his colleagues (Benzécri, 1981) already imagined the introduction of annotations although the technologies were not operational. The methods gradually improved thanks to natural language processing tools, which allowed syntactic, semantic and even prosodic aspects to be taken into account. A text could be associated with a series of descriptive characteristics, concerning different linguistic levels. In this perspective, influenced by (Biber, 1989) who aimed at inductively constructing textual typologies from descriptive traits, the field of corpus linguistics grew up (Habert et al., 1997). Let us take for examples of its application, the TypTex project (Habert et al., 2000), the characterisation of a corpus of texts according to morpho-syntactic traits by (Malrieu and Rastier, 2001; Rastier, 2011) or the attempt to articulate phonetic, morpho-syntactic, rhythmic and semantic characteristics by Beaudouin, (2002). To sum up, approaches that exploited the progress made in the natural language processing no longer limited themselves to words, but now included other levels of linguistic analysis (phonetics, syntax, semantics…). The principles of correspondence analysis and clustering are therefore now applied to much larger tables than they used to be.

The new frontier for textual data analysis is the analysis of web documents. Text was the first medium to enter into the digital world, before images, sounds or videos. It is thus quite natural that the statistical study of texts should

have started long before other contents. In France, the digitization of large sections of literature on the Frantext database combined with mathematical and statistical progress in the area of data analysis have fostered the remarkable rise of the field of textual data analysis. Today, digitalisation has reached the entirety of cultural productions and, as a recent development; more and more production is "born digital". This has opened new research questions. It is no longer possible to reduce the Web to text only, so it will be necessary to enrich the current methods with resources that appertain to the Web's particularities (multimedia, hypertextual, imbricated in reception, dynamic) and develop approaches that combine different methods, textual statistics being just one among others.

5 Bibliography

Armatte, M., (2008). Histoire et Préhistoire de l'Analyse des données par J.P. Benzécri: un cas de généalogie rétrospective. *Journl Electronique d'Histoire des Probabilités et de la Statistique*, vol. 4, p. 1–22.

Beaudouin, V. (2002). Mètre et rythmes du vers classique - Corneille et Racine. Champion, coll. Lettres numériques. Paris.

Beaudouin, V., Lahlou, S. (1993). L'analyse lexicale: outil d'exploration des représentations. CRÉDOC, Cahier de Recherche, n°48, Paris.

Benzécri, J.-P. (1968). La place de l'a priori, "Organum". In: *Encyclopedia Universalis. pp. 11–24.*

Benzécri, J.-P. (1980). *Pratique de l'analyse des données. Analyse des correspondances & classification. Exposé élémentaire*. Paris.

Benzécri, J.-P. (1982). *Histoire et préhistoire de l'analyse des données.* Paris: Dunod.

Benzécri, J.-P. (1992). *Correspondence Analysis Handbook*. New-York, Basel, Hong Kong: Marcel Dekker, Inc.

Benzécri, J.-P. et al. (1973a). *L'analyse des données. 1 La taxinomie*. Paris: Bordas.

Benzécri, J.-P. et al. (1973b). *L'analyse des données. 2 L'analyse des correspondances.* Paris: Bordas.

Benzécri, J.-P. et al. (1981). *Pratique de l'analyse des données, Linguistique et lexicologie*. Paris: Dunod.

Biber, D. (1989). A typology of English texts. *Linguistics*, vol. 27, p. 3–43.

Bourdieu, P. (1984). *Distinction. A Social Critique of the Judgement of Taste*. Harvard University Press.

Brian, E. (1986). *Techniques d'estimation et méthodes factorielles, exposé formel et application aux traitements de données lexicométriques*. Ph.D., Orsay.

Brunet, E. (1988). *Le vocabulaire de Hugo.* Paris : Slatkine-Champion.

Brunet, E. (2009). *Comptes d'auteurs - Tome 1. Etudes statistiques, de Rabelais à Gracq.* Paris : Honoré Champion.

Brunet, E. (2011). *Ce qui compte. Ecrits choisis, tome II. Méthodes statistiques*. Paris: Honoré Champion.

Diday, E., Lebart, L. (1977). L'analyse des données. *La Recherche* p. 15–25.

Fénelon, J.-P. (1981). *Qu'est-ce que l'analyse des données?* Paris: Lefonen.

Fisher, R.A. (1940). The precision of discriminant function. *Annals of Eugenics vol.* 10, p. 422–429.

Greenacre, M., Blasius, J. (2006). *Multiple Correspondence Analysis and Related Methods.* Boca Raton: Chapman & Hall/CRC.

Guiraud, P. (1954). *Les caractères statistiques du vocabulaire.* Paris: PUF.

Habert, B., Illouz, G., Lafon, P., Fleury, S., Folch, H., Heiden, S., Prevost, S. (2000). Profilage de textes: cadre de travail et expérience. In: *JADT'2000. 5èmes Journées Internationales d'Analyse Statistique Des Données Textuelles*, Lausanne, 9-11 Mars 2000.

Habert, B., Nazarenko, A., Salem, A. (1997). *Les linguistiques de corpus*. Paris: Armand Colin/Masson.

Hill, M.O. (1974). Correspondence Analysis: A Neglected Multivariate Method. *Journal of the Royal Statistical Society* vol. 23, p. 340–354.

Husson, F., Lê, S., Pagès, J. (2009). *Analyse des données avec R.* Rennes: Presses Universitaires de Rennes.

Kendall, M.G., Stuart, A. (1961). *The Advanced Theory of Statistics, Volume 2: Inference and Relationship.* Hafner Publishing Company.

Lafon, P. (1984). *Dépouillements et statistiques en lexicométrie.* Genève-Paris : Slatkine-Champion

Lahlou, S. (1992). Sialors: "bien manger"? - Application d'une nouvelle méthode d'analyse des représentations sociales à un corpus constitué des associations libres de 2000 individus. *Cahiers de recherche*. Paris: CRÉDOC.

Lahlou, S. (1995). Vers une théorie de l'interprétation en analyse statistique des données textuelles. In: S. Bolasco, A. Salem (eds), L.L. (Ed.), *JADT 1995. III Giornate Internazionali Di Analisi Statistica Dei Dati Testuali.* CISU, Roma: p. 221–228.

Lahlou, S. (1998). *Penser manger. Alimentations et représentations sociales*. Paris: PUF.

Lebart, L., Morineau, A. (1982). *SPAD: Système Portable pour l'Analyse des Données.*

Lebart, L., Morineau, A., Bécue Bertaut, M. (1989). *Spad.T: Système portable pour l'analyse des données textuelles.*

Lebart, L., Morineau, A., Tabard, N. (1977). *Méthodes et logiciels pour l'analyse des grands tableaux.* Paris: Dunod.

Lebart, L., Salem, A. (1988). *Analyse statistique des données textuelles*. Paris: Dunod.

Lebart, L., Salem, A. (1994). *Statistique textuelle*. Paris: Dunod.

Lebart, L., Salem, A., Berry, L. (1998). *Exploring Textual Data*. Dordrecht, Boston: Kluwer Academic Publisher.

Malrieu, D., Rastier, F. (2001). Genres et variations morphosyntaxiques. *TAL*, vol. 42, n° 2, p. 547-577 .
Martin, O. (1997). Aux origines des idées factorielles. *Histoire & Mesure*, vol. 12(N), p. 197–249.
Mayaffre, D. (2000). *Le poids des mots. Le discours de gauche et de droite dans l'entre-deux guerre.* Paris-Genève: Slatkine-Champion,
Muller, C. (1992). *Principes et méthodes de statistique lexicale*. Paris: Larousse, 1977, réimpression Champion-Slatkine, 1992.
Murtagh, F. (2005). *Correspondence Analysis and Data Coding with Java and R.* Boca Raton: Chapman & Hall/CRC.
Rastier, F. (2011). *La mesure et le grain.* Paris: Honoré Champion.
Reinert, M. (1983). Une méthode de classification descendante hiérarchique: application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données*, vol. VIII, n°2, 187–198.
Reinert, M. (1987). Classification descendante hiérarchique et analyse lexicale par contexte: application au corpus des poésies d'Arthur Rimbaud. *Bulletin de Méthodologie Sociologique,* vol. 13, n°1, p. 53-90.
Reinert, M. (1990). ALCESTE: Une méthodologie d'analyse des données textuelles et une application: Aurélia de Gérard de Nerval. *Bulletin de Méthodologie Sociologique*, n°26, p. 24-54.
Reinert, M. (1993). Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars. *Langage et société*, n°66, p. 5–39.
Salem, A. (1987). *Pratique des segments répétés*. Paris : Klincksieck.
Salem, A. (1995). La lexicométrie chronologique. L'exemple du Père Duchesne d'Hébert. In: *Langages de La Révolution (1770-1815)* (Actes Du 4ème Colloque International de Lexicologie Politique). Paris: Klincksieck.
Schonhardt-Bailey, C., Yager, E., Lahlou, S. (2012). Yes, Ronald Reagan's rhetoric was unique — but statistically, how unique? *Presidential Studies Quarterly*, vol. 42, n°3, p. 482–513.
Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology 15, 201–292*.
Tournier, M. (2010). Mots et politique, avant et autour de 1980. Entretien. Mots. *Les langages du politique 94, 211.*
Yvon, F. (1990). L'analyse lexicale appliquée à des données d'enquête: états des lieux, CRÉDOC, Cahier de Recherche, n°5.