History of Quantitative Linguistics in France

edited by

Jacqueline Léon
Sylvain Loiseau

2016

RAM-Verlag

Studies in Quantitative Linguistics

Editors

1. U. Strauss, F. Fan, G. Altmann, *Problems in quantitative linguistics 1*. 2008, VIII + 134 pp.
2. V. Altmann, G. Altmann, *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen.* 2008, IV+193 pp.
3. I.-I. Popescu, J. Mačutek, G. Altmann, *Aspects of word frequencies*. 2009, IV +198 pp.
4. R. Köhler, G. Altmann, *Problems in quantitative linguistics 2*. 2009, VII + 142 pp.
5. R. Köhler (ed.), *Issues in Quantitative Linguistics*. 2009, VI + 205 pp.
6. A. Tuzzi, I.-I. Popescu, G.Altmann, *Quantitative aspects of Italian texts*. 2010, IV+161 pp.
7. F. Fan, Y. Deng, *Quantitative linguistic computing with Perl.* 2010, VIII + 205 pp.
8. I.-I. Popescu et al., *Vectors and codes of text*. 2010, III + 162 pp.
9. F. Fan, *Data processing and management for quantitative linguistics with Foxpro***.** 2010, V + 233 pp.
10. I.-I. Popescu, R. Čech, G. Altmann, *The lambda-structure of texts*. 2011, II + 181 pp
11. E. Kelih et al. (eds.), *Issues in Quantitative Linguistics Vol. 2*. 2011, IV + 188 pp.
12. R. Čech, G. Altmann, *Problems in quantitative linguistics 3*. 2011, VI + 168 pp.
13. R. Köhler, G. Altmann (eds.), *Issues in Quantitative Linguistics Vol 3.* 2013, IV + 403 pp.
14. R. Köhler, G. Altmann, *Problems in Quantitative Linguistics Vol. 4*. 2014, VIII+148 pp.
15. Best, K.-H., Kelih, E. (eds.), *Entlehnungen und Fremdwörter: Quantitative Aspekte*. 2014. VI + 163 pp.
16. I.-I. Popescu, K.-H. Best, G. Altmann, G. *Unified Modeling of Length in Language*. 2014, VIII + 123 pp.

17. G. Altmann, R. Čech, J. Mčutek, L. Uhlířová (eds.), *Empirical Approaches to Text and Language Analysis dedicated to Luděk Hřebíček on the occasion of his 80th birthday*. 2014. VI + 231 pp.
18. M. Kubát, V. Matlach, R. Čech, *QUITA Quantitative Index Text Analyzer*. 2014, VII + 106 pp.
19. K.-H. Best, *Studien zur Geschichte der Quantitativen Linguistik.* 2015. III + 158 pp.
20. P. Zörnig, K. Stachowski, I.-I. Popescu, T. Mosavi Miangah, P. Mohanty, E. Kelih, R. Chen, G. Altmann, *Descriptiveness, Activity and Nominality in Formalized Text Sequences.* 2015. IV+120 pp.
21. G. Altmann, *Problems in Quantitative Linguistics Vol. 5*. 2015. III+146 pp.
22. P. Zörnig, K. Stachowski, I.-I. Popescu, T. Mosavi Miangah, R. Chen, G. Altmann, *Positional Occurrences in Texts: Weighted Consensus Strings.* 2015. II+178 pp.
23. E. Kelih, R. Knight, J. Mačutek, A.Wilson (eds.), *Issues in Quantitative Linguistic Vol. 4*. 2016. III + 231 pp.

ISBN: 978-3-942303-48-4

RAM-Verlag
Stüttinghauser Ringstr. 44
D-58515 Lüdenscheid
Germany
RAM-Verlag@t-online.de
http://ram-verlag.eu

Contents

I. Vocabulary Statistics

II. Mathematical Models

Introduction

Jacqueline Léon,
UMR 7597, Histoire des Théories Linguistiques, CNRS
Université Paris Diderot
jacqueline.leon@univ-paris-diderot.fr

Sylvain Loiseau,
Sedyl Laboratory (UMR 8202 CNRS/Inalco),
sylvain.loiseau@univ-paris13.fr

Scope of this volume

This volume give a historical account of the field of quantitative linguistics in France. It focuses about developments initiated in France implying mathematical methods or the usage and interpretation of quantitative data. It does not include material about corpora compilation, computational implementation, or formal modelization.

Quantitative linguistics in France

1. Early quantitative linguistics

Historically, quantitative linguistics has a specific status in France. It could be said that statistical studies, namely statistical studies of vocabulary, opened the way to the reception of formal languages and the computerization of linguistics in France. On the one hand, French statistical works were deeply anchored in French linguistic tradition mainly concerned by philology, etymology, dialectology, stylistics and studies on specialized vocabularies. On the other hand, contrary to the USA, the USSR and Great Britain, France significantly lagged behind for computing, logic and formal languages. In fact, the field of vocabulary statistics played a crucial role by the rivalry it had introduced between both approaches, formal and quantitatives. This is shown by the multitude of denominations naming the field and subfields the Americans have referred to as ‘computational linguistics’.[1]

The Centre Favard, also named the ‘Seminar of Quantitative Linguistics’, created in March 1960 at the Henri Poincaré Institute of the Faculté des Sciences de Paris, significantly exemplifies that issue. Under the name ‘Seminar of Quantitative Linguistics’ were brought together both the formal aspects of linguistics and statistical methods. It was an important place for training the linguists in

[1] See Cori & Léon (2002).

mathematics, logic, information theory, set theory, language theory, statistical linguistics — and more generally statistics and probabilities.

A further example is the classification introduced by Solomon Marcus during the *Séminaire International de Linguistique Formelle* that took place in Aiguille in 1968. He put forward a classification of the subfields of formal linguistics, the term which subsumed the whole set (see Desclés et Fuchs 1969). Marcus distinguished between algebraic linguistics (for instance Chomsky-Schützenberger's work on monoids) and mathematical linguistics (using Markov chains), the latter involving probabilistic linguistics and quantitative linguistics, automatic, computational and cybernetic linguistics, finally applied linguistics. It should be noted that Marcus did not include either works on formal grammars or vocabulary statistics within computational linguistics.

Yet another grouping was offered by Bernard Vauquois, a major leader in the field of machine translation in France. He grouped generative grammar and statistical studies of vocabulary on one side, and natural language processing and semantic formalisation on the other side. His reasons were probably more political than epistemological, as he aimed to ensure that Natural Language Processing be recognized by the CNRS.

It is worth noting that these various classifications do not reflect the distinctions made by Computational Linguistics as defined in 1962 by American institutions, such as the Association for Machine Translation and Computational Linguistics (AMTCL) and the ALPAC report in 1966. Computational Linguistics claimed to involve every theoretical aspect of the interaction between formal languages, linguistics and programming on the one hand, and the practical aspects of language engineering on the other hand. The whole set would be carried out by NLP in the 1970s. As can be seen, for the Americans and unlike the French, statistical studies did not pertain to computational Linguistics.

2. Current quantitative linguistics in France

How to characterize the field of quantitative linguistics in France ? Let us stress, to start with, that the term « Quantitative Linguistics » (or a translation such as *linguistique quantitative*) is not really used in French nowadays. The field is mainly termed *lexicométrie* (cf. *infra*), *linguistique de corpus* (corpus linguistics), ou *statistique linguistique* (statistical linguistics). The fact that the term « quantitative linguistics » is not widely used reflects the fact that researches tend to focus not so much on quantitative laws as on the historically-situated interpretation of quantative data in corpora.

In order to characterize these local developments amongst the various possible avenues of research in the field of quantitative linguistics, we can draw a basic typology of the different kind of quantitative linguistics. A good criteria for such a typology is the object they focus on and its degree of abstraction (Loiseau 2010): quantitative linguistic works can be divided into (i) those that seek for universal tendencies, irrespective about any particular language, at a very abstract level (ii) those that work on quantitative tendencies at the scale of a particular linguistic system, and (iii) those that work on quantitative tendencies at the scale of a genre, a discourse, or a speaker for instance.

The best known example of the first type (universal law) is the Zipf law : it is valid for any language and refers to the general economy of the language faculty. Such a universal research is the heart of the denomination of « quantitative linguistics ».

An example of the second type could be the notion of functional burdening of a phonological distinction (eg Herdan 1958)[2]. Another example could be the analysis of morphological productivity (Baayen 2009): the formula proposed for the computing of productivity index is supposed to be valid for any (fusional) languages, but it aims nevertheless at quantifying the productivity of a given morpheme in a given language. It focuses on the linguistic system.

Examples of the third type are now numerous with the development of corpus linguistics: many works try to caracterize, through quantitative properties, a genre, a discourse, a style, a variety, according to a corpus representative of that socio-/idio-lect. Methods of descriptive statistics or statistical modelling applied to corpora are most of the time aiming at describing such corpora.

Having this small typology in mind, we can try to better characterize the Quantitative Linguistics developments in France. Universalist quantitative models of linguistic data are represented by works by Mandelbrot on statistical law of the distribution of words and by information theory (Le Roux ; Léon). We can also add to that group works on the morphodynamic paradigm (Petitot and also Ploux), and Guiraud's work (Bergounioux).

However, the large bulk of works focuses on sociolect/idiolect descriptions and to the analysis of the links between discursive phenomena on one hand and historical / ideological conditionnings on the other hand. This applies to the joint works by historians and lexicologues (Mayaffre), to the *lexicométrie* school (Loiseau ; Brunet ; Longré and Mellet), to the works done under the umbrella of the TLF (*Trésor de la langue française*), a large dictionary based on a corpus of French texts (Candel).

As in several other countries, the field of Quantitative linguistics in France arose during the first half of the 20th century and develop mainly during the second half of that century. Today, the major part of the research in this field has been incorporated into an international research field and have no national peculiarity anymore. However, some subfields such as *Lexicométrie* are still mainly developed in France. A focus on texts is still strong among French scholars.

[2] if a phonemic distinction is contrasting a large number of minimal pairs this distinction cannot be withdrawn without producing a lot of homonymies ; whereas if it is contrasting few pairs of lexemes, it can be lost without producing much lexical ambiguity ; the « functional burdening » is a measure of this importance of a phonemic distinction.

Interviews with actors of the field

In the process of the preparation of this volume, several interviews have been made with actors and witnesses of the field : Robert Nicolaï (University of Nice), Micheline Petruszewycs (EHESS), Pierre Lafon (ENS Lyon), Jean Petitot (EHESS), †Maurice Tournier (ENS Lyon), Évelyne Bourion (UMR Modyco). These interviews aimed at setting up the institutional context of Quantitative linguistics in France.

Robert Nicolaï witnessed the development of quantitative linguistics in the Nice University around the proeminent figure of Pierre Guiraud. Robert Nicolai was Guiraud's student at the University of Nice in the late 1960s, early 1970s. He recalls that Guiraud's lectures focused on lexicology and semantics more than statistics. He was especially concerned by morphosemantic roots and the etymological structures of French which can be only tackled with large lexical data allowing to deal with semantic universals (see Bergounioux this volume for more details).

Guiraud created the department of General Linguistics at the University of Nice and, with Gabriel Manessy, a research group named Ideric (Institut de Recherche Interethnique et Interculturel) [Institute of Interethnic and Intercultural Research] which Nicolai directed after Guiraud's death.

The interview with Jean Petitot turned into a chapter in this volume (Petitot, Léon, Loiseau, this volume).

Interviews with Pierre Lafon, Maurice Tournier, Évelyne Bourion and Micheline Petruszewycs focused mainly on the development of the *lexicométrie* school. Micheline Petruszewycs has been the assistant of the mathematician Georges-Théodule Guilbaud (1912 - 2008). G.-T. Guilbaud was the founder of a laboratory, the « Center for analysis and mathematics for Social sciences » (*Centre d'analyse et de mathématiques sociales*) at the 6th section of EPHE (*École pratique des hautes études*). He also animated two seminars for years.[3] The Friday seminar was devoted to mathematics for social sciences and had been very influencial. It was attended by various people such as the mathematicians Pierre Achard, Bernard Jaulin, Simon Reignier, the ethnologist Robert Jaulin, and many psychologists — among them François Bresson and his students. The composer Iannis Xenakis used to attend this seminar, too. It should be said that statistics, and more generally mathematics, were well regarded by social scientists of EPHE (6th section). Guilbaud was even invited to give a course on statistics within Levi-Strauss's seminar. The Thursday seminar focused on linguistics, more specifically on lexicometry with the participation of Pierre Lafon, Annie Geoffroy, Maurice Tournier, André Salem (who studied mathematics in Moscow with Andrej Kolmogorov) as well as other members of the *Laboratoire de Lexicométrie de St Cloud*. Guilbaud introduced the hypergeometric law which helped solving some ill explicited formulations made by the St Cloud group.

[3] Thanks to Micheline Petruszewycs, we consulted the book signing sheets of these seminars showing the diversity of people who attended it.

Presentation of the contributions

The volume gathers contributions either by people that have been involved in the field of which they are giving an account, or by people specialized in the history of linguistics. In both cases, however, the same historical focus has been adopted by all contributors.

Some contributions are focussing on individuals (such as Thom, Benzécri, or Guiraud), while others are focusing on larger fields of research (chapters by Léon, Loiseau, Longré and Mellet).

The contributions have been organised in two parts. The first one focuses on vocabulary statistics. The second one gathers contributions presenting mathematical models.

The first part 'Vocabulary statistics' includes four papers on the pioneering works in the 1950-60s and three papers on contemporary research. Some emblematic personalities and projects played a significant role in these early years and it is not surprising that they were adressed in several chapters : Pierre Guiraud and Georges Gougenheim (Bergounioux, Léon) ; Benoît Mandelbrot (Léon, Le Roux) ; TLF (Brunet, Candel).

Jacqueline Léon deals with early French statistical linguistics and its institutionalisation: after presenting the role of the instigators, Mario Roques (1875-1961) and Marcel Cohen (1884-1974), she examines the three paths followed by the pioneers: (i) the teaching track with Georges Gougenheim (1900-1972) and *Le Français Élementaire* ; (ii) the stylistic track with Pierre Guiraud (1912-1983) and later Charles Muller (1909-2015) ; (iii) finally, the mathematical and Information theory track with Benoît Mandelbrot (1924-2010), René Moreau (1921-2009) and the Centre Favard. She shows that statistical studies of vocabulary contributed greatly to the changes in French linguistics that took place in the 1950-60s. As statistical studies of vocabulary, simultaneously with the first experiments of machine translation, were the first fields to be computerized, they made possible the automation of linguistics that took place in the USA ten years before.

Gabriel Bergounioux dedicates a whole chapter to Pierre Guiraud (1912-1983), one of the major pioneers of quantitative linguistics in France. He emphasizes the originality of Guiraud's approach and his role in the beginning of statistical studies on vocabulary. Guiraud published three key works on the domain, a comprehensive bibliography (1954) and two methodological essays (1954 and 1960). At first, Guiraud characterized linguistics as an observational science grounded on statistics, like sociology and economics. Later, he claimed that it was cognitive-based. His approach was both stylistic (with statistical studies of Guillaume Apollinaire's and Paul Valéry's vocabulary) and etymological. For that purpose he worked out the concept of "morpho-semantic field". Bergounioux shows how his position of outsider shed light on the conditions in which quantitative linguistics emerged in France.

Danielle Candel's chapter is a testimony on the building of the Trésor de la Langue Française, a major dictionary project (1971-1994). This project aimed at building a reference corpus providing the basis and the data for the lexicographic analyses. The corpus built for that project is then one of the early exemples of corpus built to be representative of a language, to assist the lexicographer to derive meanings from observed usages in context, and large enough in order to extract the frequencies of the lexical items. Danielle Candel shows how the quantitative approach and the use of a large-scale database is linked with many steps in the building of the dictionary.

In the chapter devoted to the ”lexicometric” school, Sylvain Loiseau tries to show the theoretical assumptions and the institutional settings that lead to the development of this very influencial line of research. The quantitative analysis according to lexicometry is aimed at providing a scientific tool for the analysis of ideological content of texts. The main methods developed in the field of lexicometry are presented, focusing on the quantitative assumption of the method specificity. Some characteristics of lexicometry are still influential in contemporary research in corpus linguistics in France: the focus on text, the search for an ideological “backstage” beneath the words, the idea that quantitative textual analysis can help providing an objectivity in the analysis of such a backstage.

The chapter by Damon Mayaffre focuses on the historical studies of corpora of political texts. This avenue of research originates in the development of the « lexicometry » approach of quantitative analysis of vocabularies in the 1970’s in France and had always been associated with that field of research. The author shows that the lexicometry approach, aiming at unravelling the social position and ideological content, and due to its focus on political texts, has interested historians from the beginning. Damon Mayaffre then offers several examples of the methods elaborated and of their usage for the historical interpretation of political texts.

Longrée and Mellet’s chapter contributes to the statistical handling of a corpus of Latin texts. The authors, as latinists and proponents of statistical studies, address the specific issue of the variability of word order in Latin which constitutes a guiding thread for quantitative linguistics in Latin. This issue raised latinists’s interest as early as the 1970s so that they worked up counts of various configurations in Latin texts. By taking over that type of work, Longrée and Mellet identify ‘motifs’ in the aim of establishing a typology of texts. Motifs associating lexical and grammatical constraints subsumed the notions of repeated segments, collocations and colligations and led to new software developments for the treatment of Latin.

Etienne Brunet deals with the history of large computerized corpora and data bases of written texts in France. The first French corpus was the TLF (*Trésor de la Langue Française*) (cf. Chapter by Candel), which, in fact, was the first computerized corpus in the world while the Brown Corpus was thought out a little later. Brunet recalls how the making of the dictionary was computer-aided, with co-occurrences and frequencies at the editors’ disposal. Examining the TLF’s successor, Frantext, he makes a distinction between a corpus (Frantext) and a base (TLF). Contrary to a corpus, a base has a fixed frame with ordered sections that items must fill by a number, a code or text. He compared these two French projects with American bases such as Encarta Encyclopedia

(1993-2009), Wikipedia, and with corpora of the French Language made outside of France, such as the German Wortschatz, the English Sketchengine, the American Google Books.

The second part of the book, ‘Mathematical models’, focuses on seminal and influent contributions in the field of the mathematical models of language. It includes three chapters on pioneering works and one chapter, by Sabine Ploux, that illustrates contemporary research. Three lines of research are represented : research on distribution laws by Benoît Mendelbrot (Ronan Le Roux), the development of a family of factorial analysis, correspondence analysis, for the distributional analysis of a language by Jean-Paul Benzécri (Valérie Beaudouin), and the development of Catastrophe theory by René Thom, a model aiming at being a mathematical tools for language modelling, reminiscent of neural networks (Jacqueline Léon, Sylvain Loiseau, Jean Petitot).

Ronan Le Roux devotes his chapter to Benoît Mandelbrot (1924-2010), another key pioneer of quantitative linguistic in France. The author shows that Mandelbrot’s study of language was mostly limited to Zipf’s law. He questions the sources of the mathematician’s significant work on Zipf’s law pointing out the discrepancy between the horizon of retrospection (Auroux 2007) he claimed and the real background of his works. In particular the scientific environment of the California Institute of Technology where he stayed in the late 1940s, the inspiring figures of Wiener and Von Neumann and cybernetics played a major role in the way he tackled Zipf’s law. Le Roux shows that Mandelbrot’s later works on the fractal paradigm were consistent with his early work on Zipf’s law, exemplifying what Le Roux identifies as ‘the transversal regime of scientific modelling’, a typical mode of scientific activity.

The chapter by Valérie Beaudouin accounts for the elaboration of "correspondence analysis", a family of factorial analysis methods, by Jean-Paul Benzécri. From the middle of the 1960s, Jean-Paul Benzécri (1932-) has introduced and developed a series of methods called “Analyse des Données” (Data Analysis) whose heart is Correspondence Analysis, a method for the analysis of multidimensional data. Valérie Beaudouin traces the intellectual project behind these methods, showing that linguistic data played a major role in the elaboration of Correspondence Analysis: Correspondence Analysis aims at supporting an inductive approach of languages, based on the exploration of corpora and the synthesis of large distributional patterns.

The next chapter is an interview with Jean Petitot by Jacqueline Léon and Sylvain Loiseau. Jean Petitot presents in great details the mathematical work of René Thom (1923-2002), and the application of this work to linguistics, of which Jean Petitot is one of the best specialists. René Thom have defined an array of concepts – singularity, structural stability, catastrophe, bifurcation – for the mathematical modelling of morphogenesis, a field pioneered by Alan Turing that studies the formation processes of complex forms, in particular those of life. One central issue of morphogenesis is the mereological problem, i.e. how totalities can be organized with constituents, relations and transformation rules between constituents, and how totalities can show an organization which is more than the sum of their constituents. This constituency problem is of course central for linguistics. The article shows how the mathematical tools built by Thom addresses these issues and shows how these propositions are related to other theoretical frameworks or approaches, such as cognitive linguistics or neural networks. Petitot (2011) shows how the theory of dynamic systems can account for categorical perception in phonology and also in syntax, where he stresses the fundamental links between vision and syntactic structures.

Sabine Ploux presents an approach aiming at modelling the polysemy and the contextual variation of the meanings of lexical items using graph theory. She first shows the limits of two other paradigms : the dynamic approach, illustrated by René Thom (cf. Chapter by Petitot, Léon & Loiseau), and the linear model, illustrated by the concordance factorial analysis elaborated by Benzécri (cf. Chapter by Beaudouin). Both use the context of lexical units to model their meaning. The former (as well as connexionnist approaches) adequatly models the structural stability of a concept or a category despite its "deformations" in context. However, it can be applied only to few lexical units. The latter is based on the whole lexicon and give a static core meaning. This approach is called today vector space models, it doesn't give access to the processes of the building of the lexical meanings, but give a static representation of lexical meanings. Ploux shows an alternative approach based on graph theory. In graphs built from large corpora, lexemes are represented as vertices and cooccurrences are represented as edges. In such graphs, the systematic structure of the lexicon can be observed, while the various acceptions can be accessed through the cliques or the communities (dense group of vertices) in the graph.

References

Auroux, Sylvain (2007) *La question de l'origine des langues* suivi de *L'historicité des sciences*, Paris: PUF.

Baayen, R. Harald (2009) "Corpus linguistics in morphology: morphological productivity", In : *Corpus Linguistics. An international handbook* A.Luedeling and M. Kyto (eds.), Berlin: Mouton De Gruyter, p.900-919.

Cori, Marcel; **Léon**, Jacqueline (2002) "La constitution du TAL. Etude historique des dénominations et des concepts", *Traitement Automatique des Langues* 43 (3), 21-55.

Herdan, Gustav (1958) "The Relation Between the Functional Burdening of Phonemes and the Frequency of Occurrence" *Language and Speech* 1(1), 8-13.

Loiseau Sylvain (2010) "Paradoxes de la fréquence", *Energeia* 2, 20-55.

The Statistical Studies of Vocabulary in the 1950-60s in France. Theoretical and Institutional Issues

Jacqueline Léon
UMR7597, Histoire des Théories Linguistiques, CNRS, Université Paris Diderot
Jacqueline.leon@univ-paris-diderot.fr

The statistical studies of vocabulary opened the way to the reception of formal languages and computational linguistics in France. They appeared before the first experiments in Machine Translation which, elsewhere in the world, started the automatization of the language sciences (Léon 2015). Several facts may explain this specific French situation. On the one hand, France significantly lagged behind in the domains of computing and formal languages, contrary to the USA, the USSR and Great Britain. On the other hand, statistics and probabilities federated the interests of both linguists and mathematicians who shared common objects, such as letters, words and texts, pertaining to the French linguistic tradition and that can be dealt with by statistical methods. Consequently statistical methods benefited from a wide interest among linguistic academic institutions where major issues on the relationship between statistics and linguistics were discussed. Among them: (i) Is frequency an intrinsic property of words? (ii) Are word frequencies a property of languages, a property of texts or a property of speech or discourse? (iii) Should statistics be regarded as mere tools or models for language?

In this paper, I will address the context in which the statistics of vocabulary appeared, especially how Mario Roques and Marcel Cohen initiated the domain. I will then examine the three different pathways followed by the pioneering statistical studies, namely stylistics, language teaching and information theory. In doing so, I will focus especially on the debates that took place about the relationship between linguistics and statistics.

1. Words, texts, letters: vocabulary studies in France in the 1950s

Linguistics in France was dominated in the 1950s by historical linguistics and philology so that texts were at the core of every linguistic activity in France. American linguistics (i.e. linguistic anthropology and distributionalism), still virtually unknown, will be introduced only in the late 1960s (Chevalier 2006).

Le Français Moderne, one of the two main linguistic journals in that period, is quite representative of that state of mind. Created in 1933 by Albert Dauzat, it succeeded the *Revue de philologie française* while keeping the same objectives: philology, etymology, dialectology, studies on specialized vocabularies, stylistics. All the authors, whether linguists or not, were involved in the big game of word dating. Debates on theoretical matters hardly existed. As Chiss and Puech (1987, p.171) recalls, *Le Français Moderne* was designed to be a journal of synthesis and vulgarization, leaving the field of general linguistics and comparatism to the *Société Linguistique de Paris* (*SLP*) and its *Bulletin* (*BSL*).

Works on specialized vocabularies were then flourishing; most of them were supervised by Robert-Léon Wagner (1905-1982), a specialist of French lexicology, professor at the Sorbonne and at the Ecole Pratique des Hautes Etudes. Among them, Matoré Georges, 1951, *Le Vocabulaire et la société sous Louis-Philippe;* Wexler Peter-J., 1955, *La formation du vocabulaire des chemins de fer en France (1778-1842);* Quemada Bernard, 1955, *Introduction à l'étude du vocabulaire médical;* Dubois Jean, 1962, *Le vocabulaire politique et social en France de 1869 à 1972. A travers les œuvres des écrivains, les revues et les journaux*; Guilbert Louis 1965. *La Formation du vocabulaire de l'aviation ;* Wagner Robert-Léon, *Les Vocabulaires français*. I, II.

Bernard Quemada (b. 1926), who created the Centre d'Etude du vocabulaire français at the University of Besançon in 1958, the Centre de linguistique appliquée in 1959 and the journal *Les Cahiers de lexicologie* in 1959, launched the series *Matériaux pour l'histoire du vocabulaire français*, the first volume of which was published in 1965.

As we will see, what will make the success of statistical studies is that the SLP will become precisely the place for the discussions involving words, texts and statistics, generalizing the interest in words hitherto vested in *Le Français Moderne* only.

2. The instigators: Mario Roques and Marcel Cohen

2.1. Mario Roques (1875-1961)

Mario Roques was a specialist of French and Roman philology as well as a specialist of Balkan languages. He was appointed Professor at the Collège de France in 1937 and gave his chair the name of «History of French Vocabulary». In 1944, he became a member of the executive Board of the CNRS (Centre National de la Recherche Scientifique) where he put into practice his taste for big collective projects (Chantraine 1961, Roques 1937, Chevalier 1990, Chevalier 2009).

It can be said that Mario Roques played a key role in setting up the first lexical data bases which will be used for statistical studies. It should be mentioned that regional atlases implementing compilations of vocabulary and results of investigations and recounts were extremely important in the beginnings of vocabulary studies. In 1955, Roques replaced Dauzat (who himself had succeeded to Gilliéron) at the head of the «Atlas linguistique de France» project. In 1933, he launched the «Inventaire de la Langue Française» which began to be operational in 1936. It consisted of a set of 6 million slips: each slip included a word and its textual context. The sheets were based on the systematic review of literary and technical texts. They were listed alphabetically by authors and by periods. The «Inventaire» was not intended for the making of dictionaries, but was intended to be a database available to all researchers.

In the early 1960s, Mario Roques and Paul Imbs decided to merge the IGLF with the TLF (Trésor de la Langue Française, a data base of the whole of modern French literature since 1789[1]). It should be mentioned that the TLF can be considered the first computerized corpus of texts as it was conceived and exploited by statistical methods a little earlier than the Survey of English Usage and the Brown Corpus (Léon 2005).

2.2. Marcel Cohen (1884-1974)

Marcel Cohen played a crucial role in the institutionalisation of statistical studies in linguistics. He was a specialist in comparative grammar of Semitic languages, in the sociology of language, historical linguistics and writing systems. He was a pupil of Mario Roques and Antoine Meillet (his supervisor) and taught at l'Ecole des Langues Orientales and l'Ecole Pratique des Hautes Etudes.

As did Mario Roques, Marcel Cohen played a crucial role in the beginnings of linguistics at the CNRS. He was a very inventive researcher and had a huge intellectual curiosity. He devoted himself early to the promotion of statistical studies in the linguist community which traditionally was hostile to mathematics[2]. In the 1930s, he reviewed Zipf's books in the BSL (Cohen, 1932, 1935, 1950). He managed to add linguistic statistics (as the third sub-theme) among the themes of the 6th International Congress of Linguists that took place in Paris in 1948. General morphology was the general theme – the two other sub-themes were linguistic terminology and linguistic survey.

In the congress proceedings (Cohen 1948), he advocated the creation of a specific commission dedicated to linguistic statistics. He argued that:

> "No one will dispute the usefulness of counts in all parts of linguistics. But it must be noted that until now little has been made in this direction… In a period of linguistic studies whose beginning does not go back far into the past, efforts were made to obtain qualitative accuracy first… Knowing that opposition works is important; but, in order to use that fact, we should know the intensity of this functioning and its importance in relation to others… Therefore linguists must turn their attention to quantitative concepts, and appropriate processes should be implemented to achieve this study"[3] (Cohen 1948, p. 83-84).

[1] See Candel in this volume.

[2] Cohen (1967) mentions Meillet's indignation in his review of Zipf's *Relative frequency as a determinant of phonetic change* (1929) in *BSL* t.31 1930 p.17.

[3] Personne ne contestera l'utilité des numérations dans toutes les parties de la linguistique. Mais il faut bien constater que jusqu'à présent peu de travaux ont été faits dans ce sens… Dans une période des études linguistiques dont le début ne remonte pas loin dans le passé, on s'est efforcé d'obtenir d'abord de la précision qualitative… Savoir qu'une opposition fonctionne est important; mais il faudrait, pour utiliser le fait, connaître l'intensité de ce fonctionnement et son importance par rapport à d'autres.… Il convient donc que l'attention des linguistes se porte vers les notions quantitatives, et que des procédés appropriés soient mis en œuvre pour la réalisation de cette étude (Cohen 1948, p. 83-84).

The aim of the commission was to recommend to linguists how to use statistics and to establish a bibliography. The last two recommendations (6 and 7) concerned the use of mathematical methods and electronic machines.

> 6. We expose the mathematical processes to be used, including the setting up of proportions, and the possibilities of use of calculating machines.
> 7. Mathematical operations should be summarized in common language for use by linguists unversed in mathematics [4].

It should be noted that when discussing Marcel Cohen's proposals at the end of his presentation, crucial questions were addressed:

> "Can relative frequency, in the absence of other formal criteria, serve as a conclusive test for the special character of certain linguistic forms?… Can relative frequency be a constitutive formal feature of language? [5]" (Cohen 1948 p.88)

In 1949 he gave a talk on linguistic statistics at the Institute of linguistics of Paris, and in 1967, he wrote a history of linguistic statistics (Cohen 1967).

3. Pioneers and horizons of retrospection

Although American, English, Russian and German works can be found in the French pioneers' horizons of retrospection (Auroux 1987, 2007), it still can be said that French works were original. Word frequency lists were well-known. Both Guiraud (1954a, b) and Gougenheim et al. (1956) mention Käding's frequency dictionary of German (1897), Thorndike (1921), Henmon (1924), and Vander Becke (1935). Moreau (1964c) discusses Vander Becke and Henmon's works on French word frequencies. They talk about Markov's chains, Zipf's law and Shannon and Weaver's Information Theory. All of them attempt to adapt Information Theory to the issues they examine.

French statistical works are anchored in earlier works that can be classified on three main axes: (i) stenography, cryptography and phonetics; (ii) teaching and (iii) distributions of word frequencies for philological, stylistic and etymological studies. A fourth source should be added: the importance of Russian works for the French linguists at that time.

[4] « 6. on exposerait les procédés mathématiques à employer, notamment pour établir des proportions, et les possibilités de recours à des machines à calculer.

7. les opérations mathématiques devraient être résumées dans une transposition en langage ordinaire à l'usage des linguistes peu versés dans les mathématiques » (Cohen 1948, pp.85-86).

[5] « la fréquence relative peut-elle, en l'absence d'autres critères formels, servir de critère probant pour établir le caractère spécial de certaines formes linguistiques ? … question sous-jacente : la fréquence relative peut-elle être un caractère formal constitutif de la langue ? » (Cohen 1948 p.88).

Cohen (1967) recalls that, apart from phoneme frequencies[6], pioneering works in word frequencies were initiated for language teaching: 1894 for English, 1897 for German, 1924 for French. Those kinds of frequency lists for educational purposes still existed in the 1950s (see for example Josselson 1953, reviewed by Cohen 1955). Therefore, it is not surprising that statistical methods had been first applied in France by *Le Français Élementaire*.

3.1 The literary track

Remember that early statistical works on texts had been undertaken on literary works : Markov (1913) on Eugene Onegin and Zipf (1949) on Joyce's Ulysses (among others). The French pioneers also chose to apply statistical methods to literary texts. Guiraud studied Paul Valéry for his PhD ; Moreau (1963a) worked on Racine's *Les Plaideurs*; Charles Muller (1967) on Corneille. In the late 1960s they started to work on political texts, then joining Discourse analysis studies (Dubois's PhD on social and political vocabulary in 1962) and The St Cloud lexicometry group worked on May 1968 leaflets[7], Cotteret and Moreau's *Le vocabulaire du général de Gaulle* was published in 1969.

3.2. The Russian connection

One of the features of French statistical linguistics is its familiarity with Russian works. According to Papp (1966), statistical linguistics was highly developed in Russia in the early 20th century. Of the three research centres he identifies in Russia, two of them involved scientists working in this area: Moscow with Markov's and Marozov's works, and the Kazan School in Petrograd, carrying out the counting of vowels and consonants in Passy's *Le Français Parlé*. There had been continuous contacts between Russian and French researchers. Papp mentions the relations between Lev V. Ščerba and the slavist André Mazon as early as 1900 (see also Rjéoutski 2011). In the 1950s, the French linguists are familiar with Soviet works (Léon 2015). Many of them were members of the French Communist Party. This was the case of Marcel Cohen who published an article in the journal *Voprozy Yazykosnaniya* «Linguistique moderne et idéalisme», in 1958 at the moment when structuralist issues were discussed among Russian linguists, and when mathematical linguistics, strictly speaking, started in the USSR.

The works from the former Soviet bloc countries were systematically reviewed in the *Bulletin de la Société de Linguistique*, especially papers from the journals *Voprosy Jazykoznanija*, *Izvestija Nauk SSSR*, and *The Prague Bulletin of Mathematical Linguistics*. Thus, in the early 1960s, the French linguists had more information on Soviet works in mathematical linguistics, in machine translation and computational linguistics than on American ones. The work of the Soviets N.D. Andrejev, V.J. Rozencveig, P.S. Kuznecov, I.A. Mel'čuk, A.A. Reformatskij, V.V.Ivanov, and those of the Czechs Petr Sgall and Lubomír

[6] The best known of which being Troubetzkoy's chapter on phonological statistics in his *Grundzüge der Phonologie* (1938).

[7] See Loiseau in this volume.

Doležel were well known among linguists in France. In the area of machine translation, the French develop mixed models inspired by both American and Russian models (Léon 2015).

Let us add that, in order to establish an institute of quantitative linguistics in France (cf. § 5 below), scientific military services made a survey of existing work in the USSR. When created in 1960, the Centre Favard edited many works of researchers from the Soviet bloc countries in its series *Documents de linguistique quantitative* published by Dunod. The book of the Romanian Solomon Marcus *Mathematical introduction to structural linguistics,* intended to initiate both linguists and mathematicians into mathematical aspects of language, was published in its French version the same year (1967), as was Gross and Lentin's introduction to formal grammars *Notions sur les grammaires formelles.* In the same series, A.V. Gladkij's *Cours de linguistique mathématique* was published in a bilingual edition (Russian and French). It is not surprising, then, that the French pioneers in statistical linguistics often mentioned Russian authors in their bibliography.

3.3. The teaching track. Georges Gougenheim and *Le Français Élementaire*

Georges Gougenheim (1900-1972) had a strong institutional position. A pupil at the École Normale Supérieure, with the degree of agrégé of grammar, he was appointed chair of History of French Language at the Sorbonne in 1957. In 1951, he was charged by the Ministry of National Education to create the *Français Élementaire* for teaching purposes. He created the Centre d'Étude du Français Élémentaire and worked on this project with Aurélien Sauvageot et Pierre Michéa et Paul Rivenc. *Le Français Élementaire* was published in 1956. Its name was changed in 1959 for *Le Français Fondamental,* considered less "schoolish" (Coste 2006).

As its title suggests, the project was to develop a basic grammar and vocabulary for teaching purposes. It was inspired by *Basic English,* but on a quite different methodological basis. In any case, contrary to the criticisms that have been made of it, there was no intention that *Le Français Élementaire* determine a limited and required content of education for elementary schools in France. It certainly "was not a syllabus for French native speakers" (Coste 2006 p.11 note 14). Actually, the project was devised as a tool for literacy for immigrants and rapid dissemination of French in the world. The aim was that the users of *Le Français Élementaire* were able to understand and speak French in given situations. *Le Français Élementaire* was based on the following principles:

(i) As Pinchon (1991) recalls, it was based on the primacy of spoken language over written language.

> "the essential force that acts on the vocabulary (as well as on grammar), is mutual understanding, the need we have to be understood by our inter-

> locutor and to understand him, to avoid ambiguities constantly renewed" (Gougenheim 1970 p.240 – quoted by Pinchon 1991, p. 270)[8].

Hence the importance to found *Le Français Élementaire* on speech data collection and to treat the data by statistical methods (that is 275 audio recordings, 1090 pages of transcript, 312,135 words (tokens), 7,995 types).

(ii) The statistical treatment was based on two main notions: a) frequency and range, and b) availability. Two lists were established: the frequent words resting on a tape recording of 300,000 words; and a list of available words. Availability is an original notion invented by the *Le Français Élementaire* team as they regard the mere opposition between frequent usual words and non frequent usual words as unsatisfactory. The interesting point for them was that there are words that are rarely pronounced in a conversation but that are nevertheless permanently available for the speakers. Most of the time, these are concrete words, such as *fork*, *chair*, *pencil*, *bus* etc., which are a major component of the vocabulary that has to be part of language teaching. As such they should be part of *Le Français Élementaire.* In order to determine which infrequent words have to be chosen, an availability degree was defined using the method of "areas of interest". Sixteen areas of interest were identified from a survey involving pupils living in different parts of France. The definitive list comprised 1475 "elementary" or "fundamental" words (1222 lexical words and 253 grammatical words) and 1900 available words.

(iii) As a third principle, they advocated the heterogeneity of vocabulary, a principle closely linked to the second one. Vocabulary is made of two irreducible sets: frequent words (grammatical words and verbs) and available words connected to areas of interest; in between there are adjectives and nouns likely to be used within various circumstances. Michea (1967) added that frequencies of concrete nouns are not only low but also non-constant, which means that words have no proper frequencies. This led him to assume a kind of heterogeneity of the vocabulary.

The project was criticized by Guiraud (1956) in his review of the book. His main criticism was methodological. He showed that the sample of 300,000 words was not sufficient to obtain the list of the most frequent words. Only the head of the frequency list was relevant and, from the 400th word on, it was no longer valid. To obtain the 800 most frequent words, the authors should have used a compilation of 2,000,000 words.

It should be noted that, in the conception of *Le Français Élementaire,* frequency cannot be a property of words, since available words are infrequent words most of the time. Availability is then regarded by the authors as a new property of language[9].

[8] «la force essentielle qui agit sur le vocabulaire (comme d'ailleurs sur la grammaire) est l'intercompréhension, le besoin que nous avons d'être compris par notre interlocuteur et naturellement de le comprendre, d'éviter les ambiguités sans cesse renaissantes».

[9] The availability of vocabulary marks one major difference between *Basic English* and *Le Français Élémentaire*. Besides, *Le Français Élémentaire* was not an autonomous

3.4. The stylistic track: Pierre Guiraud

Pierre Guiraud's role (1912-1983) was paradoxical. On the one hand, he can be considered a pioneer[10]; on the other hand, his views on language statistics were much criticized, especially by mathematicians. In addition, although a very inventive and prolific author, he remained quite marginal among the linguists. He became renowned thanks to his bibliography of linguistic statistics which was published in 1954 (Guiraud 1954a) and for which he was helped by Whatmough at Harvard. As Marcel Cohen (1967) relates in his history of statistics for linguistics, Guiraud made contact with him in 1947 when he started his PhD on Paul Valéry and just before Cohen asked the 6th Congress of Linguists to add linguistic statistics as a new theme. In his PhD – supervised by R-L. Wagner – and his book on stylistics, he intended to renew stylistics by statistical methods, on the grounds that stylistics aims at studying linguistic variations, especially deviations from the norm in a writer's style. As statistics is the science of deviations from the norm, and a writer's style is a deviation from the norm which can be defined quantitatively, stylistics can be studied by statistical methods.

Guiraud became the secretary of a specific committee created within the *International Permanent Committee of Linguists* (*CIPL*), thanks to Marcel Cohen[11]. As one of the tasks of the committee was to set up a specialized bibliography on linguistic statistics, a first version ("*a tentative bibliography*") was published by B. Trnka in 1950 with a foreword by Cohen. Guiraud took over the project and developed a more complete bibliography with the help of Whatmough's research team at Harvard and a UNESCO grant[12]. In his methodological introduction, Guiraud proposed to make a distinction between statistical linguistics and quantitative linguistics. As quantitative linguistics only makes countings, only statistical linguistics is able to analyze and interpret those countings. Guiraud gathered about 1,400 references on statistical linguistics (grouped into 10 axes), which nowadays constitutes a precious tool for historians of the language sciences. The book, published in 1954 (Guiraud 1954a) was reviewed by Marcel Cohen (1954) in the BSL. In his four books on linguistic statistics published from 1953 to 1960 and regularly reviewed in the BSL, he held the following views. Guiraud claimed that frequency is a property of languages. His position on that point evolved between 1954 and 1960. While in 1954 he alleged that "any language element can be defined by its frequency in discourse"[13], he radicalized his position when, in 1960, he underpinned it on cognitive hypotheses (see Bergounioux in this volume), and he argued that frequency is not a property of discourse but a law of language. He argued that linguistic units and elements are countable:

language, but was the first step in learning French. Finally it was based on spoken language more than written language.

[10] see G.Bergounioux's contribution in this volume

[11] Cohen always supported Guiraud, most notably at the SLP: At the meeting of April 14 1951, he announced that Guiraud had begun a PhD on linguistic statistics.

[12] Note that at that time, G.K. Zipf was also at Harvard University.

[13] « Tout fait de langue peut se définir par sa fréquence dans le discours » (Guiraud, 1954b, p.1)

> “ the linguist has the advantage of observing the facts easily identifiable and countable; these facts exhibit very stable probabilities… signs (sounds, words, grammatical marks and constructions, figures of style) are repeated with a fixed frequency in a given state of language.… This is the postulate on which the application of the method and its legitimacy are based, and more than a postulate this is a fact now so universally observed and verified that we must consider it as a language law …” [14] (Guiraud 1960, p.16).

He concluded that “language is essentially a statistical phenomenon; that is to say subjected to constant and digital laws and, as such, susceptible to quantitative definitions and interpretations”[15] (Guiraud 1960, p.16). He adds: “the frequency of the sign would be an objective attribute of language just as important as its form or meaning.”[16] (Guiraud 1960, p.17-18). In his book of 1954, he explicitly borrowed from Yule the hypothesis of random word distribution, and the distinction between vocabulary and lexicon. From this distinction, he criticized Zipf. Unlike Zipf, he did not see Zipf's law as a characteristic of the vocabulary of words in texts, but a characteristic of the lexicon of potential words in a language. In other words frequency is a property of language.

Several reviews of Guiraud’s books were published in the *BSL* between 1954 and 1963, mostly in the area of statistical linguistics (Cohen 1954, Mandelbrot 1954b, Gougenheim 1955, Gougenheim 1960, Gougenheim 1961, Gougenheim 1963). Guiraud himself wrote reviews on Herdan’s and Mandelbrot’s works (Guiraud 1957-58a, 1957-1958b). He gave five talks at the SLP, although only one of them was on information theory, and published a paper in the BSL on Martinet’s “double articulation”. From 1963 on, that is from his appointment at the University of Nice, he seemed to have no more activities within the SLP.

Thus, Giraud was not isolated during that period; he was quite accepted by the SLP, published many works in the BSL and was even considered a pioneer. However, the reviews of his last book in the field, *Problèmes et méthodes de la statistique linguistique*, were quite negative. One of his aims was to make linguists conscious of the interest of statistical methods. He insisted on not being technical and that his books were intended for linguists more than mathematicians, thus following Cohen’s recommendation (see §2.2) to translate math-

[14] « …en effet, le linguiste … a l’avantage d’observer des faits facilement identifiables et dénombrables ; ces faits par ailleurs présentent des probabilités très stables. les signes (sons, mots, marques et constructions grammaticales, tours de style) se répètent avec une fréquence fixe dans un état de langue donné. … Ceci constitue le postulat sur lequel repose l’application de la méthode et sa légitimité, et plus qu’un postulat c’est un fait désormais si universellement observé et vérifié qu’on doit le considérer comme une loi du langage. (Guiraud 1960, p.16)

[15] … le langage est un phénomène essentiellement statistique ; c’est-à-dire soumis à des constantes et à des lois numériques et susceptibles, à ce titre, de définitions et d’interprétations quantitatives. » (Guiraud 1960, p.16)

[16] la fréquence du signe … serait un attribut objectif de la langue tout aussi important que sa forme ou sa signification. (Guiraud 1960, p.17-18)

ematical language into ordinary language for the linguist's use. Yet, this did not prevent him from being criticized by mathematicians, namely Mandelbrot and Moreau[17], for his incompetence in mathematics. In his review of Guiraud's *Les caractères statistiques du vocabulaire*, Mandelbrot (1954b) pointed out many errors in the area of statistics. His 1960 book was even more criticized. Gougenheim (1961) challenged Guiraud's view that all words are homogeneous, with no intrinsic difference from their grammatical nature[18]. He objected to Guiraud's implicit refusal to acknowledge the heterogeneity of vocabulary and his view that a word has a given frequency, a position Gougenheim, and *Le Français Elémentaire* had criticized with the notion of availability (see also Michea 1967). To these criticisms may be added that of Greimas in *Le Français Moderne* (1963) in which he opposed Guiraud's view that graphic words are the only units of style at the expense of linguistic structure. This debate would be taken up by Charles Muller and Maurice Tournier in the early 1970s. As Yule and Guiraud, Muller (1967) maintains the idea of a distinction between lexicon (concerning language) and vocabulary (concerning speech). Muller sought to develop statistical linguistics instead of linguistic statistics promoted by Guiraud. Tournier (1985) criticized him for considering frequencies as properties of language, and promoted discourse as the field of exploration of statistical properties.

4. The Mathematicians and Information Theoreticians: Mandelbrot and Moreau

4.1. Benoît Mandelbrot (1924-2010)[19]

After graduating from Polytechnique, Benoît Mandelbrot spent two years (1947-1949) at Caltech (California Institute of Technology) and a year (1953-54) with Von Neumann at the Institute for Advanced Study in Princeton. From 1953 to 1971 he was regularly invited at MIT before teaching in other American universities and the *Collège de France*. He was then totally aware of the new areas that have emerged in the USA during World War II: cybernetics, information theory, electronics and computing.

In the wake of his presentations to the Academy of Sciences in 1951, Mandelbrot published one of the first articles on lexical statistics in the journal *Word* and attempted to provide a theoretical explanation of Zipf's law in order to generalize it (Mandelbrot 1954a)[20]. One of his criticisms of Zipf's law relates to the fact that Zipf took as a basis the word-forms (also called tokens or inflected forms), that is to say the words as they appear in a text, thus giving word-forms an intrinsic statistical property in a text (for a given language) regardless of their

[17] Moreau, personal communication (interview René Moreau 27 April 1999, HTAL).

[18] According to Gougenheim (1961), Guiraud held that the most frequent words have the following characteristics : they are the shortest, the oldest, the simplest morphologically and the broadest semantically.

[19] See also Le Roux's paper in this volume.

[20] See Mandelbrot (1968, p.48-51) for an explanation of his critique of the Estoup-Zipf law, in terms understandable to linguists.

use. He showed that the Estoup-Zipf law gives the same result regardless of the text where it is applied; thus the Estoup-Zipf law suggests that all texts (in the same language) are identical. But this surprising conclusion is not accurate. Rather it seems that frequency as a function of rank, depends on the text, but only to a very limited extent. The formula ceases to apply if T (the total number of words in the text) is very small or very large. Mandelbrot shows that the method can only be applied on empty forms (types, lemmas) for a given text. He proposed a generalized Estoup-Zipf law which can only be valid for a given text.

In his contribution to the book he wrote in 1957 in collaboration with Apostel and Morf, he outlined his conception of the relationship between statistics and linguistics. For Mandelbrot, the ability to make statistics for linguistics is due to the purely formal nature of language. "Formal" does not mean "logical" or "logical-mathematical" but refers to the code-like nature of language, regardless of meaning. It may be sufficient, he said, to fully utilize the tools already available, such as statistics, thus adopting the good habits that telegraph operators acquired by manipulating signals, without recourse to meaning. In statistical linguistics, the language units are to be meaningless, as the signals are for telegraph communication[21]. For Mandelbrot, the linguistic units are physical units, or, more accurately, one can choose to treat them as physical units because of some of their common properties. By analogy with statistical physics, Mandelbrot claims to treat the relationship between macroscopic language (vocabularies, taxonomic families) and microscopic language (laws of grammar and logic). Macroscopic elements, since they are numerous, can only be addressed by probabilistic methods. As Lees (1959) said in his review, Mandelbrot is interested in language as a mass phenomenon (bulk data).

The macrolinguistic laws allow him to redefine the notion of richness of vocabulary, usually a very approximate value, and an index to help intuition. Richness of vocabulary generally refers to the maximum potential number of available different words. Instead of richness of vocabulary, he put forward the notion of informational temperature for the vocabulary of a given text. The availability of vocabulary (as defined by *Le Français Elémentaire*) helps him determine the informational temperature of texts (B and 1/B). If the informational temperature is high (very close to 1), it means that the available words are used properly (even rare words are used with significant frequency). A low temperature, on the contrary, means that the words are misused (rare words become extremely rare). For example, James Joyce who has a varied vocabulary, also has a B very close to 1.

Mandelbrot criticizes Zipf’s use of Joyce’s *Ulysses*:

> “This example was a poor guide to Zipf, because this author regarded Joyce as the best sample available to him, because of the length and variety of his

[21] In a view belonging to Information Theory, Mandelbrot holds that statistical properties are those of the receiver more than the emitter: « “such things as frequency relationships are rather foreign to the emitter’s introspection, except when specially trained. Signs are believed to be far more conditioned by the corresponding meanings than by any stochastic schemes. But for the receiver, the statistical properties of discourse are extremely real.” (Mandelbrot, 1961, p.212).

> works; so he considered B = 1 as the best estimate of B for any author, while in fact this value is due to the exceptionally high potential variety of Joyce's text"[22] (Mandelbrot 1957 p.31).

Mandelbrot seeks general properties of statistics for language. To call 'linguistic' the statistical properties of words in speech (discourse), they must be systematic and general enough to be independent from what the speech is about, that is independent from its meaning and from the context in which it is produced.

The reception of Mandelbrot was rather low among French linguists, probably because of the hard side of his mathematical explanations[23]. Besides, he seemed to have been more in contact with cyberneticians, philosophers and psychologists than linguists[24]. Only late in his career, was he invited by linguists. In 1968, Martinet invited him to write a chapter on "Les constantes chiffrées du discours" in the volume on *Le Langage* which he edited in the *Encyclopédie de la Pléiade*. Before, his work was brought to the attention of linguists thanks to Guiraud (1957-58b) who wrote a review on Apostel and al. (1957) in *BSL*. In that review, Guiraud criticizes Mandelbrot's interpretation of Zipf's law in terms of thermodynamics, without acknowledging either the significance of the changes made by Mandelbrot to Zipf's law, or the difference between macrolinguistics and microlinguistics (directly inspired by thermodynamics) for the use of statistical methods in linguistics. In his book of 1960, Guiraud explains the difference between their respective positions. For Mandelbrot, distribution is a feature of the vocabulary of the text while for Guiraud distribution is a feature of the lexicon of the text (hence the language)[25]. Later, in the issue n°2 of ELA dedicated to statistics and applied linguistics, Guiraud (1963, p.38) mentions Mandelbrot's mentalist interpretation of Zipf's law[26], while Herdan criticized Mandelbrot's substitution of variables in his modification of the law:

> "Feeling that 'rank' as not a real variable and, in fact, had no linguistic meaning, Mandelbrot substituted for it first 'cout' (cost) and later 'word length' and 'occurrence frequency', always arguing as if the curve described by the formula remained identically the same in spite of the change in the variable. Now this is against even the elementary idea of co-ordinate geometry" (Herdan 1963 p.51).

[22] Cet exemple a été un mauvais guide pour Zipf, car cet auteur considérait Joyce comme étant le meilleur échantillon à sa disposition, à cause de la longueur et de la variété de ses ouvrages ; il considérait donc B=1 comme étant la meilleure estimation de B pour tout auteur, tandis qu'en fait cette valeur est due à la variété potentielle exceptionnellement grande du texte de Joyce (Mandelbrot 1957 p.31).

[23] Only in 1968, he was invited by Martinet to write a chapter on « Les constantes chiffrées du discours » in the *Encyclopédie de la Pléiade Le Langage*.

[24] See Le Roux, this volume

[25] See also Bergounioux, this volume.

[26] It should be said however that Guiraud fully recognized the preeminence of Mandelbrot in the field of language statistics. He quoted him many times in his work (Guiraud 1954a and 1960), often pretending to join his main views (the reverse was obviously not the case for Mandelbrot).

4.2. René Moreau (1921- 2009)

René Moreau (1921- 2009) was a French engineer and a military man, an alumnus of St. Cyr and of the Ecole Supérieure d'Electricité. Like Mandelbrot but 10 years later, in 1961, he wrote reports on statistical vocabulary for the Academy of Sciences, entitled "On the distribution of r-grams in French" and "On the distribution of tokens in French writing". During the French Indochina War, he was in charge of cryptography as an “officier de gendarmerie” and eventually became chief scientist at IBM France. René Moreau played a very important role in disseminating information technology, statistics and information theory among linguists. He was one of the founders of the Quantitative Linguistics Seminars of the Centre Favard in 1960 (see next section). He was one of the first users of mechanographic machines and of computers for statistical treatments of texts, at the Centre mécanographique de la Faculté des Lettres de Besançon led by Bernard Quemada. Besides, thanks to his military functions, he could use the computer of the Laboratoire Central de l’Armement. He was recruited at IBM in 1962, where he performed pioneering text analysis using computers, including his *Vocabulary of General de Gaulle* published in 1969 (Cotteret & Moreau 1969)[27].

Moreau discussed the statistical laws of Zipf, Poisson, and Herdan and their extension and explanation to linguists (1963a). In his early works, in the wake of his duties as a cryptographer, he was interested in improving cryptography with the help of frequencies of letters in a text. He thus contested the position which had been accepted since the countings made in the 19th century, that the frequency of letters in French is constant and varies little from one text to another. Cryptography, which was based on this idea, has shown that it was partly wrong: the frequencies of letters in French texts written by different authors can vary significantly. However, the frequency is statistically constant in the texts written by the same author when dealing with the same topic (Moreau 1961a and b). Moreover, he showed that, for coding, the very special distribution of series of letters resulting from Markov chains should be taken into account. These sequences of letters, called r-grams, have a structure corresponding to a particular mathematical function whose knowledge should facilitate the solution of some problems regarding the transmission of written messages.

Then he became interested in linguistic “equilibrium", a term borrowed from Zipf, who was at the centre of his conception of the relationship between statistics and linguistics. According to this view, linguistic phenomena are constantly changing. They are the result of a series of equilibria varying according to individuals and groups of people. The equilibrium of linguistic units, he said, are some function of the frequency of these units in the set where they exist: “The statistical aspect of linguistic equilibria is … one of the main objects of random

[27] “Since I was one of the few French able to work on linguistic statistical research on computer, IBM hired me as part of a group the company created in Machine Translation. It is in this context that I published all my other articles in linguistics.” (interview René Moreau 27 April 1999, HTAL).

mathematical linguistics" (Moreau 1963 b). However, neither information theory nor statistics theory are sufficient to explain the logic and genesis of the code rules. Random mathematical linguistics can only account for specific aspects of linguistic equilibria.

He showed the limits of Zipf's idea (1935) of an equilibrium between word length and frequency (which for Zipf is the key to explaining all linguistic phenomena). According to Zipf's principle of equilibrium and least effort, the more a speech element is used, the more it will tend to become simple. In other words, the more frequent a word is, the shorter it becomes.

In terms of information theory, the formulation of the principle becomes: if the frequency-cost adaptation leads to a decrease in the length of the most frequent sequences, it also leads to an increase in the length of the rarest ones. That is to say, the more frequent a word is, the shorter it is, and the rarer a word is, the longer it is.

When working with Quemada at the Centre of Lexicography of Besançon on the statistical analysis of texts, Moreau was led to extend the hypothesis of frequency-cost to the relation between the signifier and the signified. His point of reference was still the coding (language as a cryptographic code). In an ideal coding, there is a one-to-one correspondence between the signifier and the signified. However he recognized that only a few words – defined as intervals between two separators –have only one meaning, and that in most cases, context is needed to give the word its exact meaning.

Moreau (1963a, p.77) argues: "Can we assume that the link between the word and the signified is even closer to the one-to-one correspondence than the word is long?" The shorter a word is, the more meanings it has – it is the case of "*de*" and "*le*" in French - and the less information it has. The longer a word is, the closer the link between the signifier and the signified. However, he admits that the application of this principle cannot be completely generalized, particularly in relation to literary texts. He regrets that the application of this principle to Racine's *Les Plaideurs* led to the removal of many interjections, *oh! ah! he! eh!* etc., that have a semantic value much greater than what could be expected by only the two letters of their coding.

The impact of Moreau's work was less among mathematicians than among linguists. One of his main goals was to make statistical methods and information theory available to linguists. Contrary to Mandelbrot, Moreau was well integrated with linguists and worked directly with them, most notably Marcel Cohen, Bernard Quemada, André Martinet, Georges Mounin and Georges Gougenheim. Although he shared with Mandelbrot the adoption of Martinet's concept of double articulation and the idea of implementing the code-like aspects of languages, Mandelbrot (1968, p.54) denied that frequencies of words depend on frequencies of letters. Moreau also played a significant role for promoting the automatization of the language sciences. For Moreau (1963c), the analysis of equilibria makes statistics a mandatory tool for quantitative research in linguistics, as it is the only method that allows to interpret a counting scientifically. Three conditions are necessary to carry out such a project:

(i) Statistical methods should be sufficiently developed. Progress in the domain had been made only since the end of the 19th century.

(ii) Statistical methods can be validated only by processing large data sets

(iii) This treatment can be processed only by automatic calculation.

He concludes his paper by paying tribute to Marcel Cohen: “the researchers who responded to Marcel Cohen’s call after 1948 had little catching up to do: before that date, any study could only be fragmentary[28].” (Moreau 1963c, p.9))

5. The Centre Favard

The Centre Favard played a pioneering role for the development of statistical linguistics in France. The Seminar of Quantitative Linguistics of the Faculté des Sciences de Paris, also called le Centre Favard, was created at l'Institut Henri Poincaré under the leadership of Jean Favard (1902-1965), with the help of René Moreau and Daniel Hérault (1936-2009).

Initially the seminar began in a group, the CASDN (*Comité d'Action Scientifique de Défense Nationale)* founded by the Ministry of Defence after the Suez adventure to study the coding of messages by the use of statistical models. As Moreau reports, the scientific council of the CASDN included high-level scientists, each of whom was accompanied by an army officer[29]. René Moreau accompanied the mathematician Jean Favard, who was a member of the Academy of Sciences, professor at L’Ecole Polytechnique and member of the Bourbaki group. The third founder of the Centre Favard, Daniel Hérault, was also a Polytechnique graduate and one of Jean Favard’s pupils.

After Jean Favard’s death in 1965 and after the dissolution of the CASDN, Daniel Hérault became the head of the *Centre de linguistique quantitative*, and created *l’Association Jean-Favard pour le développement de la Linguistique Quantitative.* Jointly the Centre and the Association edited several series at the Dunod publishing house, including (i) lectures given at the Centre de Linguistique Quantitative, (ii) monographs in mathematical linguistics, and (iii) the *Documents de linguistique quantitative* until 1981, which, especially in the 1970s, published many works from the former Soviet bloc countries (see § 3.2 above).

The seminar was an important place for training the linguists in mathematics (André Martinet, Jean Dubois). Besides mathematics, there were classes on logic, information theory, set theory, statistics and probabilities (René Moreau and Daniel Hérault). In 1962-1963, an extra class on language theory was given by Jean Pitrat and Maurice Gross. In 1963, lectures on statistical linguistics were given by Georges Gougenheim, Pierre Guiraud and René Moreau, among others, which were published in the second issue of the journal ELA (*Etudes de Linguistique Appliquée*) created in 1963 by Bernard Quemada.

[28] « … aussi les chercheurs qui répondirent après 1948 à l’appel de Marcel Cohen n’avaient guère de retard à combler: avant cette date toute étude ne pouvait être que fragmentaire». (Moreau 1963a, p.XX)

[29] Personal communication, (interview René Moreau 27 April 1999, HTAL).

In the early 1960s both the Centre Favard and the ATALA (Association pour la Traduction Automatique et la Linguistique Appliquée) created in 1959, fulfilled the task of spreading the new theories and methods in the fields of machine translation and computational linguistics. However only the Centre Favard brought together statistical studies, formal languages and computer programming.

Conclusion

To conclude, it should be said that statistical studies of vocabulary contributed greatly to the changes in French linguistics that took place in the 1950-60s. They had a strong institutional impact. The debates on the relationship between statistics and linguistics took place essentially at the SLP, either in the communications or the book reviews. It can be said that the statistical studies of vocabulary joined the proponents of the history of language and vocabulary of *Le Français Moderne,* a journal where philology was still dominant, and the proponents of general linguistics of the SLP. Thus, they played a significant role in establishing the field of statistical linguistics institutionally in French linguistics making possible the reception of the mathematization of language that took place in the USA. The Centre Favard, with its seminar of Quantitative Linguistics, dedicated to the training in statistical studies of vocabulary and, simultaneously, in formal languages and computing, has been one of the central places for the mathematization and automatization of linguistics in France. Actually, the statistical studies of vocabulary, simultaneously with the first experiments in machine translation, was the first field to be computerized.

References

Archives *Histoire du Traitement Automatique des Langues* HTAL (UMR7597, Histoire des Théories Linguistiques, Université Paris Diderot, ENS Lyon).

Auroux Sylvain (1987). Histoire des sciences et entropie des systèmes scientifiques. Les horizons de retrospection. In: Peter Schmitter (Hrsg.*) Zur Theorie und Methode der Geschichtsschreibung der Linguistik* : 1-26. Tübingen : Narr.

Auroux Sylvain (2007). *La question de l'origine des langues* suivi de *L'historicité des sciences* Paris: PUF.

Chantraine, Pierre (1961). Discours de M. Pierre Chantraine, [...] à l'occasion de la mort de M. Mario Roques, [...] séance du 24 mars 1961. Paris: Institut de France, Académie des inscriptions et belles-lettres.

Chevalier, Jean-Claude (1990). La linguistique au CNRS, 1939-1949. *Cahiers pour l'histoire du CNRS* 9, *39-80.*

Chevalier, Jean-Claude (2009). Roques Mario Louis Guillaume. *Lexicon Grammaticorum* Tübingen: Max Niemeyer Verlag, p.1285.

Chevalier, Jean-Claude; **Encrevé, Pierre** (2006). *Combats pour la linguistique, de Martinet à Kristeva.* Lyon: ENS Editions.

Chiss, Jean-Louis; Puech, Christian (1987). *Fondations de la linguistique.* Bruxelles: De Boeck.

Cohen, Marcel (1932). Compte-rendu de George Kingsley Zipf *Selected Studies of the principle of relative frequency in language,* Cambridge, Mass., Harvard University Press, 1932. *Bulletin de la Société de Linguistique 33, 10-11.*

Cohen, Marcel (1935). Compte-rendu de George Kingsley Zipf *The psycho-biology of language, an introduction to dynamic philology.* Boston 1935, *Bulletin de la Société de Linguistique 36,8-10.*

Cohen Marcel (1948). Statistique linguistique. In: *Actes du VIe congrès international des linguistes. Rapports sur les questions historiques et pratiques mises à l'ordre du jour: 83-91,* Paris: Klincksieck.

Cohen, Marcel (1949). Sur la statistique linguistique. In: *Conférences de l'institut de linguistique de l'université de Paris:* 7-16. Paris, Klincksieck.

Cohen, Marcel (1950). Compte-rendu de George Kingsley Zipf 1949 *Human Behavior and the principle of least effort. An introduction to human ecology,* Cambridge, Mass., Harvard University Press. *Bulletin de la Société de Linguistique 46, 12-13.*

Cohen, Marcel (1954) Compte-rendu de Pierre Guiraud 1954 *Bibliographie critique de la statistique linguistique,* révisée et complétée par Thomas D. Houchin, Jaan Puhvel et Clavert W. Watkins de l'U. De Harvard, sous la direction de Joshua Whatmough. *Bulletin de la Société de Linguistique 50, 44-46.*

Cohen, Marcel (1955). Compte-rendu de Harry H. Josselson 1953 *The Russian word count and frequency analysis of grammatical categories of standard literary Russian,* Detroit, Wayne Universtiy Press. *Bulletin de la Société de Linguistique 51, 112.*

Cohen, Marcel (1958). Современная лингвистикаа и идеализм. *Voprosy Jazykoznania 2, 57-65.*

Cohen, Marcel (1967). Sur l'histoire de la statistique en linguistique. *Études de linguistique appliquée 5, 3-8.*

Coste, Daniel (2006). Français élémentaire, débats publics et représentations de la langue. *Documents pour l'histoire du français langue étrangère ou seconde* [En ligne], 36, 2006, mis en ligne le 24 août 2011, URL: http://dhfles.revues.org/1181

Cotteret, Jean-Marie; Moreau René (1969). Le vocabulaire du général de Gaulle 1958-59. *Cahier de l'Herne 21, 226-245.*

Dubois, Jean (1962). *Le vocabulaire politique et social en France de 1869 à 1972. A travers les œuvres des écrivains, les revues et les journaux,* Paris: Larousse.

Gougenheim, Georges (1955). Compte-rendu de Pierre Guiraud 1955 *La sémantique, Que-sais-je ? Bulletin de la Société de Linguistique 51, 77-79.*

Gougenheim, Georges (1960). Compte-rendu de Pierre Guiraud 1958 *La Grammaire Que-sais-je ? Bulletin de la Société de Linguistique 55, 153-154.*

Gougenheim, Georges (1961). Compte-rendu de P. Guiraud 1960 *Problèmes et Méthodes de la Statistique Linguistique* PUF. *Bulletin de la Société de Linguistique 56, 33-36.*

Gougenheim, Georges (1963). Compte-rendu de P. Guiraud 1962 *La syntaxe, Que-sais-je ? Bulletin de la Société de Linguistique 58,127.*

Gougenheim, Georges (1970), *Études de grammaire et de vocabulaire français* Paris: Picard.

Gougenheim, Georges; Michea, René; Rivenc, Paul; Sauvageot, Aurélien (1956). *L'élaboration du français fondamental. Etude sur l'établissement d'un vocabulaire et d'une grammaire de base.* Paris: Didier.

Greimas, Algirdas J. (1963), La linguistique statistique et la linguistique structurale. A propos du livre de P.Guiraud: Problèmes et méthodes de la statistique linguistique. *Le Français Moderne 31(1), 55-68.*

Guilbert, Louis (1965). *La Formation du vocabulaire de l'aviation.* Paris: Larousse.

Guiraud, Pierre (1954a). *Bibliographie critique de la statistique linguistique*, révisée et complétée par Thomas D. Houchin, Jaan Puhvel et Calvert W. Watkin. Publications du Comité de la statistique linguistique. Utrecht: Éditions Spectrum.

Guiraud, Pierre (1954b). *Les Caractères statistiques du vocabulaire*. Paris: P.U.F.

Guiraud, Pierre (1956). Compte-rendu de Gougenheim, Michéa, Rivenc et Sauvageot 1956: *L'élaboration du français élémentaire, étude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Paris: Didier. *Bulletin de la Société de Linguistique 52, 81-88.*

Guiraud, Pierre (1957-58a). Compte-rendu de Gustav Herdan, 1956, *Language as Choice and Chance. Bulletin de la Société de Linguistique 53, 22-23.*

Guiraud, Pierre (1957-58b). Compte-rendu de Apostel L., B. Mandelbrot et A. Morf, 1957, *Langage et théorie de l'information. Bulletin de la Société de Linguistique 53, 23-25*.

Guiraud, Pierre (1960). *Problèmes et méthodes de la statistique linguistique*. Paris: P.U.F.

Guiraud, Pierre (1963). La mécanique de l'analyse quantitative en linguistique. *Etudes de linguistique appliquée 2, 35-46.*

Henmon, V.A.C. (1924). *A French Word Book Based on a Count of 400.000 running Words.* Bureau of Educational Research, University of Wisconsin, Madison (Wisconsin), 1924 (ronéotypé).

Herdan, Gustav (1963). Mathematical models of linguistic distribution functions. *Etudes de linguistique appliquée 2, 47-64.*

Josselson, Harry H. (1953). *The Russian word count and frequency analysis of grammatical categories of standard literary Russian.* Detroit: Wayne University Press.

Käding J.W. (1897). *Häufigkeitswörterbuch der deutschen Sprache*, Stegliz bei Berlin.

Lees, Robert B. (1959). Review of Logique langage et théorie de l'information by Léo Apostel, Benoît Mandelbrot, Albert Morf. *Language 35(2), 271-303.*

Léon, Jacqueline (2005). Claimed and unclaimed sources of Corpus Linguistics. *The Bulletin of the Henry Sweet Society for the History of Linguistic Ideas 44, 34-48.* Reprint 2007. In: *Corpus Linguistics: Critical Concepts in Linguistics* 6 volumes, Ramesh Krishnamurthy & Wolfgang Teubert (eds.) vol. 1, London & New York: Routledge, p.326-341.

Léon, Jacqueline (2015). *Histoire de l'automatisation des sciences du langage.* Lyon: ENS Editions.

Mandelbrot, Benoît (1954a). Structure formelle des textes et communication. *Word 10(3), 1-27.*

Mandelbrot, Benoît (1954b). Compte-rendu de Pierre Guiraud 1954 *Les caractères statistiques du vocabulaire. Essai de méthodologie. Bulletin de la Société de Linguistique 50, 16-21.*

Mandelbrot, Benoît (1957). Linguistique macroscopique. In: Apostel-Mandelbrot-Morf (éds.) *Logique, langage et théorie de l'information: 1-78.* Paris: PUF.

Mandelbrot, Benoît (1961). On the theory of word frequencies and on related Markovian Models of Discourse. In: Jakobson, R. (ed.), *Structure of Language and its Mathematical Aspects.* Proceedings of Symposia in Applied Mathematics, vol XII, 190-219. Providence, Rhode Island : American Mathematical Society.

Mandelbrot, Benoît (1968). Les constantes chiffrées du discours. In: A. Martinet (ed.). *Encyclopédie de la Pléiade Le Langage: 46-57.*

Markov, Andrej A. (1913). Exemple d'une étude statistique d'un texte extrait de 'Eugene Oneguine' illustrant les probabilités liées. *Bulletin de l'Académie Impériale des Sciences de St Pétersbourg: 153-162.* Traduction Française: Comité d'action scientifique de défense nationale T/R/-427-561 [Archives HTAL].

Matoré, Georges (1951). *Le Vocabulaire et la société sous Louis-Philippe.* Genève: Droz. Réédition: Slatkine, 1967.

Michéa, René (1967). La relation rang-fréquence et la structure statistique de la langue parlée. *Bulletin de la Société de Linguistique 62, 9-14.*

Moreau, René (1961a). Sur la distribution des r grammes en français. In: CRAS (CR de l'Académie des Sciences) séance du 24 mai 1961.

Moreau, René (1961b). Sur la distribution des unités lexicales dans le français écrit. In: CRAS (CR de l'Académie des Sciences) séance du 27 novembre 1961.

Moreau, René (1963a). Sur la distribution des formes verbales dans le français écrit. *Etudes de Linguistique Appliquées 2, 65-88.*

Moreau, René (1963b). Les équilibres linguistiques: conférence donnée au Palais de la découverte, le 30 mars 1963.

Moreau, René (1963c). Linguistique statistique et calcul automatique. *IBM Point 6, 6-12.*

Moreau, René (1964). Initiation à la méthode statistique en linguistique. *Cahiers Vilfredo Pareto Vol. II, 3, 103-124.*

Muller, Charles (1967). *Etude de statistique lexicale: le vocabulaire du théâtre de Pierre Corneille.* Paris: Larousse.

Papp, Ferenc (1966). *Mathematical Linguistics in the Soviet Union*, London, The Hague, Paris: Mouton.

Pinchon, Jacqueline (1991). Georges Gougenheim (1900-1972). Traditionalisme et Modernité. In: Huot, Hélène (ed.), *La grammaire française entre comparatisme et structuralisme 1870-1960*. Paris: Armand Colin, p. 257-311.

Quemada, Bernard (1955). *Introduction à l'étude du vocabulaire médical*. Paris: Les Belles Lettres.

Quemada, Bernard (1965). *Matériaux pour l'histoire du vocabulaire français*, vol 1. Paris: Les Belles Lettres.

Rjéoutski, Vladislav (2011). André Mazon et les relations scientifiques franco-soviétiques (1917-1939). *Revue des études slaves 82(1), 95-113*.

Roques, Mario (1937). *Travaux et publications de Mario Roques*. Macon: Impr. de Protat.

Thorndike, E.L.; Lorge, I. (1944). *The Teacher's word book of 30.000 words*. New York: Bureau of Publications, Teacher's college, Columbia University.

Tournier, Maurice (1985). Sur quoi pouvons-nous compter? réponse à Charles Müller. In: Mélanges H. Naïs, numéro spécial de *Verbum* 8, 481-492.

Trubetzkoy, Nicolaï S. (1976/1938). *Grundzüge der Phonologie*, tr. française Jean Cantineau (1976), *Principes de phonologie*. Luis Jorge Prieto ed., préface d'André Martinet. Paris: Klincksieck.

Vander Becke, George E. (1935). *French Word Book*. New York: The Macmillan Company.

Wagner, Robert-Léon (1967-70). *Les Vocabulaires français*. I, II, Paris: Didier.

Wexler, Peter-J. (1955). *La formation du vocabulaire des chemins de fer en France (1778-1842).*, Genève: Droz.

Zipf, George Kingsley (1935). *The Psycho-Biology of Language*, Cambridge Mass.: Harvard University Press.

Zipf, George Kingsley (1949). *Human Behavior and the principle of least effort. An introduction to human ecology*. Cambridge, Mass.: Harvard University Press.

How Statistics Entered Linguistics: Pierre Guiraud at Work. The Scientific Career of an Outsider

Gabriel Bergounioux

University of Orléans

0. Introduction

Looking back, Pierre Guiraud (1912-1983) stands out conspicuously from the rest of the French academic world. His career, his work and his chosen topics pioneered a novel conception of how computation could be applied to linguistics. This approach was not understood in his time by French academics, perhaps due to the fact that he was the only humanities scholar to venture into a field that had been largely pre-empted by mathematicians (see Hérault & Moreau 1967), even though, motivated by natural language processing, mathematicians focused on parsing rather than on statistics, as did Maurice Gross for example in the same issue (Gross 1967). Of course, one has to take into account both the internal hierarchy in mathematics, where statistics were ranked low on the scale amid Bourbaki's logicist conceptions, and the desire to differentiate computer science in its early stages from electronics. As a matter of fact, despite Guiraud's copious production (eighteen books) in the famous paperback encyclopaedia collection "Que sais-je?", he never wrote one on the topic he knew so well, quantitative linguistics.

1. A short biography

Pierre Guiraud was born in Sfax (Tunisia) on September 26th 1912 and died on February 2nd 1983. His mother quickly divorced and when she died in Paris, a few years later, the young orphan was raised by two aunts in Genolhac, a small village located in Gard (south of France). He moved to secondary school in Alès and was awarded his undergraduate degree (*licence de lettres*) in Montpellier in 1934. He held a position as a teacher in Aubusson (Creuse) and Chatellerault (Vienne). Lacking the requisite qualifications (*agrégation*) to be a secondary school teacher in France, he accepted a position abroad as French language assistant in Chisinau (Romania) in 1939. Meanwhile, he joined the British Intelligence Service where he was promoted, at the end of the war, to the rank of colonel and received the D.S.O. for his action. When Chisinau and all the territory east of the River Prut (eastern Moldova) were occupied by the Soviet Union in June 1940, in accordance with the German-Soviet Pact signed in August 1939, Guiraud was repatriated to Bucharest (Romania) where the Vichy government had set up a secondary school. The "lycée français" was closed in June

1941 when Romania entered the war on the side of the Axis powers. From 1943 to April 1944, Guiraud was employed as a French language teacher in Hungary where he acted as a spy for the United Kingdom. Back in Bucharest, he was immediately arrested by Antonescu’s police. In August 1944, Marshal Antonescu was toppled and, as the country joined the Allies, Guiraud was released and returned to France.

Since his initial academic studies did not allow him to obtain a position in higher education in France, he took up a position as a lecturer at the University of Swansea at the end of the 40s where he prepared his doctoral thesis (a Higher Doctorate, or “doctorat d’état”, involving much more extensive research than a current PhD) to apply for a position as professor. He became a professor at Groningen (the Netherlands) and, following a reform of the legal framework in France, at Nice (1964) and also taught as a visiting professor at Bloomington in the same years. He spent the remainder of his academic career at the University of Vancouver until his retirement (August 1978) in France.

As neither a former student of the École Normale Supérieure, nor an “agrégé”, Guiraud was considered an outsider as were, in those days, A. J. Greimas or Roland Barthes (on this topic, see the interview with Greimas in Chevalier & Encrevé 2006), and he failed when he applied for a chair at the Sorbonne, despite an attempt to portray himself as a follower of Charles Bruneau by dedicating his thesis to him. At that time, Bruneau held the only French Language chair, established for Ferdinand Brunot (Bruneau's former teacher) at the beginning of the 20th century. But Bruneau had not kept pace with new trends in linguistics and Guiraud’s remoteness was not on his side, despite the encouragement of Robert-Léon Wagner (1905-1982), an acknowledged grammarian of the Sorbonne and the École Pratique des Hautes Études.

2. Linguistics in France: a policy of containment towards statistics

For a long time, the French syllabus in the universities was dominated by literary studies but nonetheless made a B.A. dependent on acquiring specialized knowledge. Undergraduate studies were divided in four parts: the least important one (aka "the 4th certificate") because it was a technical one and not an aesthetic one, was the "certificate of grammar and philology (= Old French)", of which a small part was devoted to stylistics. This organization had been decided during the 1870s, the starting point of an academic structure designed on the German model, a process completed in 1896 and retained until it was updated in the 1960s (Bergounioux 1998).

At first glance, it seems that Guiraud missed his aim three times in his career:

(i) When he tried to renew the stylistics studies of his time by means of statistics, during the 50s and 60s, an approach that was deemed unacceptable before the major reforms of higher education;

(ii) When he proposed a new deal in linguistics where stylistics and semantics would play a leading role. Despite the special orientations given by

Benveniste and Martinet in general linguistics, by Ducrot in semantics and by Jakobson, Mounin, and Ruwet in poetics, he remained outside the scope of the new trends, however;

(iii) Lastly, when he studied etymology in connection with semiology. As he was more interested by Lazare Sainéan's, Lucien Tesnière's, or even Gustave Guillaume's working hypotheses, he remained isolated, far from the functionalist and generativist schools then prevailing.

Nevertheless, when defending his doctoral thesis, Guiraud had the opportunity to adopt a stance on stylistic questions, in particular on an internationally renowned poet, Paul Valéry (1871-1945). But his method was original. Although he had signed a contract with the Éditions du Seuil to write an academic literary study entitled *Valéry par lui-même*, in his thesis *Langage et versification d'après l'œuvre de Paul Valéry* (1953) [Language and versification based on Paul Valéry's work] he did not deal with any biographical topics but devoted himself entirely to formal questions of literary work, in particular metrics and sound symbolism.

The positioning of this research differed from the approach of mathematicians who favoured logical formalisms in which poetic and lexical studies were discarded in favour of syntax and phonology. Nor did Guiraud's recourse to the enumeration of tokens tally with the survey conducted during this period for the definition of "Français Fondamental" [Basic French] (Gougenheim *et al.* 1956/1964). Although both these initiatives appeared to converge, almost to the year, in introducing word counts into language sciences, the differences between them are very great. First, Basic French concerned non-literary language. Based on an oral survey, it focused exclusively on spoken, even colloquial French. Second, as its objective was the teaching of French, especially French as a foreign language, this led to the preparation of dictionaries and textbooks published by an educational publisher (Didier). Guiraud, in contrast, undertook a very ambitious analysis of an author who is notoriously difficult to understand. His study was published in the highly ranked collection "Linguistique" of the Société de Linguistique de Paris. A significant fact, pointed out by the lexicographer Alain Rey, was that:

> From his beginnings, by his very conception of syntax and stylistics, and his constant interest in quantifiable formal features – Guiraud was one of the main introducers of language statistics in France – he sought to reconcile and articulate the essential forces that are at work in language and more broadly in semiosis (Rey 1985: 48).

In the introduction to his doctoral thesis, Guiraud justified his approach as follows:

> I must now say a word about the method. I had always thought it would be interesting to count all the components of a text until all the possible combinations had been exhausted (...). As I progressed in this direction I

> rapidly acquired the certainty of being on the right track. It seemed to me more and more that every style corresponds not to a purely quantitative definition but rather to a standard deviation from a norm (...). In summary, three guides should help the reader to navigate through this essay: (...).
> 3° a statistical analysis of these problems; and the claim that literary expression and style are "standard deviations" which justify our analytical method. (Guiraud 1953: 15-17)

3. The use of statistics: seeking scientific certainty in the humanities

While the end of the introduction to the thesis was addressed to all the lovers of pure literature who would not appreciate the book, Guiraud first explained how he was led to use statistics:

> The analysis of my predecessors' innumerable studies, however, suggested some doubt about the value of my original project, as most of these studies seemed to me very fragile. The analysis of a standard deviation presupposes the establishment of a standard and a measurement system. Soon I felt lost in the complexity and mystery of numbers and turned for a time to mathematics. This research resulted in two studies currently in press: one is a bibliography of statistical linguistics that contains an analysis of nearly two thousand books and papers on the topic and a discussion of the applications of statistics to problems of language; the other is an attempt to analyze the statistical characteristics of vocabulary. I tried to address the issue with as much mathematical rigor as I could. I provide – from a theoretical viewpoint in the first study, and a pragmatic one in the second – the qualitative value, the limits and the conditions of application of statistics in the analysis of language (Guiraud 1953: 16).

It is no small paradox that counting was required by the analysis of poetry, not in terms of the number of syllables as usual but in terms of words or of phonemes. In the summary of the book statistics are mentioned, apart from the introductory and the concluding parts, in the following chapters:

Ch. II “Rhythm”
Statistical study of the frequency of mute *e* which proves that this rate is abnormally high for some poets (...) Valéry has the highest frequency of mute *e* among all our poets (58 sq.)
Ch. IV “Rhyme”
Statistical analysis of rhyming dictionary (108-109)
Identical rhymes. Statistical review (115-117)
Frequency of isometric rhyming words (124)
Ch. VII Extension of meaning
Valéry’s high frequency of derived words (179-180)

While the importance accorded to statistics may seem slight, there are numerous other accounts in percentages and a roster of statistical tables enumerates 17 frequency distributions.

Although the main innovation of Guiraud's doctorate was the use of statistics, among the two hundred items listed in the bibliography there are only two explicit references, both of them to Zipf's work. One is under the heading "Phonetic and phonological system of the French language" where the book *The Psycho-Biology of Language* (Zipf: 1935) is incorrectly cited as "*Psychology of Language*", the other under the heading "Vocabulary and syntax: parts of speech" in which "Human behavior and the principle of least effort" (1949) is mentioned.

4. An example of literary study in the light of statistics: Apollinaire (1953)

In 1953, Guiraud published (in French) his *Index of the Vocabulary of Symbolisme I. Index of the Words of Alcohols by Guillaume Apollinaire*. In his foreword, Wagner draws attention to the difficulties which had arisen with phonetics and semantics and he emphasizes the results obtained by linguistic statistics applied to literature, and which complement the survey conducted by Gougenheim *et al.* on spoken French. Quoting Eluard, Wagner highlights the specificity of stylistic devices in modern poetry, even if he regrets the lack of a table of rhymes.

As Wagner points out, the original idea behind this program was shared by a few linguists:

> Fortuitously and independently, without knowing each other, Mr Pierre Guiraud and I were following the same path. A chance encounter led us to work together; first, to correct our mutual prejudices. Statistics can be off-putting and it took me some time to convince Mr P. Guiraud that his tables and his calculations could find, so to speak, a literary application. After discussing matters on an equal footing, I can say – I believe in both our names –, that as long as there are more indexes, they will from now on more conveniently meet the needs of readers for whom they have been written. (p. III-IV)

The book is a short, 29-page monograph, with half a page to explain how the lemmatisation had been done, one page for theme-words and one more for key-words, and one and a half pages for POS distribution. The remainder of the book is an alphabetical list of words with an asterisk preceding words which are not on Van der Beke's list (1929). Unsurprisingly, these words are proper names, poetic words (Apollinaire had a special liking for them, some of which are unknown even to French readers, such as *dulie* or *sistre*), non lemmatised words and compound expressions. Nevertheless, one can note that Van der Beke had omitted *ciseaux* (scissors), *médicament* (drug), *voisin* (neighbour) and… *vocabulaire* (vocabulary).

5. Guiraud as a reference in statistical linguistics: counting and techniques

So, forsaking literary studies as they had been practised previously, Guiraud adopted a quantitative approach. During this period, he published a series of indexes to prepare the ground for an inventory of the vocabulary of the Symbolist poets (1953-1954 and 1960a) and, with the assistance of Robert W. Hartle, of Jean Racine's tragedies (the general title of the series was "Great seventeenth-century French dramatists", but in fact only Racine was analyzed), with the support of R.-L. Wagner. The data obtained by such painstaking and tedious compilations did not result in a lot of papers. A compilation of nine of them (Guiraud 1969) out of a total of thirty gives a single reference in the table of contents to "statistics", in the chapter: "Language and style: form".

One year later, in an anthology co-authored with P. Kuentz, statistics was again mentioned only in passing. Guiraud just quoted a text by Dolezel when presenting the statistical theory of poetic language (1970: 62-4) before introducing his own work (1954a) on the opposition between *theme-words* (the words most frequently used by an author) and *keywords* (the words whose frequency deviates from the normal range in an author) (1970: 222-4).

At the same time, he conducted a comprehensive and up-to-date database of bibliographical references (1954b) as a result of the decision taken at the sixth Congrès International des Linguistes [International Congress of Linguists] in Paris, to establish a committee for linguistic statistics to investigate what has been published. For this second title in the series, Guiraud supervised a team comprising Joshua Whatmough, Thomas D. Houchin, Jean Puhvel, and Calvert W. Watkins, all from the Department of Comparative Linguistics at Harvard University. While it is strange that one of the most inventive and creative linguists of his generation spent ten years as a researcher compiling a bibliography and counting tokens in literary texts (even if some of these tasks were done by his wife), we can consider that it is the price he had to pay to compensate for his lack of academic qualifications.

Meanwhile, Guiraud wrote a short methodological essay of 116 pages, entitled "The statistical characteristics of vocabulary", dedicated to R.-L. Wagner and published in 1954. Two thirds of the book are devoted to "The distribution of words", the last third to "The lexicon of poetry". The last part applies the theoretical principles outlined in the early chapters and it is exemplified by the Symbolist poets' vocabulary. Quoting Henmon (1924) at the very beginning, Guiraud followed in the footsteps of pioneering studies and complied with the guidelines of the "Français Fondamental" program, with which he was never associated: in the bibliography of Gougenheim *et al.* (1964), for example, Guiraud is referred to only once, versus ten references to René Michéa on related topics.

Now let's look at the first paragraph of the foreword, entitled "Language and numbers":

> Any language event can be defined by its frequency in discourse; between this frequency and all its psycho-physical characteristics, constant and strict relationships are established. Linguistics, which studies the elements of sounds and their mutations, the structures of grammatical forms, the meanings of words and the mechanism of changes which transform them, generally ignores one of their most important and most significant features: frequency. (Guiraud 1954a: 1)

If we have a closer look at this excerpt, we can see that there are two differences with the philological tradition and also with Saussure's theory embraced by Wagner and also by Guiraud. Instead of the *langue/parole* (language/speech) distinction, Guiraud employed the word *discours* (discourse) which was not commonly used in French linguistics at the time (it was to become widespread in the 1960s). Admittedly, he was influenced by English terminology. Moreover, he did not confine himself to lists of words but he included in his work the three main linguistic domains (phonology, morpho-syntax and semantics) and the two approaches, synchronic and diachronic. The use of statistics was therefore both an improvement in the definition of the scientific object of study (*discours* instead of *parole*) and an advancement of the method.

The book was primarily intended for linguists even if it established a link between lexicography and stylistics. Thus after a presentation of Zipf – reiterated in a short paper to the *BSL* (Guiraud 1955b) – he devoted a few pages to Yule (1944) in order to preserve the relationship to literary studies, but apparently this attempt at conciliation convinced neither linguists nor professors of literature. A conclusion to this research resulted in (Guiraud 1960b) where he tried to go beyond the aims of a method, by taking into account the difficulties entailed by using statistics.

Problems and Methods of Statistics in Linguistics (1960)

Except for three subsequent papers, this book was Guiraud's last contribution to the topic. A brief foreword outlines the plan, divided in two parts: five chapters deal with “method”, and seven chapters with “problems”, most of which are reprinted or revised articles. Chapter one lists ten areas to which linguistic statistics can be applied: (i) methodology; (ii) phonetics (= phonology); (iii) metrics and versification; (iv) indexes and concordances; (v) lexical distribution and frequencies; (vi) semantics; (vii) morphology; (viii) syntax; (ix) child language; and (x) philology. This broad coverage makes it clear that the implementation of statistics can reorganize linguistics at large.

A wide variety of areas are itemized and the key authors are mentioned. In methodology, following Herdan (1956) and Miller (1951), Guiraud enumerates the following authors:

> While our field may claim the patronage of the greatest names in linguistics, Whitney, Reinach, Riemann, Gaston Paris, Saussure, Troubetzkoy, it was not before the 40s that it became aware, thanks to Zipf, Yule, and Ross, of the possibilities of an analysis based on a rigorous methodology. Until then we had quantitative linguistics but which could not be called statistical linguistics (Guiraud 1960b: 6).

In chapter 2, “Postulates and limits of the method”, Guiraud characterizes linguistics as an observational science grounded on statistics, like sociology or economics:

> Linguistics is the typical statistical science; while statisticians are well aware of that, most linguists are still unaware of that fact. This is because the separation between literary and scientific disciplines limits the number of researchers who can address aesthetic issues using fairly complex mathematics (...). (*Ibid.*: 15)

He further assumes that there is a cognitive substructure underpinning this phenomenon:

> [These facts] allow us to imagine language as a sum of the mental images that exist objectively in the speaker's brain in the form of marks or engrams in memory. What is more, it can be plausibly argued that each sign is present together with its frequency. In this way, there are as many engrams as the number of times that the word has been received and the frequency of the sign, far from being an accident of speech, is an objective attribute of the language that is just as important as its form or its meaning. Under this assumption – which is confirmed more strongly every day – any speech or text can be considered to be a sample of a linguistic state that reflects its numerical structure as well as the possibilities of its semantic performances. (*Ibid.*: 17-18)

In the original text, there are two occurrences of “ingramme” instead of “engramme”, a word coined by the German psychologist Richard Semon in 1904, and translated into English and French (Larousse dictionary, 1932). This probably means that this odd spelling is patterned after the American one, perhaps after Miller’s books. Then, Guiraud says, five difficulties are encountered: (i) the qualitative dimension of language; (ii) the distortions of measurements performed on speech, not on language; (iii) the heterogeneity of data; (iv) the complexity of language, and (v) the size of the problem, which is an obstacle to data processing. On the last point, Guiraud predicts an increasing use of electronic machines and he mentions, as an example, what was being carried out at MIT.

Chapter three is a re-issue of (Guiraud & Wagner 1959) with an unexpected psychological incursion into characterology (probably inspired by McCormick (1920) more than by Le Senne (1945)):

> The real problem is the characterology of the language. That is to say, we must begin by defining a method similar to the method of anthropology or of graphology, a kind of linguistic bertillonnage [from the Bertillon system]. It is questionable whether this is possible. (*Ibid.*: 27)

Three core issues are discussed: the genealogical relationship of languages (without considering linguistic typology), linguistic chronology and, with respect to literature, authorship attribution. Even though the aim assigned to statistics is to take linguistic tasks beyond description and classification to a science of causes, the paper concludes with a definition of the general principles of quantitative stylistics.

Chapter four, “Statistical analysis (how to describe)”, is a presentation for dummies (i.e. linguists) of statistical method, especially the use of tables. Chapter five, “Statistical analysis (how to interpret results)”, is a continuation of the previous chapter, distinguishing between quantitative linguistics and statistics in linguistics, in contrast to Grammont’s claims (1923). Chapter six, “Language and information” is a re-issue of an article first published in the *Journal de Psychologie* (1958). The seventh chapter, “Estoup-Zipf Equation and information substrate in verbalisation” links statistics and information and quotes the stenographer Jean-Baptiste Estoup’s proposal (1912), as a precursor to Zipf, and Mandelbrot (1961) on the statistical interpretation of data.

Chapter eight, “Estoup Zipf Equation and statistical characteristics of vocabulary” begins with two considerations regarding word status (“word definition is not relevant in practice”) and mental projection (“the vocabulary of a text reflects the mental lexicon from which it has been drawn”). On the second point Guiraud expresses a difference of opinion with Mandelbrot:

> Mr Mandelbrot thinks that distribution is a characteristic of the vocabulary of the text and has a constant slope for this text. I think that the distribution is a characteristic of the lexicon of the text, that is to say a characteristic of all the words from the memory storage of which the words of the text are derived. (*Ibid.*: 87)

Sampling requires particular attention to the number of words, especially for pedagogical purposes (there are recurrent references to Gougenheim *et al*. 1956), since there is an inverse relationship between the frequency of a word and the quantity of information that may be deduced from it.

Chapter nine, “Distinctiveness structure and statistical distributions of phonological systems”, correlates the distinctive features with the frequency of phonemes, in an attempt to compare the viewpoints of Zipf and Martinet or Haudricourt. The linguistic changes that have taken place from Latin to modern French are scrutinized, a reflection pursued in chapter ten, on the effects of loanwords: “Loanwords and phonological balance”, written as a tribute to Walther von Wartburg and first published in the *Zeitschrift für Romanische Philologie* (1958). Foreign words, by introducing distortions in the phonotactic and statistical distributions, allow the assignment of semantic values to a certain number of sound concatenations, for example, says Guiraud, “KA- has a negative connotation in many words; B- contributes to creating many onomatopoeic words” (*Ibid*.: 123). This suggestion will guide his further enquiries into phono-semantism.

Chapters eleven and twelve conclude this book, dealing with “The evolution of Rimbaud’s style and the chronology of *Illuminations*” and “The phonetic structure of verse”, i.e. stylistics and metrics. There is neither a conclusion, nor a bibliography.

This book is in some respects the acme of Guiraud’s work on statistical linguistics. Compared with the ten subdivisions of the initial enumeration (see above), we can note that methodology takes the lion’s share (chapters I to V). Overall, phonetics is covered in chapters IX to X, metrics and versification in chapters XI to XII (placed at the end of the book, in spite of the fact that they were at the beginning of the list), indexes and lexical distribution in chapter VIII, semantics in chapters VI to VII, basically grounded on information theory. There is no part devoted specifically to the other topics (morphosyntax, child language, philology) and no special discussion of language training or didactics which pioneered the work in this field.

In assessing the points at issue, besides an open-mindedness with respect to new trends in psychology (characterology) and mathematics, Guiraud returned to his initial subject of interest: literature; but he pointed in the direction of two new topics, word characterization and semantics.

6. How to quantify what is uncountable? From disambiguation to metaphor

A recurring problem in the field of linguistic statistics could be worded as follows: how can one count lexical units or tokens which are identical in appearance (the same character strings) but that fall into different categories? For example, rather than lemmatisation, which requires additional processing, Guiraud dealt with the question of French *locutions* (fixed expressions or chunks) in his book on the theme (1960c). In a phrase, each word, defined as a cluster of letters between two blanks, should not be counted separately but as a whole, as a macro-unit. So the same token can be classified in two different ways. The same problem occurs with homonyms, especially homographs, which must be distributed under different headwords.

This question was first approached by Guiraud through the example of slang (1956a) and the concept of “morpho-semantic field”, coupled with etymology (*BSL*, 1956b), a path undertaken much earlier in *Valéry* (1953) about sound symbolism (131-150). This transfer of an infra-lexical semantic level is developed, for the first time, in a systematic and comprehensive way, in “The morpho-semantic field of the root T.K.” (*BSL*, 1963c) and later in *Le Français Moderne* (1966). In 1967, in his masterpiece *Structures étymologiques du lexique français*, Guiraud synthesizes the findings and deals with issues relevant to the etymological structure of the French lexicon. The observed regularities induced a form of statistical determinism and thereby, the idea that it was possible to predict meaning on the basis of a purely phonetic assessment. Some basic combinations of phonemes (consonants mainly) in specific fields based the principles of etymology on particular sound sequences, by means of a consonant frame. Unlike conceptual metaphor, the sounds organize the content. So, he shares the views of other authors, ranging from Le Senne’s and Berger’s characterology to Lacan’s conceptions, on psychoanalytic issues.

Over the years, Guiraud's thinking on the role of statistics in linguistics had evolved. By the late 1960s, he no longer envisioned the statistical approach as a merely quantitative computation but as an intuitive recognition of the link between the distribution of the letters in a text, or in a list of words, and its global signification. To a certain extent, it was still a matter of quantitative linguistics but it was no longer a matter of statistical linguistics. And even in stylistics, when Guiraud attempted to follow in the footsteps of his predecessors and continued to build on the heritage left by Bruneau, after having distanced himself from Marouzeau or Cressot because he was a lot more interested in Bally's and Spitzer's work, the time had now come for analysts such as Barthes, Kristeva, Todorov or Genette to prevail.

Conclusion

Despite his position as leader in the field of statistical linguistics, and his pioneering work, Guiraud never received the recognition he deserved. Working far from Paris, even outside France until 1964, without the academic qualifications expected of a professor at the Sorbonne, he was trapped by his inability to respond to changing circumstances. Linguistics and literature, that he had always attempted to reconcile, had become two distinct and quite antagonistic domains in the universities and his broad professional network seemed to be out-dated at a time when new linguistic schools sprang up. His sole contribution to Martinet’s guidebook “Language” in the famous “Encyclopédie de la Pléiade” is truly symbolic: “The secondary functions of language”.

There was no room for him in French linguistics in this period. Neither before the 50s for academic reasons, nor during the 50s and 60s, when the confrontation between Benveniste and Martinet had split the field into two factions, nor since the 60s when the generativists (Ruwett), the harrissians (Dubois), the “énonciativistes” (Culioli) and the semanticians (Ducrot) discussed guidelines for phonology, syntax and semantics, not for lexicology or statistics. Even poetics was, at the time, controlled by Jakobson, Ruwet, and Mounin and prosody by Meschonnic or Roubaud. Although Guiraud dedicated his 1967 book to “Hjelmslev, Guillaume, Jakobson, Benveniste, and Martinet”, he remained alone, without any successors. Through this position of outsider, however, his professional career sheds light on the conditions in which French quantitative linguistics emerged.

References

Guiraud, P. (selected bibliography)

(1953). *Langage et versification d'après l'œuvre de Paul Valéry*. Paris: Klincksieck.

(1953-1954). *Index du vocabulaire du symbolisme,* avant-propos de R.-L. Wagner. Paris: Klincksieck.

[1. Index des mots d'« Alcools » de G. Apollinaire; 2. Index des mots des poésies de P. Valéry; 3. Index des mots des poésies de S. Mallarmé; 4. Index des mots des « Illuminations » d'A. Rimbaud; 5. Index des mots des « Cinq grandes odes » de P. Claudel; 6. Index des mots des « Fêtes galantes », de « La Bonne chanson » et des « Romances sans paroles » de P. Verlaine]

(1954a). *Les Caractères statistiques du vocabulaire*. Paris: PUF.

(1954b). *Bibliographie de la statistique linguistique*. Utrecht-Anvers: Spectrum.

(1954c). L'évolution statistique du style de Rimbaud et le problème des *Illuminations. Mercure de France 322, 201-234.*

(1954d). *La Stylistique*. Paris: PUF.

(1955a). *La Sémantique*. Paris: PUF.

(1955-1964). *Index du vocabulaire de la tragédie classique*. Paris: Klincksieck.

(1955b). A propos des caractères statistiques du vocabulaire et de l'équation de Zipf. *Bulletin de la Société de Linguistique de Paris LI(1), 236-239.*

(1956a). *L'Argot*. Paris: PUF.

(1956b). Les champs morpho-sémantiques. Critères externes et critères internes en étymologie. *Bulletin de la Société de Linguistique de Paris LII(1), 265-288.*

(1958). Langage, connaissance et information, *Journal de Psychologie Normale et Pathologique, juillet-septembre, 302-318.*

(1958). Emprunts et équilibre phonologique. *Zeitschrift für romanische Philologie, 74(1-2), 78-88.*

Guiraud, P.; Wagner, R.-L. (1959). La méthode statistique en lexicologie. *Revue de l'Enseignement Supérieur 1, 154-159.*

(1960a). *Index du vocabulaire du symbolisme*. Paris: Klincksieck. [7. Index des mots d'« Une saison en enfer » de Rimbaud]

(1960b). *Problèmes et méthodes de la statistique linguistique*. Paris: PUF.

(1960c). *Les locutions françaises*. Paris: PUF.

(1963a). La mécanisation de l'analyse quantitative en lexicologie. *Etudes de Linguistique Appliquée 2, 33-46.*

(1963b). Structure des répertoires et répartitions fréquentielles des éléments de la statistique du vocabulaire écrit. *Communications et Langage, 37-48.*

(1963c). Le champ morpho-sémantique de la racine T.K. *Bulletin de la Société de Linguistique de Paris LIX(1), 135-155.*

(1965). Diacritical and statistical models for languages in relation to the computer. In: D. Hymes (ed.), *The Use of Computers in Anthropology [1962] 235-254.* La Haye: Mouton.

(1966). De la grive au maquereau : le champ sémantique des noms de l'animal tacheté, *Le Français Moderne (octobre), 280-308.*

(1967). *Structures étymologiques du lexique français*. Paris: Larousse.

(1969). *Essais de stylistique*. Paris: Klincksieck.

Guiraud, P.; **Kuentz, P.** (1970). *La Stylistique : Lectures*. Paris: Klincksieck.

Secondary sources

Augé, P. (1932). *Larousse du XX^{e} siècle*, Paris: Larousse.

Bergounioux, G. (ed.) (1998). Un siècle de linguistique en France. 1. Institutions et savoirs. *Modèles linguistiques XIX(2).*

Bouton, Ch. (ed.) (1985). Hommage à Pierre Guiraud. *Annales de la Faculté des Lettres et Sciences Humaines de Nice 52* .Paris: Les Belles Lettres.

Chevalier, J.-C. ; Encrevé, P. (2006). Combats pour la linguistique, de Martinet à Kristeva: essai de dramaturgie épistémologique. Lyon: ENS Editions.

Estoup, J.-B. (1912). Gammes sténographiques. Paris: Institut steno-graphique.

Gougenheim, G.; Michéa, R.; Rivenc, P.; Sauvageot, A. (1964). L'Élaboration du français fondamental. Paris: Didier. [First printed as L'Élaboration du français élémentaire, 1956]

Grammont, M. (1923/1936). Le vers français. Ses moyens d'expression, son harmonie. Paris: Champion.

Gross, M. (1967). Linguistique et documentation automatique. *Revue de l'enseignement supérieur 1-2, 47-55.*

Henmon, V.; Allen, Ch. (1924). A French Word Book based on the Count of 400,000 Running Words. Madison: Bureau of Educational Research, University of Wisconsin.

Hérault, **D.; Moreau, R.** (1967). La linguistique quantitative. *Revue de l'enseignement supérieur 1-2, 113-127.*

Herdan, G. (1956). Language as choice and chance. Groningen: P. Noordhoff N. V.

Le Senne, R. (1945). Traité de caractérologie. Paris: PUF.

Mandelbrot, B. (1961). On the theory of word frequencies and on related Markovian models of discourse. Structure of language and its mathematical aspects. Providence: American Mathematical Society.

McCormick, L. H. (1920). Characterology: An Exact Science Embracing Physiognomy, Phrenology and Pathognomy, Reconstructed, Amplified and Amalgamated, and Including Views Concerning Memory and Reason and the Location of These Faculties Within the Brain, Likewise Facial and Cranial Indications of Longevity. New York - Chicago: Rand McNally.

Miller, G. A. (1951). *Language and Communication*. New York - London: McGraw Hill.

Rey, A. (1985). [Obituary]. In: Ch. Bouton (ed.) *Annales de la Faculté des Lettres et Sciences Humaines de Nice 52: 47-49*. Paris: Les Belles Lettres.

Ross, A. S. C. (1950). Philological Probability Problems. *Journal of the Royal statistical Society 12 (B): 19-59*.

Semon, R. (1904) *Die Mneme als erhaltendes Prinzip im Wechsel des organischen Geschehens*. Leipzig: Engelmann.

Vander Beke, G. E. (1929). *French Word Book*. New York: Macmillan.

Yule, G. U. (1944). *On the Statistical study of Literary Vocabulary*. Cambridge: CUP.

Zipf, G. K. (1935). *The Psycho-Biology of Language*. Cambridge (Mass.): Harvard University Press.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge (Mass.): Addison-Wesley.

Pioneering Statistical Applications to the *Trésor de la Langue Française* Dictionary

Danielle Candel
Histoire des Théorie Linguistiques, UMR 7597, CNRS
Université Paris Diderot
Danielle.candel@univ-paris-diderot.fr

The aim of this chapter is to recount how quantitative tools and methods were first used in lexicography. For this, we consider the concepts put forward during the early 1950's and 60's and their application to the *Trésor de la langue française* project. These developments belong to the ensemble of quantitative methods which provide computational and statistical tools that can be used to analyze large data sets which would otherwise be described in an empirical fashion. Among these methods, descriptive statistics is particularly well suited to linguistics (Embleton 2001). While statistics was known to the Babylonians in the 3rd century B.C. and used to determine the relative positions of the sun, moon and planets, it became a real discipline, in the 20th century, with many applications in a range of fields. In linguistics this was beautifully exemplified by Morris Swadesh in his analysis of the separation of pairs of languages.

Quantitative linguistics relies on countings. Computerizing helps dealing with large data sets, automatically treated, and aimed at developing objective analyses and conclusions. Statistics, for Moreau (1962), is a scientific method of observation – and not a new linguistics in itself –, which allows us to examine a subset instead of working on the whole corpus.

Since much of the present chapter is concerned with a French dictionary, it is interesting to evoke the *Dictionnaire français latin* by Estienne (1539), which is the first having French as one of its languages. Its success can be attributed to the fact that the author was the official printer of the kingdom and had the required technological know-how. Four centuries later the *Trésor de la langue française* dictionary (*TLF)* managed by the Centre national de la recherche scientifique (National Centre for Scientific Research, CNRS) is also an official undertaking and here too, the technological know-how is important. However, what is new is its quantitative approach and large-scale database[1].

The prominent example of organizing a Treasury of the French Language as a database and of constructing the corresponding dictionary (1971-1994) constitutes an interesting case study. It is shown in what follows how statistical methods were used to design it, assist in its composition and, in parallel, develop a framework for further research in quantitative and lexicological extensions.

This chapter begins with an overview of the development of statistical ideas in the field of linguistics during the late 1950's and early 1960's. It is shown that considerable discussions and interactions were taking place, allowing a remarkable growth of ideas and projects (section I). This formed the basis for the *TLF* project (section II and III). The writing of the dictionary and specific

[1] See Pruvost 1997.

applications of statistics are described in section IV, while section V underlines some critical points and section VI discusses the follow-on to the *TLF*.

I. Before the *TLF* project: a solid ground in the 1950's and 1960's

It is timely to look back at the period preceding the publication of the first volume of the *TLF* dictionary.

It is important to show that the new lexicography, the focus of this chapter, results from the conjugated action, over a period, of a set of individuals and a variety of research groups. Almost everyone is at some point in the centre of action of the "quantitative" innovation that marks the *TLF*.

The role played in quantitative linguistics by some of them is discussed in other chapters in this book (see G. Bergounioux's study on Guiraud, or Léon's one on Roques, Guiraud, Gougenheim and Moreau). Yet to be complete, it is natural to briefly mention their contributions to the initiation of the dictionary project and see how the idea came out of exploiting a huge database with quantitative techniques and these grew out from the combined efforts of different specialists.

I.1. Objective measures vs. introspection

It is clearly stated that objective linguistic measures are needed, for instance when building basic vocabularies such as the *Français élémentaire* project as soon as 1954. Authors are oscillating between functional (linguistic) and descriptive (statistical) options, but they are aiming at objective evaluations more than intuitive ones and for that start to use frequency analyses (Quemada 1974).

I.2. A manual inventory

The *Inventaire général de la langue française* (*IGLF*) begins in 1936, due to Mario Roques - a scholar in line with of Ferdinand Brunot. As early as 1932 he had begun compiling handwritten data: the *IGLF* comprises ultimately six millions "records". Mario Roques seems to be one of the first to innovate in the field of lexicography in France by systematically collecting textual data in order to provide descriptive examples of usage in French written texts for a future dictionary. He is helped in 1936 by the *Front populaire* government of Léon Blum in officializing the *IGLF*. Chevalier (2006) qualifies this undertaking as "a precursor to what is now known as a database"[2]. This project will be pursued later on by Quemada in Besançon (Chevalier 2006, Martin 1969).

I.3. Active scholars

A few productive educational and research centres are contributing to these new

[2] This is the official translation in English, chosen nowadays by the French « Académie des inscriptions et belles lettres » (« AIBL »).

orientations in descriptive linguistics of French, in particular in lexicography: Strasbourg, with Paul Imbs, Charles Muller, Bernard Pottier, Robert Martin; Besançon, where Bernard Quemada welcomes an impressive number of colleagues; Nancy, with Paul Imbs and also Robert Martin, and later joined by B. Quemada. It is also natural to include Nice, where Gérard Moignet as well as Pierre Guiraud are active[3]; and Paris of course, for a number of them, as for Robert-Léon Wagner[4].

As indicated previously, Georges Gougenheim (Gougenheim & al. 1964) opened new perspectives in the 1950's with the elaboration of the basic French vocabulary, a project supported by Unesco. His *Français fondamental* – the new name of the *Français élémentaire* –, was created from word frequency lists in a spoken French corpus and was designed to help teach French to adults or children. Klinger & Véronique (2006) conclude that Gougenheim's work is still up to date from linguistic and didactic viewpoints, in terms of a grammar of the oral language, "interaction" and constitution of oral corpora.

Guiraud is interested in combining lexical statistics and a structural approach. Considering the huge analysis carried out by Wartburg for the *Französisches Etymologisches Wörterbuch* (*French etymological dictionary*, *FEW*), he shows that studying the history of a large number of words leads to permanent "models" able to explain whether words are successful or not. Guiraud likes to innovate, for instance when creating the Applied Linguistics division at the University of Nice; he lets colleagues explore the new field opened by him[5]. At his time, Muller himself becomes influential in the field. As usual for innovations, designations are not quite fixed and Muller speaks of *lexical statistics*, or of *linguistic statistics*, or *quantitative lexicology*, or *quantitative linguistics* or *lexicometry*. Muller prefers the fourth term to the first one for its broader sense (Brunet undated, 2009, Gougenheim, 1954, 1960, 1967).

I.4. Productive institutions

Pioneering research centres develop quantitative linguistics around the 1960's. First among them, the Strasbourg *Centre de philologie romane* begins its activities in 1955. In 1958 Quemada creates the *Centre d'étude du vocabulaire français*[6] and the *Laboratoire d'analyse Lexicologique* at university of Besançon. Besançon also hosts René Moreau's group on statistical research. The *Centre de recherche et d'étude pour la diffusion du français* (*CREDIF*, Paris), directed by Gougenheim, was born in 1959. In 1960 Imbs creates the *Centre de recherche pour un Trésor de la langue française* in Nancy which, at that time, is supposed to gather an inventory of 250 million words as concordances and text-files (*fiches-textes*), and to produce a Treasury of the French language. This centre first run by Paul Imbs takes the name of *Institut national de la langue française*

[3] And later Étienne Brunet.

[4] And later Maurice Tournier, then Pierre Lafon.

[5] See Bergounioux in this book.

[6] See for instance Rondeau 1968 p. 84.

(*INaLF*) when directed by Bernard Quemada who arrives in Nancy in 1977. Institutions are organizing conferences which will considerably help develop research and more generally the field of lexical statistics and lexicology.

I.5. Some influential symposia in linguistics, lexicography and statistics: 1957-1964

Many scholars attend the Strasbourg 1957 symposium *Lexicologie et lexicographie française et romanes. Orientations et exigences actuelles*[7]. Pierre Guiraud is one of them and on his way to publish his fundamental book *Problèmes et méthodes de la statistique*. People do not really speak of *statistiques* but focus on *relevés* ("records") and *dénombrements* ("countings"). The word *informatique* ("computer science") is not yet widely used at that time. It might be useful, in the twenty first century, to go back to the definition of *mécanographie*. It may be defined, like in *TLF*, by *Utilisation de techniques, de machines et de supports (cartes et bandes perforées) destinés à mécaniser le traitement de l'information*, that is: "Using techniques, machines and materials (cards and punched tapes) for mechanizing information processing". It was first defined in 1947, as shown in *TLF*, by *emploi des machines à calculer, ou comptables, des machines servant à trier et classer les documents*: "use of calculators or accounters, machines for sorting and filing documents". There was some confusion at that time between *mécanographie* and *informatique*. The distinction is already well understood by Bernard Quemada and Robert-Léon Wagner. A few years later, *statistique* sounds apparently more attractive to non scientists than *informatique* and the Besançon center helps its development, so do Gougenheim, later Guiraud and Muller. Gougenheim explains the kind of progress statistics is going to bring once complete indexes are available, provided that lexical research jumps from the most artisanal stage to the industrial one. As a result, the first publication from the Nancy centre deals with statistics as do the four parts of the *Dictionnaire des fréquences* (1971). The first volume of the *TLF* dictionary is another demonstration of it. The project of having a thesaurus is first discussed during the conference. Imbs participates, together with Paul Robert, the editor of the forthcoming *Grand Robert de la langue française* dictionary, Josette (Rey) Debove, Alain Rey, and many other language specialists such as Jean Fourquet, Henri Frei, Georges Gougenheim, Rudolf Hallig, Louis Hjelmslev, Jean Martinet, Bernard Pottier, Jean Dubois, Charles Bruneau or Georges Matoré. Despite some scepticcism, the results Quemada presents convince the CNRS decision makers that only mechanization (one does not yet say "computerization") can help complete, within a reasonable time, a large-scale project such as the *Trésor de la langue française* dictionary. The draft of *Trésor de la langue française* project was born (Brunet 2011, Chevalier 2006, Muller 1971, Quemada 1995).

A symposium held in Besançon in 1961, *Colloque sur la mécanisation de recherches lexicologiques*, brings together specialists for automatic document-

[7] See *Lexicologie et lexicographie françaises et romanes. Orientations et exigences actuelles* (1961).

ation, lexicologists and lexicographers, all dealing with mechanical tools. The aim is to study the best ways of using these devices and the recent electronic machinery. Little by little, punched card machines, which permitted collecting about 20 million punched cards, are replaced by electronic systems. The mechanical data file initiated at the end of April 1959 has more than 4 millions cards in June 1961. The difficulty of distinguishing, for instance, homographs or orthographical variants (future lemmatization) is uncovered. Quemada proposes a new vision, in which he imagines researchers working with microfilms and optical reading systems at home. The automatic grammar used at that time is a basic one, where frequency is taken into account (such a minimal level of automation is also the one exploited in the *Français fondamental* project) (Pruvost 2000, Quemada 1962a, 1962b).

The 1964 Strasbourg symposium *Statistique et analyse linguistique* is also to be mentioned. In the aftermath of this symposium and dwelling on their early results, Muller and Pottier (1966) advocate the innovation of quantifying and counting in order to derive language descriptions. Numerical issues in statistical methods become important in lexicology but one also notices some hesitation on the value of statistics and its relevance in relation with specific research directions.

Concluding on the 1957 and 1961 meetings, Pottier first identifies signs of nascent French linguistics. Second, he finds that the project carried out by Imbs is to write an enormous dictionary with machines in order to build a huge corpus while Quemada, in Besançon, foresees the possibilities of machines: future computers will extend capabilities in a remarkable way (Chevalier 2006).

Finally, Quemada is well known for having initiated the use of the computer for studying and processing modern languages and for having helped develop automatic procedures for natural language processing (Zampolli 1990). For Busa (1991), he is the man of the “new lexicography”.

I.6. New journals

So many new data are obtained and results ready for publication are getting so numerous that Quemada creates new journals: in the 1960’s the *Cahiers de lexicologie* and *Études de linguistique appliquée* and later, in 1966, *Langages*. This pursues the work undertaken by Ferdinand Brunot in *Histoire de la langue française*, when, in the XXth century, vocabulary becomes a matter of fundamental interest (Chevalier 2006).

II. Towards the *Trésor de la langue française* (*TLF*)

In the program conceived by CNRS, equipments and tools are central.

II.1. Specific equipments and quantitative tools

As a result of the conferences, Quemada prepares the program of computerizing the documentation helping prepare for the Treasury[8]. He understands the importance of mechanographical and electronic equipment in order to gain efficiency in large-scale projects and to readily obtain lexical statistics more effectively than those derived manually. Mechanographical operations begin as early as 1958, raising issues in lexicography automation, derivation of vocabulary frequency lists, frequency indexes and reverse indexes[9]. These devices are used to directly print the code's indications, allowing direct reading by researchers themselves. Matoré reports that, at that time, using available electronic devices and magnetic tapes, allows a tremensdous progress in archiving so that a hundred million words memorized on tapes only need five cabinets instead of 1540 of the older type!

Great corpora are necessary for descriptive methods and for distributionalism. Results are obtained with punched card machines and electronic equipment. Linguistic applications are developed in Europe, in the USSR and in the United States from 1951 on. In 1963, Quemada participates in the selection of the technical equipment in Nancy where the famous large Gamma BULL 60 computer, with the highest performance of that time, is to be installed as early as 1964 (Chevalier 2006). The Gamma BULL 60, created in 1958 by the "Compagnie des Machines Bull", is the first multitasking computer and one of the first to employ multiple processors. It features several input and output units: magnetic drums, tape, card readers, card punches, printers, paper tape readers, paper tape punches, and a terminal. In total, 20 units were produced.. In 1969, Robert Martin notes that the computer is able to treat about 100 000 occurrences (words) in seven hours (Martin 1969).

Starting in 1969, a sizable amount of data is gathered in Nancy for the *TLF*: 1000 "texts", 80% of them belonging to literature, are transformed into concordances and indexes. In the Nancy research centre, a "text" counts conventionally 100 000 occurrences, as shown below. The proportion of literary texts from the nineteenth century is 41,51 %, and one has 58,48 % of the texts from

[8] It may be necessary to clarify the sense of two words which are sometimes leading to ambiguities. First, there is a confusion concerning the word « treasury ». It may be used for designating (1) the rich amount of texts gathered in order to make lexicographical studies, or (2) the place where this work is done, that is the research laboratory of CNRS in Nancy, which was called at one time « Centre de recherche pour un Trésor de la langue française », finally (3) the result of the project dealing with this lexical treasury, that is the dictionary *Trésor de la langue française* itself. Second, it seems that there is a confusion too about the name « Frantext », which is (1) the name of the computerized database used for writing the dictionary *Trésor de la langue française* and (2) this database in its enriched form: one should distinguish these two states and be aware of this variation in time.

[9] This is based on a range of specialized machines including *bill-feed, duplicatrice* ("duplicating machine"), *interclasseuse* ("collator"), *lecteur de bandes* ("magnetic tape drive"*), machine à bandes perforées* (or *télébande*) ("perforated tape machine", "teletape"), *machine imprimante* ("printing machine"), *perforateur, perforatrice* ("keypunch operator), *poinçonneuse-récapitulatrice* ("punching machine"), *positionneuse* ("placement-machine"), *reporteuse-traductrice* ("transfer-interpreter"), *reproductrice* ("duplicator"), *reproductrice-duplicatrice, tabulatrice* ("punched-card tabulator"), *traductrice* ("translator"), *trieuse* ("card sorter"), *vérificatrice* ("verifier").

the twentieth century. It was decided that only 20% of the selected corpus would be non "literary" texts. The word list of the dictionary is then given by the automatic and statistical treatment of these texts, as shown below.

Up to 1985, no other language is subjected to such a large-scale computerization and such a performance is at that time a world first in this field (Mitterand & Petit 1962, Quemada 1959, Quemada [ed.] 1962, Quemada 1995).

II.2. Indexes

An historical critical review of indexes, covering the nineteenth century (and also including previous periods) to Quemada's work in Besançon, mentions among others P. Guiraud, B. Pottier, J. Rey-Debove or R.-L. Wagner (Wagner 1967 and 1970) as well as C. Muller, J. and C. Dubois, G. Gougenheim, Algirdas-J. Greimas, L. Guilbert. It also refers to journals like *Computers and the Humanities*, dealing with studies of the Bible to author's concordances and the mechanical revolution, including a typology of indexes and concordances (Brackenier 1972).

The first modern work concerning lexicographical frequency is related to pedagogical issues (as explored by the *Laboratoire d'Analyse statistique des langues anciennes* at the University of Liège in Belgium: see Longrée and Mellet's chapter in this book). The quantitative treatment of a corpus yields a vocabulary index, a list of words and their frequency, and a table of the frequency distribution (*tableau de distribution de fréquence*). Each word has an actual frequency (*fréquence réelle*) inside a given corpus and one wishes to know if this information can be used to deduce the frequency of occurrence in the part of the corpus which has not yet being analyzed. The real question concerns the frequency stability of a lexical item rather than the frequency itself, a question considered in the *Français fondamental* (see Léon's chapter in this book). Stability is naturally opposed to variety. (Muller 1964)

Statistics may be applied to linguistics, beginning from a textual unit or a lexical one to full vocabulary quantification, indexes, concordances and frequency distributions. Examining the various kinds of initial indexes due to Guiraud, Quemada, and Wexler (in Great Britain), Muller (1962) explains the advantages of Guiraud's lemmatizing method, distinguishing homograph forms.

Going through the history of statistics, Muller considers that even if the French school of statistics is sometimes thought to be in Strasbourg, its source is Besançon, recognizing that Quemada started this venture during the 1957 symposium. This was followed by some first results in 1962, which themselves initiated quantitative analysis (thanks also to the former work of Yule, Guiraud and Herdan). Muller highlights the importance at that time of the new mechanographical tools in Besançon[10].

In retrospect one may conclude that theory and practice have been, in this field, constantly and successfully linked (Muller 1962, 1963, 1964, 1965, 1968, 1970, 1978, 1979, 1981).

[10] Later replaced by computers – as the Gamma 60 in Nancy.

III. Preparing for writing the dictionary itself

The 1960's mark the *Trésor de la langue française* project launch.

For the first time, the construction of a huge textual corpus is initiated, aimed at preparing the future French language dictionary managed by CNRS.

III.1. Technical infrastructure in the Nancy laboratory

The electronic research material is described by Imbs (1971a, 1971b) as being similar to what is commonly used in the natural sciences. There are at first 30, and later 46, perforating machines. Two punching workshops with forty-two punch machines, of the "Flexowriter Friden" type, converting texts into punched tapes. The Gamma 60 computer is operational in 1964, with equipment for "photocopic reproduction"[11], photocopying, microfilms, and finally a pin and binding workshop. The data are stored on perforated tapes before being transformed into magnetic tapes. The computer is able to gather the variants of a verb form and the different inflected forms, and to muster them correctly.

About a hundred and thirty people are working at preparing the dictionary. In the steering committee members, one finds Antoine, Gougenheim, Matoré, Quemada, Wagner[12] (Imbs 1971a, 1971b, Martin 2000, Matoré 1968).

III.2. Tools for the *TLF* project

Among the manual data sets, the "Inventaire général de la langue française" (IGLF) needs to be mentioned again, besides some other manual collections of data, such as a collection of neologisms gathered manually, or copies of general and specialized dictionary articles.

Computerizing is the great innovation of this dictionary with concordances, shorter or longer contexts and binary groups **(***groupes binaires***)** indicating the most frequent semantic co-occurrents of a word, *i.e.* the association of two "semantic words", separated or not by "functional words".

A *Frequency dictionary* (*Dictionnaire des fréquences* 1971), due to Robert Martin and Roland Vienney, offers a corpus of over 70 million occurrences, in about 3 500 pages. It deals exclusively with literary texts (from 1789 to 1964), the ones treated on the Gamma Bull 60 computer.

The four parts composing this *Frequency dictionary* are:

(1) The alphabetical frequency tables of *circa* 70 000 lemmatized forms, in eleven columns, giving indications by century (nineteenth and twentieth centuries) and even half centuries (1789-1849, 1850-1879, 1880-1918, 1919-1964), by absolute frequency and relative frequency, and by principle textual

[11] The French term being « appareils de reproduction photocopique ».

[12] Although being quite appreciative, Matoré (1968) wonders whether the idea of a « Trésor » may already be belonging to a past stage of science.

genres (such as prose or verse). By "relative frequency" is meant the quantitative distribution in, for instance, either prose or verse;

(2) A table of decreasing frequencies, in six columns, grouping information by frequency classes, the words being alphabetically ordered; here, a list of 20 000 hapax is also given;

(3) A table of frequency variations delivering information as precise as a decade;

(4) A table, in four columns (verbs, nouns, adjectives and grammatical words), giving information about homographs; more than 4000 homographs have been listed in the Gamma Bull 60 computer.

Let us give some examples, as in Tables 1 and 2.

Table 1a.
Alphabetical frequency table (detail)

term	class number	absolute frequency	relative frequency		
			century	1st half-century	2nd half-century
SYSTÉMATISER	2981	31	94	49	166
	3066	55	146	98	174
TECHNIQUE	2890	122	373	259	555
	1723	1546	4105	1074	5930
	classes	occurrences		occurrences	occurrences
XIXth century:	3012	32.663.549		20.066.761	12.596.788
XXth century :	3121	37.653.685		14.148.234	23.505.451

Table 1b.
Alphabetical frequency table (detail)

term	class number	absolute frequency	relative frequency					
			prose	verse	prose poem	soliloquy	dialog	rest
SYSTÉMATISER	2981	31	101			15		158
	3066	55	150			118		185
TECHNIQUE	2890	122	399			202	164	511
	1723	1546	4223	153	223	1478	833	5517
		classes	occurrences	occurrces	occurrences	occurrences	occurrences	occurrences
XIXth century:		3012	30.503.002	1.906.722	253.825	6.426.549	7.278.980	18.958.020
XXth century :		3121	36.553.411	651.864	448.410	6.761.651	5.518.025	25.374.009

Table 2
Homograph table (detail)

homograph form	occurrences	Grouping term							
		verb	%	noun	%	adjective/participle adjectival noun substantivized participle	%	diverse	%
TUE	2 319	TUER	83			TU	6	A TUE-TÊTE	11
VAGUES	3 570	VAGUER	-1	VAGUE	19	VAGUE	+ 80		
VALSES	94	VALSER	-1	VALSE	+ 99				

A textual database has been constructed in order to begin writing the dictionary, just as was programmed at the 1957 Strasbourg conference. This selected *TLF* corpus of computerized written texts is an inventory of French vocabulary, in the form of cumulative indexes or concordances. It is the reference database for written French, for a period ranging from 1520 to 1964: with 350 authors, 1000 titles, 70 million words, it is known for being at that time the richest textual database. It gathers 586 "texts", considering that each "text" would count about 100 000 words. They belong to an ensemble of 416 literary works from the nineteenth century and 586 ones from the twentieth century. Furthermore, it has been decided that the corpus would contain a literary part and a technical part as well. But the amount of technical texts was to reach 20% of the whole corpus. These values stated were in conformity with the definition of the project in 1957 but later on, it was possible to enrich the database so that it could serve new research programs, as shown below (Brunet 1978 [2011], Imbs 1971a, Marchello-Nizia 2004, Pruvost 2000)[13].

Muller suggests using the frequency analysis to help detect, in the hapax corpus and for given periods and types of language, the more productive and the best accepted examples of "language creativity" (Muller 1973).

But in about 1968-1969, everything is ready for the *TLF* dictionary to be written.

IV. Organizing and writing the *TLF* articles

The writing period of the *TLF* dictionary is considered in this section mostly with an insider's view, based on my experience as a member of the Parisian team.[14]

IV.1. Frequency analyses

It is worth noting first that the dictionary includes a quantitative element which is generated automatically and is appended to the article in addition to the lexicographer's work itself. The *TLF* being the first dictionary with a frequency rate for each entry, the frequency class is given for each word of the whole corpus (1789-

[13] See § VI.2.

[14] I was a member of the Parisian team from 1975 to 1994, the year of publication of the last volume, vol. 16. I also strongly participated in volume 17 – which the last director of INaLF never brought to publication. Some of the pieces of information given are based on my personal experience. Bernard Quemada followed Paul Imbs as a director. Imbs had been the director from 1961 to 1979 – partly from 1975 on. Quemada organized the arrival of a "Parisian team of lexicographers" in 1975. The laboratory was renamed *Institut de la langue française* (1977) – later *Institut national de la langue française*. Quemada became the editor of the dictionary in 1978, from vol. 8 on (the CNRS Laboratory becomes in 2001 *Analyse et traitement informatique de la langue française, ATILF*). About the *TLF* writing, see also Candel (ed.) 1990.

1964), the absolute one, and the relative one, per century. Each relative frequency is referred to a fictive ensemble of 100 million occurrences. For frequencies ranging from 1 to 200, only the absolute frequency is given. For larger frequencies, the relative frequencies by half centuries are also included. One should mention that there is no frequency information if the entry has not been extracted from the official corpus, but added to it.

The statistical treatment of the textual corpus delivers to the dictionary writers several types of data. They enrich the *dossier de mots*, a "word-folder" comprising photocopies of pre-existing dictionary articles for each word to be treated. Regarding the automatic extraction of data, it provides the dictionary author with *feuilles concordances* or "concordances sheets" and *fiches-textes* or "text sheets".

The "concordances sheets" offer three lines of concordances, the second line presenting the entry word. "Concordances" are introduced by the book reference and sub-reference; a number permits finding the original "text sheet". This is the most important document given to the lexicographer, who may ask, if necessary, for the corresponding "text sheet".

The "text sheet" is formed by an ensemble of eighteen lines, where the eight lines in the middle compose the text itself – the lines above and below are there just as a complement.

Even if these documents are much less available in the case of grammatical entries, and just non-existent in the case of prefixes or suffixes, they are quite helpful otherwise. Nevertheless, in 1969, Robert Martin notes that, with a magnetic tape, the Gamma 60 treatment is rather slow.

Another device helps the lexicographer: the *groupes binaires* or "binary groups", which represent a rather frequent sequential association of two "semantic words" (as opposed for instance to "grammatical words"). It is obvious that out of a corpus of six or seven millions occurrences, most of the "binary groups" with a frequency higher than six are linked to semantic motivation and not just to chance (Martin 1969).

What follows is more precisely focused on the lexicographer's work.

IV.2. The lexicographer's position

The *TLF* articles are derived from situated analyses. As a matter of fact, in a way, the dictionary may be considered both as having some common features with a corpus-based dictionary (as it essentially relies on a corpus) and also with a corpus-driven one (as the examples are exclusively given by the corpus – for the distinction between both approaches, see Léon 2008). But, after all, with such a large amount of data, the corpus might also be considered as being used to confirm expectations, and assumptions. It may appear to correspond to intuitive knowledge, just as in the case of a "corpus-based" approach. As the lexicographer has to follow the corpus, his feelings need not be taken into account. This cautious framework provides the unity of the dictionary. Finally, as the corpus is extensive, the lexicographer is in an excellent position but quantitative data remain the determining factor.

IV.3. Instructions to the lexicographers

The composition of the dictionary relies on quantitative parameters. A precise balance is used to specify the number of quotations the lexicographer can select per entry. Concordances and binary groups allow one to quickly discern the most frequent constructions which should be documented.

Guidelines are given to the *TLF* team concerning the words to be introduced into the dictionary, countable guidelines, generally depending on precise conditions and answering the following questions:

(a) Which entries need to be treated ?
(b) How many entries have to be treated ?
(c) How many examples should be included per article ?
(d) What size should be selected for each example ?

IV.3.1. Which entries need to be treated?

Instructions depend on the coverage of the treated word.

The word is attested in the electronic databases, corresponding to literature texts (designated like the computer itself "Gamma 60"), or in the supplement of technical texts ("20%"), or in the manual database ("IGLF"). An entry is accepted if the word features:

(a) over 100 occurrences
(b) between 100 and 10 occurrences
if it is present in the *Dictionnaire de l'Académie* + in recent general dictionaries
(c) less than 10 occurrences
– in case of a root-word:
if it is either present only in the "20%" base + one general dictionary at least, or in the "Gamma 60" + "IGLF littéraire"
– in case of a word morphologically related to a root-word:
if it is a hapax or a word of only 2 occurrences, it may be accepted under specific conditions.

Words lacking in the databases may be accepted if they are present for instance:
(a) in certain editions of the *Dictionnaire de l'Académie* provided that they are not too old,
(b) in certain reference dictionaries if textual examples are available,
(c) in the lexicographer's "linguistic awareness", provided that examples are available in two contexts due to two different authors.

Sub-entries may have a different status. Hapax, for instance, have to be introduced by the marker "Remarque". As to the special case of adverbs in "–ment", between 1 and 10 occurrences they have to be introduced by the label "Remarque"; between 10 and 100 occurrences, by "Dérivé". Above 100 occurrences, the word becomes an autonomous entry.

IV.3.2. How many entries have to be treated?

The rule is, for each lexicographer, to treat about 180 synchronic articles per year, provided that the following countings are observed:

105 words with 1 to 10 occurrences
36 with 10 to 100
26 with 100 to 1 000
10 with 1 000 to 10 000
1 or 2 with 10 000 to 100 000 occurrences.

In a note concerning the second part of *TLF* vol. 14 (February 1986), the lexicographer (specialized for the synchronic part) is reminded that he has commited himself to treat:

(a) 68 to 70 words with 1 to 500 occurrences
(b) 2 to 3 words with 501 to 1 000 occurrences
(c) 3 to 4 words with 1 001 to 5 000 occurrences
(d) 4 to 5 words with 5 000 to 20 000 or more occurrences.

Here are two examples: the noun *similitude* counted more than 5 000 occurrences and the verb *sembler* featured more than 20 000 occurrences.

IV.3.3. How many examples should be included per article ?

Occurrence classes	Examples per entry
1 (case of words being monosemic)	1 example or "énoncé(s) réduit(s)" (one short example)
2 to 500 occurrences	4 examples
501 to 1 000	5 to 9
1 001 to 3 500	9 to 12
3 501 to 8000	13 to 19
8 001 to 236 000	20 to 135
236 000	135

(proposal of Feb. 1979: Radermacher 2004)

IV.3.4. What size should be selected for each example?

The author of the article has to choose between frequent syntagms, short quotations or longer quotations, all depending on the number of occurrences in the database. An important issue during the writing of the dictionary is the balancing between frequency and representativeness, or the lexicographer's linguistic awareness. Frequency is constantly respected and linguistic facts are as well. If it is a case of weakness in the quantitative tools, the lexicographer may intervene.

IV.3.5. Repeated expression of the importance of statistics and other quantitative matters

The writing of the dictionary and the recurrent questions and recommendations from the dictionary staff reminds the lexicographer of the importance of quantitative guidelines.

For instance, at a time when it was still allowed to give "detached examples" (*exemples détachés*) – which was not the case anymore when writing the end of the dictionary –, it was decided that such a "detached example" should not exceed five lines. These *exemples détachés* were supposed to represent only a quarter of all examples.

In the file prepared for an expert meeting in 1979[15], it is recommended that the number of authorized quotations per article be reconsidered, so that each volume would typically contain about 35 000 signed quotations.

The same year[16], it is also noted that 60% of a volume (words with less than 5000 occurrences) would be reviewed internally and then externally; 40% (that is about 400 words with more than 5000 occurrences) would receive two external reviews.

Before a meeting in 1980 between writers and two categories of reviewers (*réviseurs*, "reviewers", experts from outside, and *relecteurs*, "readers" from the lab)[17], the following question was raised about statistical matters. Is it a good idea to use as a criterion for external expertise the fact that the word being treated reaches more than 5000 occurrences? The response was that the importance of an article is not always proportional to the number of occurrences.

The lexicographers were supposed to consider also all kinds of short syntagms or short noun phrases. If this group of short noun or verb phrases took only two or three lines, they were to appear immediately after the definition. If they were more numerous, they had to appear in a special paragraph with the introductory mark "SYNT."

In 1985[18], the issue was to save room. "Detached examples" were no longer possible for monosemic words presenting less than 100 occurrences. And even if three quotations were still allowed inside paragraphs representing big sections, like those in sections "A" and "B" of the article, it was highly recommended to keep two of them only.

Re-reading the texts and recommendations, we now see the formula "*si la fréquence le justifie* ("if frequency justifies it"), sometimes like a leitmotiv.

V. Some critical points

Problems arise when a discipline is innovating or evolving, as seen previously. Some points have been already addressed, like scepticism from the scientific

[15] Nancy, 22-23 November 1979, p. I.18.

[16] Nancy, 24 February 1979, p. II.2.

[17] Nancy, 13-14 November 1980, p.1.

[18] Nancy, 11 February 1985.

community for new developments (I.5.), hesitations in naming a new field such as linguistic statistics (I.3.), distinctions between objective evaluations and more intuitive ones (I.1. and IV.2.), questions concerning the notions of stability and variety (II.2.), or attempts towards predicting “language creativity” (III.). A dual issue such as lemmatizing, briefly mentioned in II.2., will be discussed in what follows, before evoking again the use of statistics. But first, let us highlight some inadequacies observed while writing the dictionary.

V.1. Some inadequacies, zero occurrence frequency and necessary adjustments

Some modifications had to be made while writing the dictionary to deal with inadequacies in the instructions given to the lexicographers. Of course, the choices are linked to the corpus. But the strict statistics-driven rules make it even less plausible to treat a term or a “frequent syntagm” featuring a frequency of zero.

This will be shown by two examples. The first case is provided by the adjective “coquin, coquine”. When the letter C- is treated, the lexicographers are not allowed to look for illustrative examples outside the official corpus. The expression “petite coquine” is quite usual (“coquin” meaning “mischievous” and “playful”, or what can only be ineffectively translated, as “little rascal”). But as there is no example in the corpus, the colleague, breaking the rule, has to comply with other dictionaries’s examples, such as the one taken from the Dubois dictionary:

> **coquin, ine**
> (…) *Par antiphrase, fam.* et *cour.*
> **1.** [En parlant d'un enfant] Enfant espiègle. *C'est un aimable petit coquin* (*Ac.* 1798-1932). *Eh bien, petit coquin, me dit-il d'un air assez affable, que me veux-tu?* (Andrieux ds *Lar. 19e*). *Petite coquine, tu étais cachée derrière la porte!* (DUB.) (…).

Another case deals with an article I was rewriting with Robert-Léon Wagner. At that period, about 1978 or 1979, it was common to hear on French broadcasts a French-speaking clock, giving the “exact time”: “au troisième top il sera exactement…” (“at the third stroke it will be exactly …”). It seemed indispensable to both of us to include an example of this usage, but none is to be found; there is no written sentence, in the whole computerized corpus, of the required phrase. We decided to “invent” a – written – frequent usage example:

> **exactement**
> (…) [Correspond à *exact* C; avec l'idée d'une précision (plus ou moins) rigoureuse]
> b) [Correspond à *exact* C 3; avec une idée de rigueur excluant l'approximation ou *indiquant une égalité parfaite (de mesures)*]
> *Au quatrième « top » il sera* ***exactement*** *sept heures quarante-deux minutes* (Horloge parlante).

V.2. About language for special purpose

The initial goal for the Treasury is to describe general language. Nevertheless, in the very beginning, it seriously takes into account language for special purposes (Quemada 1959). From a quantitative and statistical point of view, this seems logical as technical language, "language for special purpose", and terminology, includes a much larger number of words than general language. But, unfortunately, it appears that specialized vocabulary is not considered to be of central importance when the *TLF* dictionary is designed. The computerized corpus is quite reduced and not representative of the twentieth century scientific and technical fields (the "20%"). Finally, one may add words, as proposed by Bernard Quemada, and some modern quotations as well. More generally, as the *TLF* is being written over more than thirty years, added items are accepted under strictly defined conditions, whenever the corpus, officially ending in 1964, no longer satisfies the lexicographer[19] (Quemada 1980).

V.3. The importance of lemmatization and controversial views

The important thing is to distinguish ambiguity and polysemy. As a matter of fact, the machine is able to contribute "disambiguation" and to produce a lemmatized index. Taking this into account, the machine is able to produce frequency classes. Muller notes that Guiraud actually did not use the term "lemmatization", although he was regularly "lemmatizing": Zipf was himself a "formalist" but Guiraud considers the word as a lexical unit, without distinguishing different meanings of a word, nor its different forms (see Léon in this book).

Positive comments were made by Muller concerning Guiraud's choice of lemmatization, as seen above. Brunet's opinion evolves in the same direction – he was the first to use *TLF's* data and the new computerized tools for studying a large textual corpus. Nevertheless, Geffroy & Lafon (1982) discuss this view. First, they criticize the way the Nancy corpus was composed, the one used by Brunet, in its chronological partition and also in its partition into textual genres (prose, verse etc.). But specific controversy remains also about lemmatization. "Formalists" are opposed to "lemmatizers", the former believing that form indexes are better than lemmatized ones, the latter arguing that lemmatized results are more easily "readable". What people criticize in lemmatization is a loss of information, due to the fact that an objective piece of information is replaced by a subjective one. Another objection by Geffroy & Lafon is that Brunet was not allowed to use directly Nancy's textual corpus and that he had to make reference only to data derived from the frequency dictionary of the Nancy Research Center. So, lacking the source texts, he had to conduct a combined disambiguation/ lemmatization/grammatical coding, using the frequency dictionary entries and not the original textual data themselves. Geffroy & Lafon definitely state that lemmatization is a loss of information. On the other hand, as

[19] See also Candel 1992b, Candel (ed.) 1994.

they add, they appreciate for instance Brunet's work on Rousseau's *Emile* in 1979.

Quemada (1973) explains the double criterion of the word frequency and of the word textual environment, summarizing the spirit of the *Trésor de la langue française* and of the lexicographers' work (Muller 1985). Finally, indexation induces a loss of information, while lemmatization induces only a light one. There is a gain, and this level of lemmatization is adopted for the *TLF* database and automatically applied, thanks to machine-dictionaries, grouping inflected forms and variants.

Lemmatizing is naturally a big issue for the *TLF* dictionary and its environment. What has been called "the INaLF categorizer", due to Jacques Maucourt and Marc Papin and named after them, is a segmenter, a categorizer and a lemmatizer as well. It is designed from rules tested on a large number of literary works and allows annotating large textual corpora. It specifically provides detailed segmentation of texts and features thousands of compound words, grammatical units or lexical ones, out of the nomenclature of the *TLF* dictionary. The part of speech is based on an electronic dictionary reference, derived from the nomenclature of the *TLF* dictionary. Disambiguation of homographs is obtained thanks to an electronic reference dictionary, by means of several thousand rules, established by testing the discriminating contexts in several hundred works. When discriminating context does not exist, it is proposed that the most common realization speech should be followed by a question mark. Many language difficulties are addressed: disambiguation poly-categorical grammatical words, the processing of hyphens, numbers, etc. The system allows associating a lemma to each categorized form[20].

V.4. What about using more statistics?

Fifty years after the publication of *TLF* volume 1, Brunet regrets that the Nancy statistics potential was not exploited to a sufficient level and that statistics, finally, had not been as fruitful in linguistics as in other social sciences, like sociology, psychology, human geography, economy or political sciences. Language facts, examples, and quotations are more commonly taken into account than frequency. Linguistic statistics should have kept more links with other quantitative investigations (Brunet 2011).

VI. Beyond the writing of the *TLF* dictionary

For the first time in the history of lexicography, quantitative methods and computer science have been decisive with, as a result, the sixteen volume dictionary (220/305 mm, 1000 to 1400 pages each volume), written from 1968 to 1993, and published from 1971 to 1994. The dictionary offers 100 000 words, 270 000 definitions, 430 000 examples, 350 million characters and it is the biggest dictionary ever written for the French language. The *TLF*, a successful combination

[20] www.elda.org/fr/proj/euromap/panofr/tools/220.html accessed 5 June 2015.

of countings and usages**,** is characterized by a continuous balance between quantitative systematics and usage varieties (Martin 2000). The period following the dictionary composition, first called "*l'après-TLF* ", is quite rich too.

VI.1. The dictionary

When the computerized version of the *TLF* dictionary is initiated[21], in the 1990's, the paper dictionary is not yet finished; one difficulty is that the first volumes have to be re-recorded (this is done thanks to the National Library BNF), and re-analyzed, in order to be able to reuse the information types for the electronic version. This implies hierarchizing and coding before marking the text, as well as lemmatizing. The *TLF* has to be seen (a) as a hypertext from the user's point of view, (b) as a lexicological base from the lexicographer's point of view, (c) as a knowledge base from the linguist's point of view. Finally Pierrel (2010) reports about the tremendous success of the online version of the dictionary. This new version naturally allows all kinds of quantitative research. The decisions for a computerized *TLF* are presented during the 1995 conference in Nancy "Autour de l'informatisation du Trésor de la langue française". Quemada compares this 1995 meeting to the 1957 one held in Strasbourg, when the decision was first taken to create the *Trésor de la langue française*. Quemada proposed in 1957 to introduce mechanographical techniques for lexicographical and lexicological analyses, and in 1995, it is again proposed to use contemporary technologies in order to help develop lexicography (Martin 1994, 1996 and 2000; Piotrowski 1996)[22].

VI.2. The Frantext database

The *Trésor de la langue française* was written thanks to a treasury of texts, which became the Frantext database. The treasury and its treatment permitted multiple enterprises: first from the huge database, it was possible to identify the more frequent words in it. Second, this allowed one to examine the usual environment of these words and, therefore to highlight the special environments of these words.

VI.2.1. An evolving universal tool

The Internet makes it possible to consult what has been identified as an exemplary linguistic database: Frantext (Brunet 1999). It is unique in terms of length, data homogeneity, and universal accessibility. Jacques Dendien's query

[21] The future *TLFi*. See Dendien 1996, Pierrel 2010 and Pierrel *La preface du du Trésor de la Langue française informatisé,* *www.atilf.fr/IMG/pdf/La_preface_du_TLFi_par_Jean.pdf*

[22] The day Jacques Dendien asked what kind of query I might imagine some day if the dictionary were computerized, I proposed to ask for the number of entries marked "Mechanics" in use since 1870 and still in use: a query that – at that time – sounded rather far-fetched.

program Stella (*Système de traitement électronique en ligne et libre accès*) allows complex queries, and hyper-navigation throughout the interconnected databases. As a matter of fact, the remarkable database used for the dictionary is continually evolving. It is instructive to compare two different periods, for example 1988 and 2014. In 1988 (Martin 1988), the data make 600 000 printed pages, 300 000 different word forms, more than a hundred million words, and over a billion characters, as well as some other data such as those collected by Quemada in Besançon between 1956 and 1969. There are a total of 160 million quotations out of four centuries of French literature and of an ensemble of 20% scientific and technical texts, all this corresponding to about 900 authors. Altogether, this represents 2 600 "texts" from the years 1600 to 1969. It is interesting to note that the plan was even aiming at a specific distribution concerning the number of items per decade – actually about 6 million words. When a *TLF* dictionary writer needs a quotation, he may get one of up to 300 words. But the situation changes whenever one jumps into the twenty-first century's Frantext database. In 2014[23], the database Frantext counts 4 609 distinct "texts", running between the tenth and the twenty-first centuries; that makes 277 377 496 words, and this is more than twice as much as in 1988.

VI.2.2. An international collaboration

One should first mention the collaboration between the Centre national de la recherche scientifique (CNRS, French government), and the University of Chicago, starting in 1982. The ARTFL Project, the "American and French Research on the Treasury of the French Language", is a cooperative enterprise of the Laboratoire INaLF (Institut national de la langue française) – today ATILF (Analyse et Traitement Informatique de la Langue Française) – of CNRS and the Division of the Humanities and Electronic Text Services (ETS) of the University of Chicago. This consortium-based service provides its members with "access to North America's largest collection of digitized French resources".[24]

There is also a strong collaboration between Quemada and the Pisa research group on computational lexicography: Quemada enters the Scientific council of the Istituto di Linguistica Computazionale of Pisa in 1980 and becomes its president in 1988. Beginning in 1992 and thanks to strong links between Zampolli and Quemada, the INaLF laboratory, close to finishing the *TLF* dictionary and improving its computerized database, is asked to join the Network for European reference corpora project (NERC), then the Preparatory Action for Linguistic Resources Organization for Language Engineering project (PAROLE), part of the Language Engineering "Multilingual Action Plan"[25]. The aim is to produce corpora and lexica for the major European Union languages, with at

[23] _www.frantext.fr, accessed 16 June 2015.

[24] https://artfl-project.uchicago.edu/ accessed 13 May 2015.

[25] The INaLF researchers associated were Pierre Lafon, Danielle Candel and Patrick Paroubek. See also Candel 1992a.

least 20 million words, a lexicon of 20 000 lemmas for about twenty languages etc. This quantitatively balanced corpus is to be encoded with the Corpus Encoding Standard, following EAGLES recommendations and it is specified that statistical functions, i.e. the word frequency counts, should be offered (Calzolari & al. 1995, Sinclair 2004, Zampolli & Calzolari 1996). Nevertheless, I remember that during a 1992 NERC session, a change in Sinclair's attitude was noticed[26]. Sinclair was clearly feeling less confident in statistical procedures and explained that he was strongly advocating the use of a "monitor corpus", a kind of open corpus, which would be, therefore, constantly and usefully evolving. The quantitative issues are a permanent subject of discussion. In one of its 1997 Reports, the *INaLF* focuses on the text database FRANTEXT, the computerization of the *TLF* dictionary, and the European PAROLE Program (INaLF 1997).

As a result, in this period "beyond the *TLF*", on the one hand, the computerized *TLF* dictionary, became a model for the French *Dictionnaire de l'Académie* and on the other hand, the CNRS know-how in creating and exploiting corpora served to advance European projects.

Concluding remarks

The objective of this chapter was to trace the early elements of the rich history of quantitative lexicography. Moving from statistics to linguistics may be seen as a fundamental process: a kind of "translation" process (Altmann 2009). Data are provided to a statistical model, the statistical result is interpreted and is translated back to linguistics to be linguistically interpreted. For Muller, after Guiraud**,** linguistics is typically a statistical science – what statisticians know quite well and what most of the linguists still ignore. At this point in time, one may wonder whether this ignorance has either diminished or perhaps disappeared. The 1950's and 1960's are when the concepts were laid out and when the technical means became available and started to be used to perform record counting in a corpus directed at lexicology or lexicography.

One landmark during that period is the early design and successful completion of the *TLF* dictionary, a remarkable undertaking managed by CNRS. Looking back at the past of this project and its collateral developments and examining a time period of over a half century, one can sense the progress accomplished which was worth recounting. Begun in the fifties' with rudimentary mechanographical equipment, the technical progress in computer hardware and software have led to an online dictionary and a connectable enriched database which can be easily accessed through the web. Many scholars cited in this chapter (such as Imbs, Quemada, Martin) participated, in a way, in the development of this innovative dictionary and related database, a large-scale application of lexicographic statistics.

For the first time in lexicography, informatics and quantitative methods were combined in a worthwhile effort. While these tools were quite effective and placed lexicographers in an excellent position, there was still much room for the

[26] Probably the one Léon (2008: 27) highlighted.

lexicographer's metalinguistic awareness for a kind of intuitive threshold of word usage. It is however clear that much insight could be gained from the statistical tools exploited in the *TLF* and in this sense, this dictionary illustrates the usefulness of statistical linguistics.

References

Lexicologie et lexicographie françaises et romanes. Orientations et exigences actuelles (1961). Actes du colloque de Strasbourg, 12-16 nov. 1957. Paris: Éd. du CNRS. (Coll. Internat. du CNRS. Sc. Hum.).

Académie des inscriptions et belles lettres, Institut de France http://www.aibl.fr/introduction-76/missions-77/, accessed 14 April 2015.

Académie française (1994 –). *Dictionnaire de l'Académie française,* 9^{e} éd. http://atilf.atilf.fr/academie9.htm, accessed 27 April 2015.

Altmann, Gabriel (2009). Texte und Theorien. In: Christian Delcourt; Marc Hug (eds.), *Mélanges offerts à Charles Muller: 37-45*, CILF et université de Strasbourg. Paris: CILF.

Auroux, Sylvain; Koerner, E.F. Konrad; Niederehe, Hans-Josef; Versteegh, Kees (eds.) (2001). *History of the Language Sciences — Geschichte der Sprachwissenschaften. Histoire des sciences du langage — An International Handbook on the Evolution of the Study of Language from the Beginnings to the Present* 2, *1998-2005* Berlin, New-York: de Gruyer.

Brackenier, Roland (1972). Index et concordances d'auteurs français modernes. Étude critique (1^{e} partie). *Travaux de linguistique 1972-3, 1-43*.

Brunet, Étienne (1978/2011). L'analyse statistique du Trésor de la langue française. *Le français moderne 46-1, 54-66*. (Republished in: Étienne Brunet, Céline Poudat, Ludovic Lebart (eds.), *Ce qui compte* II, *Méthodes statistiques*. Paris: Champion, 45-59)

Brunet, Étienne (1999). Ce que disent les chiffres. In: Chaurand, Jacques (ed.), *Nouvelle histoire de la langue française: 673-727*. Paris: Seuil.

Brunet, Étienne (2009). Muller le lexicomètre. In: Delcourt, Christian/ Hug, Marc (eds.), *Mélanges offerts à Charles Muller: 99-119*. Paris: Conseil international de la langue française.

Brunet, Étienne (2011). Plaidoyer pour la statistique linguistique. In: Étienne Brunet, Céline Poudat, Ludovic Lebart (eds), *Ce qui compte* II, *Méthodes statistiques: 311-329*. Paris: Champion.

Brunet, Étienne (undated). C[ompte] r[endu] de Ch[arles] Muller, *Langue française. Débats et bilans. (Recueil d'articles: 1986-1993)*. Paris: Éd. Honoré Champion, 1993, 246 p.
*ancilla.unice.fr/~brunet/pub/*muller*.html accessed 17 may 2014*

Busa, Roberto (1991). Préface. In: Cignoni, Laura & Peters, Carol (eds.), *Computational lexicology and lexicography, Special issue dedicated to Bernard Quemada, 2, Linguistica Computazionale, vol. VII, IX-XIII* Pisa: Giardini Editori e stampatori.

Calzolari, Nicoletta; Baker, Mona; Kruyt, Johanna G. (eds.), Zampolli, Antonio (coord.), (1995). *Towards a network of European reference corpora,* *Report of the NERC Consortium feasability study, Linguistica Computazionale*, vol. XI. Pisa: Giardini Editori e stampatori.

Candel, Danielle (ed.) (1990). *Autour d'un dictionnaire, le* Trésor de la langue française, *témoignages d'atelier et voies nouvelles, Dictionnairique et lexicographie* 1, INaLF, CNRS, INaLF, Didier érudition.

Candel, Danielle (1992a). *Rapport sur les besoins exprimés par les utilisateurs virtuels de corpus linguistiques français*, NERC-WP2-31, INaLF, CNRS Paris, 24 p.

Candel, Danielle (1992b). Vers une base de données textuelles spécialisée. In: Quemada, Bernard (ed.), *Frantext – Autour d'une base de données textuelles, témoignages d'utilisateurs et voies nouvelles, 341-350. Dictionnairique et lexicographie* 2, INaLF, CNRS, Paris***:*** Didier Érudition,

Candel, Danielle (ed.) (1994). *Français scientifique et technique et dictionnaire de langue*, Coll. *Études de sémantique lexicale* 2, CNRS, INaLF, Paris: Didier Érudition.

Centre de recherche pour un Trésor de la langue française, Nancy (1971). *Dictionnaire des fréquences* (*Table alphabétique, Table des fréquences décroissantes, Table de répartition des homographes*).

Chevalier, Jean-Claude; Pierre Encrevé (2006). *Combats pour la linguistique, de Martinet à Kristeva*. Lyon: ENS Éditions.

Cignoni, Laura; Peters, Carol (eds.) (1990). *Computational lexicology and lexicography,* Special issue dedicated to Bernard Quemada, 1, *Linguistica Computazionale*, vol. VII. Pisa: Giardini Editori e stampatori.

Cignoni, Laura; Peters, Carol (eds.) (1991). *Computational lexicology and lexicography,* Special issue dedicated to Bernard Quemada, 2, *Linguistica Computazionale*, vol. VII. Pisa: Giardini Editori e stampatori.

Delcourt, Christian; Hug Marc (eds.) (2009). *Mélanges offerts à Charles Muller*, CILF et université de Strasbourg. Paris: Conseil international de la langue française.

Dendien, Jacques (1996). Le projet d'informatisation du *TLF.* In: Piotrowsky, David (ed.). *Lexicographie et informatique – Autour de l'informatisation du* Trésor de la langue française, Actes du colloque international de Nancy (29-31 mai 1995), INaLF. Paris: Didier Érudition, 25-34.

Dictionnaire des Fréquences (1971). Études statistiques sur le vocabulaire français. Vocabulaire littéraire des XIXème et XXème siècles. (4 vol.). CNRS. Centre de recherche pour un trésor de la langue française. Nancy: Didier.

Embleton, Sheila (2001). Quantitative methods and lexicostatistics in the 20th Century. In: Auroux & al. (eds.), *History of the Language Sciences — An International Handbook on the Evolution of the Study of Language from the Beginnings to the Present 2, 1998-2005*. Berlin, New-York: de Gruyer.

Geffroy, Annie; Lafon, Pierre (1982). L'insécurité dans les grands ensembles. Aperçu critique sur Le vocabulaire français de 1789 à nos jours d'Etienne Brunet. *Mots 5, 129-141*.

Gougenheim, Georges (1960). Problèmes et Méthode en lexicologie – La statistique linguistique et l'histoire du vocabulaire. *Cahiers de lexicologie 1960-2, 32-40.*

Gougenheim, Georges; Michea, René; Sauvageot, Aurélien; Rivenc, Paul (1964). *L'élaboration du français fondamental 1^er^ degré*. Paris: Didier.

Guiraud, Pierre (1954). *Les caractères statistiques du vocabulaire*. Paris: Presses universitaires de France.

Guiraud, Pierre (1960). *Problèmes et méthodes de la statistique linguistique*. Paris: Presses universitaires de France.

Guiraud, Pierre (1967). *Structures étymologiques du lexique français*. Paris: Larousse.

Herdan, Gustav (1966). How can quantitative methods contribute to our understanding of language mixture and language borrowing ?. In: Muller and Pottier (eds.), *17-39*.

Imbs, Paul (1971a). «Préface», *Trésor de la langue française. Dictionnaire de la langue du XIXe et du XXe siecle (1789-1960),* IX-XLVII. Paris: CNRS – Gallimard.

Imbs, Paul (1971b). «Rapport de M. Paul Imbs», Le trésor de la langue française. *Bulletin de la société de linguistique* (*BSL*) *66(1), 85-106*.

Institut National de la Langue Française, CNRS (INaLF) (1997). USR 705 – *Rapport d'étape*. Nancy: CNRS-INaLF.

Klinger, Dominique; Véronique, George Daniel (2006). La grammaire du *Français fondamental*: Interrogations historiques et didactiques. *Documents pour l'histoire du français langue étrangère ou seconde.* 36. http://dhfles.revues.org/1189. Accessed 18 June 2015

Lafon, Pierre (1997). Le programme européen 'Parole'. I *Rapport d'étape*, *47-49*. Nancy: CNRS-Institut National de la Langue Française, USR 705.

Léon, Jacqueline (2008). Aux sources de la 'Corpus Linguistics': Firth et la *London School*. *Langages 171, 12-33.*

Marchello-Nizia, Christiane (2004). Linguistique historique, linguistique outillée: les fruits d'une tradition. *Le français moderne 1, 58-70.*

Martin, Éveline (1988). Frantext, la base de données textuelles du français de l'INaLF. *Bulletin de l'EPI* (Enseignement Public et Informatique) *52, 184-200.*

Martin, Éveline (ed.) (1994). *Les textes et l'informatique.* Coll. *Études de sémantique lexicale,* CNRS, INaLF. Paris: Didier Érudition.

Martin, Robert (1969). Le Trésor de la langue française et la méthode lexicographique. *Langue française 2, 44-55.*

Martin, Robert (1994). Dictionnaire informatisé et traitement automatique de la polysémie. In: Martin, Éveline (ed.), 1994, *Les textes et l'informatique*, Coll. *Études de sémantique lexicale*, CNRS, INaLF, 77-114. Paris: Didier Érudition.

Martin, Robert (1996). Introduction. In: Piotrowski, David (ed.), *Lexicographie et informatique – Autour de l'informatisation du* Trésor de la langue française, Actes du colloque international de Nancy (29-31 mai 1995), INaLF, 9-21. Paris: Didier Érudition.

Martin, Robert (2000). Le Trésor de la langue française. In: Antoine, Gérald; Cerquiglini, Bernard (eds.), *Histoire de la langue française 1945-2000, 969-979*. Paris: CNRS Editions.

Matoré, Georges (1968). *Histoire des dictionnaires français*. Paris: Larousse.

Mitterand, Henri ; Petit, Jacques (1962). Index et concordances dans l'étude des textes littéraires. *Cahiers de lexicologie 3, 160-176.*

Moreau, René (1962). Au sujet de l'utilisation de la notion de fréquence en linguistique. In: Quemada, Bernard (ed.). « Actes du *Colloque international sur la mécanisation des recherches lexicologiques* – Réalisations récentes et nouveaux équipements, Besançon, juin 1961 », *Cahiers de lexicologie 3, 140-175.*

Muller, Charles (1962). Les index de vocabulaire. *Bulletin des Jeunes Romanistes IV(2), 9-14.*

Muller, Charles (1963b). Le MOT, unité de texte et unité de lexique en statistique lexicologique. *Travaux de linguistique et de littérature romanes 1, 155-173.*

Muller, Charles (1964). Réflexions sur la fréquence des mots. *Praxis 1964-1, 23-29.*

Muller, Charles (1965). Quelques principes de linguistique statistique: expérience, calcul et prévision. *Bulletin des jeunes romanistes 11, 7-14.*

Muller, Charles (1968). La statistique lexicale. *Langue française 2, 30-40.*

Muller, Charles (1970). Sur la mesure de la statistique lexicale. *Études de linguistique appliquée nouvelle série 1, 20-46.*

Muller, Charles (1971). Fréquence des signifiés ou fréquence des signifiants. *Études de linguistique appliquée nouvelle série 2, 74-87.*

Muller, Charles (1973). Le *Trésor de la Langue Française* et la statistique lexicale. *Travaux de linguistique et de littérature IX(1), 85-95.*

Muller, Charles (1978). Fréquence des mots et statistique lexicale. *Le français moderne 1, 1-5.*

Muller, Charles (1979). *Langue française et linguistique quantitative — Recueil d'articles.* Genève: Slatkine.

Muller, Charles (1981). La linguistique quantitative et le centre de Philologie Romane. *Bilans et perspectives.* 1980-1981, Fascicule 25-26, 95-102. Strasbourg: Centre de Philologie romane, Université de Strasbourg.

Muller, Charles (1985). Lemmatisation et informatisation. In: Bouton, Charles P.; Brunet, Étienne; Calvet, Jean-Louis (eds.), *Hommage à Pierre Guiraud*, Université Simon Fraser, Canada, Université de Nice, INaLF – CNRS, *285-291.* Paris: Les belles lettres.

Muller, Charles; Pottier, Bernard (eds.) (1966). *Statistique et analyse linguistique – Colloque de Strasbourg (20-22 avril 1964).* Paris: Presses universitaires de France.

Pierrel, Jean-Marie (2010). À la découverte d'un trésor: le mariage de l'informatique et de la lexicographie au service de la valorisation de la langue française. Manuscrit auteur, *Conférence de l'Académie Stanislas, France (2010),* 18 p.

Pierrel, Jean-Marie, *La Préface du TLFI*, *www.atilf.fr/IMG/pdf/La_preface_du_TLFi_par_Jean.pdf* accessed 15 june 2014.

Piotrowski, David (ed) (1996). *Lexicographie et informatique – Autour de l'informatisation du* Trésor de la langue française, Actes du colloque international de Nancy (29-31 mai 1995), INaLF. Paris: Didier Érudition.

Piotrowski, David (1996). Opérations hypertextuelles et formes lexicographiques. In: Piotrowski, David (ed.), *Lexicographie et informatique – Autour de l'informatisation du* Trésor de la langue française, Actes du colloque international de Nancy (29-31 mai 1995), INaLF: 319-336. Paris: Didier Érudition.

Présentation du Trésor de la Langue française informatisé, http://www.cnrs.fr/cw/fr/pres/compress/atilf/tlf.htm accessed 13 may 2015

Pruvost, Jean (ed.) (1995). *Les dictionnaires de langue – Méthodes et contenus*, La Journée des dictionnaires 1994, Cergy Pontoise Université, Centre de recherche Texte/Histoire.

Pruvost, Jean (1997). Avant-propos. In: Pruvost, Jean (ed.). *Les dictionnaires de langue française et l'informatique – La Journée des dictionnaires 1995*, Cergy Pontoise Université, Centre de recherche Texte/Histoire, 7-28.

Pruvost, Jean (ed.) (1997). *Les dictionnaires de langue française et l'informatique – La Journée des dictionnaires 1995*. Cergy Pontoise: Cergy Pontoise Université, Centre de recherche Texte/Histoire.

Pruvost, Jean (2000). *Dictionnaires et nouvelles technologies*. Paris: Presses universitaires de France.

Quemada, Bernard (1959). La mécanisation dans les recherches lexicologiques. 1. Les Inventaires de vocabulaire. Moyens et methods. *Cahiers de lexicologie 1, 7-46.*

Quemada, Bernard (1962a). Les travaux du laboratoire d'analyse lexicologique. In: Quemada, Bernard (ed.). Actes du *Colloque international sur la mécanisation des recherches lexicologiques* – Réalisations récentes et nouveaux équipements, Besançon, juin 1961, *Cahiers de lexicologie 3, 58-63*.

Quemada, Bernard (1962b). L'inventaire lexicographique en vue d'un Thesaurus national. In: Quemada, Bernard (ed.), Actes du *Colloque international sur la mécanisation des recherches lexicologiques* – Réalisations récentes et nouveaux équipements, Besançon, juin 1961 ». *Cahiers de lexicologie 3, 119-133.*

Quemada, Bernard (ed.) (1962). Actes du *Colloque international sur la mécanisation des recherches lexicologiques* – Réalisations récentes et nouveaux équipements, Besançon, juin 1961. *Cahiers de lexicologie* 3.

Quemada, Bernard (1973). Bilan des applications de l'informatique aux études lexicologiques. *Meta* 18-2: 87-102.

Quemada, Bernard (1974). Remarques de méthode sur une recherche d'indices d'utilité du vocabulaire. *Le français dans le monde 103, 18-24.*

Quemada, Bernard (1980). Foreword. In: *Trésor de la langue française* (1971-1994) *TLF, Dictionnaire de la langue française du 19^{e} et du 20^{e} siècle* (CNRS), Vol. 8: VII. Paris: Klincksieck, Gallimard.

Quemada, Bernard (1995). « Entretiens sur le thème : 'Des mots aux dictionnaires' », suivi de « Questions et remarques ». In: Pruvost, Jean (ed.), *Les dictionnaires de langue – Méthodes et contenus. La Journée des dictionnaires, 25-42*. Cergy Pontoise Université: Centre de recherche Texte/Histoire.

Radermacher, Ruth (2004), *Le "Trésor de la Langue Française". Une étude historique et lexicographique.* Thèse, Strasbourg.

Rondeau, Guy (ed.) **(**1968**)**. *Linguistique et mathématique*. Textes des communications données lors du Colloque de linguistique mathématique, Sherbrooke, 3 novembre 1967. Montréal: Presses de l'université de Montréal.

Sinclair, John (2004). *Developing Linguistic Corpora: a Guide to Good Practice – Corpus and Text — Basic Principles*, Tuscan Word Centre. http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm, © John Sinclair 2004. Accessed 13 April 2015.

Trésor de la langue française (1971-1994). *TLF, Dictionnaire de la langue française du 19^{e} et du 20^{e} siècle* (CNRS). Paris: Klincksieck, Gallimard.

Wagner, Robert-Léon (1967). *Les vocabulaires français*. Paris: Didier.

Wagner, Robert-Léon (1970). *Les tâches de la lexicologie synchronique, glossaires et dépouillements. Analyse lexicale*. Paris: Didier.

Zampolli, Antonio (1990). Foreword. In: Cignoni, Laura; Peters, Carol (eds.), *Computational lexicology and lexicography – Special issue dedicated to Bernard Quemada*, 1, *Linguistica Computazionale* VI, XIII-XV. Pisa: Giardini Editori e stampatori.

Zampolli, Antonio; Calzolari, Nicoletta (1996). LE-PAROLE: Its history and scope. *The ELRA Newsletter 5-6.*

Lexicométrie: A Linguistic School in France in the 1960s-1980s. History, Theories and Methods

Sylvain Loiseau
Sedyl Laboratory (UMR 8202 CNRS/Inalco)
sylvain.loiseau@univ-paris13.fr

0. Introduction

Lexicométrie is a discourse analysis school founded in France during the 1960s that focuses on the study of vocabulary. It developed statistical methods for the analysis of the lexicon and, in particular, statistical methods to describe the contexts of words in corpora.

In a narrow sense, *lexicométrie* refers to a linguistic 'school' defined by several theoretical and institutional features, such as (1) an object (political discourse and political texts), (2) a theory (language is highly political and struggles between ideologies also take place in discourse), (3) an hypothesis (statistical analyses of the lexicon may help to uncover and to study objectively the ideological opinions hidden in the discourse), (4) a method (statistical procedures designed for the analysis of the contexts of words) and (5) an institution, the "École Normale Superieure de Saint Cloud", a graduate school and a laboratory founded at this institution. This narrow sense corresponds to the early period, from the mid-1960s to the mid-1980s.

Lexicométrie has been seminal up to today in France, mainly in the field of text analysis and amongst linguists working on French. It is mentioned in several chapters of this volume (Beaudouin, in connection with the field of data analysis, Mayaffre, in connection with the analysis of text by historians). It continues to influence the way corpus linguistics is currently developing in France. The term itself is still in use, with a broader meaning than that of a school: *lexicométrie* today refers more loosely to a methodology for the interpretation of the content of textual corpora using software designed for the end-user and implementing the statistical methods developed earlier for the analysis of words in context. The term is not limited to linguists but is also used in other fields (mainly history, but also sociology and literary studies). In this broader meaning, *lexicométrie* is less closely linked with a specific theory of language and a specific hypothesis, but some features of the *lexicométrie* school remain influential, such as the focus on text corpora, on the interpretation of content, and on vocabulary.

Lexicométrie in the narrow sense is strictly speaking not a *linguistic* theory. Its relation with linguistic frameworks (e.g. discourse analysis, sociolinguistics, distributionalism) is eclectic[1] and its theoretical content more general and political, placing stress on the relation between the lexicon and power (see section 5 below). *Lexicométrie* is not related to the modelling or computation of quantitative data *per se*, and is also distinct from the use of quantitative data in the applied field of NLP. Rather, the focus was on the interpretation of frequencies in order to analyze the cultural content embedded in the lexicon. Some original and carefully designed statistical models were proposed (see section 9). Due to the seminal influence of *lexicométrie*, there is no equivalent of "Quantitative Linguistics" in France (the term *linguistique quantitative* is seldom used). Apart from pioneers in the field of quantitative analysis (see Bergounioux, this volume; Léon, this volume), there has been no field of research in France concerned with the analysis of quantitative constants in linguistic data, such as the Zipf law. Except in applied domains such as NLP, all the uses of statistics in linguistics are made with reference to the sociological and historical context of the texts under scrutiny (cf. section 4).

The history of *lexicométrie* is interesting for several reasons. The *lexicométrie* school is representative of some of the features of scientific and intellectual life in France in the wake of the 1968 uprising. It is also representative of the impact of an institution (the elite graduate school ENS St-Cloud) on intellectual life. Its long-term effect on the practices of some subfields of linguistics in France needs to be documented. The scientific aims and the underlying ideology of *lexicométrie*, while they now seem to belong to the past of linguistic ideas, continue nevertheless to influence research in contemporary France in the field of discourse analysis applied to corpora of the French language.

In retracing the history of the *lexicométrie* school, many accounts by the actors themselves can be called upon (e.g. Tournier 1969, 1976, 2010; Bonnafous/Tournier 1995). Of course, reflexive texts by those involved may not always give a faithful account.

This contribution aims at outlining the context of *lexicométrie* and its original objectives as well as presenting the main methods developed in the field of lexical statistics.

1. *Lexicométrie* is lexicology

The development of *lexicométrie* is rooted in a long-standing avenue of research in France that focused on the lexicon – or rather the vocabulary – as a window onto historical and cultural issues. *Lexicométrie* pursued the project of ideological analysis of the lexicon and the development of statistical analyses

[1] Such multidisciplinary frameworks were used by several research groups in France in the post-structuralist context of the 70s and 80s. For instance, another research group, led by Michel Pêcheux, called "automatic discourse analysis", included researchers from several disciplines. It worked on the automatic analysis of political discourse. Unlike the *lexicométrie* group, they were reluctant to use statistical methods, which imply an unequivocal link between signifier and signified (Pêcheux *et al.* 1982; Pêcheux 1969; Léon 2010).

was not conceived as a departure from that project but rather as an aid in achieving exhaustivity and systematicity. *Lexicométrie* used statistics as a method, while its descriptive aims remained the lexicon, focusing on political texts. The name of the laboratory emphasizes this continuity: originally called "lexicologie politique" (1966-1975), it became "lexicométrie politique" (1975-1980) and then "Lexicologie et textes politiques" (1980) (see section 2).

The tradition continued by *lexicométrie* started with Ferdinand Brunot (1860-1938). Brunot was a professor of the history of the French Language at the Sorbonne and an influential scholar. He contributed to the development of the field of French lexicology, "writing the history of French society while writing the history of its vocabulary" (Chevalier 2006: 205; cf. Tournier 2010: 213). He showed that the French revolution had a strong effect on the lexicon and started a systematic, fine-grained study of the lexicon at various periods in the history of the French language. Several landmark studies followed Brunot (their titles often included both "vocabulary" and "society" or "social"), for instance Matoré (1908-1998), who introduced the notion of the "mot-témoin" (word as a witness) in *Le vocabulaire et la société sous Louis-Philippe,* published in 1951[2] (he also published an influential *La méthode en lexicologie*, cf. Tournier 2010: 222) or Jean Dubois with *Le Vocabulaire politique et social en France de 1869 à 1872* in 1962 (cf. Tournier 2010: 214). For the "new lexicology" of the 1950s, "lexical data, in particular changes in the lexicon, were assumed to allow one to study social or ideological reality" (Chevalier 2006: 227). These studies focused on the *vocabulary*, a set of lexemes actually used by speakers in a given socio-historical context, rather than on *lexicon*, viewed as an abstract set of lexemes that do not match any actual usage. Another important influence was the work by Mario Roques who "was in France the defender of an anti-lexicographical ideology" (Chevalier 2006: 212). *Lexicométrie* followed in their footsteps.

The director of the laboratory in which *lexicométrie* was first developed in the 1960s was Robert-Léon Wagner (1905-1982)[3]. While supporting younger researchers' interest in the use of quantitative methods for the exhaustive analysis of corpora, he was himself reluctant to engage in quantitative studies. He was a follower of Brunot (Tournier 2010: 213-214), and contributed several books on the description of vocabularies and the history of the French language such as *Les vocabulaires français*, 1970, an introduction to the study of the vocabulary.

A comparison can be made with the contextualist school in the UK. Both the contextualist school and *lexicométrie* followed an avenue of research rooted in pre-quantitative work (Firth's work for the former, studies of vocabularies for the latter). Moreover, both of these avenues of research were independent of quantitative analyses: Firth was not a proponent of quantitative methods (Tognini-Bonelli 2001) and similarly a scholar such as Wagner was not interested in quantitative measurement of the vocabulary, but he helped the new team to emerge (Tournier 2010: 214).

[2] Followed by Matore 1985, 1988.

[3] Cf Chevalier & Encrevé 1984: 71. Wagner had already been a member of the team created to devise a "basic French" in 1952, also at the Saint-Cloud graduate School (Chevalier & Encrevé 1984: 79-80; cf. Léon, this volume).

2. The institutional development of *lexicométrie*

As already mentioned, *lexicométrie* was first developed in a laboratory hosted at a elite higher education establishment, the École normale supérieure de Saint Cloud. Work on the quantitative analysis of the vocabulary of political discourse started there in 1964 under the direction of R.-L. Wagner. After three years of informal investigations, an official laboratory was created on 1st January 1966 at ENS St-Cloud, named "Lexicologie politique" headed by R.-L.Wagner (1905-1982) (Tournier 1976: 11-12; Tournier 2010: 214-215; Tournier 1969; Bonnafous/Tournier 1995).

A year later, in 1967, it was decided to focus on the quantitative and machine-assisted analysis of vocabularies (Tournier 1969) under the impetus of young researchers such as Maurice Tournier (1933-2014). The aim was to make machine-readable (mechanographic) versions of political corpora on which an exhaustive quantitative analysis of the lexicon could be based. Four families of methods were used at that time (Tournier 1976: 13): (1) classical statistical studies of word frequencies (such as Guiraud, Muller or Herdan: see Bergounioux, this volume and Léon, this volume); (2) factorial correspondence analysis (see Beaudouin, this volume); (3) distributionalism *à la* Harris and, lastly, (4) the study of co-occurrences in order to analyze the context of words. The latter approach was the main original contribution of the laboratory (see below, section 9), and the one for which *lexicométrie* developed original statistical methods.

In 1975, the laboratory became a joint laboratory of the CNRS and the ENS[4] and was renamed "Lexicométrie politique" (Tournier 2010: 215; Bonnafous & Tournier 1995), still under the direction of R.-L. Wagner. After a decade of developing the quantitative analysis of vocabulary, the laboratory finally included the word *lexicométrie* in its name. In 1977, M. Tournier became head of the laboratory (Bonnafous & Tournier 1995). In 1980, the laboratory changed its name again, to "Lexicologie et textes politiques" (Tournier 2010: 220-221). During the following three years, twelve PhD theses were defended by young members of the laboratory (Tournier 2010: 221) and a journal was launched (*Mots*, cf. below section 7).

In 1992, the laboratory[5] was named "Lexicométrie et textes politiques" (Bonnafous Tournier 1995), under the direction of A. Geffroy and P. Lafon.

According to Tournier (2010: 217), the word *lexicométrie* appeared for the first time in unpublished texts produced by the laboratory (cf. below) in the early 1970s. The word appeared in the title of Tournier's thesis (Tournier 1975): *Un vocabulaire ouvrier en 1848. Essai de lexicométrie* (*The Vocabulary of Workers in 1848. An Essay in Lexicometry*)[6].

The *lexicométrie* school is representative of some of the features of the scientific and intellectual life in France in the aftermath of the 1968 uprising. First of all, most people involved were not trained as linguists in the current

[4] L.A. 246

[5] UMR 9952 of the CNRS (INaLF) and ENS de Fontenay Saint-Cloud

[6] "Lexicométrie" is also found in Matoré (1953: 82), but with a slightly different meaning and there seems to be no connection.

sense. They were trained in literary studies or the history of the French language. At that time, only a few linguistics departments existed (the first linguistics department was created at Nanterre University in 1968). Wagner was an historian of the French language. Other people were trained in French literature (such as Maurice Tournier or Benoît Habert). Moreover, the ENS institutions did not favour disciplinary specialisation and new disciplines, but rather general intellectual skills irrespective of "technical" disciplinary frameworks. Hence, the theoretical basis of *lexicométrie* was not a linguistic one, but a more general theory of language as related to power (see below, section 5).

Moreover, from the outset, *lexiométrie* had close ties with other fields interested in the interpretation of texts (mainly history, see Mayaffre, this volume). Early contributors such as Guilhaumou were historians. Mathematicians also joined the team in order to develop statistical tools (André Salem, Pierre Lafon). The *lexicométrie* team had few linguists at the beginning.

The management of the laboratory is illustrative of the behaviour of academics in an institution such as ENS St-Cloud and of their political orientation. The laboratory tried to adopt a very collective management: decisions were made as collectively as possible and discussions were supposed to be conducted irrespective of academic rank. (Tournier 2010: 14). Another illustration of the intellectual context of these years is the decision not to foreground any particular individuals but to focus on the collective nature of the research (Tournier 2010: 217). A long collective work on political flyers during the 1968 uprising was published under the title *Des tracts en mai 68* (Political flyers during May 1968) with no proper names on the cover. Inside the book, the names of the six authors are given in alphabetical order, with no distinction between full professor and research engineer.

3. *Lexicométrie* and historians

Beside Dubois' thesis, another important book that influenced the development of *lexicométrie*, according to Tournier (2010: 15) was a book by the historian Antoine Prost (1933-), on the vocabulary of the electoral campaigns in 1881, 1885, and 1889[7].

Due to its focus on the lexicon as a way to study past ideological opinion and change, "vocabulary studies" as described above (and *lexicométrie*) were close to history. As stressed by Chevalier (2006: 208), the direction given by Brunot to the field of lexicology was close to the interests of the field called "new history" in France at that time, and that focused on long-term trends in historical change through the use of large bodies of data. The new historian Lucien Febvre[8] published *Civilisation: the word and the idea*[9] following the

[7] Prost A. (1974) *Vocabulaire des proclamations électorales, 1881, 1885, et 1889*, Paris, PUF, Publications de la Sorbonne.

[8] Lucien Febvre (1878-1956) was the founder with Marc Bloch (1886-1944) of the "École des Annales", that focused on long-term as opposed to event history and included economic and social data.

principle laid down by Brunot (Chevalier 2006: 211). The historian A. Prost stressed that the work by Dubois (1962) "initiated a renewal of interest among historians for language records" (Prost 1974: 6). Symmetrically, the linguist Dubois stated (1962: 192): "due to the so-called 'sociological methods', lexicology does not belong to linguistics; it is a branch of history. The research focuses on keywords or metaphorical figures that characterize and that, to a certain extent, illustrate the way of thinking at a given period of time."[10] In sum, due to its focus on political discourse (cf. next section), its interest in interpreting ideology, and the use of quantitative methodologies, *lexicométrie* was very close to history.

In his 1974 book, that was seminal in the development of *lexicométrie*, Antoine Prost used the factorial analysis proposed by Benzécri (cf. Beaudouin, this volume). The same year, Prost also edited the proceedings of the first conference on political lexicology (ENS St Cloud, 1968; cf. Tournier 2010: 218 and note 31).

Collaboration between historians and practitioners of *lexicométrie* was frequent. For instance, Jacques Guilhaumou (born in 1948), after a PhD thesis in history, became a member of the Saint-Cloud laboratory from 1982 to 1992. Damon Mayaffre, also an historian (see Mayaffre, this volume), is a member of a linguistics laboratory. *Lexicométrie* has been thoroughly presented in manuals written by historians (such as Robin 1973, Claire & Lemercier 2008). Mayaffre (this volume) deals with this point more fully. Some researchers in *lexicométrie* published in historical journals[11].

Some "New historians" shared with the practitioners of *lexicométrie* an object (political discourse), a method, and also a tendency to be scientific, with the aim of providing an exhaustive analysis of vocabulary (see below, section 6).

4. Object: political discourse

Another key feature of *lexicométrie* is its focus on political discourse. During the 1960s, encouraged by R.-L. Wagner, the focus of the group shifted toward popular (spontaneous) texts (Tournier 2010: 214). The four studies that were very influential according to Tournier (2010: 218 and 1998) during the development of *lexicométrie* (Prost 1974, Cottert / Moreau 1969, Roche 1971 and Dubois 1962) are all analyses of political discourse. A look at the theses and at the main corpora built in the laboratory may help in drawing a big picture of the kind of data analysed. First, a lot of work in the laboratory was devoted to

[9] Febvre L. (1930) *Civilisation. Évolution d'un mot et d'un groupe d'idées*, Paris, Renaissance du livre.

[10] "Avec les méthodes dites sociologiques, la lexicologie n'appartient plus à la linguistique ; elle n'est qu'un sous-chapitre de l'histoire: on détermine des mots-clefs ou des métaphores 'caractéristiques' qui, en quelque sorte, symbolisent une mentalité et une époque".

[11] Annie Geffroy (2013) "Les manuscrits de Robespierre", *Annales historiques de la Révolution française*, 2013/1 (n° 371).

building corpora of all the revolutions in France since 1789 (apart from the July Revolution in 1830):

- 1789 (French Revolution) (Robespierre and Marat corpora by Annie Geffroy[12]; the *Père Duchesne* journal by Jacques Guilhaumou),

- 1848 (French Revolution of 1848) (Tournier 1975, thesis),

- 1870 (Paris Commune) (Ida Porfido *La mise en scène du Peuple de la Commune par Jules Vallès*, PhD thesis, 1995, Paris 3 University),

- 1968 (May 1968 events) political flyers of May 1968 (Gabrielle Muc).

Other studies and theses focused on political movements:

- the Dreyfus Affair (J.-P. Honoré, *Le discours politique et l'Affaire Dreyfus. Étude des vocabulaires, 1897-1900*, 1982, thesis, Paris 3 University)

- French trade unionism: CGT, CFTC, CFDT, FO (Josette Lefèvre; B. Habert, *Les Résolutions générales des congrès de la CFTC-CFDT de 1945 à 1979*, 1982, thesis, Paris 3 University; M. Launay, *Le syndicalisme chrétien en France (1885-1940)*, 1981, thesis, Paris 1 University);

- Communists (Denis Peschanski, *Discours communiste et 'grand tournant'*, 1981, thesis, Paris 1 University);

- Socialists (S. Bonnafous, *Les motions du Congrès de Metz du PS : processus discursifs et structures lexicales,* 1980, thesis, Paris 10 University)

- Pétain (Gérard Miller, *Les Pousse-au-jouir du maréchal Pétain*, 1975, thesis, Paris).

Other authors, most of whom were involved in the current history of French political ideas, were also studied: Malebranche (Majid Sekhraoui), Rousseau (Michel Launay), Montesquieu (Jean-Marie Goulemot), Voltaire (Georges Mailhos), Auguste Comte (Raymond Lallez (PhD thesis 1980)), Hugo (Nelly Danjou), Stendhal (Jean-Marie Gleize), Vallès (Ida Porfido), Eluard (Marie-Renée Guyard), surrealists (Danièle Bonnaud-Lamotte, Henri Béhar, Jean-Luc Rispail) (cf. Tournier 2010: 214).

Other theses applied the methodology to political situations in other countries (Miloud Bel Cadi *Le mot démocratie dans le discours électoral de 1977 au Maroc: analyse des réseaux sémantiques,* 1986, thesis, Paris 3 University; Irène Rabenoro *Le vocabulaire politique malgache pendant les évenements de Mai 1972*, 1995, thesis, Paris 3 University; B. Kadima-Tshimanga, *L'univers sociopolitique de l'*Évolué congolais *entre 1955 et 1981. Une étude du vocabulaire de La Voix du Congolais*, 1983, thesis, Paris 3).

A large collective undertaking was the analysis of the flyers that flourished during the 1968 uprising. This work was the opportunity to develop complex procedures including sampling, translating into machine-readable format, categorising, etc. The first outcome of this work was in Demonet et al. (1975, 1978) and fed work up to Tournier 2007. This large editorial project was also one of the founding principles of the journal *Mots* (see below) (Tournier 2010). The work

[12] Geffroy A., 1974, " Formes de base et formes spécifiques dans le discours robespierriste", *Cahiers de lexicologie*, no 2[5], p. 96-116.; Geffroy A. 1980, "Trois successeurs de Marat pendant l'été 1793. Analyse lexicométrique des spécificités", *Mots. Les langages du politique,* no 1, p. 167-187.

was characterized by very strict procedures that delayed publication of the results for many years, “a delirium of excessive methods” (“Un délire de méthodes excessives”) according to Tournier (2010: 216).

This focus on political movements and leaders was another reason why *lexicométrie* established strong ties with historical research.

5. Language and power: the political interpretation of frequency

As mentioned above, the *lexicométrie* school was rather eclectic with respect to theoretical frameworks. Tournier (2010) mentions as sources of inspiration Discourse Analysis (Pécheux 1969; Dubois 1962; cf. Léon 2010, 2014), Harris distributionalism, sociolinguistics, natural language processing (cf. Tournier 2010: 217-218), but also psychoanalysis. Distributionalism *à la* Harris was used for the analysis of corpora, but as a method rather than as a theoretical foundation.

All in all, *lexicométrie* is not a linguistic theory. However, it did include a theory about language. This theory is not strictly a linguistic one; rather, it is a holistic view of language as wrought by political struggles. In the *lexicométrie* framework, political discourse, but also language in general, is viewed as a locus of political conflict. Speakers' ideologies are reflected in their productions and language is a place where speakers and social groups struggle or negotiate for the dissemination of their own ideology. Dubois (1962: 49) claimed that “the lexicon reflects the economic, social and political relations that link the various social classes”[13], and Prost (1974: 15) enquired “at what point does a variation in frequency cease to be due to chance and require a political explanation? ”[14]

Words or meanings are seen as real objects exchanged in discourse. Some claims for such a theory are in the following quotes, which may sound rather exotic today: “all the research done at the Saint-Cloud laboratory tended to prove, through statistics as well as through discourse analysis, that politics is a struggle for the stabilisation or the destabilisation of the language, where there are only false truces on the value [of words]” (Bonnafous & Tournier 1995: 69)[15]; “If we have to explain the place of lexicometric research in a theory of

[13] “Le lexique, objet de cette étude, traduit les rapports d’ordre économique, social et politique qui existent entre les diverses classes de la société.”

[14] “ […] à partir de quel moment une différence de fréquence cesse-t-elle d’être imputable au hasard et autorise-t-elle une interprétation politique ? L’idée fondamentale de la comparaison, en effet, est que toutes choses (public, sujet, genre littéraire) étant égales par ailleurs, si l’orientation politique des locuteurs n’avait aucune influence sur leur vocabulaire, ils puiseraient de façon aléatoire à l’intérieur du stock lexical, en l’on devrait constater entre les fréquences d’emploi des différents vocables des écarts minimes, explicables par le seul hasard. Si les écarts sont accentués, une interprétation politique est justifiée […]”

[15] “Que la politique soit une lutte pour la stabilisation ou la déstabilisation langagières, où n'existent que de faux armistices sur des valeurs de langue, tous les types de recherche pratiqués au laboratoire de Saint-Cloud le montrent, qu'il s'agisse de constats statistiques ou d'analyses discursives.”

politics as conflict and a theory of language as discord, we can say that the goal of lexicometric research is to analyse, through the comparison of text corpora, how the terms that are used (exchanged) in the public space, when power is at stake, reflect struggles for the ownership of symbols that are at play in the interaction"[16] (Bonnafous & Tournier 1995: 69).

6. The quest for a scientific and semi-automatic interpretation of political discourse

The bulk of work in the *lexicométrie* school was devoted to the interpretation of (political) texts with the help of quantitative data. Tension arose and was constantly at work between the quest to be scientific (hence the use of quantitative data) and the aim of analyzing meaning in context and the ideological content of political discourse.

6.1. A quest to be systematic and objective

The development of quantitative methods in *Lexicométrie* was due to the search for a more "objective" and "scientific" interpretation of political texts. This aim can be seen as echoing, at the same period, other attempts to turn social science into a predictive science through theories such as Marxism (in the field of linguistics, Marxists such as Michel Pêcheux were very influential and interested in automatism in discourse analysis: Pêcheux 1969; Pêcheux et al. 1982; Léon 2010, 2014). Both distributionalism and exhaustive quantitative analysis were seen as a means to achieve more objectivity in the field of the history of ideologies and the analysis of textual data. The quest for rigorous quantitative methods was perhaps seen as a way to escape political subjectivity in the analysis of political texts.

Another strong motive for the use of formal and quantitative methods may have been the institutional context. It was only in 1968 that the first linguistics departments were created in French higher education (see above) and most of the first generation of linguists was trained in literature and eager to distinguish themselves from literary methods and objects of study.

This quest to be scientific and objective was stressed by R. Robin (Guilhaumou - Maldidier - Robin 1994): "rigorous method would allow us to break the vicious circle of presuppositions, insinuations and implications that were pervasive in the debates about the French Revolution."[17] In his pioneering 1962 work, Dubois also argued strongly that the exhaustive quantitative analysis

[16] "S'il faut situer la recherche lexicométrique à l'intérieur d'une théorie conflictuelle du politique et d'une théorie dissensuelle de la langue, nous dirons qu'elle est chargée d'examiner, à partir de corpus de textes soumis à comparaison, comment les termes échangés dans l'espace public autour des enjeux de pouvoir rendent compte des luttes d'appropriation ou de dépossession symboliques qui se jouent dans le lieu même de l'échange."

[17] " […] des méthodes rigoureuses nous permettaient de sortir du cercle vicieux des présupposé, des sous-entendus et de l'implicite à l'œuvre dans les débats concernant la Révolution française."

of the lexicon of texts would result in greater objectivity: “A comprehensive survey makes it possible to objectively list the oppositions, identities and phrasal units that form the structure of the lexicon. In fact, this Cartesian principle, and its effectiveness in science, has never been questioned, but it has long been held that it was physically impossible to apply it in the field of vocabulary.” (Dubois 1962: 192)[18]

This quotation stresses the role of exhaustivity in achieving objectivity. Tournier (2010: 216) mentions the conditions of the survey that was published as (Demonet *et al.* 1978): “We were fundamentalists of “scientificity” and of the distancing necessary for analyses of political vocabulary: building ‘neutral’ and ‘representative’ corpora, categorizing textual forms and complete counts, comparing statistics, sorting probabilities, building contexts and repeated segments [see below], identifying indices of repetition, […] indices of proximity, […] lexicogrammes, connexion trees... We wanted to ‘cool down’ research and to reduce interpretation to a minimum, while focusing on the implementation of sampling techniques, sorting and processing.”[19]

Another influential source for the use of exhaustive counting by quantitative methods was the work by Gougenheim in building a “Basic French” (Léon, this volume). He stressed the methodical and exhaustive observation of word distributions rather than picking on epiphenomena (Tournier 2010: 213).

6.2. The interpretation of frequencies

Rather than being related (only) to general laws, observed frequencies were interpreted as reflecting cultural, social or political historical positions.

Exhaustivity in the counting of lexical phenomena is one of the central features of *lexicométrie*. It governs the choice of analysing corpora comprising complete texts that have been carefully sampled rather than analysing frequency as a general property of the language.

As *lexicométrie* is based on the idea of the exhaustive and comprehensive tabulation of frequency, it shows little interest in the idea of general frequencies, related to the language itself, and not related to a text, a corpus, a norm (Lafon

[18] “Le relevé exhaustif permet de dresser objectivement la liste des oppositions, des identités et des unités syntagmatiques qui forment la structure du lexique. En fait, ce principe cartésien n’a jamais été mis en cause, non plus que son efficacité dans le domaine scientifique, mais on a longtemps jugé qu’il était matériellement impossible de l’appliquer dans le domaine du vocabulaire.”

[19] “Intégristes de la ‘scientificité’ et de la distanciation nécessaires aux analyses du vocabulaire politique (construction de corpus ‘neutres’ et ‘représentatifs’, dépouillement des formes textuelles et comptages exhaustifs, concordances et contextes, comparaisons et parentages statistiques, tris en probabilité, construction automatique des cooccurrences et des ‘segments répétés’, indices de répétivité, de cadences d’apparition, de proximité, de ‘figement’, probabilisation des fréquences et des cofréquences, lexicogrammes, arbres de connexion...), nous avons trop voulu ‘refroidir’ la recherche en ne suggérant qu’un minimum d’interprétations et en axant la publication sur les techniques d’échantillonnage, de dépouillement et d’analyse mises en œuvre.’”

1980: 127) or usage (Tournier 1980: 194), understood as "a system of habits that trigger the use of words in context."

Hence, *lexicométrie* does not seek to uncover a general law of frequency, such as the Zipf law, that determines a general quantitative profile, irrespective of social determination.

It is worth mentioning that these frequency analyses were not intended to sum up the meaning and the content of discourse. Meaning was distinguished from word frequency. Through comparison, frequencies provided a diagnostic tool of positions and strategies, but were not conceived of as summarizing the content: "The meaning in context is too subtle to be modelled. A lexicometric study has to start with the mere gathering, the comparison and the opposition of occurrences of lexical units. Questions about meaning have to be asked later, through the analysis of the statistical results. This analysis is qualitative and focused on the context, but it is no longer guaranteed by statistics. Statistical analyses indeed must be done on stable units. The content of a word is not stable, only the graphical form is. This is why the *lexicométrie* laboratory chose to work first on textual surface forms and not on the contents or the referents of the words. It is a textualist orientation, both arbitrary and pragmatic." (Bonnafous & Tournier 1995: 69)

Frequency, however, did receive a great deal of attention. Since researchers in the lexicometric school were mainly interested in the (ideological) content of texts, and in a kind of analysis that pre-dated the use of quantitative methods, it was not easy to define rules for frequency analysis. In many articles, some of the researchers tried to give a political meaning to frequency, even cautiously. In an article devoted to the question of the interpretation of frequencies, Tournier (1980) for example, in a style that is, again, hard to understand today, explained that: "in an interaction, there is no place for a class-specific language (most of the time speakers have received the same language training); but, together with the slight drift of the referents beneath the words, can we speak about class-specific frequency? Frequency and co-occurrence may be systems whose rhythm, irrespective of the logical content, produces not meaning but contagion"[20]

7. The *Mots* journal

The journal *MOTS* was created in 1980. The name MOTS ("words") was written in capital letters during the early years of the journal and stood for "Mots, Ordinateur, Textes, Société" (words, computers, texts, society). The title reflected, again, the strong commitment of *lexicométrie* to lexicology. It followed the journal "Travaux de lexicologie et de lexicométrie politique" (created in

[20] "Au sein d'une situation d'échanges, il n'y a pas place pour des langues de classe (les émetteurs disposent le plus souvent d'un bagage formellement semblable) ; mais, combinés aux glissements insidieux des référents sous les dits et les compris, n'existerait-il pas des usages fréquentiels de classe? Fréquences et co-fréquences ne constituent-elles pas des systèmes dont la rythmique, au-delà du texte logique, produirait non un sens mais une contagion?"

1976). The journal was initially intended to help publish the work of the laboratory on 1968 flyers (Tournier 2010).

8. The statistical methods developed

8.1. Quantitative methods

As stated above, *lexicométrie* is first and foremost a lexicology interested in analysing social values embedded in the use of the lexicon. *Lexicométrie* uses quantitative methods in order to assist interpretation and to turn it into something that aims to be more "objective". Thus, statistics is an interpretative tool, not a model of the language or even of textual phenomena and, as already stated, *lexicométrie* was not interested in general quantitative laws in the lexicon such as the Zipf law.

Two mathematicians were particularly influential, Jean-Paul Benzécri (born in 1932, cf. Baudouin, this volume) and Georges-Théodule Guilbaud (1912-2008). Several members of the laboratory attended Guilbaud's courses. From 1955, he was Directeur d'étude (senior researcher) at the "École pratique des hautes études" (a graduate school). He founded a research group on mathematics for social sciences and influenced several researchers in the social sciences (social choice theory, kinship systems in anthropology (with Claude Levi-Strauss), psychoanalysis (with Jacques Lacan), as well as a composer interested in mathematics such as Iannis Xenakis). He advocated the use of the hypergeometric distribution in the lexicometric method called *specificités* (cf. below). Two journals were seminal: *Cahier de l'analyse des données* (edited by Benzécri) and *Mathématiques et sciences humaines* (edited by Guilbaud).

Due to the function of statistics in *lexicométrie*, linguists using this framework were not trained in statistics themselves. The statistical methods were developed by engineers or mathematicians, mainly two of them: Pierre Lafon (trained at the Institut polytechnique de Grenoble) and André Salem (trained in the USSR). A third engineer, Majid Sekharoui, was a computer scientist. Lafon eventually headed the laboratory in the 1990s. The three of them completed their PhD theses during their stay in the laboratory:

- Pierre Lafon, *Automatisation des dépouillements et études statistiques sur le vocabulaire*, 1981, Paris 3;
- André Salem, *Méthodes de la statistique textuelle*, 1993;
- Majid Sekhraoui, *Concordances : histoire, méthodes et pratique*, 1983**.**

Others contributed to *lexicométrie* slightly later, such as the statistician Ludovic Lebart trained by Benzécri or the computer scientist Serge Heiden.

Several handbooks were written in order to spread the use of these methods: Lafon 1984, Salem 1987, Lebart & Salem 1988, Fénelon 1981, or articles such as Habert 1985, Fiala 1994.

These methods were particularly designed for the analysis of the context of words, i.e. the "co-occurrence clusters in which the units of a text preferentially function" (Tournier 1976: 13) as opposed to other methods available at that time, such as correspondence analysis. They were, in the eyes of their promoters, an "original innovation" (Tournier 1976: 13).

These methods were presented in several papers in the early issues of the journal *MOTS* in the 1980s, but they had been defined and developed earlier: software developed by Pierre Lafon was available as early as 1967 (Tournier 2010: 217). At that time, an expensive computer was purchased at the ENS St-Cloud graduate school.

Some years later, the methods were implemented in end-user software (such as Lexico by André Salem), that made them available to a larger number of linguists. Methods such as *specificités* and *segments répétés* are still largely in use today. The former can be used to assess the attraction of a word for a part of the corpus, or for a word with another word. They are similar to various other measures of word attraction defined in other approaches. The latter is used to extract frequently repeated sequences of words in a corpus. They are close to n-grams.

Below, some of the methods are briefly presented. One of them, *spécificités,* is outlined in more detail in the following section.

8.2. Segments répétés

The work on repeated segments started from questions about the forms in the corpus: how can the text be segmented into lexical units? Due to practical considerations, texts were segmented into lexical units using their (graphical) surface form and white spaces as separators. Obviously, homographs were not distinguished and the inflected forms of the same lexical unit were not conflated, which was one of the limitations of *lexicométrie* (Lafon & Salem 1983: 172). This confusion between *signifiant* and *signifié* was particularly criticised by a competing theoretical approach, "AAD" (Analyse automatique du discours", Pêcheux 1969; Léon 2010; Pêcheux *et al.* 1982).

Since technical limitations impeded the use of lemmatisation, another research avenue was explored, the identification of multi-word lexical units (frozen idioms). Rather than considering the various occurrences of *état* as one lexical unit, identifying complex units such as *état d'esprit* (state of mind) and *capitalisme monopolistique d'état* (state monopoly capitalism) (Lafon & Salem 1983: 172 sqq) was another way of strengthening the link between graphical and semantic units.

A sequence is a suite of (graphical lexical) forms not separated by a punctuation mark.

A biform, triform, etc. is a sequence of two, three, etc. forms. A repeated biform is a sequence of two forms with a frequency equal to or greater than two.

In building the inventory of the repeated segments of a corpus, a segment is listed only if it is not always included within another, longer, segment. For instance if there are five occurrences of *l'ensemble des*, but all of these occurrences are part of a longer segment, *l'ensemble des travailleurs*, the five shorter segments are not listed.

Repeated segments may be extracted and sorted by length, by frequency or alphabetically. This type of information lies halfway between concordance lines and a word index (Lafon & Salem 1983: 176). Repeated segments may also be used for the actual segmentation of the text. As stressed above, distinguishing *état d'esprit* and *capitalisme d'État* helps strengthen the link between graphical forms and semantic units. However, it is not advisable to distinguish too many different multi-word units. For instance, if *unité d'action*, *unité*, etc. are distinguished, a pertinent association between the graphical unit *unité* and the corresponding semantic unit is, again, lost. One method was to consider only the case of words specialised in the context of a multi-word unit. Thus, if *ouvrière* (labour) is found only after *classe* (*social class*) (giving *classe ouvrière*), then the segmentation *classe ouvrière* is adopted.

Repeated segments give a partial account of the combinations of units on the syntagmatic axis, that are otherwise ignored by *lexicométrie*: texts are treated as a "bag of words" in most statistical procedures. It is close to the *n-gram* procedure used in corpus linguistics. This method is still often implemented in current software developed in France for corpus linguistics and general accounts have been proposed (Salem 1987).

8.3. Co-occurrence

Lafon (1981b) proposed a method for assessing the frequency of the co-occurrence between two forms as well as the mean distance between the two forms. Among the three methods assessed in (Lafon 1981b), the best one used an indicator based on the probability of the frequency of the co-occurrence in the contexts of the sentences. The difficulty in building a "co-occurrence distance" indicator was that the two values – co-occurrence association and mean distance – could not be conflated. This is because the method mixed two kinds of phenomena: distant co-occurrences in the sentence on the one hand, and frozen idioms on the other hand. This led to the development of the method of *segments répétés* (see above), better suited for the identification of frozen idioms.

8.4. Rafale

Rafale (Lafon 1981a) is a method developed for the analysis of sequentiality, while other methods focused on frequency and on the text as a "bag of words". The method aimed at quantifying to what extent the occurrences of a lexical phenomenon are regularly distributed over the text, or, on the contrary, are concentrated in some parts. The lexical forms whose occurrences are concentrated are said to appear in *bursts* (*rafales*). This method was used in *La parole syndicale* (Bergounioux *et al.* 1982)

8.5. Correspondence analysis applied to chronological corpora

André Salem (1988) analyzed the case of the factorial analysis of correspondence when applied to corpora of texts that are chronologically ordered (such corpora are termed "chronological textual series"). When analyzed with correspondence analysis, the texts of such corpora appeared ordered on a parabola on the factorial map: the second factor is a quadratic function of the first. This effect is called the Guttman effect, and is the manifestation of data that are fundamentally one-dimensional. Such corpora are one-dimensional due to the high number of forms that evolve chronologically (whose frequency decreases or increases along the chronological axis). Salem contributed to the definition of correspondence analysis when applied to chronological textual series.

8.6. The *Spécificités* method

"*Spécificités*" (specific forms) is a statistical method for selecting which forms (lexical types) are specific of a subcorpus. A subcorpus is called a "part" of the corpus. A form may be positively or negatively specific. A positively specific form is a form "attracted" or "over-used" in the part, whereas a negatively specific form is a form "repelled" or "under-used" in the part. For each form in a corpus, the specificity is calculated. The forms that are below a given threshold and that are not specific are said to be "commonplace". This is the method that had the greatest influence[21].

8.7. The construction of the specificities measure

In the presentation of the method, Lafon (1980) first recalled that absolute frequency is not a reliable indicator of the strength of the link between a form and a part, since obviously, a form that is very frequent in a part may not be specific to this part if it is also very frequent in the whole corpus.

A better modelling of the "surprise" associated with the number of occurrences of a form in a part, given its number of occurrences in the whole corpus, is needed. Lafon chose the hypergeometric distribution as a model of this surprise. This choice was suggested by the mathematician G.-T. Guilbaud (cf supra). The hypergeometric distribution models a sampling without replacement. It calculates the probability of drawing a given number of white balls from an urn that contains a given proportion of white balls during a sampling without replacement of a given number of balls. Let's consider an urn that contains 50 white balls and 50 black balls. We assume that 20 balls are sampled without replacement from the urn. What is the probability of having 0, 1, …, 20 white balls in the sample? Hypergeometric distribution calculates the probability of these different outcomes using combinatorial calculus. The most expected outcome is intuitively known: it is to have the same proportion of white balls in the sample as in the whole urn, hence 10. But the calculation of a given probability requires combinatorial calculus (Fig 1.)

[21] It has been implemented in an R package, WAM (for Word Association Measures, https://r-forge.r-project.org/projects/lexicalstat/).

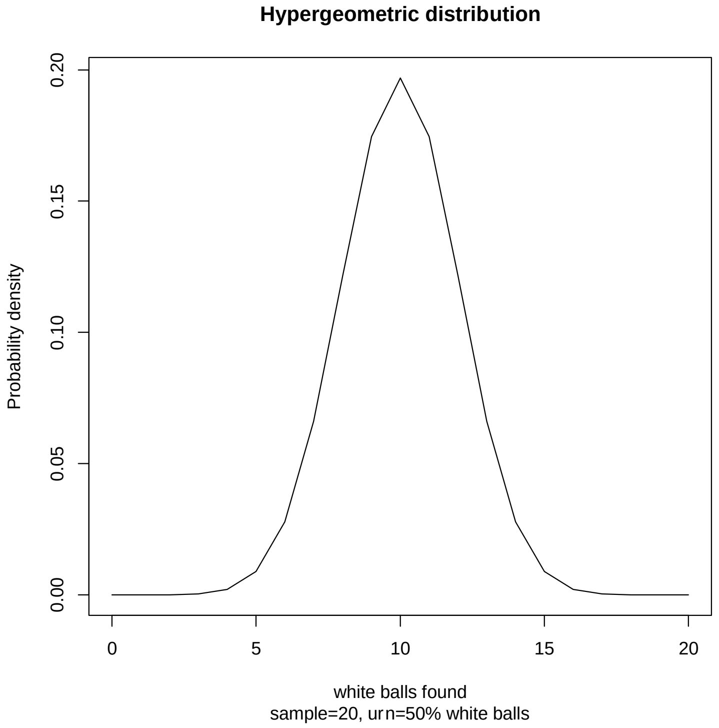


Figure 1: hypergeometric distribution (probability density function)

The model of the urn may be applied to the situation found when trying to calculate specificities of a form in a given part. The corpus is the urn, the part is the sample, and the occurrences of the form are the white balls. Hence, the hypergeometric distribution can be used to calculate the probability associated with the number of occurrences of a form found in a part.

Let us consider the type *peuple* ("people" as in "we, the people") in a corpus of speeches by the French Revolution leader Robespierre. We will focus on the first speech. Is *peuple* over-used, under-used, or commonplace in this part? To answer this question, we need to consider:

- the total frequency F of *peuple* in the corpus, and the total number N of occurrences in the corpus in order to calculate the proportion of white balls in the urn;

- the number of occurrences *n* in the first speech, in order to know the size of the sample.

From these parameters, we can calculate the probabilities associated with the various possible frequencies of *peuple* in the first speeches (Fig 2):

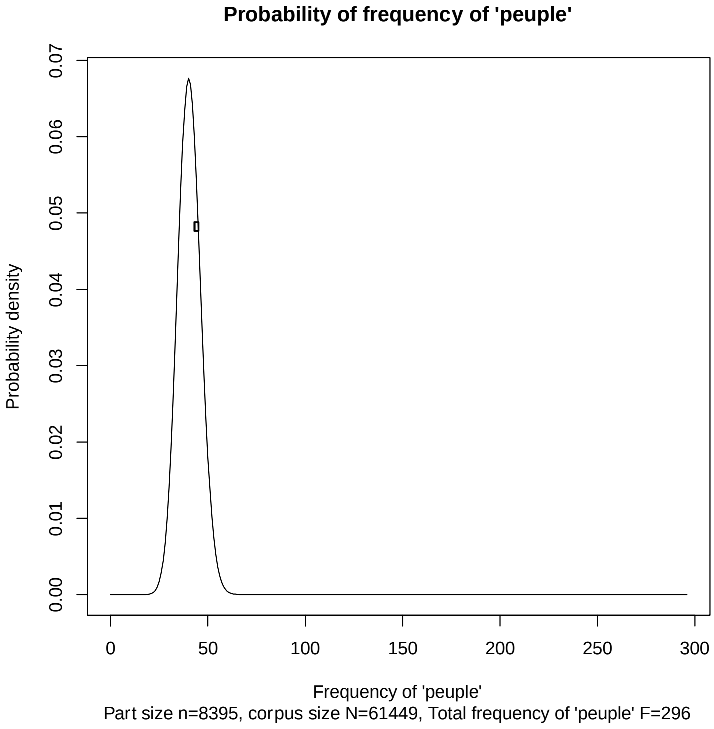


Figure 2. Frequency of « *peuple* »

The frequency of *peuple* in the first speech may vary between 0 and 296, the total frequency of *peuple* in the whole corpus. The mode (the most probable outcome) is 40 occurrences; we found 45 occurrences of *peuple* in this part in fact (plotted as a dot on Fig. 2). The form is therefore used more than expected, with a probability of 5.457782e-02.

However, the raw probability of this exact outcome is not really what we are interested in. What we want to know is the probability of "to be so diverging from the mode", i.e. the cumulative probability of the events with an equal or a greater deviation from the mode. In order to define the specificities indicator, the following computations have still to be done. First, we take the cumulative probability:

- If the observed frequency is less than the expected frequency, we compute the sum of the probabilities for a frequency lesser than or equal to the observed frequency (Prob($X \leq k$)) – i.e., the cumulative probability.

- If the observed frequency is greater than the expected frequency, we compute the sum of the probabilities for a frequency greater than the observed frequency (Prob($X > k$)) – i.e., the cumulative probability for the upper tail of the distribution.

Again, we are not interested in the probability of "drawing exactly 45 occurrences of *peuple*" but in whether "there are at least 45 occurrences of *peuple* in this part" (Lafon 1980: 141).

Second, we use the log of the probability, rather than the probability itself, in order to focus on differences in order of magnitude rather than on epiphenomenal differences.

Last, we add a negative sign if the actual frequency is lower than the mode, and a positive sign if the actual frequency is greater than the mode.

Finally, the graph of the specificities indicator can be plotted with the parameters used below (Fig. 3). The values of the specificity function are plotted for all the possible frequencies of *peuple* in one of the speech by Robespierre. The actual frequency of *peuple* in that speech is shown by a dot:

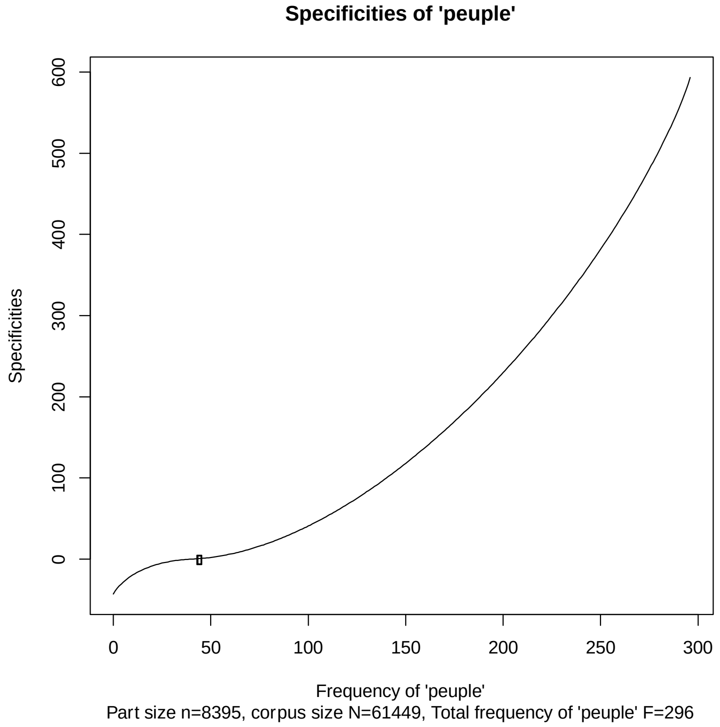


Figure 3. Specificities of « *peuple* »

However, we still don’t know if *peuple* is really a (positive) “specific”, i.e. an attracted form. It may be a “commonplace” form, whose departure from the mode is not significant.

8.8. Computing specificity indicators for all the forms of a part

In order to select (positive as well as negative) specific forms of a part we need to calculate the specificity indicator for all the forms of a part and fix a threshold. The threshold may be an absolute value (for instance, specificities > 2) or a number of forms (for instance, we will consider the first 20 specific forms).

Table 1

The first 20 (strongest) positive specificities for Robespierre's first speech (“*Sur la situation politique de la République*”) in a corpus of ten speeches.

Lexem	Sub-corpus frequency	Corpus frequency	Specificity
france	36	70	29.73
puissances	13	19	15.81

politique	19	57	8.23
autriche	9	19	7.11
français	27	109	6.07
a	61	319	4.84
other	4316	30786	4.61
faire	21	87	4.60
projet	7	19	3.94
conquête	3	5	3.15
croit	3	5	3.15
d	91	565	2.28
au	55	325	2.25
de	464	3173	2.18
ennemis	30	165	2.11
guerre	13	64	1.68
république	35	207	1.61
br1ssot	5	19	1.56
courage	5	19	1.56
ceux	13	67	1.56

Table 2

The first 20 (strongest) negative specificities for Robespierre's first speech ("*Sur la situation politique de la République*") in a corpus of ten speeches.

Lexem	Sub-corpus frequency	Corpus frequency	Specificity
que	94	803	-2.19
bourdon	0	20	-2.19
doute	0	20	-2.19
montagne	0	20	-2.19
dans	55	493	-2.21
hommes	9	109	-2.28
homme	5	73	-2.37
patriotisme	3	56	-2.46
représentants	4	66	-2.61
crime	4	71	-2.90
vertu	5	83	-3.16
ils	43	419	-3.17
justice	3	73	-4.41
publique	2	65	-4.90
fait	2	69	-5.24
nationale	6	124	-6.15
convention	6	126	-6.30
on	33	393	-6.39
patriotes	2	81	-6.79
patrie	6	153	-9.20

Through this list of specific forms, the first speech, which deals with foreign policy and war against other countries, can be compared to the fifth, which deals with domestic issues:

Table 3
The first 20 (strongest) positive specificities in Robespierre's fifth speech ("sur les principes de morale politique") in a corpus of ten speeches.

Lexem	Sub-corpus frequency	Corpus frequency	Specificity
faut	24	70	12.05
gouvernement	34	120	11.52
vertu	24	83	8.68
de2	103	563	8.23
ou	62	305	8.10
est	119	673	7.89
peut	24	91	7.13
par	86	499	5.23
la	405	2788	4.81
peuple	53	296	4.25
il	97	605	3.66
que	124	803	3.40
but	13	58	2.85
a	52	319	2.55
cause	13	61	2.36
politique	12	57	2.28
le	194	1351	2.25
pour	88	595	1.74
tyrannie	13	72	1.46
les	291	2123	1.40

Table 4
The first 20 (strongest) negative specificities in Robespierre's fifth speech ("sur les principes de morale politique") in a corpus of ten speeches.

Lexem	Sub-corpus frequency	Corpus frequency	Specificity
existence	0	19	-2.02
fabre	0	19	-2.02
généreux	0	19	-2.02
genre	0	19	-2.02
laches	0	19	-2.02
main	0	19	-2.02
projet	0	19	-2.02
puissances	0	19	-2.02

bourdon	0	20	-2.09
montagne	0	20	-2.09
ce	37	365	-2.10
convention	10	126	-2.24
un	57	545	-2.35
crimes	3	59	-2.42
et	196	1708	-2.43
français	8	109	-2.43
plus	30	317	-2.55
nationale	9	124	-2.80
comité	6	103	-3.50
other	3708	30786	-19.85

Let us now sum up. The calculation of the specificities measure requires four arguments:

- T: the total number of occurrences in the corpus;
- t: the total number of occurrences in the part;
- F: the total frequency of the form under scrutiny in the corpus;
- f: the frequency of the form under scrutiny in the part under scrutiny.

This indicator can be computed for all the forms of a part and then for all the parts of a corpus in order to compare them.

The method is contrastive, as are most of the methods used in *lexicométrie*. No absolute frequency is used: the parts are compared with the whole corpus, not with another indicator of frequency.

8.9. Using the specificities indicator to study word association

The specificities measure can be used to analyze associations between words as well as an association between a word and a part. The measures of association between words are most often represented using a 2 * 2 contingency table:

Table 5
The parameters of word association computation

	word A	¬ word A	Total
word B			
¬ word B			
Total			

The part (or subcorpus) above can be made of all the contexts of word B. In that case, specificity is used as a word association measure. The values of this contingency table can be computed using the four parameters given previously:

Table 6
Word association computation parameters and specificity parameters

	word A	¬ word A	Total
subcorpus	f	n - f	n
¬ subcorpus	F - f	N – K – (n – k)	N – n
Total	K	N-K	N

Spécificités has frequently been used as an indicator of association between words (e.g. Gréa & Haas 2015, Gréa to appear).

9. Interpretation

Lafon stressed that the use of a probability distribution does not mean that we are doing an hypothesis test. The threshold is not a statistical significance threshold but a practical choice. The specificities indicator is used in order to compare the forms of a part and to assess which one is more attracted than the other: “this procedure may look like hypothesis testing. But the point here is not to accept or to reject a hypothesis, with a given risk of error. Here, the null hypothesis is never expected to be true. That’s why we will use as threshold the *n* first specific forms rather than a given p-value.” [22] (Lafon 1980: 141)

The probability distribution is not used as a significant test because it is not a probabilistic model of reality. Lexical data do not follow a hypergeometric distribution: “in our view, the statistical model, which is completely distinct from the linguistic realm, is unable to represent it accurately. That’s why we do not expect the statistical model to give an approximation of the frequency distribution in the parts of a corpus. It is inconceivable to employ the model to predict the content of a fragment or to predict the sub-frequency of such and such a form in a part of a corpus […] the statistical model is distinct from the linguistic realm. For us, it is nothing more than a tool for measuring the forms that diverge the most from it, so that we give a rigorous account of that reality”[23] (Lafon 1980:

[22] “Cette procédure rappelle celle habituellement nommée ‘test de signification’. Mais il ne s'agit pas ici d'accepter ou de rejeter une hypothèse, en prenant un risque de se tromper. Ici, l'hypothèse d'équiprobabilité n'est jamais contestée. C'est pourquoi nous fixerons le seuil plutôt d'après le nombre de formes qu'il permet de sélectionner que pour sa propre valeur.” (Lafon 1980: 141)

[23] “A notre avis, le modèle statistique, radicalement séparé du fonctionnement linguistique, est, en effet, tout à fait inapte à représenter celui-ci. C'est pourquoi nous ne lui demandons pas de nous fournir une approximation de la distribution des formes à travers les fragments d'un corpus. Il n'est pas imaginable d'employer le modèle pour prévoir la composition d'un fragment ou pour prédire la sous-fréquence de telle ou telle forme dans une partie du corpus […]. Le modèle statistique est de nature totalement étrangère à la réalité linguistique. Il n'est pas autre chose pour nous qu'un instrument de mesure permettant de détecter les formes qui justement s'éloignent le plus de lui, afin de donner une description précise de cette réalité. ” (Lafon 1980: 164)

164; cf. also Lafon 1981a: 186-187; cf. Kilgarriff 2005 for a similar position in the context of corpus linguistics).

10. Conclusion

Lexicométrie is linguistic school that emerged during the 1960s and 1970s committed to the task of uncovering political representations through the quantitative analysis of the words and speeches of speakers taken as representative of social or political positions. The emergence of this school took place in a particular context. First, investigating the relation between the linguistic properties of discourse and political positions was a popular avenue of research at the time, also illustrated by the “Automatic Discourse Analysis” school (Pêcheux 1969). Second, the development of statistical tools for the analysis of digitized corpora was in its infancy, and *lexicométrie* pioneered, in France, practical methods for the quantitative analysis of corpora and the use of computers in linguistics for quantitative research [cf. also Candel, this volume]. A tradition was established that has proved to be influential up to now in France in the design and use of software for the analysis of textual corpora.

This linguistic school can be compared with contextualism. The two schools converge in that both are rooted in pre-quantitative traditions that were sceptical about the use of quantitative methods: vocabulary study on the one hand and Firth on the other. Both worked on the lexicon and they shared some questions about frequency, the focus on discourse rather than on the language system, and the focus on the context. They diverged however in many assumptions, such as the political orientation of *lexicométrie*, its quest for a (quantitatively founded) scientific approach, and its strong multi-disciplinary approach.

References

Bergounioux A., Launay M.-F., Mouriaux R., Sueur J.-P., Tournier M. (1982). *La parole syndicale. Etude du vocabulaire confédéral des centrales ouvrières françaises. 1971-1976.* Paris: PUF.

Bonnafous S., Tournier M. (1995). Analyse du discours, lexicométrie, communication et politique . *Langages 29/117, 67-81.*

Chevalier, J.-Cl., Encrevé, P. (2006). *Combats pour la linguistique, de Martinet à Kristeva. Essai de dramaturgie épistémologique*. Lyon: ENS Éditions.

Chevalier, J.-Cl., Encrevé, P. (1984). La création de revues dans les années 60: matériaux pour l'histoire récente de la linguistique en France. *Langue française 63, 57-102.*

Cotteret J.-M. & Moreau R. (1969) *Le vocabulaire du général de gaulle*, Paris : Armand-Colin.

Demonet M., Geffroy A., Gouazé J., Lafon P., Mouillaud M., Tournier M. (1978 [1975]) *Des tracts en mai 68.* Paris: Champ libre.

Dubois J. (1962). *Le vocabulaire politique et social en France de 1869 à 1872.* Paris: Larousse.

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

Fiala, P. (1994). L'interprétation en lexicométrie. Une approche quantitative des données lexicales. *Langue française 113-122.*

Fiala, P., Lafon, P. (eds.) (1998). *Des Mots en liberté. Mélanges Maurice Tournier.* Fontenay-aux-Roses : ENS Éditions.

Gréa, P. & Haas, P. (2015). Mode de N et type de N, de la synonymie à la polysémie. *Langages 197, 69-98.*

Gréa, P. (to appear) « Inside in French », *Cognitive Linguistics*, 28, 1.

Guilhaumou J., Maldidier, D., Robin, R. (1994). *Discours et archive.* Liège: Mardaga.

Habert, B. (1985). L'analyse des formes spécifiques: bilan critique et proposition d'utilisation. *Mots* 11. 127--154.

Kilgarriff, A. (2005). Language is never ever ever random. *Corpus Linguistics and Linguistic Theory 2(1), 263-276.*

Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. Mots 1, 127-165.

Lafon, P. (1981). Analyse lexicométrique et recherche des cooccurrences. *Mots 3, 95-148.*

Lafon, P., Salem, A. (1983). L'inventaire des segments répétés d'un texte. *Mots 6, 161-177.*

Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots 1, octobre, 127-165.*

Lafon, P. (1981a). Statistiques des localisations des formes d'un texte. *Mots 2, mars, 157-188.*

Lafon, P. (1981b). Analyse lexicométrique et recherche des cooccurrences. *Mots* 3, *octobre, 95-148.*

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

Lebart, L., Salem, A., Berry, L. (1997). *Exploring Textual Data*. Boston: Kluwer.

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

Lemercier, C., Zalc, C. (2008). *Méthodes quantitatives pour l'historien.* Paris: La découverte.

Léon, J. (2014). Questioning the idea of 'Founding Text': Harris's Discourse Analysis and French Analyse du discours. In: Kasevich, Vadim, Yuri A. Kleiner and Patrick Sériot (eds.), *History of Linguistics 2011. Selected Papers from the 12th International Conference on the History of the Language Sciences (ICHoLS XII), Saint-Petersburg, 28 August - 2 September 2011* : 117-125. Amsterdam: John Benjamins Publishing Company.

Léon, J. (2010). AAD69. Archéologie d'une étrange machine. *Semen 29, 79-98.*

Matoré, G. (1951/1967). *Le vocabulaire et la société sous Louis-Philippe*. Genève: Droz/Slatkine.

Matoré, G. (1985). *Le vocabulaire et la société médiévale*. Paris: PUF.

Matoré, G. (1988). *Le vocabulaire et la société du XVIe siècle*. Paris: PUF.

Matoré, G. (1973). *La méthode en lexicologie*. Paris: Didier.

Pêcheux, M. (1969). *Analyse Automatique du Discours.* Paris: Dunod.

Pêcheux, M., Léon, J., Bonnafous, S., Marandin, J.-M. (1982). Présentation de l'analyse automatique du discours (AAD69): théories, procédures, résultats, perspectives. Mots 4(1), 95-123.

Prost, A. (1974). *Vocabulaire des proclamations électorales de 1881, 1885 et 1889.* Paris: PUF.

Robin, R. (1973). *Histoire et linguistique.* Paris: Armand Colin.

Roche, J. (1971). *Le style des candidats à la présidence de la République (1965, 1969). Étude quantitative de stylistique,* thèse. Toulouse : Privat.

Salem, A. (1987). *Pratique des segments répétés.* Paris: Klincksieck.

Salem, A. (1988). Approches du temps lexical. Statistique textuelle et séries chronologiques. *Mots 17, octobre, 105-143.*

Salem, A. (1993). *Méthodes de la statistique textuelle.Paris:* Université Paris 3.

Sekhraoui, M. (1983). *Concordances. Étude de vocabulaire.* Saint-Cloud: ENS Saint-Cloud.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work.* Amsterdam/ Philadelphia: Benjamins.

Tournier, M. (1969). Le centre de recherche de lexicologie politique de l'E.N.S. de Saint-Cloud. *Langue française 2(2), 82-86.*

Tournier, M. (1975). *Un vocabulaire ouvrier en 1848. Essai de lexicométrie*. Saint-Cloud : Publication de l'Ecole Normale Supérieure.

Tournier, M. (1976). La constitution du L.A. 246 au CNRS. Pp. 11-30. *Travaux de lexicométrie et de lexicologie politique. Bulletin du L.A. 246, 1.*

Tournier, M. (1980). D'ou viennent les fréquences de vocabulaire? *Mots 1, 189-212.*

Tournier, M. (1981).Spécificité politique et spécificité lexicale. *Mots.* 2(1), *5-10.*

Tournier, M. (ed.) (1982). *Actes du 2e Colloque de lexicologie politique, 15-20 septembre 1980.* Paris : Klincksieck, 3 vols.

Tournier, M. (2010). Mots et politique, avant et autour de 1980 Entretien. *Mots. Les langages du politique 94, 211-33.*

Quantitative Linguistics and Political History

Damon Mayaffre
BCL Laboratory (UMR 7320 University Nice Sophia Antipolis
University/CNRS); DamonMayaffre@wanadoo.fr

Introduction

In France, quantitative linguistics has found fertile terrain in political History: unexpectedly at first, then privileged to the point of developing into a tradition or even an autonomous sub-discipline.

This chapter reflects on this non-natural French scientific reality, now several decades old, which has managed to take on academic and institutional characteristics (laboratories, journals). But it will proceed in this way with a two-fold simplification.

First, with regard to quantitative linguistics, we intend to cover only quantitative *corpus* linguistics, quantitative *discourse* analysis or quantitative *text* linguistics. Phonologists or syntacticians also sometimes use computerized statistical approaches, but in quantitative linguistics we will focus here on one particular field, with a strong identity, which in France is called *textometry, logometry, textual statistics* or *(statistical) analysis of textual data.*

With regard to History and then political History, there has also been a narrowing of the point of view. The history of politics cannot be reduced to the mere study of words, speeches, programmes or ideologies. Still, we will ignore here work done, for example, on party structures, electoral sociology or institutional workings, to concentrate solely on historical studies that place the centrality of language in political activity, and that consider language production (constitutional texts, meeting discourse, press articles, propaganda leaflets, etc.) not only as political witnesses from a particular era, as a source or a medium towards an historical reality that remains to be discovered, but as historical actors *per se*, and, in consequence, as an object of study in its own right: political language as an object of History.[1]

When, why, how and for what benefits, has political speech become a privileged subject of quantitative linguistics in France, in a gestational phase at first and then ultimately in full development? And what are the data treatment methods - descriptive statistics and exploratory statistics - and the software available on the scientific market - Hyperbase, Iramuteq, Lexico or TXM - which have been successfully applied to this subject, and which have themselves been enriched in return? It is these questions that this contribution attempts to answer,

[1] A true scientific program, taken up by several generations of scholars, and also the subject of this contribution, "discourse as an object of history" was the manifesto published in 1974 by the pioneering historians: Jacques Guilhaumou, Antoine Prost and Régine Robin, associated with the linguist Denise Maldidier: *Langage et idéologies. Le discours comme objet de l’histoire,* Paris, Les éditions ouvrières, 1974

first by tracing the history of the interdisciplinary encounter between (quantitative) Linguistics and (political) History [Part One], then by highlighting contemporary developments under the influence of the digital revolution and the development of digital humanities [Part Two], and finally by illustrating the topic with concrete results and convincing machine outputs [Part Three].

1. History and Linguistics: beyond misunderstandings

1.1. Early Engagements (1950s, 1960s and 1970s)

In France, it is undoubtedly with *Les caractères statistiques du vocabulaire* and then *Problèmes et méthodes de statistique linguistique* by Pierre Guiraud, published in 1954 and 1960, that quantitative linguistics gained its founding books.[2] While the French linguist, a great reader of Zipf, perhaps at that time aimed to create a mathematical model of Language, statistical methods still in their infancy were very quickly applied to clearly identified corpora of discourse, thus seeking less to describe the linguistic System, this being impossible to sample or "represent" in all its complexity, than some of its discursive realisations: not an absolute corpus, that is, but detailed corpora that constitute norms of their own; not an absolute frequency or frequency *in Language,* but statistics and semantics endogenous to duly problematized corpora of discourse.

With Pierre Guiraud (1954, 1960) and Charles Muller (1968, 1973), it was first and foremost literary works - in this case the classical theatre of Corneille - which served as a corpus for study, thus opening up a long tradition for literary theorists, which, through the work of Etienne Brunet, became monumental over time [see Brunet 1981, and recently his collections of articles: Brunet 2009, 2011, 2015]. But beside literature, which remained very present and which benefitted from the creation of the *Trésor de la Langue Française,*[3] the field of political discourse quickly emerged as a productive and fertile ground for statistical linguistics.[4]

Indeed, a decisive factor is that in 1967, the Ecole Normal Supérieure de Saint-Cloud, in association with the CNRS, created an interdisciplinary laboratory for "Lexicology and political texts", whose object of study was political

[2] For an even older archaeology, see the contribution of Jacqueline Léon in this volume.

[3] *Le Trésor de la langue française* (The Treasury of the French Language), undertaken by Paul Imbs in Nancy in 1957, and then by Bernard Quemada at the Institut National de la Langue Française (National Institute of the French Language), constituted one of the largest and earliest works of digital input of texts on paper. Fifty years before GoogleBooks, and with certain unrelated philological precautions, its purpose was to enter the entirety of French literature, as well as writings from other origins.

[4] In detail, we note with Lemercier and Zalc [2008] that within the field of history this connection was not foreseeable and could even seem paradoxical, as political history traditionally cultivated a qualitative approach (of great events, great men, secret or "diplomatic" documents, etc.), whereas only economic and social history had learned to develop quantitative approaches.

language, and in which computer-assisted methods were developed, taking a quantitative approach to discourse.[5] From that time on, and for several decades thereafter, the Saint Cloud laboratory was at the forefront – in any discipline – of the development of French lexicometry. Computer scientists, mathematicians, linguists (Robert Léon Wagner, Maurice Tournier, Pierre Lafon, André Salem, Benoit Habert, etc.) rubbed shoulders there for 30 years, and the historians of discourse were the driving force there, both in modern history and in contemporary history (Annie Geffroy, Jacques Guilhaumou, Michel Launay, Denis Peschanski, etc.).

It is thus from the Saint Cloud laboratory that the institutionalisation of the term *lexicométrie* (lexicometrics) occurred, with a first journal *Travaux de lexicométrie et de lexicologie politique*, published in 1976, which in 1980 became the journal MOTS (*Mots/Ordinateurs/Textes/Sociétés*) and then *MOTS, Les langages du politique*, which is still flourishing today. It was also in this laboratory that certain major software developments were born, like LEXICO, whose modern version, implemented by André Salem, remains widely used in France today; and again it was in this laboratory that many of the most significant applications of lexicometry, and thus on political language, were published, with a first decisive book appearing in 1975: *Des tracts en mai 1968,* with the work of Maurice Tournier (1993, 1997 2002), with studies on trade union discourse and socio-political vocabulary.[6] Again, alongside the lexicometric exploitation of literature in Besançon (Quemada), Nancy (Imbs), Nice (Brunet), Strasbourg (Muller) or Liège for Latin literature within the LASLA (Evrard), we owe many inventions, improvements or statistical applications to Saint Cloud (Tornier), from the calculation of specificities to calculations of co-occurrences. In parallel to the first issues of the journal *MOTS*, in which the main lexicometric algorithms can be found, two Ph.D. dissertations from Saint Cloud were critical in this regard: the thesis of Pierre Lafon, defended in 1980, in which the foundations were laid [Lafon 1984]; and the thesis of André Salem, published in part in collaboration with the statistician Ludovic Lebart, which remains today, through its various editions and translations into several languages, a core knowledge base for Human and Social Sciences (HSS) researchers in this field (Lebart and Salem 1994).

1.2. History and Linguistics: a shared heritage

It must be said that the 1960s, 1970s, and 1980s lent themselves to the flourishing of quantitative linguistics in the field of history and political discourse.

[5] We thus find traces of this in April 1968 at a symposium published by *Cahiers en Lexicologie* nos. 13 and 14, 1968 and 1969.

[6] In this overview, it should be added that in addition to Saint Cloud, Jean Dubois defended a decisive thesis in Nanterre on *Le vocabulaire politique et syndical en France* (Political and trade union vocabulary in France). Flourishing at first, the school of Nanterre then died out, probably due to the absence of an historical background. The theses defended there intended to establish a socio-linguistics but found no echo in the historical community (see below).

In the twentieth century, History, as a centuries-old discipline within HSS, became aware, early on, of the epistemological gains to be made by conversing with linguistics, which was still young but which had been rapidly expanding since the advent of Saussure.[7] Beyond the initial contacts with a Marc Bloch or a Lucien Febvre (Febvre 1953), and beyond Alphonse Dupront's decisive statements about historical semantics (Dupront 1969), and particularly beyond the importance that structuralism and (therefore) Linguistics took on for all of post-war HSS, let us recall that work on textual archives is the very definition of historical work (as opposed to the pre-historic work of palaeontologists), and this necessarily made the community of historians sensitive not only to classical philology but also to the modern language sciences. The major work, unequalled to date, of Régine Robin, *Histoire et Linguistique,* published in 1973, became a structuring element for several generations of French historians. While the author certainly regretted the relational difficulties between historians and linguists, she established an ambitious trans-disciplinary research programme in which History and Linguistics were to cross-fertilize one another, and maintain a non-ancillary relationship. And in this fundamental work, a large portion, devoted to a better future, was dedicated to the first lexicometric results, supported by the new approach from Saint Cloud (Robin 1973. Chap 5 and 6 and Appendices): with strong theoretical postulates, *Histoire et Linguistique* thus declared, and gave concrete illustrations of, the merits of the quantitative treatment of the first large digital corpora, which the community of historians now had at their disposal with the arrival of computers in research laboratories.

History in particular, as one of the Humanities aspiring to the status of a science, had an old methodological tradition that it is not necessary to recall here: from the pre-War methodological school to the Annales school of the inter-war years, passing, generally, through the rigours of the Marxist approach, historians had long claimed to go beyond a mere impressionistic narrative of past events to establish controlled methodological protocols to deal with their sources and objects. In this respect, therefore, the historical sciences were ripe, in the 1960s and 1970s, the period of interest to us here, to welcome the kind of formalized and mathematical methods that computer-assisted quantitative linguistics was striving for. For example, serial history, with its tutelary figure Ernest Labrousse, had just established the principle that quantitative data processing made it possible to go beyond the anecdotal to achieve something representative and structural. While serial history was at that time applied primarily to the economic domain, naturally more sensitive to figures, the transposition to politics was envisageable: in a way, a corpus of political texts could be considered as a *series* within which word frequency and vocabulary regularities and irregularities could be described and interpreted. And Régine Robin, to mention just one example, explained her methodological detour towards lexicometrics by the need to process a corpus composed of a series of a hundred Books of Grievances (1789), something not accessible to human memory. In her wake, finally, all the historians who use lexical statistics and computer tools to process their (large) textual

[7] Notwithstanding that Saussure's Course in General Linguistics was known quite late in France.

corpora fundamentally share this quantitativist posture, concerned as they are with exhaustivity, systematicity, representativity, and seriality.

1.3. Correspondence factor analysis: an immediately cutting-edge practice

For its part, quantitative linguistics experienced a major and very rapid enrichment in France, something in which the political historian was both a participant and a driving force. Alongside the primarily descriptive field of lexical statistics, which was efficient but elementary (lexical frequencies, vocabulary specificity scores, calculation of co-occurrences), and which originated, as we have seen, with Guiraud, which underwent its decisive improvement with Muller, and which blossomed in Saint Cloud in the 1950s, 1960s and 1970s, the mathematician Jean-Paul Benzécri proposed, in his lectures on *l'Analyse des données et reconnaissance de formes (Data Analysis and Pattern Recognition)* (1965), a form of exploratory multi-dimensional statistics[8] which would revolutionise the French methodological panorama for decades: correspondence factor analysis (Benzécri 1972).

This is not the place to settle the debate over the purely French or Anglo-Saxon origins of a method that, beyond the earlier mathematical presuppositions, needed the computerised tools of the second part of twentieth century to thrive. We will only note that, in France, it was the historian of political discourse, Antoine Prost, who was the first in the Human and Social Sciences to use it on textual data in a pioneering work, *Vocabulaire des proclamations électorales de 1881, 1885, 1889,* written in 1970 and published in 1974. From that time on, political historians, and more generally the whole of French lexicometry, systematically used exploratory multidimensional statistics, which is still implemented today at the heart of all French software on the market (DTM, Hyperbase, Iramuteq, Lexico, TXM, etc.). Thanks to this method, under the direction of Antoine Prost, and in the Saint Clous laboratory, the historian Denis Peschanski, for example, was able to defend a thesis applied to a chronological corpus of communist speeches; 40 years after Antoine Prost, our own books on contemporary French presidential speeches (Mayaffre 2004 and 2012) still owe much to correspondence factor analysis.

The methodological stakes were high and two-fold:

First, faced with textual corpora crossed by multiple socio-historical variables (chronology, the political identity of the speaker, his institutional status, and the conditions of speech itself), the historian could unravel and prioritise the extra-linguistic constraints weighing on the discourse, or more precisely, test these variables in an exploratory way, that is to say, without projecting working hypotheses onto the corpus that are too strong. For example, at the outset, Antoine Prost tested the political affiliations of all French MPs during the initial legislatures of the Third Republic by looking at whether or not unbiased statist-

[8] We are referring here to the title of the book by Ludovic Lebart, a statistician and disciple of Benzécri: L. Lebart *et al., Statistique exploratoire multidimensionnelle,* Paris, Dunod, 1995 (reprinted 1998 and 2000).

ical processing allowed the grouping together, by virtue of a shared vocabulary, of deputies from the left or the right; and as the results of the enquiry were positive, the outlying deputies, transgressing the norms of the established political groups, could be identified; and the vocabulary responsible for these similarities and deviations was identified. In History, this method and this type of exploration found a natural field in chronological corpora, which André Salem (1991) called in an ad hoc manner "chronological textual series". In such cases, when it comes to comparing texts from a single source (same political party, same speaker, same press organ) but produced at different and regular times (every week, every year, every decade), correspondence analysis makes it possible to reveal a chronology endogenous to the corpus - and not projected into it by the historian - by identifying ruptures and continuities in the discourse. (See *infra* Part III).

Next, AFC (correspondence factor analysis) made it possible to widen and increase the field of observation of the original lexical statistics. Up until now usually reduced to the individualised distribution of the single unit, in the context for example of a calculation of *specificities*, lexical statistics offered only a partial and fragmented view of the text. With AFC there is a large number of units, up to the entire vocabulary of the text, which are all considered at the same time, allowing us to work out their relationship and their organisation; their statistical relationship and organization, of course. The idea then arose that it was the text in its entirety and its full complexity, if not its meaning, which could thus be addressed.

In other words, to repeat the two virtues heretofore mentioned, AFC allows the political historian to process complex matrices: tables (x rows and y columns), with a series of the texts in the columns that we want to compare based on their chronological, political and generic relationships, and in the rows the entire body of vocabulary that is to be assessed (as illustrated in Part III).

1.4. Beyond words, discourse: an essential epistemological position

Be that as it may, the emerging relationship between (political) History and (quantitative) Linguistics had as a cause and would also have as a result an acute awareness of the complexity of the discourse object; an acute epistemological awareness that makes the lexicometric analysis under discussion here, and which can be considered as a preferred method for historian-linguists analysing discourse, cannot be confused, in France, with the American content analysis that had been proposed a few years earlier by Lasswell and Lazarsfeld; in France, at that time, a very strong problematisation of both text and discourse lay behind the computerised and statistical processing of corpora.

In France, reflections on the discourse object during these baptismal years immediately took the form of an interdisciplinary scientific effervescence and intellectual adventure almost without equal, which it is beyond the scope of this paper to describe, all the while being at its centre: the French school of discourse analysis (Mazière 2005), to which Critical Discourse Analysis (CDA) does not hesitate to claim membership today.

Although we have talked about the birth of the Saint Cloud laboratory both for its importance in the development of textual statistics and for its important role in the historical approach to political language - in a word, as the meeting place *par excellence* between quantitative linguistics and political history - we should mention at the same time the work of Jean Dubois and the School of Nanterre which, although it eventually disappeared, played a decisive role for two decades, and which was responsible for the propagation of discourse analysis in France.

Leading on from lexicology, that is to say, the claim that the lexicon is a structured whole and must be understood in use or in context, in the early 1960s Jean Dubois campaigned for a trans-phrastic linguistics, and found a model to follow in the work of Z. Harris, which he had had translated in the journal he had just created (Langage 1969).

Ultimately, this founding and original discourse analysis did not stand the test of time, perhaps for two reasons that are directly concerned with the subject of the present contribution: first, because the quantification of linguistic phenomena did not play a strong enough role, to the benefit of a more formal linguistic approach (distributionalism), of which Jean Dubois himself foresaw the limits; and second because the proposed method, due to its linguistic complexity, could not be assimilated by the (political) historians, whose role as a driving force we have just emphasised, even though the corpora being processed were highly political and historical:[9] in fact, the theses of Marcelessi on the Congress of Tours (1971) or of Maldidier (1970) on the war in Algeria, while important, were not given any consideration by the historical community (without being given much consideration by linguists either): interdisciplinarity, which was so promising elsewhere, was unfortunately a general failure in this case, with only a few rare exceptions.

Still, the important thing is: Jean Dubois, with his thesis (Dubois 1962), encouraged the emergence of a linguistics of discourse in France - particularly of political discourse - that is to say, on the one hand, (i) a linguistics that would not be limited to the morpheme or the sentence but would address trans-phrastic organisation, which is something far more complex; that is to say, on the other hand (ii) a linguistics of usage or of a socio-historically situated language and soon of one that is ideologically constrained. Although composed primarily of linguists, *the School of Nanterre established discourse as an object of history*: the "I" of texts ceased to be the "I" of the grammatical subject to become the "I" of the historical or ideological subject of the historian; vocabulary could only be grasped in (historical) context. The structures of discourse, which were to be updated, were of course subject to linguistic constraints, but also to social and ideological constraints.

[9] It should not be necessary to emphasise here that what is globally called "discourse analysis" was at that time above all an analysis of political discourse. We recall in particular that Jean Dubois and his followers were under Communist discipline at a time when political militancy was an integral part of intellectual training.

1.5. Beyond discourse, ideology

The 1969 issue of *Langage* constitutes the official birth certificate of discourse analysis for linguistics: a birth certificate, let us repeat, that is paradoxical insofar as everyone still wants to lay claim to it but nobody uses the distributional method today. But French discourse analysis has transcended this birth in linguistics significantly and has developed in France under the influence of three master thinkers, Althusser, Foucault and Pecheux, who offered a comprehensive Marxist or "Freudian-Marxist" approach (Rastier 2001) of discourse. And in an incredibly contracted chronology, we find ourselves at the end of the 1960s.

With these thinkers, it is not only the relationship that words have with discourse that is posited, but the relationship between the subject, language and ideology which takes on a central role and ultimately becomes the keystone of the French School of discourse analysis; and as for Michel Pêcheux, we note in passing that his epistemological proposals are coupled with a strong methodological proposal of Automatic Analysis of Discourse (AAD), which, although it does not explicitly rely on statistics, took the visionary step of using digital technology (Pêcheux 1969).

It is perhaps the idea of the non-transparency of discourse that governs discourse analysis and that political historians were able to seize upon at the outset; and it is this lack of transparency that demands the development of a more effective methodological protocol than simple reading.

While the meaning of a sentence can be formally attained by linguistics, the meaning of discourse is not obvious or explicit, and texts are never transparent. A discourse can say more than it says explicitly; no production of language is ever obvious, and language is always penetrated with ideology.

In psychoanalytic terms - since Freud and Lacan are concerned here too - *manifest content,* which is accessible through normal reading, may mask a more complex *latent content,* and the *psychological subject* - the assumed "me" of the speaker - can reveal a deeper *psychoanalytic subject*. In Marxist terms above all - for Marxism overshadows all of this nascent discourse analysis - political discourse explicitly exposes a programme or a thought, but also betrays and constructs, at a deeper level, an ideology - a general relationship to the world - that the analyst must discover under the deceptively evident material or linguistic surface of the corpora.[10]

Although one could criticize a majority of the historian researchers of the 1970s for a certain naivety in their approach to discourse and to linguistic material (Robin), this basic critical stance of discourse analysis, if not this hermeneutic posture, could not fail to seduce historians, for whom making texts and archives "speak" in order to reconstruct the past constitutes the heart of their profession.

In this context, it would obviously be reductive to assert that quantitative methods are the only imaginable means to render historical interpretations ob-

[10] We recall in particular that the key concept is the "discursive training" by which the speaker was constrained to express himself.

jective, but it was clear at that time that computers and statistics were an effective and accessible lever for historians, even if it meant, according to Régine Robin, instrumentalising them.

In addition, beyond the historian community, there is the whole of French-style discourse analysis which considered lexicometry, which we can appropriately call logometry (logos = discourse; metry = measurement), as an essential method, as evidenced by its prominence in the textbooks that were an authority on the subject in France throughout the 1980s (Maingueneau 1976, 1987), and the many articles published over the decades in the journal *Mots*, in *Histoire et Mesure*, and in *Lexicométrica*. And if one had to choose just one major text for historians, it would be the contribution of Antoine Prost in 1988, an intelligent plea for using quantitative linguistics in political history.

1.6. History and computational linguistics: a delayed marriage (1990-2000)

The years 1990-2000 were marked by a significant epistemological retreat; a retreat that could perhaps be generalised for all the HSS with the collapse and non-replacement of such models of systematic thought as Marxism, structure-alism, generativism, Freudianism, etc.; in any case, there was a significant retreat concerning interdisciplinary exchanges between History and Linguistics, and also concerning methodological acuity in political history.

A young historian such as Eric Anceau, for example, who retraces the historiography of political history in the late twentieth century and who campaigns ambitiously for a total political history that could converse with other disciplines, not only mentions lexicometry very little but also only pays scant attention to the dialogue between History and Linguistics (Anceau 2012): the ideals of the 1970s seem to have run their course.

Admittedly, Régine Robin herself was pessimistic from the start, showing the extent of the "misunderstanding" (Robin, 1973, Chapter 1: *Le malentendu*) between historians and linguists. And while her book had the impact described above, she made a negative assessment of the whole enterprise in 1986, using the expression "the continuing misunderstanding"; and eventually Régine Robin, her little remaining support in France coming only from Jacques Guilhaumou, preferred geographic exile in Canada and disciplinary exile in Sociology.

However, even during a period unfavourable to interdisciplinary dialogue and to methodological precautions, and for political history marked by the return of "battle history", of coffee-table biographies and of anecdotal history, we still find traces of other initiatives undertaken.

Besides the major and theoretical work of Jacques Guilhaumou, to which we will return, it is perhaps the work of Jean-Philippe Genet, a medievalist at the Sorbonne, that is the most remarkable because of his tenacity. In addition to numerous publications (Genet 2012, Genet and Lafon 2003) and the transmission of a scientific posture to the younger generation (Sébastien Benjamin Déruelle, Stéphane Lamassé, etc.), in 1988, together with some fellow historians, he created the association *Histoire et Informatique*, this being the French section of The Association for History and Computing, which had been founded a year earlier. This venture was limited to neither political history nor lexicometry, and was only partially successful, but the soil for a history of political discourse assisted by lexical statistics thus remained under cultivation.

Similarly, during the same decade the political scientist Dominique Labbé published several books on Communist language, on the discourse and vocabulary of François Mitterrand (Labbé 1990), on governmental discourse (Labbé and Monière 2004), and numerous articles on the rhetoric of de Gaulle and on French trade union discourse. Our own thesis, published under the title *Le poids des mots* in 2000, was also part of this tradition, using logometrics to interpret the discourse of the left and the right in the inter-war period, and to update the discursive battle with reversed front lines between a national right suddenly converted to the spirit of Munich and an internationalist left converted to national defence against fascism (Mayaffre 2000).

Finally, and in addition, with respect to textual statistics the 1990s saw the French textometrics community, including analyses of French and foreign political discourse (Bécue, Bolasco, Labbé, Marchand, Monière), organise and internationalise around the biannual Analysis Days for Textual Data (*Journées d'Analyse de Données Textuelles,* or JADT), and around certain journals like *Lexicométrica, Corpus* or *Histoire et Mesure.*

It is clearly through this internationalisation of the community, and the sharing of digital textual resources and community software tools, that the present-day conjuncture must be understood.

2. The current turn of digital humanities

Far from the grand explicative systems such as Marxism, Freudianism, or Structuralism that form the backdrop for the interdisciplinary rapprochement described by Régine Robin and Antoine Prost between (political) history and (quantitative) linguistics in 1960-1980, the current scientific landscape is marked by certain significant elements in the relationship that connects researchers in the human sciences - and particularly historians - with the textual or the linguistic.

The two most important factors, which are essential to the daily practice of researchers in the 21st century, are the digital revolution and the hermeneutic turning point that has occurred in HSS; and to these two points we may add, from a technical point of view, the popularisation and development of lexicometric tools and functionalities to the point where nobody can ignore their existence: such as, for example, searches by keyword in internet search engines, or the processing of co-occurrences.

2.1. The universal digital archive

After the invention of language, which enabled Man to be human, that of writing, which brought him into History (versus pre-history), and that of the printing press, which swung us into modernity, everything seems to indicate that our civilisation is experiencing a 4th major cultural and epistemological revolution: the digital revolution (Goody 2000 and Darnton 2009).

Day after day, our Gutenberg society is being transformed indeed into a digital society; the keyboard replaces the pen and the screen replaces paper; text becomes hypertext; reading becomes hyper-reading; modernity becomes hyper-modernity.

When it comes to the disciplines covered in this paper (Linguistics and History), the evolution is a major one.

For example, faced with the mass of data and its accessibility by a simple click, introspective linguistics such as generativism concedes a new relevance to corpus linguistics. Corpora of hundreds of billions of words are now immediately accessible to researchers, like Google Books (Brunet, Vanni 2014): universal corpora can therefore now support universal grammar.

In history - and in political history - the revolution is equally significant. Long constrained by the scarcity of existing or materially available sources, the historian is now faced with an almost infinite archive: the web. Old collections are digitized and accessible from home, especially for medievalists and modernists. Above all, for specialists in contemporary history, new collections are emerging daily, immeasurably rich but which can be taken in instantly, such as collections of media items or political speeches.

In other words, the digital revolution has reinvented the textual archive, and the historian and the linguist are being fundamentally questioned once again about their basic skills, at the crossroads of the two disciplines. The interdisciplinarity between Linguistics and History, which cooled in the years 1990-2000, seems to us to be in need of reviving.

Finally, and more prosaically, the immensity of these resources raises once again, and rather mercilessly, the issue of quantitative data treatment and the contribution of statistical linguistics or computational linguistics: computer science and statistical processing, which might once have appeared as a luxury or an option, have now become a necessity.

2.2. Digital hermeneutics

After the turn taken by linguistics, it is now the turn being taken by hermeneutics, since the end of the 20th century that seems to be marking all of the HSS disciplines. Because the explicative systems mentioned above have been partly abandoned, it is now less a question of explaining than of trying to understand, that is to say, to interpret. The whole world is to be interpreted; the archive, the meaning, the corpus are all objects of interpretation.

In France, at the border between history and linguistics, it is perhaps the philosophical figure of Ricoeur that has been most significant in this interpretative turn: history is a narrative, and the narrative is a shaping of the world through language and interpretation. The work of the historian-linguist Jacques Guilhaumou has played a decisive role here in discourse analysis; and as a former researcher at Saint-Cloud, he is no stranger to the development of French lexicometry. In 2006, Jacques Guilhaumou firmly established the language dimension of political events, and argues for an hermeneutical posture in the face of an historical record that is necessarily textual: historical events or facts are always presented to us to examine and to understand in their dual material and linguistic nature. Similarly, the writings of François Rastier play an important role at the beginning of the 21st century in France, by defining the text as a place of "interpretative pathways". Meaning is never given by the text but rather constructed through reading, that is to say, through the reader's interpretation. And François Rastier stresses the importance of methodological protocols meant to signpost or even encompass these pathways going beyond literary intuition. In this context, and in two successive books, he stresses the contribution of digital technology (Rastier 2001) and quantitative methods (Rastier 2011) in the development of "new observables" in linguistics, which are invisible on paper but visible on a tablet, like so many objectifiable interpretative elements.

2.3. Computer performance and software popularization

Beyond this double epistemological situation (the digital revolution and the interpretative turning), everyday practices have also evolved very quickly at the beginning of the 21st century. The tools that now instrumentalise our reading of texts (search engines, keywords, word clouds, etc.) are found everywhere, in science and in society.

The two preconditions for the data treatment of quantitative linguistics, which still had to be justified in the 1980s, namely tokenisation and indexing, are at the basis of the big search engines like Google: all researchers and citizens use them without even knowing it. The lemmatisation and morpho-syntactic tagging that allow automatic entry into text with linguistic units that are better established than graphic words have also become necessary, at least for experts. As for the frequency-based approach, it also appears indispensable in the face of the big Web data.

In France, the lexicometric, textometric, and logometric software that generally came into existence in the 1980?s is multiplying, being freshened up, and is putting 30 years of statistical expertise into a modern ergonomic form. An historic programme such as Hyperbase, for example, is now in its 10th version in 2016 and is being distributed on the Web in a "light" version [http://hyperbase.unice.fr/ hyperbase/]. New software is appearing with an open source logic such as TXM and Iramuteq. Beyond that, the general public is becoming familiar with networks of words and co-occurrence graphs. Especially during election periods, candidates' speeches are often decrypted in the media on the basis of a lexicometric approach.

Finally, the institutions in charge of research are measuring the interest of a field in full expansion, and ANR and Equipex projects are financing software development in the field.[11]

[11] One of the major projects is the Equipex MATRICE (2010-2020, 2.6 million Euros; dir. D. Peschanski), which funds the development of the TXM software.

3. Applications

Since the early work of Jean-Marie Cotteret and René Moreau on the vocabulary of de Gaulle (Moreau and Cotteret 1969), the work of Antoine Prost on electoral proclamations of the Third Republic (Prost 1974), of Régine Robin on Cahiers de doléances (Notebooks of Complaints) of 1789 (Robin 1974), or of Saint-Cloud on the tracts of May 1968 (Demonet et al. 1975), studies of political history making use of quantitative linguistics have been numerous in France, and it would be presumptuous to claim to summarize them all here. We can reduce them to just four – necessarily arbitrary – headings, referring the reader to a rich multi-decadal bibliography of dozens of books.

3.1. Men and words (vocabulary specificity scores)

Whether one perceives it as a simple subject – in the Marxist sense of the term – or as a charismatic leader, the political historian has always been preoccupied with people in the polis; and this biographical concern naturally meets the concern of the linguist or speech analyst, for whom, fundamentally, there cannot be any language without speech, nor speech without speakers.

Quantitative linguistics and political history have thus been concerned with describing and interpreting the production of individual speakers,(Presidents, First Ministers, etc.), but also collective speakers (parties, unions, press publications, etc.) whose words have made society.

While literary lexicometry has been sensitive to the calculation of lexical richness to describe the style of authors (Brunet 2009), political lexicometry has widely used, as a major tool, the calculation of specific vocabulary (Lafon 1984)[12] to describe the discourse of socio-political players.

In the necessarily contrastive corpus (multiple speakers), the goal is to identify the words (or other linguistic units) that statistically characterise a particular speaker. So, out thousands of possible examples, in the French presidential corpus since de Gaulle and the beginning of the Fifth Republic, we could point out the *specificities* of Nicolas Sarkozy [Table 1].

Table 1
Specific vocabulary of Nicolas Sarkozy (2007-2012)

Specificities	Frequency in Sarkozy (2007-2012)	Frequency in the corpus (1958-2014)	Scores
Ça (That)	663	1153	+33
On (you, one)	2524	13,961	+27

[12] Now firmly established, the calculation establishes the probability of a word having the frequency k in a text: *Let T* = size of the corpus, *t* = size of the text, *f* = frequency of the word in the corpus, *k* = frequency of the word in the text, *prob (x = k) =*

Crise (Crisis)	368	1077	+22
Pas (not)	3501	23,099	+20
Je veux (I want)	355	1206	+19
Vouloir (to want)	1047	5504	+18
Ne (not)	4128	28,765	+17.5
Travail (Work)	415	1628	+17.5
Je (I)	4810	34,542	+17
Demonstrative pronouns	5866	43,723	+16
Banque (Bank)	130	307	+15
Ce (This, It)	4136	30,237	+15
Pronoun+adverb+ verb	2252	15,606	+13.5
Immigration	51	122	+9
Policier (Policeman)	37	70	+9
Délinquant (Delinquent)	25	33	+8
Moi (Me)	448	2784	+8

And behind this statistical list, ranked here in order of precedence and according to an elementary index, it is the overall position on the political right, called neo-populist, of Nicolas Sarkozy that we have been able to interpret. For example, the statistical preponderance of the verbal group, "I want" participates in the construction, through speech, of the figure of the leader or charismatic authority; the over-use of the popular forms "on" (you, one) or "ça" (that) seems to participate in the demagogic relaxation of speech addressed to the greatest number; the repetition of the syntactic structure [pronoun+adverb+verb], which in French always has a negative aspect ("je ne veux..." – I don't want"; "il ne faut..." – "There mustn't"; "vous ne pouvez..." – "you can't", etc.), plays a part in the establishment of a Caesar who grumbles and thunders in his speech, etc.

Similarly, Pascal Marchand has systematically described the vocabulary of all the French first ministers in their general policy speech since the establishment of the Fifth Republic (Marchand 2007) and we now know, in political history, the statistical and lexical features of the speech of people like Michel Debré in 1959, Raymond Barre in 1976, Jospin in 1997, Valls in 2014, etc.

St. Cloud, to which we owe this calculation of *specificities* - a calculation, we repeat, that is widely used in France - in more ideological works, strove to characterise the vocabulary of Communist speech (versus bourgeois speech) of the interwar period through the collective speakers represented by L'Humanité or Cahiers du Bolchévisme;[13] to characterise also the speech of the right (versus the speech of the left); and to characterise the speech of the CFDT (versus the CGT) (Demonet *et al.* 1978; Peschanski 1988; Tournier 1993). Etc.

[13] Obviously, in a chronological corpus, the calculation of particularities can be used to distinguish a specific point in time (e.g. one year) during a period (a decade, for example). See below, 3.2.

To conclude, we will anticipate a little: thanks to the correspondence factor analysis described below, a synthetic view of the most remarkable *specificities* can be produced. For example, in the 1958-2014 presidential corpus, the ten main characteristics of each president are distributed on the graph as follows (Figure 1).

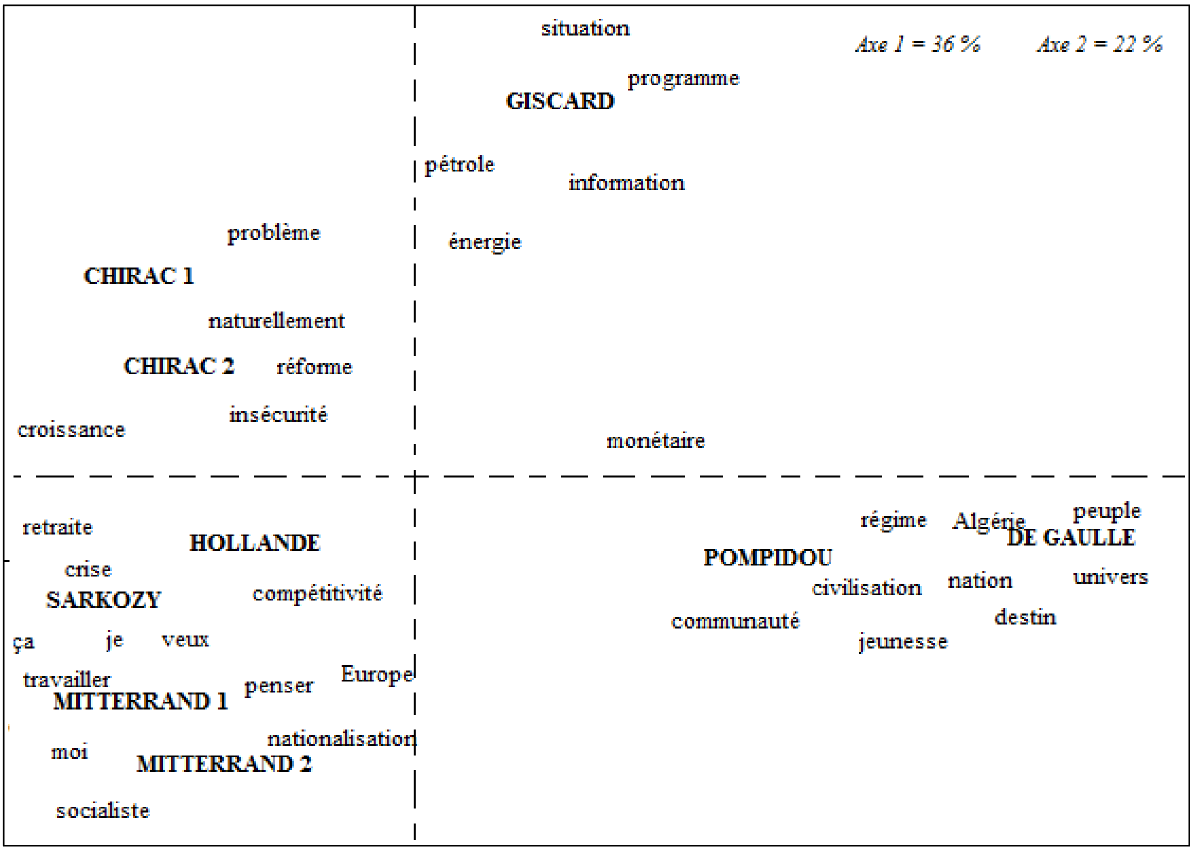


Figure 1. Factor map of the first *specificities* of the presidential corpus (1958-2014)

3.2. Discourses and periods (correspondence factor analysis)

Thanks to statistics, the characterisation of the vocabulary of the texts in large corpora takes on a particular acuity for the historian in the context of diachronic corpora that Andre Salem defined as *chronological textual series* (Salem 1991). Thus we can show that over long periods, a form of lexical continuity takes shape, and that given this continuity the discrepancies observed allow us to update a chronology of political events that is sometimes unexpected for the historian, and endogenous to the corpus.

It is through the *Descriptive Multivariate Statistical Analysis* (Lebart et al. 1995) and the *Correspondences analysis*, developed in France by (Benzécri 1973) from the 1970s on, that the most convincing results have been achieved. For example, an examination of the entire vocabulary of the Communist leader in France, Maurice Thorez, between 1930 and 1939 allowed us to attribute a new dating to the Popular Front (Figure 2).

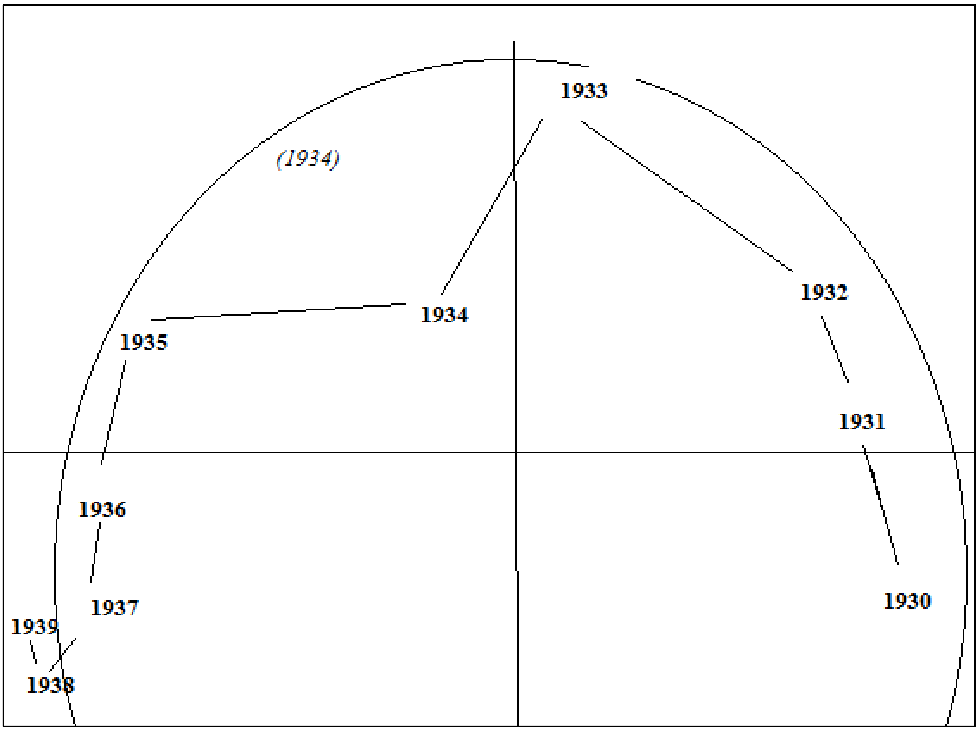
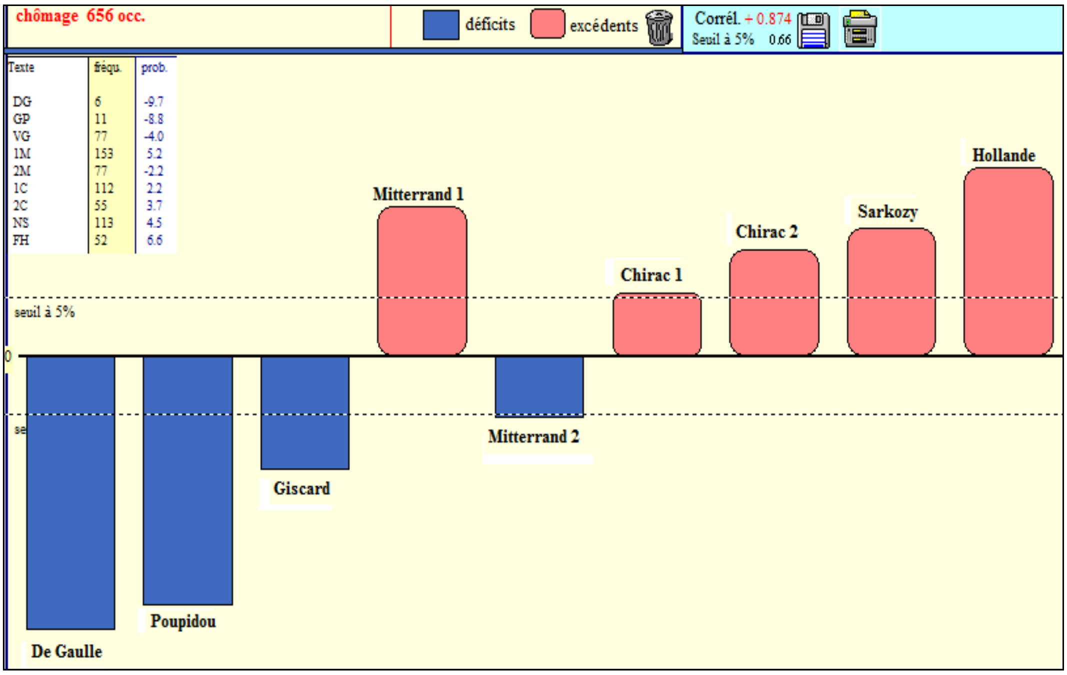


Figure 2. Correspondence factor analysis (Thorez corpus 1930-1939)

Maurice Thorez's speech changes early on and from as early as 1932-1933 mobilises Jacobin vocabulary (as opposed to traditional Bolshevik vocabulary) that foreshadows the Popular Front; a progressive and early evolution only broken by the year 1934, which appears as atypical in the corpus and on the graph, in moving away from the ideal parabola that the Guttman effect produces on chronological corpora.

Beyond this type of general chronological study, other more specific indices allow us to understand the temporal logics that run through the corpus, such as the calculation of the chronological correlation (Brunet 1981: 401-406) applied to each unit and which allows us to identify the most striking progressions and regressions. For example, during a period of 60 years, in the French presidential corpus (1958-2014), "unemployment" (chômage) is the word that has progressed the most, and the most regularly, with an index of 0.874 (Figure 3)



Texte	fréqu.	prob.
DG	6	-9.7
GP	11	-8.8
VG	77	-4.0
1M	153	5.2
2M	77	-2.2
1C	112	2.2
2C	55	3.7
NS	113	4.5
FH	52	6.6

Figure 3. Chronological distribution of "unemployment" (chômage) in the presidential corpus (1958-2014)

3.3. Text classification (intertextual distance and tree analysis)

Directly linked to the previous concerns of characterisation, logometry seeks to *classify* texts according to their origin: historical origins (as previously), political or ideological origins, and obviously, generic origins; this is classification on the sole basis of the linguistic materials used, both the words and also the grammatical or syntactical combinations.

Many indices of intertextual distance (or distance between texts) or lexical connection (Muller 1973; Labbé and Labbé 2004, etc.) have been developed by statisticians and used by the historian.

Dominique Labbé especially and Denis Monière have provided measure and represented under a tree form the existing distance between all the Throne Speeches in Canada, representing 128 speeches between 1867 and 2010 (Labbé and Monière 2004; Labbé and Monière 2014).

The calculation and this tree representation, implemented for example in Hyperbase [10.0 2016], allow us to classify texts by comparing chronology and politics, as in the study that we made of Chirac's and Jospin's speech between 1997 and 2002 (Figure 4).

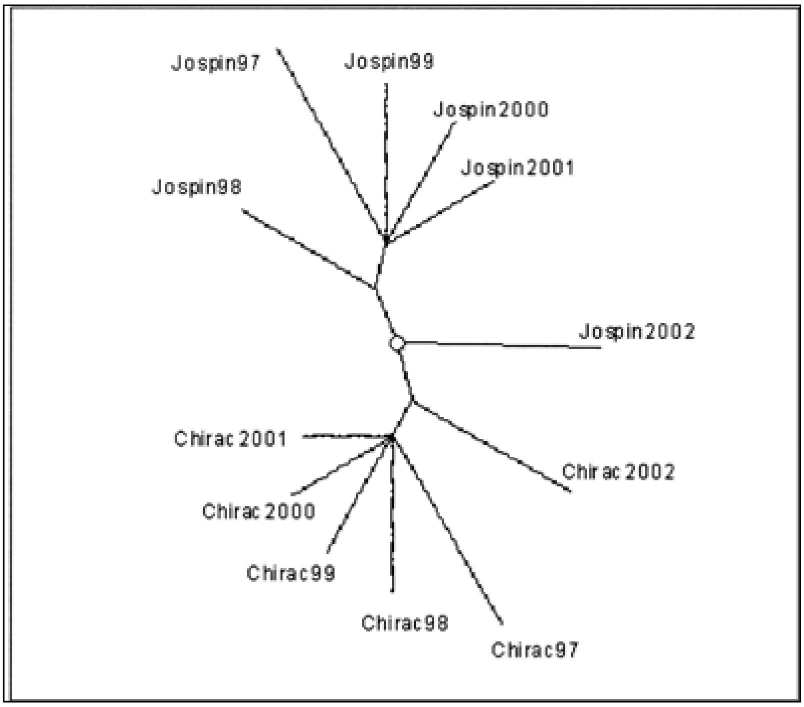


Figure 4. Tree representation of the Chirac / Jospin intertextual distance (1997-2002)

Roughly speaking, the tree distinguishes, at the top and the bottom, two versions of speech that correspond very well to the two speakers (Chirac / Jospin), and the respective chronology of each speaker has been updated. In this approach, the political historian will note the coming together over the years of the President and the First Minister (shorter branches on the tree), and the central and indeterminate position of Lionel Jospin's speech in 2002, as if his discursive identity had disappeared as a consequence of the election year. In fact, for many observers, Jospin's electoral discourse in 2002 was inaudible until his electoral defeat at the second round of the presidential election (Mayaffre 2004b)

Similarly, the intertextual distance on the presidential corpus allows us, for example, to see that François Hollande barely stands out from his predecessor at the Elysée (Figure 5), particularly because Sarkozy and Hollande use identical vocabulary in response to the economic crisis ("bank","debt","growth", etc.).

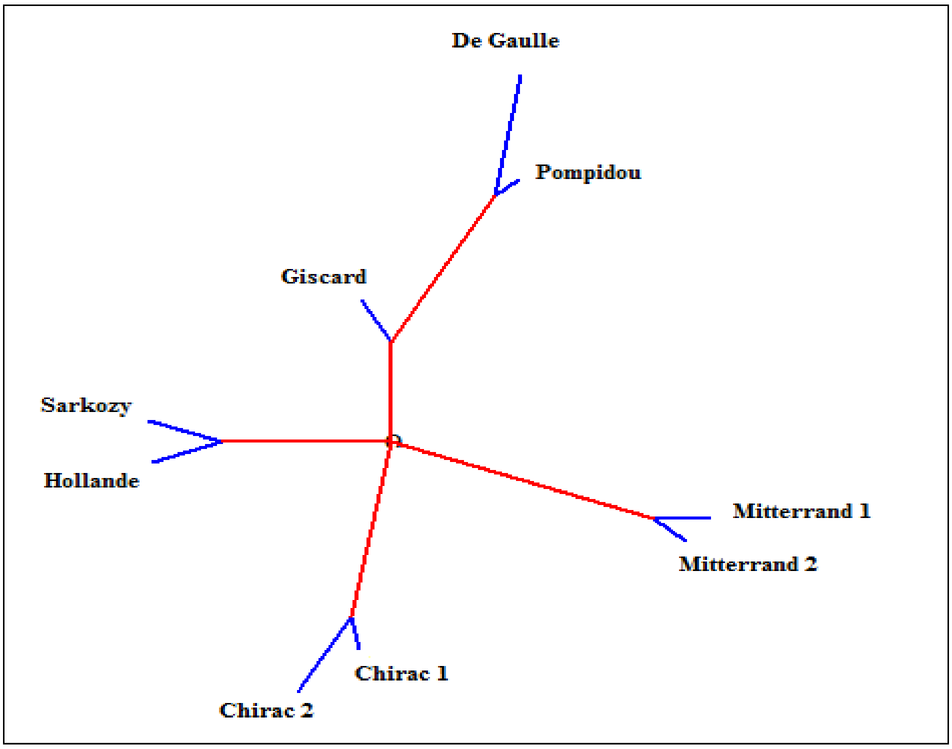


Figure 5. Tree representation of the intertextual distance in the presidential corpus (1958-2014)

3.4. Ideology and corpus semantics (cooccurrence processing and representation)

Finally, there remains the most interesting field of quantitative linguistics as applied to political history: the description of the thematic organization of texts and, hence, of the programmes developed or even the ideologies defined as coherent verbal expressions of the world in speech.

Statistical works on co-occurrence date back almost identically to the development of lexicometry in France, as (Mayaffre 2014) reminds us. And they are now undergoing a particular development, notably in favour of networks.

According to a strong presumption of the linguistics of the corpus, the meaning of words must be established not by recourse to the dictionary, but endogenously from the corpus, by the study of how it is used in context.

But the context of a word A can be defined minimally as its co-occurrence B: when A and B co-occur, A and B mutually contextualise each other. Generalising this theme, we will define the meaning of a word as the sum of its co-occurrences.

Thus it is possible to calculate from a pole word the preferred attractions which make up its lexical and semantic universe. For example, by systematically calculating the co-occurrence of "work" in the corpus of de Gaulle and Sarkozy, we have shown that the two presidents used the term in very different ways, in a Marxist sense for de Gaulle and a Hegelian sense for Sarkozy (Table 2).

Table 2
Co-occurrences of "work" in the corpus Sarkozy compared to de Gaulle

SARKOZY		**DE GAULLE**	
Words	**Deviations**	**Words**	**Deviations**
réhabiliter (regenerate)	+8.48	technique	+4.79
fruit	+6.13	rendement (productivity)	+4.34
effort	+5.95	production	+3.93
merit	+5.94	information	+3.90
partage (sharing)	+5.41	capital	+3.60
revalorisation (increase, adjustment)	+5.11	échelle (scale)	+3.50
libérer (to free)	+4.76	personnel	+3.24
possibilité (possibility)	+4.64	déplacement (shift)	+3.23
durée (duration)	+4.63	emploi (job, employment)	+3.22
valeur (value)	+4.61	commission (commission)	+3.21
récompense (reward)	+4.32	responsabilité (responsibility)	+3.19
formation (training)	+4.18	jeune (young)	+3.18
réhabilitation (rehabilitation)	+4.16	société (society)	+2.91
taxer (to tax)	+4.14	professionnel (professional)	+2.87
vivre (to live)	+3.85	intérêt (interest)	+2.84
création (creation)	+3.59	œuvre (piece of work)	+2.82
récompenser (to reward)	+3.49	direction	+2.76

This elementary treatment can be complicated by the study in particular of second-level co-occurrences (co-occurrences of co-occurrences). And several representations can be imagined, like the graphs proposed by the Hyperbase software (Figure 6)

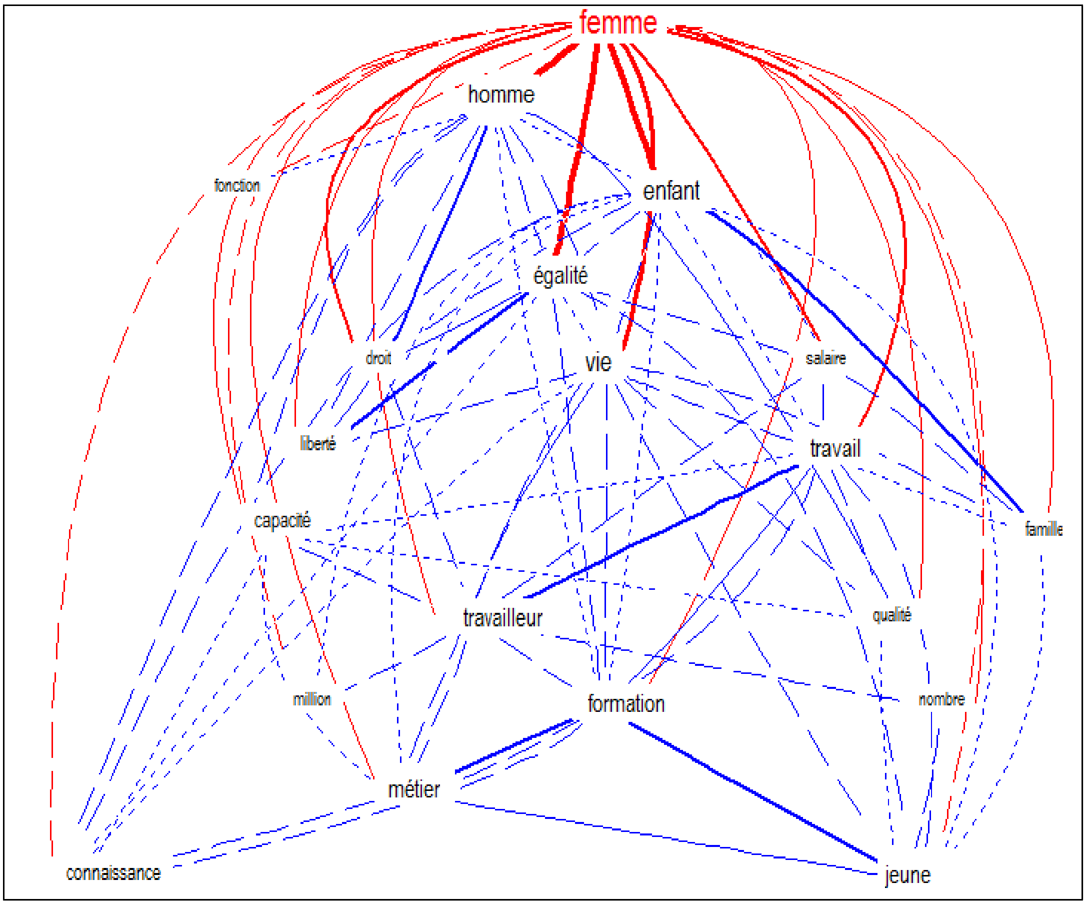


Figure 6. Graph showing multiple co-occurrences from the word "woman" in the presidential corpus (1958-2014) (Hyperbase 10.0 - 2015)

Finally co-occurrence processing allows us to consider the entire text and to highlight speech isotopies. Jean-Marie Viprey (1997) has thus proposed vector representations of co-occurrence matrices Words X Words. Today, software such as Gephi (http://gephi.github.io/) can take into account - from the same matrix - the entire lexical network that a text constitutes (Figure 7).

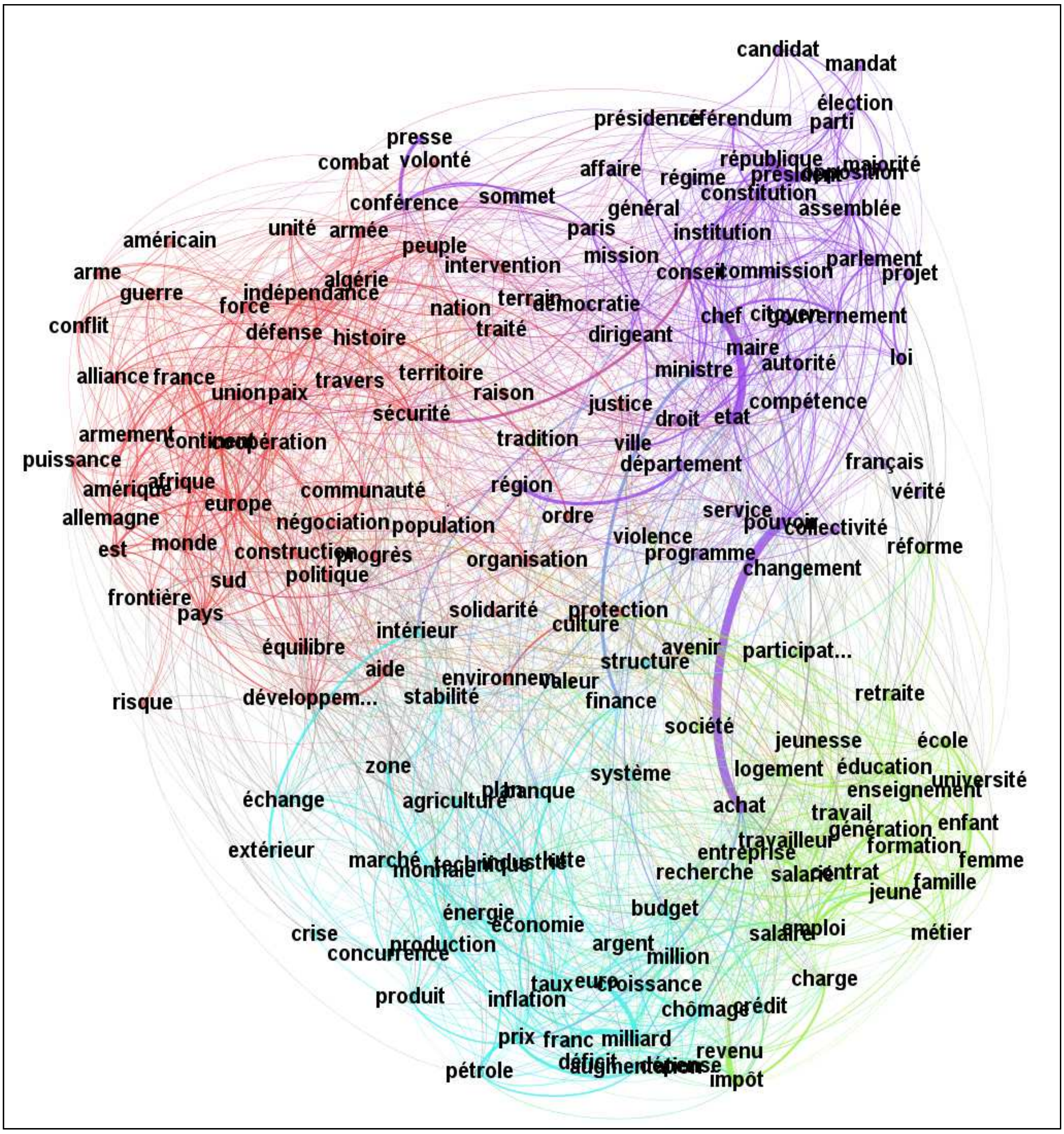


Figure 7. Network of the presidential corpus (1958-2014)

4. Conclusion

Just as there can be no language without history, there can be no history without language, without archives, without speech. Corpus linguistics naturally turns to the historical sciences to contextualise critically the selection of texts collected in a corpus. And history, in turn, naturally turns to linguistics to grasp the intrinsic linguistic nature of its sources and objects.

In France, the historian was from the outset interested in the linguistic turn, and from the 1950s to 1970s his attention was attracted to the development of corpus linguistics aided by computers and statistics. Today, in a wider international movement, digital humanities are re-examining history-linguistic interdisciplinarity, which had been at one time set aside, in the light of the digital revolution that is giving a new shape to concepts as essential as the (digital) archive, the (digital) corpus or the web. Finally, this same digital revolution, in its most technical and most recent aspects, is democratising statistical and computer approaches to the text: processing software, often developed in open source (not to mention simple search engines or hypertextual processing) are become necessary tools for the historian to deal with an ever-expanding digital archive.

So the history we have tried to trace and illustrate in this paper, which takes its national origins in the 1960s in the works of Pierre Guiraud, Jean Dubois and Maurice Tournier from a linguistic point of view, in the works of Régine Robin and Antoine Prost from an historical point of view, in the works of Jean-Paul Benzécri and Charles Muller from a statistical point of view, or the work of Althusser, Pêcheux and Foucault from a philosophical point of view, will no doubt very soon appear as a pre-history. And the aspiration towards interdisciplinary thinking, sometimes disappointed in the past, will become a complete reality.

References

Althusser, Louis *et al.* (1965). *Lire le Capital*. Paris : Maspero.

Anceau, Éric (2012) Pour une histoire politique totale de la France contemporaine, *Histoire, économie & société* 2 (31e année), *111-133.*

Benzécri, Jean-Paul (1973). *L'analyse des données, 2. L'analyse des correspondances*. Paris: Dunod.

Berelson, Bernard (1952). *Content analysis in communication research.* Glencoe: The Free Press.

Bloch, Marc (1939). *La société féodale*. Paris: A. Michel.

Brunet, Etienne (1981). *Le Vocabulaire français de 1789 à nos jours,* 3 tomes. Genève-Paris: Slatkine-Champion.

Brunet, Etienne (2009). *Comptes d'auteurs. Études statistiques de Rabelais à Gracq*. Paris: Champion.

Brunet, Etienne (2011). *Ce qui compte. Méthodes statistiques*. Paris : Champion.

Brunet, Etienne; Vanni, Laurent (2014). Goofre Version 2. *JADT 2014, Proceedings of the 12th International Conference on Textual Data Statistical Analysis conférence invitée*, édité par E. Néé, M. Valette, J.-M. Daube et S. Fleury, Paris: Inalco-Sorbonne nouvelle, pp. 1-14.

Brunet, Etienne (2015). *Au bout du compte. Questions linguistiques.* Paris: Champion

Cahiers de Lexicologie (1968 et 1969), n°13/II et n°14/I: Formation et aspects du vocabulaire politique français, XVIIe-XXe siècle ». [Actes du colloque organisé en avril 1968 par le Centre e Lexicologie politique de l'ENS de Saint-Cloud].

Cottrel, Marie; Deruelle, Benjamin; Lamassé, Stéphane; Letrémy Patrick (2012). Lexical recount between Factor Analysis and Kohonen Map: mathematical vocabulary of arithmetic in the vernacular language of the late Middle Ages. WSOM *AISC 198, 255-264.*

Cotteret, Jean-Marie; Moreau, René (1969). *Le vocabulaire du général de Gaulle*. Paris: A. Colin.

Darnton, Robert (2009). *The Case for Books: Past, Present, and Future*. New York: NY Public Affairs.

Demonet, Michel; Geffroy, Annie; Gouaze, Jean; Lafon, Pierre; Mouillaud, Maurice; Tournier; Maurice (1978/1975). *Des tracts en Mai 68. Mesures de vocabulaire et de contenu*. Paris: Champ libre (1re édition: Presses de la FNSP).

Dubois, Jean (1962). *Le vocabulaire politique et social en France de 1869 à 1872*. Paris: Larousse.

Dupront, Alphonse (1969). Sémantique historique et histoire. *Cahiers de lexicologie 15, I-II*.

Febvre, Lucien (1953). *Combats pour l'histoire*. Paris: Colin, pp. *147-244*.

Foucault, Michel (1966). *Les Mots et les Choses. Une archéologie des sciences humaines*. Paris: Gallimard.

Genet, Jean-Philippe (2012). *Langue et histoire*. Paris: Publication de la Sorbonne.

Genet, Jean-Philippe; Lafon, Pierre (2003). Des chiffres et des lettres : quelques pistes pour l'historien, *Histoire et Mesure XVIII(3/4)*.

Goody, Jack (2000). *The Power of the Written Tradition.* Washington/ London: Smithsonian Institution Press,

Guilhaumou, Jacques (2006). *Discours et événement. L'histoire langagière des concepts,* Besançon: Presse Universitaires de Franche-Comté.

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

Guiraud, Pierre (1960). *Problèmes et méthodes de la statistique linguistique*. Paris: PUF.

Kaal, Bertie; Marks, Isa; Van Elfrinkhof, Annemarie (eds.) (2014). *From Text to Political Positions.* Amsterdam: John Benjamins.

Labbé, Dominique (1990). *Le vocabulaire de François Mitterrand*. Paris: Presses de Sciences Po.

Labbé, Dominique; Monière, Denis (2004). *Le discours gouvernemental. Canada, Quebec, France*. Paris: Champion.

Labbé, Dominique; Monière, Denis (2014). Un siècle et demi de discours gouvernemental au Canada. Contribution de la lexicométrie à l'histoire politique. In: JADT 2014, *Proceedings of the 12th International Conference on Textual Data Statistical Analysis*, edited by E. Néé, M. Valette, J.-M. Daube et S. Fleury. Paris: Inalco-Sorbonne nouvelle, pp. 485-494.

Labbé, **Cyril; Labbé**, **Dominique** (2003). La distance intertextuelle. *Corpus* [En ligne], 2 | 2003, mis en ligne le 15 décembre 2004, consulté le 31 octobre 2014. URL : http://corpus.revues.org/31.

Lafon, Pierre (1984). *Dépouillements et statistiques en lexicométrie*. Genève: Slatkine.

LANGAGES (1969), n°13: L'Analyse du discours.

Lasswell, Harold et al. (1949). *Language of politics*. New York: G. Stewart.

Lebart, Ludovic et al. (1995). *Statistique exploratoire multidimentionnelle*. Paris: Dunod.

Lebart, Ludovic; Salem, André (1994). *Statistique textuelle*. Paris: Dunod.

Lemercier, Claire; Zalc, Claire (2008). *Méthodes quantitatives pour l'historien*. Paris: La Découverte.

Maldidier, Denise (1970). *Analyse linguistique du vocabulaire de la guerre d'Algérie d'après 6 quotidiens parisien* (thèse dir Jean Debois, Paris X-Nanterre)

Marcellesi (1971). *Le congrès de Tours (1920). Etude sociolinguistique*. Paris: Le Pavillon.

Maingueneau, Dominique (1976). *Initiation aux méthodes de l'analyse du discours*. Paris: Hachette.

Maingueneau, Dominique (1987). *Nouvelles tendances en analyse du discours*. Paris: Hachette.

Marchand, Pascal (2007). *Le grand oral. Le discours de politique générale de la V^{e} République*. Bruxelles: Editions De Boeck Université.

Mayaffre, Damon (2000). *Le poids des mots.* Paris: Champion.

Mayaffre, Damon (2004a). *Paroles de président. Jacques Chirac (1995-2003) et le discours présidentiel sous la V^{ème} République.* Paris: Champion.

Mayaffre, Damon (2004b). Analyse logométrique de la cohabitation Chirac/Jospin (1997-2002). Explication de la défaite de Lionel Jospin à l'élection présidentielle de 2002. In: G. Purnelle, C. Fairon, A. Dister (eds.), *JADT 2004. Le poids des mots: volume II, 785-792.* Louvain: Presses universitaires de Louvain. [Hal : http://hal.archives-ouvertes.fr/hal-00554802]

Mayaffre, Damon (2012). *Mesure et démesure du discours. Nicolas Sarkozy (2007-2012).* Paris; Presses de Sciences Po.

Mayaffre, Damon (2014). Plaidoyer en faveur de l'Analyse de Données co(n)Textuelles Parcours cooccurrentiels dans le discours présidentiel français (1958-2014). *JADT 2014, Proceedings of the 12th International Conference on Textual Data Statistical Analysis conférence invitée*, édité par E. Née, M. Valette, J.-M. Daube et S. Fleury, Paris: Inalco-Sorbonne nouvelle, pp. 15-32. [http://lexicometrica.univ-paris3.fr/jadt/jadt2014/01-ACTES/01-JADT2014.pdf]

Mazière, Francine (2005). *L'analyse du discours. Histoire et pratiques*. Paris: Puf.

Mitkov, Ruslan (ed.) (2009). *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press.

Muller, Charles (1968). *Initiation à la statistique linguistique*. Paris: Larousse.

Muller, Charles (1973). *Initiation aux méthodes de la statistique linguistique*. Paris: Hachette

Pêcheux, Michel (1969), *Analyse automatique du discours*, Paris: Dunod.

Peschanski, Denis (1988). *Et pourtant ils tournent. Vocabulaire et stratégie du PCF (1934- 1936)*, Paris: Klincksieck, Publications de l'INALF.

Petit-Dutaillis, Charles (1947). *Les Communes françaises*. Paris: A. Michel.

Prost, Antoine (1974). *Le Vocabulaire des proclamations électorales, 1881, 1885, 1889*, Paris: PUF, Publications de la Sorbonne.

Prost, Antoine (1988). Les mots. In: René Rémond (ed.), *Pour une histoire politique: 255-287*. Paris: Seuil.

Robin, Régine (1973). *Histoire et Linguistique*. Paris: Colin.

Robin, Régine (1986)..Histoire et Linguistique: le malentendu continue. *Langages 81, 121-128*.

Salem, André (1991). Série textuelles chronologiques. *Histoire et Mesure 6, 149-175*.

Tournier, Maurice (1993). *Des mots sur les grèves. Propos d'étymologie sociale (I)*. Paris: Publications de l'INALF.

Tournier, Maurice (1997). *Des mots en politique. Propos d'étymologie sociale (II).* Paris: Publications de l'INALF.

Tournier, Maurice (2002). *Des sources du sens. Propos d'étymologie sociale 3*. Lyon: ENS Éditions.

Viprey, Jean-Marie (1997). *Dynamique du vocabulaire des Fleurs du mal*. Paris: Champion.

The Contribution of Latin to French-Language Quantitative Linguistics: From Lemmatisation to Grammaticometry and Textual Topology

Dominique Longrée, Sylvie Mellet***

*Univ. de Liège. LASLA, 4000 Liège, Belgique
**Univ. Nice Sophia Antipolis, CNRS, BCL, UMR 7320, 06300 Nice, France
dominique.longree@ulg.ac.be, Sylvie.Mellet@unice.fr

0. Introduction

In this overview of quantitative linguistics in France, we focus on works involving Latin corpora. Our contribution points out that statistical handling of digitalized Latin texts is an original and important addition to quantitative linguistics studies, and we investigate the epistemological foundations of this addition. To this end, we go beyond the boundaries of France and look to Belgium, because the development of quantitative studies devoted to Latin texts is a Franco-Belgian achievement, and is based almost entirely on resources produced, beginning in 1961, by the Laboratory for the Statistical Analysis of Ancient Languages (LASLA) at the University of Liège[1]. We first emphasize the role of lemmatisation, and show how this simple operation of abstraction and regrouping allows other more or less complex analysis units to emerge. We then discuss the importance that variability of word order in Latin has assumed with regard to research issues and approaches; finally we discuss software advances and certain necessary adaptations involving digital research methods and quantitative handling made necessary by specific approaches to Latin corpora.

1. From lexicometry to grammaticometry

Preparing Latin corpora for textual data analysis or quantitative linguistics is a particular operation, since the researcher is immediately confronted with the problem of lemmatisation. In the first place, Latin is an inflectional language for which a lexicometry based on graphic forms is problematic. This does not mean that the question of lemmatisation has not been the object of a lively debate about methods for analyzing French or other living languages that are not inflectional (or not as inflectional) [2]. But in Latin, the other alternative – the one that consists in focusing on graphic forms – appears at first glance to be more limiting and

[1] http://www.cipl.ulg.ac.be/Lasla/

[2] We recall especially the revealing title of an article by Etienne Brunet: “Qui dit lemme, dilemme attise”. The explicit rejection of lemmatisation by M. Tournier (“La *lemmatisation* ne résout rien et empire tout”) dates from 1985, but the effort toward lemmatisation of Latin texts by LASLA goes back to the early 1960s.

restrictive, even paralyzing. The workaround consisting in using chains of characters in order to collect all the forms of one lexeme is quite ineffective, since inflection has a much greater impact on the variability of forms: as regards verb conjugation, nominal inflection includes 6 cases and 3 gender types. Most importantly, inflection can have a considerable effect on the forms of radicals. [3]

The automatic production of indexes – the objective of the Latin lexicometry pioneers – was directly in line with the philological tradition[4] and presupposed a particular form of the organization of data. Entries in the index were lemmas, arranged alphabetically, and under each lemma, forms were arranged in a fixed morphological order. Here is an example with the lemma DICO2 (for the verb *dico* / *dicere* of the third conjugation, differentiated from the verb *dico* / *dicare* of the first conjugation by the index 2). This example says that there are 183 occurrences of this lemma in the text among which one form *dico* (first singular person) in the 8th place of the 19th sentence of the 12th chapter of the 5th book of the work; and 3 occurrences of the form *dicis* (second singular person) with their references according to the same reference system as previously:

```
183   DICO2
      dico
         5, 12, 19, 8
      dicis
         3, 12, 19, 8
         3, 15, 13, 6
         3, 21, 2, 8
         …
```

The efforts at lemmatisation made indispensable by such a conception of indexes were therefore ahead of their time, and were inevitably accompanied by morphosyntactic analysis, the results of which could usefully, and with little additional work, be recorded in order to be directed toward other purposes. It should be noted that the granularity of grammatical labelling used by LASLA is very fine indeed. This is another direct result of the principles that governed the compilation of indexes; since it is only from a precise and complete description of the form that we can determine, automatically, its position under the lemma that is also its entry in the index.

Since this morphological information was already in computer memory and easily available, why not use it for other purposes? The first case of this was pedagogical. But researchers also found very interesting new units of analysis in this information: why not study their frequency and distribution? Thus it was that, beginning in the mid-1960s, Étienne Évrard, one of the founder of LASLA,

[3] Cf. Mellet 1996; Purnelle, 1996; Mellet 2002a; Mellet, Sylvie & Purnelle, Gérald, 2002.

[4] You can see a list of LASLA publications at: http://www.cipl.ulg.ac.be/Lasla/publications.html. But see also the early publications by Etienne Brunet and the entire collection, "Travaux de Linguistique Quantitative" from Slatkine, which collects mainly vocabularies and indexes.

began to ask questions about the stability of grammatical categories in Latin texts[5] and about the possibility (or lack thereof) of transposing, in respect of these categories, general laws discovered regarding the distribution of vocabulary. And about ten years later, studies based on grammatical categories began to appear, whether it was a matter of characterizing the writings of an author or a particular work,[6] or of studying the use, distribution and characteristics of a grammatical construction, or more broadly a complete sub-system such as the system of subordinates in Latin.[7]

Otherwise, owing to the inflectional character of the language, word order is less dispositive in Latin as regards the identification of the syntactic functions of various syntagms that make up a sentence. It appears to be more "flexible" than it is in languages such as French or English, but it is no less significant at other linguistic levels (semantic, pragmatic, stylistic, etc.). Thus Latinists early on took an interest in this question, and proceeded to work up counts of various configurations, although in the beginning these remained intuitive and only approximate. In the pioneer work of J. Marouzeau, *L'ordre des mots dans la phrase latine*,[8] there are many vague expressions of this sort: "in many cases", "quite a few examples", "most examples", etc. The first systematic counts appeared with the thesis of F. Charpin, published in 1977.[9] They had as much to do with the order of syntactical constituents as with sequences of identical endings or with chains of pre-accentual sequences. The utility of such counts became apparent to Latinists, and beginning in 1978, J. Perrot[10] emphasized the necessity of statistical enquiries concerning the "norms" for the arrangement of "meaningful material", "comparable to those that have been produced for phonic material": these enquiries were made really practical only through the existence of computerized data bases.[11]

An equally important contribution made in studies on word order in Latin can be found in research on recurrent "formulas". Recognizing such formulas is of particular interest in the linguistic or stylistic characterization of literary works. In an article written in 1989, G. Purnelle[12] presented research concerning "verbal groups [that are] syntactically homogeneous and repeated, whose con-

[5] Évrard 1966. It is true that in the year in which this communication was delivered (conference in 1964), Robert Martin and Charles Muller published "Syntaxe et analyse statistique. La concurrence entre le passé antérieur et le plus-que-parfait dans *La Mort le Roi Artu*"; but most of the counts were obtained manually.

[6] See Fleury 1978; based on observation of a positive specific deviation for the verb *dico* "to say" in the *Satires* of Persius, this paper looked at tenses, modes and persons in terms of which the verb was conjugated, and at its most frequent constructions. See also Delatte 1979, which observes the use of grammatical categories in Ovid's *Héroïdes* and introduces in this context the notion of binary chains of two labels – a notion which we refer to below.

[7] See Delatte, Govaerts & Denooz 1978.

[8] Marouzeau, 4 vol. 1922-1953.

[9] Charpin 1977.

[10] Perrot 1978.

[11] Charpin 1989a and 1989b in which the author completes the counts presented in his thesis.

[12] Purnelle 1989.

stituent elements are contiguous or nearly so in the text". This study did indeed take into account the work of A. Salem and the Saint-Cloud Laboratory concerning repeated segments, but differed in two respects. First, G. Purnelle distinguishes "recurrent verbal groups" whose identification does not at all depend on literary, stylistic or semantic considerations, from "formulas" that corresponded to "a still more homogeneous group which makes up a true fixed expression, if not in terms of the entire language, then at least in terms of the author language or of the genre of the work itself, and which functions as a single semantic entity". In addition, in Latin, a formula can have not only inflectional variations, but also inversions of the order of its constituents, or insertions of terms that can, naturally, be expansions of the formula, but which can also have nothing to do with it. Under these conditions the notion of "repeated segments" can serve as a model, but the analysis must eventually go beyond it. G. Purnelle thus offers a method that is based on the annotations contained in LASLA files, which aims at taking into account variations in morphology and word order; Purnelle suggests the possibility of developments that would take into account the distance in the text separating each occurrence of a given formula, in order to "distinguish actual formulas from what is only a simple repetition of an expression recently employed in the text, which is borne in mind by both the author and the reader". Thus the way was opened for studies of textual dynamics. The programme of research proposed in the article was not immediately carried out by the author, but would be carried forward by others a few years later.

2. Evolution and adaptation of tools and methods

2.1 Software tools

In order to reach these objectives which had been quite specific for them for a long time (constitution of lemmatized alphabetical indexes, study of morpho-syntactic categories, research on word order), Latinists had to create or adapt tools and methods that would later benefit the larger community of researchers in textual data analysis.

Well before the appearance of the first automatic taggers, LASLA designed a semi-automatic lemmatiser that took apart each textual form, comparing it with a lexicon of radicals and affixes, and then provided the philologist with a list of all possible analyses (assignment to a lemma and complete morpho-syntactical description). The philologist then had to choose the correct analysis. This produced files in which each textual form was associated with several different kinds of information:

1. its lemma, such as it appears in the reference dictionary;

2. an index allowing people to distinguish between different homograph lemmas, or to mark proper nouns and the adjectives derived from them;

3. the precise reference, and accordingly the position of the form in the text;

4. a complete morphological analysis in an alphanumeric format;

5. for verbs, syntactical information allowing researchers to distinguish between the predicates of a main clause or of a subordinate clause, and in the latter case, to connect the predicate to its subordinating word.

The first files of LASLA were thus presented as follows:

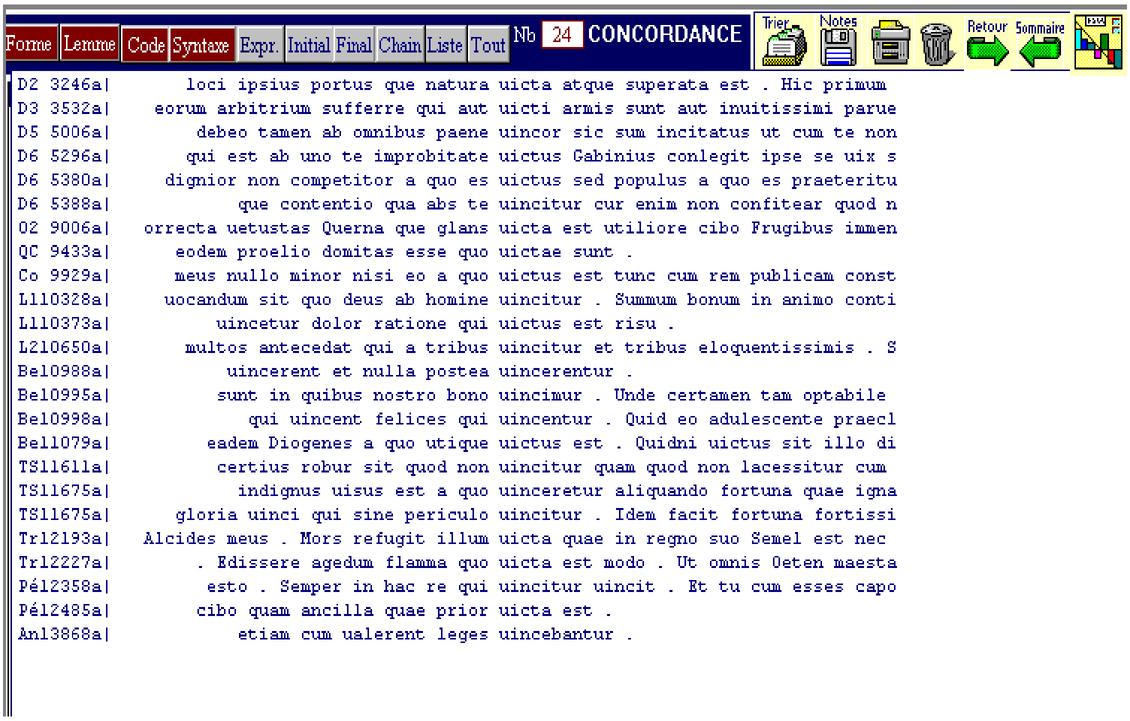
Lemma	**Index**	**Form**	**Reference**	**Morphological Tag**
LIBERTAS		LIBERTATEM	41 001 0001 007 007	13C00
ET	2	ET	41 001 0001 008 008	81000
CONSVLATVS		CONSULATUM	41 001 0001 009 009	14C00
LVCIVS	N	L.	41 001 0001 010 010	12A00
BRVTVS	N	BRUTUS	41 001 0001 011 011	12A00
INSTITVO		INSTITUIT	41 001 0001 012 012	53C14

In this table is pictured an excerpt of the LASLA file corresponding to the following sentence: *Libertatem et consulatum L. Brutus instituit*. To each form of the sentence (third column) is first associated the lemma, with an index which removes a possible ambiguity (for example ET2 = coordinating conjunction "and" while ET1 = adverb "too"). The 4th column gives the reference of this form in the book: here *libertatem* appears in the chapter 41, in the first paragraph of the chapter, in the first sentence of the paragraph; it is the 7th word of the paragraph and the 7th word of the sentence. *Et* is the 8th word of the same sentence and the same paragraph, *consulatum* is the 9th one and so on. Finally, the last column gives an alphanumeric tag: for *libertatem* 13C00 means substantive of the third declination, singular accusative; for *instituit* 53C14 means verb of the third conjugation, third singular person, perfect indicative.

In some cases, the number of data points associated with a single form could go as high as ten. Thus, for a participle such as *regnante*, the following data are given: reference, lemma, part of speech, conjugation type, voice, case, number, mode, tense and gender. Finally, one would eventually be able to determine if the form is the predicate of an ablative absolute (participial proposition).

With the development of personal computers in the early 1990s, the utility of creating software tools that could manipulate all this information became apparent. In the beginning, it was a matter of concordance programmes that could produce not only alphabetical lists of all instances of forms, but also all the forms of a lemma or all the forms associated with one or more given grammatical categories, or with a particular syntactical annotation. In a short time, software developed in order to manipulate the data in the files of LASLA, Estela and Opera Latina first, and Hyperbase-Latin following, allowed the creation of concordances on the basis of a complex search combining the research of a lemma associated with one or more grammatical or syntactic categories. For example, the concordance-maker can provide a contextualized list of the occurrences of the verb *uincere* "to win" only in passive forms in a relative subordinated clause.

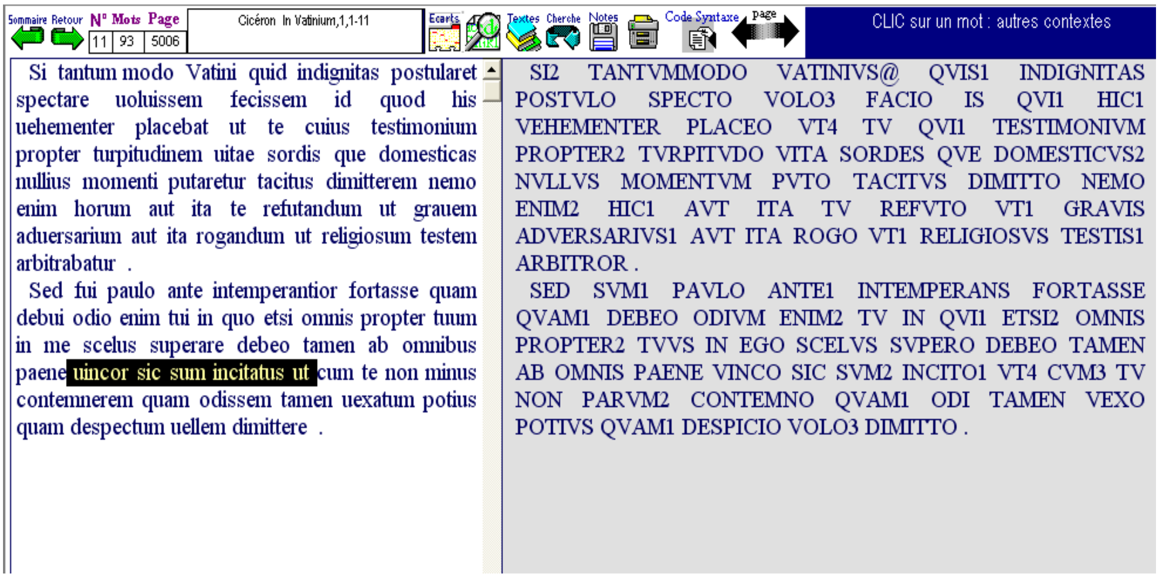
Thanks to the way the data are prepared and structured, and to the resulting enrichment of texts, these concordance makers were able to take into account the multidimensionality of textual data, well before S. Fleury and A. Salem perfected their concept of a "Trameur".



```
D2 3246a|      loci ipsius portus que natura uicta atque superata est . Hic primum
D3 3532a|    eorum arbitrium sufferre qui aut uicti armis sunt aut inuitissimi parue
D5 5006a|        debeo tamen ab omnibus paene uincor sic sum incitatus ut cum te non
D6 5296a|      qui est ab uno te improbitate uictus Gabinius conlegit ipse se uix s
D6 5380a|     dignior non competitor a quo es uictus sed populus a quo es praeteritu
D6 5388a|           que contentio qua abs te uincitur cur enim non confitear quod n
O2 9006a|   orrecta uetustas Querna que glans uicta est utiliore cibo Frugibus immen
QC 9433a|       eodem proelio domitas esse quo uictae sunt .
Co 9929a|       meus nullo minor nisi eo a quo uictus est tunc cum rem publicam const
L110328a|     uocandum sit quo deus ab homine uincitur . Summum bonum in animo conti
L110373a|          uincetur dolor ratione qui uictus est risu .
L210650a|       multos antecedat qui a tribus uincitur et tribus eloquentissimis . S
Be10988a|           uincerent et nulla postea uincerentur .
Be10995a|          sunt in quibus nostro bono uincimur . Unde certamen tam optabile
Be10998a|             qui uincent felices qui uincentur . Quid eo adulescente praecl
Be11079a|        eadem Diogenes a quo utique uictus est . Quidni uictus sit illo di
TS11611a|          certius robur sit quod non uincitur quam quod non lacessitur cum
TS11675a|            indignus uisus est a quo uinceretur aliquando fortuna quae igna
TS11675a|       gloria uinci qui sine periculo uincitur . Idem facit fortuna fortissi
Tr12193a|   Alcides meus . Mors refugit illum uicta quae in regno suo Semel est nec
Tr12227a|         . Edissere agedum flamma quo uicta est modo . Ut omnis Oeten maesta
Pé12358a|        esto . Semper in hac re qui uincitur uincit . Et tu cum esses capo
Pé12485a|         cibo quam ancilla quae prior uicta est .
An13868a|             etiam cum ualerent leges uincebantur .
```

Figure 1. Concordance of passive forms in the lemma *uincere* "to win" in a relative subordinated clause, in the whole Latin corpus of LASLA

An illustration of this premature concern for multidimensionality can be found in the simultaneous display, by Hyperbase, of a single textual sequence of both forms and lemmas corresponding to them, or of both forms and morpho-syntactic codes associated with them. Such an illustration can be found as well in the simultaneous display, in the dictionary, of forms, lemmas, and morpho-syntactic codes.



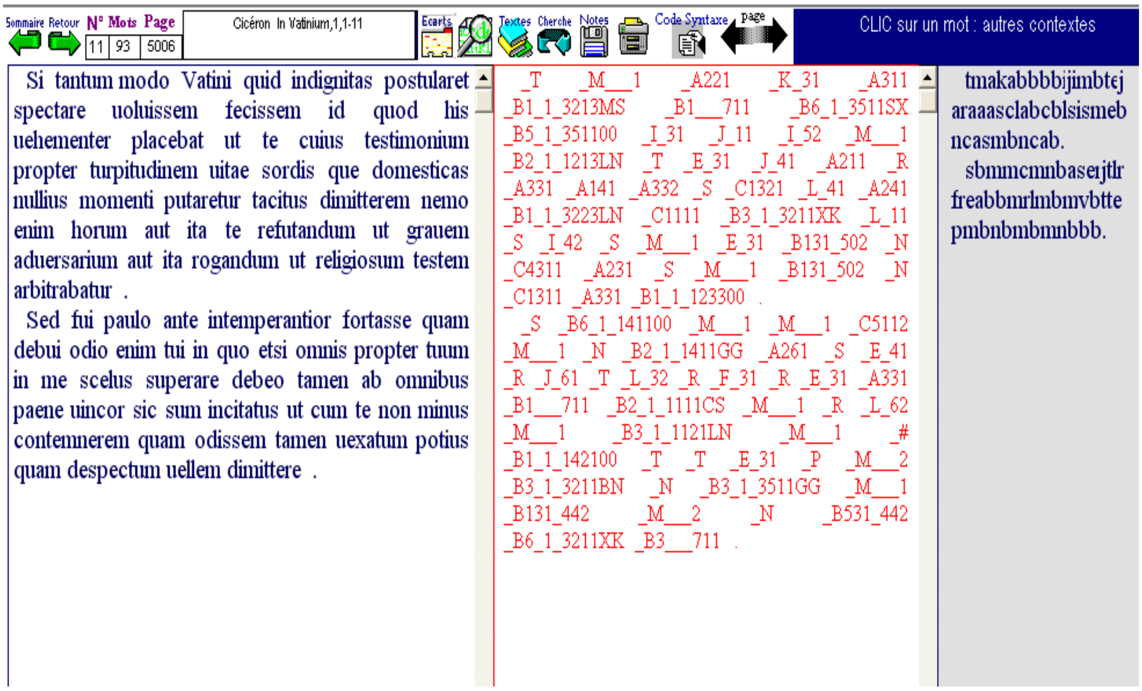
Si tantum modo Vatini quid indignitas postularet spectare uoluissem fecissem id quod his uehementer placebat ut te cuius testimonium propter turpitudinem uitae sordis que domesticas nullius momenti putaretur tacitus dimitterem nemo enim horum aut ita te refutandum ut grauem aduersarium aut ita rogandum ut religiosum testem arbitrabatur .

Sed fui paulo ante intemperantior fortasse quam debui odio enim tui in quo etsi omnis propter tuum in me scelus superare debeo tamen ab omnibus paene uincor sic sum incitatus ut cum te non minus contemnerem quam odissem tamen uexatum potius quam despectum uellem dimittere .

SI2 TANTVMMODO VATINIVS@ QVIS1 INDIGNITAS POSTVLO SPECTO VOLO3 FACIO IS QVI1 HIC1 VEHEMENTER PLACEO VT4 TV QVI1 TESTIMONIVM PROPTER2 TVRPITVDO VITA SORDES QVE DOMESTICVS2 NVLLVS MOMENTVM PVTO TACITVS DIMITTO NEMO ENIM2 HIC1 AVT ITA TV REFVTO VT1 GRAVIS ADVERSARIVS1 AVT ITA ROGO VT1 RELIGIOSVS TESTIS1 ARBITROR .

SED SVM1 PAVLO ANTE1 INTEMPERANS FORTASSE QVAM1 DEBEO ODIVM ENIM2 TV IN QVI1 ETSI2 OMNIS PROPTER2 TVVS IN EGO SCELVS SVPERO DEBEO TAMEN AB OMNIS PAENE VINCO SIC SVM2 INCITO1 VT4 CVM3 TV NON PARVM2 CONTEMNO QVAM1 ODI TAMEN VEXO POTIVS QVAM1 DESPICIO VOLO3 DIMITTO .

Figure 2. Parallel reading of a text excerpt, set up as a string of forms and a string of lemmas



Cicéron In Vatinium,1,1-11

Si tantum modo Vatini quid indignitas postularet spectare uoluissem fecissem id quod his uehementer placebat ut te cuius testimonium propter turpitudinem uitae sordis que domesticas nullius momenti putaretur tacitus dimitterem nemo enim horum aut ita te refutandum ut grauem aduersarium aut ita rogandum ut religiosum testem arbitrabatur .

Sed fui paulo ante intemperantior fortasse quam debui odio enim tui in quo etsi omnis propter tuum in me scelus superare debeo tamen ab omnibus paene uincor sic sum incitatus ut cum te non minus contemnerem quam odissem tamen uexatum potius quam despectum uellem dimittere .

```
 _T   _M__1    _A221   _K_31    _A311
_B1_1_3213MS    _B1__711    _B6_1_3511SX
_B5_1_351100   _I_31   _J_11   _I_52   _M__1
_B2_1_1213LN   _T   _E_31   _J_41   _A211   _R
_A331  _A141  _A332  _S  _C1321  _L_41  _A241
_B1_1_3223LN   _C1111   _B3_1_3211XK   _L_11
_S  _I_42  _S  _M__1  _E_31  _B131_502  _N
_C4311   _A231   _S   _M__1  _B131_502   _N
_C1311  _A331  _B1_1_123300  .
 _S  _B6_1_141100  _M__1  _M__1  _C5112
_M__1  _N  _B2_1_1411GG  _A261  _S  _E_41
_R  _J_61  _T  _L_32  _R  _F_31  _R  _E_31  _A331
_B1__711  _B2_1_1111CS  _M__1  _R  _L_62
_M__1      _B3_1_1121LN      _M__1      _#
_B1_1_142100   _T   _T   _E_31   _P   _M__2
_B3_1_3211BN   _N   _B3_1_3511GG   _M__1
_B131_442      _M__2      _N      _B531_442
_B6_1_3211XK  _B3__711  .
```

tmakabbbbijimbtej araaasclabcblsismeb ncasmbncab.

sbmmcmnbaseıjtlr freabbmrlmbmvbtte pmbnbmbmnbbb.

Figure 3. Parallel reading of a text excerpt set up as a string of forms and as a string of morphosyntactic codes

Latinists' interest in this multidimensional approach to Latin texts explains, for the most part, why Hyperbase, which was adapted to Latin at a relatively early date, later became one of the first software programmes that simultaneously took into account lemmas and grammatical categories for corpora of French, English or Portuguese texts: it was only necessary to wait until trainable automatic taggers were able to furnish dependable morphosyntactical information for these languages.

Apart from concordance makers, software for manipulating Latin textual data rapidly came to include functionality based on statistical calculations such as the calculation of chi-square, reduced variations, specificities. These functions have proven particularly useful in order to characterize the texts in terms of their use of grammatical categories, and also in order to gain better understanding of the function of these categories, which from that point on could be grasped in terms of the specificity of their distribution in context.[13] This was one of the objectives of S. Mellet in her thesis devoted to the imperfect indicative in Latin.[14] Other examples can be found in the article by C. Bertrand on verbal forms and the structure of phrases in the *Historia Augusta*[15] or in the article by D. Renard on the parts of discourse used by various characters in the *Satyricon* of Petronius.[16]

[13] See Evrard & Mellet 1998.

[14] Mellet 1987.

[15] Bertrand 1982.

[16] Renard 2000.

2.2 Statistical methods

In a first approach, the statistical treatment of grammatical categories seems to be able to use the methods and calculations that are applied to lexicons (forms or lemmas). As we have just noted, one can easily and accurately calculate the grammatical specifications of a text, or of any other part of a corpus. One can go further still in taking account of the effect of grammatical functions in the calculation of the specific co-occurrents of a keyword.[17] Finally, one may integrate grammatical categories into all methods of multidimensional calculation, and on this grammatical basis handle data matrices in order to extract from them a graphic representation that is then submitted for interpretation to a linguist. Factorial analyses of correspondences or tree analyses made on the basis of the distribution of grammatical categories in different areas of a corpus are quite expressive and often succeed in corroborating certain classifications (according to authors, to genres, to a chronology, etc.) that in turn appear in the results of lexicometric treatments. But they bring to this a healthy independence, in relation to the thematics of works. They also offer complementary elements of analysis that make more subtle classifications possible, as we were able to demonstrate as early as 1987, and several times thereafter.[18] Another pioneer in this area has been D. Biber,[19] who experimented in English with “grammaticometric” techniques, and was able to demonstrate their interest for linguistics. These methods were applied to different kinds of corpora in different languages.[20] However Latinists have retained the distinction of working with very fine grammatical categories thanks to the initial material they had to work on. The contribution of the quantitative analysis of grammatical categories is particularly valuable when a corpus groups together works that share a general theme, a vocabulary and certain conventional motifs: in fact the use of grammatical categories is more likely to be independent from this thematic and semantic framework, and to escape the control of the writer: therefore they give access to deep and intrinsic characteristics of the author’s style. This is one of the benefits used in the thesis of Caroline Philippart de Foy, devoted to an *Étude d’un corpus de traductions médiolatines d’origine grecque*,[21] the analyses of which have allowed us to globally characterize different groups of translations, initially defined on the basis of historical and philological sources, to confirm the pertinence of this classification without masking the heterogeneous aspects of each group, and to suggest some definite attributions to particular authors, or at least to suggest a school of translation, in the case of orphaned works. Used for purposes having to do with the characterization of works and not just for classification, these multidimensional methods provide solid complementary information, amounting to significant added value when complementing lexical analysis.

[17] Longrée & Mellet, 2012

[18] Mellet, 1987; Mellet 1998; Mellet 2002b; Longrée 2004; Longrée 2005.

[19] Biber, 1988.

[20] See for example Kastberg 2006; Loiseau, Poudat & Ablali 2006.

[21] Philippart de Foy 2008.

Just the same, “grammaticometricians” are quickly led to question the pertinence of applying the statistical tools of classical statistical linguistics to grammatical categories. Grammatical categories have distributional specificities that force us to rethink using the statistical methods used in classical lexicometry. On one hand it is rare for a major grammatical category to be completely absent in a text. We cannot calculate intertextual distances according to models that are set up in terms of the presence or absence of a variable in different texts that are compared. It is necessary to work with frequencies, and to develop new algorithms for this purpose. Also, the enumeration of this new type of variables produces matrices that sometimes have columns that contain very few data points, but which nonetheless contain important information in the eyes of the philologist or the expert in stylistics (for example, the use of the infinitive of narration used by historians). It is thus necessary to recover this information in the calculation of distance, even though the number of instances is much too low to allow for its being analyzed in terms of classical statistical analysis. One of the adaptations suggested by S. Mellet and X. Luong[22] was to make the computation depend upon numerical values not corresponding to the number of the instances of each category in each text of a corpus, but to a numerical ordering according to this number of instances: the matrix of initial data is converted so that it assigns to each text a number in an order that represents its rank with respect to the use it makes of various grammatical categories being examined. This classification produces first a pre-order when it gives rise to cases of equal standing; this pre-order can be transformed into a classification of “middle ranks”. In every case, we are working with homogeneous data distributed over a reduced scale, which can be submitted to a simple Euclidean calculation of distance (all forms of weighting are useless here). Results obtained in terms of works and classifications are very satisfying, in that they reveal groupings that are coherent but not completely obvious in terms of philological knowledge. It should be noted that this method does not appear to have been used by others since it was published.

Thus we see that lemmatisation and tagging of Latin texts allows us to escape from a illusory naturalness of data, and to create new analysis through a double process of abstraction and construction of the object studied. This approach reaches its highest point thanks to the conceptualization of a complex object – the motif – in the new epistemological framework of textual topology.

2.3 Textual topology

In this process of construction of the object of study, it appears that just counting the appearances of a form, a lemma or a code, taken in isolation, is not enough to give an account of the specific textual dynamics of a work. The recurrent succession of certain sequences of items and the configuration of morpho-syntactical sequences belonging to certain works appear as particularly pertinent elements of analysis. Thus D. Longrée and X. Luong in 2003 published an initial article on sequences of verb tenses: identified after a reduction of the text to a

[22] Luong & Mellet 2003.

chain of the morphological codes of predicates of main clauses, these sequences have been chosen as a parameter for the characterization of Latin historians' writings.[23] This first attempt at integrating the ordered linearity of a text through a quantitative treatment foreshadowed the later development of methods, which under the aegis of the famous "beyond the bag of words" (2005-2006),[24] were no longer content to apply to texts the traditional statistical method of the Polya urn scheme. In fact, when we work on an author's style or on the structure of a work, we quickly see the necessity of taking account – even in the context of a quantitative treatment – of the organization of a syntagmatic axis grasped at one and the same time in terms of short range (in repetitive sequences of a single form or of a single grammatical structure, and in the breaks in these sequences) and of long range (in the distribution of the studied units across the different parts of a text).[25] Such a manner of apprehending textual structure led J.P. Barthelémy, D. Longrée, X. Luong to S. Mellet to explore the possibility of a topological modelling of texts.[26] Over short ranges, the aptitude of a grammatical category to be systematically associated with other categories from a syntagmatic point of view or to favour certain collocations can be apprehended through studies of *voisinages* (neighbourhoods).[27] Over long ranges, the distribution of a sequence according to parts of the text (introduction, narration, commentary, conclusion, etc.) can be analyzed through a method of cutting the texts into fixed or variable sections,[28] and its rhythm of appearance can be analyzed through the method of the calculation of "bursts" ("rafales").[29]

Taken together, these approaches allow us to get beyond the stage at which texts are considered as simple ensemble-type structures[30] and to take into account the form of the text as a whole and in terms of its parts. The notion of topological space applied to texts has been theorized: the text becomes an ensemble of points, each of which has a family of neighbourhoods, and can thus be studied through the concepts and tools we borrow from mathematical topology – more precisely, discrete topology.

As we have deepened our study of neighbourhood structures, we have become aware that some of these have properties which make them textual objects that are particularly worthy of study: they are multidimensional (they associate lexical and grammatical constraints), ordered and recurrent, and they possess a textual function (structuring or characterizing). We have given the

[23] Longrée & Luong 2003, and also Longrée & Luong 2005; Longrée & Mellet 2007.

[24] From the name of the Workshop of the 28th Annual International Conférence of ACM SIGIR.

[25] It is interesting to note that this context is also one in which methodological work was developed on co-occurrences, whose methodological point of view is not unrelated to our purpose. Cf. Mayaffre 2008a and 2008b.

[26] Mellet & Barthélemy 2007; Barthélemy, Longrée, Luong & Mellet 2009.

[27] In the mathematical sense of the word; see Longrée, Luong & Mellet 2004

[28] Longrée, Luong & Mellet 2004 and 2006

[29] Lafon 1981. For an application, see for example Lenoble 2006, especially pp. 479-493.

[30] Longrée, Luong, Juillard & Mellet 2007.

name of "***motifs***" to these structures, which constitute elements of the modelling of a text as a topological space [31] and that formalize in a more systematic manner the properties that Gérald Purnelle had attributed to "formulas". The property of recurrence makes them good candidates for treatment by means of textometric tools. Motifs allow us to automatically characterize either the various parts of a text, or the different texts of a corpus.[32] Otherwise, thanks to the articulation between its basic schematic form and its textual functionality (which contributes in an important way to the stability of its recognition), a "motif" can feature variants (permutation of two elements; commutation within a paradigmatic series; insertion, expansion or erasure; inflectional variation). Thus the study of motifs returns us to the problem of word order, which is a guiding thread for quantitative linguistics in Latin. Subsuming the notions of repeated segments, collocations and colligations, it permits an enlargement of the domain of phraseology and constitutes a contribution – relatively unexpected – from Latin to the disciplinary field that is generally devoted to the terminology of living languages.[33] Another interesting relationship involving motifs takes place in connection with another domain of Natural Language Processsing (NLP), the domain of *data mining*, as soon as this begins to integrate sequential constraints into its methodology.[34] Finally, the notion of motif also allows openings toward psychology, inasmuch as psychologists might use it as a tool for analyzing the verbal production of subjects under examination, or insofar as it might function as a particularly complete representation of lexical associations whose cognitive functioning is thus modelled. In this area, work is underway.

Naturally, the constitution of a new unit of analysis, on the one hand, the taking into account of the topological dimension of texts, on the other hand, have led to new software developments, most often in collaboration with Latinists. The functionality of Hyperbase-Latin, and also that of Hyperbase-français have been considerably enriched in recent years. TXM has developed, especially under the influence of the reflection engaged in by B. Pincemin on the necessary modelling of texts.[35] And the designers of textometric software have discussed and collaborated with specialists in NLP in order to develop non-supervised research tools for motifs, for example, the online program for the extraction of sequential motifs, that is, SDMC, "Sequential Data Mining under Constraints". [36]

This state of the art and this assessment of the research progress show that the contribution of classical languages to quantitative linguistics is based first on

[31] Longrée, Mellet & Luong 2008; Mellet & Longrée 2009; Mellet & Longrée 2012.

[32] Gohy & Martin Leon 2012; Magri & Purnelle 2012.

[33] Longrée & Mellet 2013. This study contributes, once again, to the research in this area the strong multi-dimensionality and precise labelling of numerical Latin data, and also enhances the approaches developed, for example, in Biber 2009 or Grezka & Poudat 2012.

[34] Quiniou, Cellier, Charmois & Le Gallois 2012.

[35] Pincemin 2008; Pincemin, Heiden, Lay, Leblanc & Viprey 2010; Heiden, Magué & Pincemin 2010.

[36] Béchet, Cellier, Charnois, Crémilleux & Quiniou 2013.

the relative anteriority of computerized and tagged corpora for these languages, and consequently on the lines of questioning they prematurely supported. This longevity of a line of research in textual data analysis has been accompanied by a strong methodological consideration, linked to the specific quality of Latin data. The pathways opened up have not always been followed up or investigated by others, but in a certain number of cases, convergences have given rise to particularly fruitful collaborations, especially in order to comprehend the textual structure from a global point of view and to develop the software necessary for following up this global approach.

References

Barthélemy, Jean-Pierre ; Longrée, Dominique ; Luong Xuan ; Mellet, Sylvie (2009). Représentations du texte pour la classification arborée et l'analyse automatique de corpus : application à un corpus d'historiens latins. *Mathematics and Social Sciences* (47ème année) *187, 3 : 107-121.*

Béchet, Nicolas; Cellier, Peggy; Charnois, Thierry; Crémilleux, Bruno; Quiniou, Solen (2013). SDMC: un outil en ligne d'extraction de motifs séquentiels pour la fouille de textes. In: *Actes de la Conférence Francophone sur l'Extraction et la Gestion des Connaissances (EGC'13)*, Toulouse 2013 [see HAL web-site: http://hal.archives-ouvertes.fr/hal-00817074].

Bertrand, Cécile (1982), « L'Histoire Auguste : formes verbales et structure des phrases dans la *Vita Hadriani* et la *Vita Heliogabali* », *RELO, 18, 59-79.* [http://promethee.philo.ulg.ac.be/RISSHpdf/annee1982/CBertrand.pdf]

Biber, Douglas (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *IJCL 14(3), 275-311.*

Brunet, Étienne (2000). Qui lemmatise, dilemme attise. *Lexicometrica* 2. [see *Lexicometrica* web-site: http://lexicometrica.univ-paris3.fr/article/numero2/brunet2000.PDF].

Charpin, François (1977). *L'idée de phrase grammaticale et son expression en latin*. Lille – Paris: H. Champion.

Charpin, François (1989a). Étude de syntaxe énonciative: l'ordre des mots et la phrase. In: G. Calboli (ed.), *Subordination and other topics in Latin, Proceedings of the third Colloquium on Latin linguistics, Bologna, 1-5 April 1985, 503-520* (Studies in Language Companion Series, 17), Amsterdam – Philadelphia: John Benjamins.

Charpin, François (1989b). Les finales homonymes dans le discours latin. *Revue, Informatique et Statistique dans les Sciences humaines, 25, 65-108.* [http://promethee.philo.ulg.ac.be/RISSHpdf/Annee1989/Articles/FCharpin.pdf].

Delatte, Louis (1979), « Recherches statistiques sur les *Héroïdes* XVI et XVII d'Ovide », *RELO, 14 (2), 1-61.* [http://promethee.philo.ulg.ac.be/RISSHpdf/annee1979/02/LDelatte.pdf/].

Delatte, Louis; Govaerts, Suzanne; Denooz, Joseph (1978). *L'ordinateur et le latin. Techniques et méthodes, morphologie, syntaxe, lexicologie,* *stylistique*, Liège, LASLA [(http://promethee.philo.ulg.ac.be/LASLApdf/Lordinateuretlelatin.pdf].

Évrard, Étienne (1966). La fréquence des phénomènes grammaticaux est-elle constante? In: *Actes du premier colloque international de linguistique appliquée (Nancy, 26-31 octobre 1964).* Nancy: PUN (« Annales de l'Est »), *157-162.*

Évrard, Étienne; Mellet, Sylvie (1998). Les méthodes quantitatives en langues anciennes. *LALIES 18 : 111-155*.

Fleury, Philippe (1978). « Essai d'exploitation de données fournies par des moyens informatiques sur les *Satires* de Perse », *RELO, 14 (3), 45-70.* [http://promethee.philo.ulg.ac.be/RISSHpdf/annee1978/03/PFleury.pdf]

Gohy, Stéphanie; Martin, Leon Benjamin (2012). Détection automatique des textes épistolaires du corpus néo-égyptien: méthodes exploitant la récurrence de motifs discriminants. In: Anne Dister, Dominique Longrée, Gérald Purnelle, *JADT 2012, Actes des 11e Journées internationales d'analyse statistique des données textuelles,* Liège, *487-500*. [see *Lexicometrica* web-site*:* http://lexicometrica.univ-paris3.fr/jadt/jadt2012/tocJADT2012.htm)

Grezka, Aude; Poudat, Céline (2012). Building a database of French frozen adverbial phrases », *in Proceedings of LREC 2012, 685-692*. [see LREC web-site: http://www.lrec-conf.org/proceedings/lrec2012/pdf/1020_ Paper.pdf].

Heiden, Serge; Magué, Jean-Philippe; Pincemin, Bénédicte (2010) TXM: Une plateforme logicielle open-source pour la textométrie-conception et développement. In: Sergio Bolasco, Isabella Chiari, Luca Giuliano (eds.), *JADT 2010, Statistical Analysis of Textual Data - Proceedings of 10th International Conference*. Rome Edizioni Universitarie di Lettere Economia Diritto. [see *Lexicometrica* web-site*:* http://lexicometrica.univ-paris3.fr/jadt/jadt2010/allegati/ JADT-2010-1021-1032_025-Heiden.pdf].

Kastberg, Sjöblom Margareta (2006). *L'écriture de J.M.G. Le Clézio. Des mots aux thèmes*. Paris: Honoré Champion.

Lafon, Pierre (1981). Statistiques des localisations des formes d'un texte. *Mots 2, 157-187.*

Lenoble, Muriel (2006). *Le passif impersonnel du type uenitur chez les historiens latins (César, Salluste et Tacite). Essai méthodologique, quantitatif et descriptif*. Unpublished dissertation, Facultés Saint-Louis, Bruxelles.

Loiseau, Sylvain; Poudat, Céline; Ablali, Driss (2006). Exploration contrastive de trois corpus de sciences humaines. In: Jean-Marie Viprey (ed.), *JADT 2006, 8èmes Journées internationales d'Analyse statistique des Données Textuelles,* Besançon. [see *Lexicometrica* web-site*:* http://lexicometrica.univ-paris3.fr/jadt/jadt2006/PDF/II-056.pdf]

Longrée, Dominique (2004). Une approche statistique de la concurrence entre démonstratifs chez les historiens latins (César, Salluste, Tacite). In: C. Bodelot (éd.), *Anaphore, cataphore et corrélation en latin.* Clermont: Presses Universitaires Blaise Pascal, (Collection « Erga », Recherches sur l'Antiquité, 6), *157-178.*

Longrée, Dominique (2005). Temps verbaux et spécificités stylistiques chez les historiens latins: sur les méthodes d'analyse statistique d'un corpus lemmatisé. In: G. Calboli (ed.), *Papers on Grammar, IX, 2, Latina Lingua !, Proceedings of the Twelfth International Colloquium on Latin Linguistics: 863-875.* Roma.

Longrée, Dominique; Luong Xuan (2003). Temps verbaux et linéarité du texte: recherches sur les distances dans un corpus de textes latins lemmatisés. *Corpus* 2 (« La distance intertextuelle »), *119-140.* [see : Revues.org: http://corpus.revues.org/33].

Longrée, Dominique ; Luong Xuan (2005). Spécificités stylistiques et distributions temporelles chez les historiens latins: sur les méthodes d'analyse quantitative d'un corpus lemmatisé. In: G. Williams (ed.), *La Linguistique de Corpus: 141-152.* Rennes : P.U.R. (Rivages Linguistiques).

Longrée, Dominique ; Luong Xuan; Juillard, Michel; Mellet, Sylvie, (2007). The concept of Text Topology. Some applications to Verb-Form Distributions in Language Corpora. *Literary and Linguistic Computing 22(2), 167-186.*

Longrée, Dominique; Luong Xuan ; Mellet, Sylvie (2004). Temps verbaux, axe syntagmatique, topologie textuelle : analyses d'un corpus lemmatisé. In : Gérald Purnelle, Cédric Fairon, Anne Dister (eds), *JADT 2004, Le poids de mots, Actes des 7e Journées internationales d'Analyse statistique des données textuelles.* Louvain-la-Neuve, *743-752.* [see *Lexicometrica* web-site*:* http://lexicometrica.univ-paris3.fr/jadt/jadt2004/pdf/JADT_071.pdf].

Longrée, Dominique; Luong Xuan; Mellet, Sylvie (2006). Distance intertextuelle et classement des textes d'après leur structure: méthodes de découpage et analyses arborées. In: Jean-Marie Viprey, Claude Condé, Alain Lelu, Max Silberztein (eds.), *JADT 2006, 8èmes Journées internationales d'Analyse statistique des Données Textuelles: 643-654.* Besançon : Presses universitaires de Franche-Comté [see *Lexicometrica*, web-site*:* http:// lexicometrica.univ-paris3.fr/jadt/jadt2006/PDF/II-057.pdf].

Longrée, Dominique; Mellet, Sylvie (2007). Temps verbaux et prose historique latine: à la recherche de nouvelles méthodes d'analyse statistique. In: G. Purnelle, J. Denooz (eds), *Ordre et cohérence, en latin: 117-128.* Genève: Droz.

Longrée, Dominique; Mellet, Sylvie (2012). Asymétrie de la cooccurrence et contextualisation. Le rôle de la flexion casuelle dans la structuration des réseaux cooccurrentiels d'un mot-pôle en latin. *Corpus 11, 91-128.* [see Revues.org : http://corpus.revues.org/2230].

Longrée, Dominique; Mellet, Sylvie (2013). Le motif: une unité phraséologique englobante ? Etendre le champ de la phraséologie de la langue au discours. *Langages 189, 65-79.*

Longrée, Dominique; Mellet, Sylvie; Luong Xuan (2008). Les motifs: un outil pour la caractérisation topologique des textes. In: *JADT 2008, Actes des 9èmes Journées internationales d'Analyse statistique des Données Textuelles*. Vol. *2, 733-744.* Lyon: Presses de l'ENS. [see *Lexicometrica* web-site*:* http://lexicometrica.univ-paris3.fr/jadt/jadt2008/pdf/longree-luong-mellet.pdf].

Luong Xuan; Mellet, Sylvie (2003). Mesures de distance grammaticale entre les textes. *Corpus* 2 (« Les distances intertextuelles »), *141-166.* [see Revues.org: http://corpus.revues.org/34].

Magri, Véronique; Purnelle, Gérald (2012). Mot à mot, brin par brin: les suites [Nom préposition Nom] comme motifs. In: Anne Dister, Dominique Longrée, Gérald Purnelle (eds.), *JADT 2012, Actes des 11e Journées internationales d'analyse statistique des données textuelles: 659-673,* Liège. [see *Lexicometrica* web-site*:* http://lexicometrica.univ-paris3.fr/jadt/jadt2012/tocJADT2012.htm].

Mayaffre, Damon (2008a). Quand 'travail', 'famille', 'patrie' co-occurrent dans le discours de Nicolas Sarkozy. Etude de cas et réflexion théorique sur la co-occurrence. In: Serge Heiden, Bénédicte Pincemin (eds.)*, JADT 2008, 9es journées internationales d'analyse statistique des données textuelles: vol. 2, 811-822.* Lyon: Pul, [see *Lexicometrica* web-site*:* http://lexicometrica.univ-paris3.fr/jadt/jadt2008/pdf/mayaffre.pdf].

Mayaffre, Damon (2008b). De l'occurrence à l'isotopie. Les co-occurrences en lexicométrie. *Sémantique & Syntaxe 9, 53-72.*

Marouzeau, Jean (1922). *L'ordre des mots dans la phrase latine.* Vol. I. *Les groupes nominaux*. Paris: Champion.

Marouzeau, Jean (1938). *L'ordre des mots dans la phrase latine.* Vol. II. *Le verbe*. Paris: Champion.

Marouzeau, Jean (1949). *L'ordre des mots dans la phrase latine: Les articulations de l'énoncé*. Paris: Champion.

Marouzeau, Jean (1953). *L'ordre des mots en latin*. Volume complémentaire. Paris: Champion.

Martin, Robert; Muller, Charles (1964). Syntaxe et analyse statistique. La concurrence entre le passé antérieur et le plus-que-parfait dans *La Mort le Roi Artu. Travaux de Linguistique et de Littérature* 2: 1-27.

Mellet, Sylvie (1987). *L'imparfait de l'indicatif en latin.* Louvain – Paris: Peeters.

Mellet, Sylvie (1994). Logiciels d'exploitation de la banque de données de textes latins du L.A.S.L.A. *Revue, Informatique et Statistique dans les Sciences humaines 30, 91-108.*
[http://promethee.philo.ulg.ac.be/RISSHpdf/Annee1994/Articles/SMellet.pdf]

Mellet, Sylvie (1996). Les atouts de la lemmatisation. In : G. Moracchini (ed.) *Actes du Colloque international «Bases de données linguistiques: conceptions, réalisations, exploitations»: 309-316* (Corte 11-13 octobre 1995), Univ. de Corse / Univ. Nice Sophia Antipolis.

Mellet, Sylvie (1998). Les tragédies de Sénèque vues à travers Hyperbase. In: S. Mellet ; M. Vuillaume (eds.), *Mots chiffrés et déchiffrés, Mélanges offerts à Étienne Brunet: 255-271*. Paris: Champion.

Mellet, Sylvie (2002a). Lemmatisation et encodage grammatical: un luxe inutile? *Lexicometrica*, [see *Lexicometrica* web-site*:* http://lexicometrica.univ-paris3.fr/thema/thema1/spec1-texte2.pdf].

Mellet, Sylvie (2002b). La lemmatisation et l'encodage grammatical permettent-ils de reconnaître l'auteur d'un texte. *Médiévales* 42 (« Le latin dans les textes »), *13-26.*

Mellet, Sylvie; Barthélemy, Jean-Pierre (2007). La topologie textuelle: légitimation d'une notion émergente. *Lexicometrica* n°spécial, 12 pages. [see *Lexicometrica* web-site: http://lexicometrica.univ-paris3.fr/numspeciaux/special9/mellet.pdf].

Mellet, Sylvie; Longrée, Dominique (2009). Syntactical motifs and textual structures. *Belgian Journal of Linguistics 23, 161-173* (« New Approaches in Textual Linguistics »).

Mellet, Sylvie; Longrée, Dominique (2012). Légitimité d'une unité textométrique: le motif. In: Anne Dister, Dominique Longrée, Gérald Purnelle (eds), *JADT 2012, Actes des 11e Journées internationales d'analyse statistique des données textuelles: 715-728.* Liège [see : *Lexicometrica* web-site*:* http://lexicometrica.univ-paris3.fr/jadt/jadt2012/Communications/Mellet,%20 Sylvie%20et%20al.%20-%20Legitimite%20d%27une%20unite%20 textometrique.pdf]

Mellet, Sylvie; Purnelle, Gérald (2002). Les atouts multiples de la lemmatisation: l'exemple du latin. In: A. Morin; P. Sébillot (eds.), *JADT 2002(*2), *529-538, 6èmes Journées internationales d'Analyse statistique des Données Textuelles,* Saint-Malo: Irisa et Inria.

Perrot, Jean (1978). Ordre des mots et structures linguistiques. *Langages 50, 17-26.*

Pincemin, Bénédicte (2008). Modélisation textométrique des textes. In: Serge Heiden, Bénédicte Pincemin (eds), *JADT 2008, Actes des 9es Journées internationales d'Analyse statistique des Données Textuelles, vol. II, 949-960.* Lyon: Presses Universitaires de Lyon, [see *Lexicometrica* web-site*:* http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2008/pdf/pincemin.pdf].

Pincemin, Bénédicte; Heiden, Serge; Lay, Marie-Hélène; Leblanc, Jean-Marc; Viprey, Jean-Marie (2010). Fonctionnalités textométriques: Proposition de typologie selon un point de vue utilisateur. In: Sergio Bolasco, Isabella Chiari, Luca Giuliano (eds.), *JADT 2010, Statistical Analysis of Textual Data -Proceedings of 10th International Conference.* Rome: Edizioni Universitarie di Lettere Economia Diritto. [see *Lexicometrica* web-site*:* http://lexicometrica.univ-paris3.fr/jadt/jadt2010/allegati/JADT-2010-0341-0354_023-Pincemin.pdf].

Philippart de Foy, Caroline (2008). *Hagiographie et statistique linguistique: étude d'un corpus de traductions médiolatines d'origine grecque,* thèse non publiée de l'Université Nice Sophia Antipolis.

Purnelle, Gérald (1989). Recherche automatique de groupes verbaux récurrents et de formules dans les fichiers latins lemmatisés. *Revue, Informatique et Statistique dans les Sciences humaines, 25, 157-191.* [http://promethee.philo.ulg.ac.be/RISSHpdf/Annee1989/Articles/GPurnelle.pdf]

Purnelle, Gérald (1996). Utilisation d'une banque de données des textes latins lemmatisés et analysés. Problèmes spécifiques aux données linguistiques. In: G. Moracchini (ed.) *Actes du Colloque international «Bases de données* *linguistiques : conceptions, réalisations, exploitations», 295-307* (Corte 11-13 octobre 1995), Univ. de Corse / Univ. Nice Sophia Antipolis.

Quiniou, Solen; Cellier, Peggy; Charnois, Thierry; Legallois, Dominique (2012). Fouille de données pour la stylistique : cas des motifs séquentiels émergents. In: Anne Dister, Dominique Longrée, Gérald Purnelle, *JADT 2012, Actes des 11e Journées internationales d'analyse statistique des données textuelles: 821-833*. Liège. [see *Lexicometrica* web-site*:* http://lexicometrica.univ-paris3.fr/jadt/jadt2012/tocJADT2012.htm].

Renard, Denis (2000). Les parties du discours chez les personnages du *Satiricon*. In: Martin Rajman, Marie Decrauzat, Jean-Cédric Chappelier (eds.), *JADT 2000, 5èmes Journées internationales d'Analyse statistique des Données Textuelles.* Lausanne. [see *Lexicometrica* web-site*:* http://lexicometrica.univ-paris3.fr/jadt/jadt2000/pdf/55/55.pdf].

On Very Large Corpora of French

Etienne Brunet
BCL Laboratory (UMR 7320 Nice Sophia Antipolis University/CNRS

The first to imagine an automatic (or mechanographical) processing of a large textual corpus is an Italian Jesuit, Roberto Busa (who passed away in 2011; he was nearly one hundred years old). Father Busa liked to tell of a visit he made in 1949 to the headquarters of IBM. In the anteroom leading to the office of Thomas J. Watson, the founder, had provided a sign touting the power and speed of the company: “For emergencies, it's already done. For miracles, it is ongoing.” Father Busa brandished the sign under the director’s nose and as if he believed in miracles, he got it in the form of a thirty-year sponsorship which lead to the *Thomisticus Index* in 56 volumes, large format, bound in leather.

In France, a few years later, thanks to the support of René Moreau, the director of scientific development of IBM-France, Bernard Quemada and the researchers from Besançon initiated the *Centre for the study of French vocabulary* by extending an earlier undertaking initiated by Wagner and Guiraud as early as 1953 and dedicated to the establishment of the *Vocabulary Index of Classical Theater*.

Using the same method, the Rector Paul Imbs started the lexicographical project that would become the TLF (Trésor de la Langue Française) and where no example was to be found that was not dated and signed. The technique provided only examples, references and statements. At a time when the text input could only be manual, the creation of a large corpus required substantial resources, a long time and much effort, especially as the text input could not be conceived without grammatical correcting and corpus enrichment, related and expensive operations that were self-evident without recourse to the word *lemmatization*.

The word *corpus* itself was then a rare and almost new Latinism to designate the crude product of such undertakings. Looking back fifty years later, one can see in Figure 1 the various fates of some terms associated with the study of language and the remarkable extension of the word "corpus", which participates with a slight delay, in the explosion of linguistics in the 1960s, but without being affected by the decline observed from 1980. Note that this figure comes from the *Google Books corpus* which accounts for 100 billion words pertaining to the French domain and which we will discuss later in this study[1]. The combination of the two words is itself evolving (in the right part of Figure 1): the corpus tends to break free from the linguistic tutelage in favour of an association with the text, while conversely linguistics tends to link its fate to the corpus in the expression *Corpus Linguistics* that acquires a sudden favour in 2000.

[1] As absolute frequencies are very unequal, comparison was facilitated, without distorting, by increasing the lowest by a factor 2 or 3 (or 200 for the rare word *lemmatisation*).

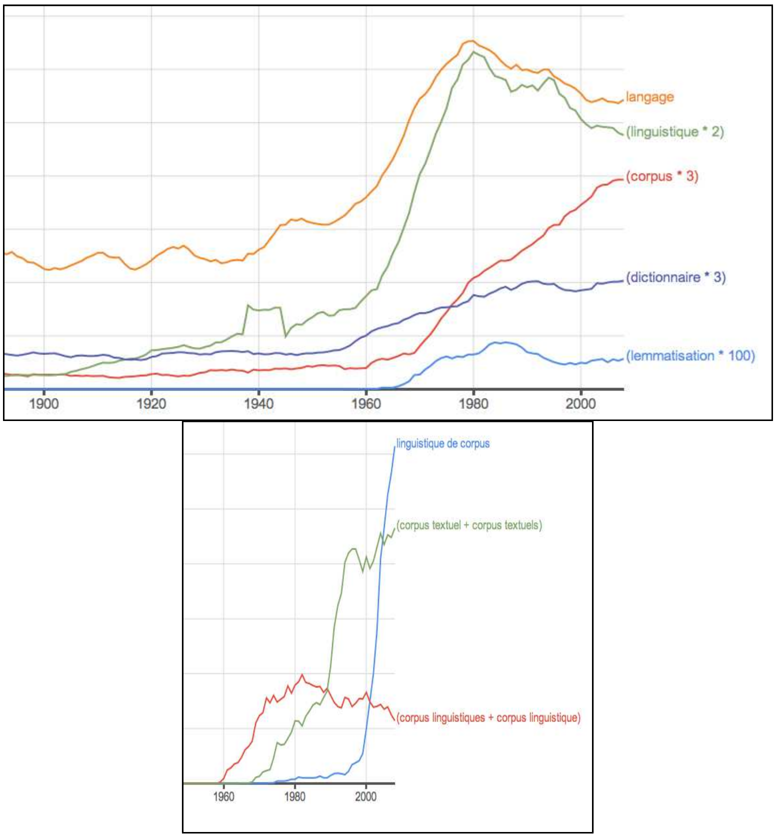


Figure 1. The evolution of some words associated with *corpus* since 1900

If the proper linguistic exploitation of corpora took time before asserting itself, it was probably due to their low availability. To judge the grammaticality of an utterance, the innate or acquired intuition of language seemed a sufficient guarantee and, at a time when the Internet did not exist, it was faster to make a handout than to question existing databases.

The TLF data have led, even in the early stages, to some outside services, mainly concordances and indexes, delivered on printed paper[2]. But sales of digital texts were exceptional, especially as the copyright slowed their spread[3]. The gestation of the TLF lasted a long time and during the 1970s, the data processing hardly crossed Nancy borders. And without being shelved, the project

[2] To prevent ingratitude, we should acknowledge that we benefited without restraint from the Nancy sources which were widely available to us as index or frequency dictionaries. If the text was not transferred directly to magnetic tapes, at least it could be reconstructed from the index and allow monographs, like Giraudoux, Proust, Zola and Hugo established in the 1980s.

[3] This brake still exists even for copyright free texts. CNTRL resource center provides only a sample of 500 texts out of the 4000 *Frantext* ones, half of which escapes the copyright (address of CNRTL: http://www.cnrtl.fr/corpus/). It is barely more than the 300 texts which were selected for the cdrom DISCOTEXT twenty years ago.

remained the prerogative of the TLF editors. Outside, some impatience accompanied this great project advancing slowly, absorbing a considerable part of research funding. In 1978, the corpus was there, almost untouched, and we could write in *Le français moderne*: “The largest world linguistic data base is French. Available. Untapped. And almost unexplored. This vast forest that covers two centuries, 350 authors, 1000 titles, 70 million words, awaits its Livingstone or Stanley[4]. "

I. National projects

1. The National Library

Concerning French, it would be natural to turn to the French National Library, which is rich in 14 million documents including 11 million books on the Tolbiac site. This would be comparable to *Google Books* offer, if access was similarly electronic. Unfortunately the number of documents accessible on the Internet, mainly in the *Gallica* base, is far from reaching that figure. We certainly have access to the catalog and a sophisticated choice of metadata parameters allows a selection as accurate as you want. But the text itself is often unavailable to the internet user. And when the text is transferred, it is usually readable only in image mode. The transcription in text mode, which is sometimes proposed, is often the raw output of the optical scanner, with mention of the probable error rate. When the rate falls below 99%, that means that a character out of a hundred is questionable or wrong, that is a word out of twenty (the average length of a word being five letters). Naturally the success rate decreases as one moves away in the past, old documents suffering from ravages of time and often offering unusual fonts. These defects are common to every corpus or base founded on automatic reading of documents, but they are more significant in *Gallica* because past centuries are less under-represented. Yet we see in the results only the correct reading of the proposed word, because it does not come to mind to look for erroneous readings. Although *Gallica* is a long-standing base and widely predates *Google Books*, its extension does not have the same scope, thus limiting its statistical interest. The number of usable documents in text mode is limited to 200,000 while its US rival offers millions. And the proper statistical information is minimized, with only the mere mention of the frequency of the desired word. It is difficult with so few elements to establish a curve, let alone a table[5].

[4] *Le français moderne*, 46/ n°1, Editions d’Artrey, Paris, 1978, pp 54-66.

[5] *Google Books* does not offer more quantitative information and, similarly, is happy to indicate the number of documents concerned by the query. But that number is of another order of magnitude and proper statistical analysis is performed by a derivative site *Culturomics*, which has no counterpart in the BNF.

2. FRANTEXT

In reality, the most reliable texts of *Gallica*, aside from newer ones transmitted by publishers in digital form, are those coming from the Frantext legacy. Those owe nothing to scanning, whose invention in 1974 by Ray Kurzweil is after the initial capturing, carried out by keyboardists on perforated tape. This manual input, duly revised and corrected for fifty years, resisted all changes of systems or supports, passing unhindered from the perforated tape to magnetic tape and disk, and finally to all types of memories available today.

To that reliability of texts, even when they are older editions, Frantext adds many other virtues: a balance between eras, allowing comparisons and providing a solid basis for analysing the evolution of the language; covering a wide chronological span of five centuries of publication; a desired homogeneity of texts whose choice is governed by specific criteria, concerning genre and language level; consistency in the services offered to the scientific community, the same software being kept unchanged for twenty years on the Internet[6]; a moderate increase and a controlled enrichment of data ensuring compatibility with the previous treatment. In brief, in the original draft of the *Treasury of the French Language* as in the derivatives *TLFI* (the digitalized version) and *Frantext*, there is a clear understanding of objectives and a precise definition of the means that have made the French project a model. Now one feature of this model interests us: it is the part played by statistics. From the beginning, TLF reserved for each article a final section where the word's frequency in the whole corpus is noted, but also in the subsets formed by time and genre. Throughout the making of the dictionary, the editors had at their disposal, besides concordances, encrypted information concerning frequencies and co-occurrences[7] and attached to written forms, lemma, parts of speech, expressions and structures. Most documentary and statistical functions that made the success of Frantext were already operational in local mode on the Nancy site or even, in distributed mode, on national networks (Transpac or Minitel) that preceded the Internet. They were also used in the CD *Discotext* produced and distributed in 1984. But it was in 1998, with *Frantext on Internet*, that proper statistical research was greatly facilitated. The use remained primarily documentary and numerical results were quite modest. For, in order not to frighten the literary populations, *Frantext* often merely provides for percentages or relative frequencies. But it is easy to deduce the actual frequencies, calculate variances, and build curves, distribution tables and multivariate analyzes, by opposing texts to each other, or authors or genres or epochs. The statistical treatment not being fully supported by *Frantext*, the user needs additional and specialized programs.

[6] This software, named *Stella*, was achieved by an exceptional engineer, Jacques Dendien. It regulates Frantext but also TLFI.

[7] The name "binary groups" had been given to these co-occurring records, sorted by grammatical categories.

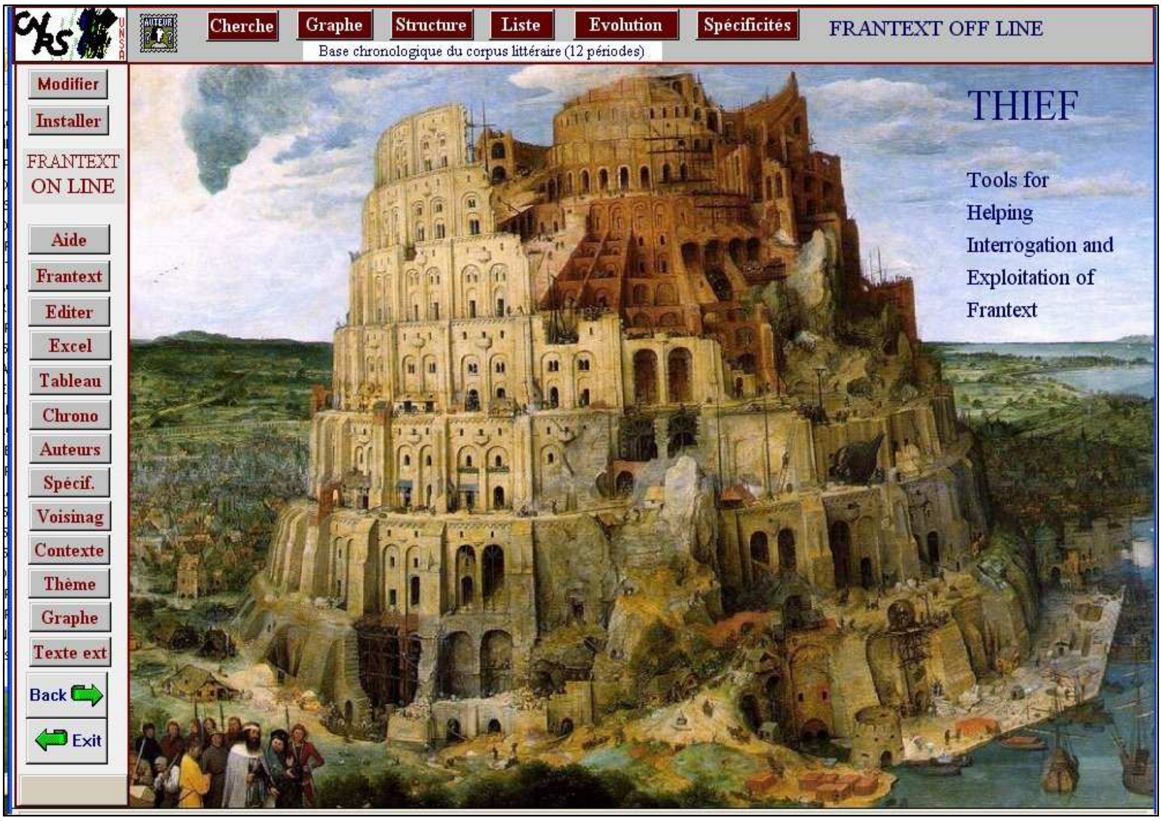


Figure 5. Statistical analysis of Frantext. *THIEF* database

Our base THIEF (Helping Tools for Interrogation and Exploitation of Frantext) addresses this need by offering the usual array of statistical tools and applying them to Frantext data, whether preloaded or downloaded on demand. In the first case, the data are frozen in the state they were in in 1998, a corpus of 117 million words divided into 12 time slices, from 1600 to 1990. This allows us to see the evolution of the literary language[8] for four centuries and discover lexical and linguistic properties for each period. One works then without connection and without text, on recorded frequencies reachable by the buttons on the top margin of the main menu (Figure 5). Actually the functions spread over the left margin deliver direct access to *Frantext* in its current state. The user is then connected to *Frantext* and can define his working corpus, according to various criteria (title, author, genre, time) and can extract any frequency or textual data he wants. Once saved in a file, the results are taken over by the software THIEF to deliver histograms (by period or author), tables, factor analysis, co-occurrence graphs, etc. However whatever its reputation and merits, Frantext has limited prospects. Fifty years of history overshadow its future. This is due in part to the timidity of its statistical apparatus: simply distributing the text as word lists or number series, at a time when images have invaded the Internet, means depriving oneself of the immediate readability specific to the graphical representation. The most serious handicap is the data: we praised its reliability and homogeneity, but it merely represents a single use of French: high language level, literary and classic. It is the French that is learnt in school textbooks or books you read in libraries. It is

[8] Technical texts were excluded, to give more coherence to the corpus.

not the French one speaks or is used in everyday life, in newspapers and the media. It bears testimony to the culture, not the reflection of current events. The catalogue is now expanding by adding more recent production: it has currently 4000 references and 270 million words. But the BNF weighs ten times more; *Google Books* is a thousand times more and its pace of growth is much faster. Finally, *Frantext* remains constrained by an agreement made with publishers, which limits the size of extracts and forces the user to a prior subscription. This subscription can be justified if it is to communicate a text or a copyrighted extract. But we do not see the legal legitimacy if it concerns quantitative information from the text, whether or not in the public domain. Furthermore, *Frantext* has not fully retained the intermediate solution which would be to export the text, at least the copyright free text, offering it for download so that the users may apply any statistical and computing processing of their choice. This distribution function has been outsourced to a subsidiary organization, the CNRTL, whose current catalogue is too small.

II. Bases and corpora. Encyclopedia

The notion of corpus became extensive and now tends to designate any set of texts likely to be submitted to statistical and computational processing. In principle, one should distinguish structured data, such as a library catalogue, from those that are not, and where the text will scroll continuously. It would be appropriate to call the first “bases” reserving the term “corpus” for the second[9]. Thus *Frantext* is clearly a corpus, while the *TLFI* (or computerized TLF) is a base. The criterion that differentiates them is the presence or absence of a fixed frame having ordered sections that items must fill in one way or another, by a number, a code or text. But the opposition is not absolute: firstly a corpus is usually partitioned into multiple texts and comparing each of which can receive qualifications or metadata: title, author, date of publication, genre, register already constitute an external structure that can continue internally with chapters, acts or scenes, collections or more generally components, disjoint or nested, of textual content[10]. In addition, opinion and market surveys, next to boxes where there appears various coded information (occupation, age, sex, education, income, etc.), often give way in form to a free section where respondents express their opinion without any binding directive. To treat this part of the survey, specialized software then uses the same tools used in the processing of corpora. It even happens that structured database information may be processed directly by ignoring or blanking out every structure and tag[11]. Now that most of the historical

[9] This distinction is what prompted the name "Bases, corpus and language" to the laboratory where this research was conducted.

[10] The normalization of textual data is greatly facilitated by the standard TEI guidelines (Text Encoding Initiative) and XML tags

[11] This unscrupulous scanning is often practiced by automata that scour the Internet. More or less coarse filters chop up the site pages to draw the best pieces, usually cutting off the head and tail.

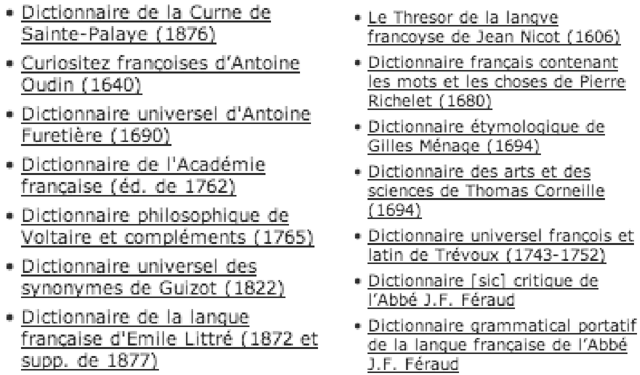
dictionaries are available on the net or on CD or DVD[12], one could flatten the text and treat it as a corpus. But the statistical significance of such an operation seems low, since there is nothing to define partitions that can be compared. One cannot oppose the words that begin with A to those beginning with B. At most, one might isolate some elements of the structure, such as entries, definitions, examples or quotations, synonyms, areas of application[13].

The enterprise is justified more easily when it concerns an encyclopedia, first because the inclusion of proper names gives the corpus a space-time dimension which a language dictionary is lacking and also because an ontology of knowledge and disciplines takes shape more accurately. We will give two examples borrowed from the editorial news.

1. Encarta

Indeed *Encarta encyclopedia* is no longer relevant since this cultural product, launched in 1993 by Microsoft, ended its existence in 2009, Microsoft having withdrawn it from the market in the face of *Wikipedia*'s dominance[14] in the global network and the *Encyclopaedia Universalis* on the French market. Benefiting from a personal contract with Microsoft in 2000, we had access to the full text of *Encarta*, which Microsoft wanted to submit to our software *Hyperbase*. As was expected, the lack of partitions reduced the interest in this undertaking. However, a function remains that can be applied to any corpus delivered in one piece and based on co-occurrences. This function can be limited to one word, by observing its lexical environment, that is identifying its closest terms, but it can also extend to the whole corpus. A separation technique and progressive refining allows one to decant and isolate lexical themes or constellations that structure the corpus[15]. Applied to the text of *Encarta*, the

[12] For example, here is the rich catalogue of Redon editions, to which are added Diderot's *Encyclopedia*, various editions of the *Dictionary of the French Academy* and the *Larousse Universal Dictionary of the Nineteenth Century*.



- Dictionnaire de la Curne de Sainte-Palaye (1876)
- Curiositez françoises d'Antoine Oudin (1640)
- Dictionnaire universel d'Antoine Furetière (1690)
- Dictionnaire de l'Académie française (éd. de 1762)
- Dictionnaire philosophique de Voltaire et compléments (1765)
- Dictionnaire universel des synonymes de Guizot (1822)
- Dictionnaire de la langue française d'Emile Littré (1872 et supp. de 1877)
- Le Thresor de la langve francoyse de Jean Nicot (1606)
- Dictionnaire français contenant les mots et les choses de Pierre Richelet (1680)
- Dictionnaire étymologique de Gilles Ménage (1694)
- Dictionnaire des arts et des sciences de Thomas Corneille (1694)
- Dictionnaire universel françois et latin de Trévoux (1743-1752)
- Dictionnaire [sic] critique de l'Abbé J.F. Féraud
- Dictionnaire grammatical portatif de la langue française de l'Abbé J.F. Féraud

[13] Specialized research in proxemy was published by Bruno Gaume from the definitions of French verbs ("For a cognitive ergonomics of electronic dictionaries" in *Document numérique*, 2004/3 (Vol.8), pp.157-181).

[14] At that time *Encarta* represented only 1% of Internet queries against 97% for Wikipedia. But it is true that *Encarta* users used the CD more readily. It is in this form that this encyclopedia still continues its career, even though marketing has stopped.

[15] The algorithm used is the *Alceste* software is one.

decomposition process delivers a spectrum of ten colours of which Figure 6 details the nuances[16].

1 (science) *vitesse onde rayon énergie surface métal électricité gaz particule atome*, etc.
2 (littérature) *roman œuvre auteur publier écrire poésie poème récit écrivain*, etc.
3 (géographie) *région nord sud département ouest habitant plateau vallée population massif côte*, etc.
4 (arts) *film cinéma carrière cinéaste théâtre scène acteur réalisateur peintre*, etc.
5(guerre) *guerre armée troupe militaire allemand Allemagne britannique allié force accord conflit offensive camp soviétique*, etc.
6 (politique) *président république élection gouvernement ministre politique député parti socialiste républicain*, etc.
7 (pensée) *philosophie dieu philosophe pensée connaissance Christ science esprit idée vérité évangile sociologie pratique âme histoire foi*, etc.
8 (société) *droit loi économique public juge entreprise salarié justice tribunal état privé social*, etc.
9 (histoire) *roi empereur empire fils Charles royaume pape Louis duc trône prince Henri Angleterre dynastie*, etc.
10 (patrimoine) *siècle musée église cathédrale château gothique ancien chapelle édifice Notre-Dame monument*, etc.

Figure 6. The disciplinary spectrum of *Encarta*[17]

As can be seen, the editorial board of Encarta hardly deals with geography, circumscribed in section 3, whether physical or human. History is best treated, whether actors of the past, especially kings or monuments that bear witness of the time (themes 9 and 10). But the distinction is made between the old and the contemporary: the theme of war makes clear reference to the world wars (theme 5). And a distinction is made in the arts between the literary tradition (theme 2) and modern performing arts, especially cinema (theme 4). It would be interesting to make a comparison with a similar editorial company and we think of the *Grand Larousse Encyclopédique du XIXe* and the *Encyclopaedia Universalis*. But in both cases, there is no honest way to obtain the full text of these bases, which can be searched word by word but not as a whole. Like most bases available online or in DVD, they answer all questions except those relating to themselves.

2. Wikipedia

However there is one base which reveals its secrets ingenuously: **Wikipedia**. There is no need to describe this cooperative encyclopedia; everyone uses it daily. What is less known is the capacity to download its content through the site http://dumps.wikimedia.org/, or more easily through the site REDAC offering exploitable resources from Wikipedia (http://redac.univ-tlse2.fr/corpus/

[16] The software (IRAMUTEQ) preventing us from treating all the articles at one time, we merely treated a sample of 2 million word-occurrences, nearly a random tenth of the whole set.

[17] The name of the list designated in brackets results from the interpretation of the elements of the list, which are much more numerous than those we supply, for lack of space.

wikipedia.html). A first approach, which is indirect, is to identify products and to note to which discipline they relate. While writing Wikipedia articles is free, it is customary to indicate, at the end of the article, to which categories and portals they can be linked, for any aspect of the content. As most articles refer to several descriptors or keywords simultaneously, one can, noting these associations or co-occurrences, map the disciplines represented in Wikipedia.

The software programme *Iramuteq*, that was used for studying *Encarta*, provides, in Figure 7, an unexpected distribution where one can hardly recognize the traditional ontology of knowledge and activities. Two areas are particularly highlighted: on the left (magenta) the performing arts, around the *cinema*, television and music, and at the bottom (dark blue) the circle of *players* who compete for the ball. The divisions of the geographical area complete the triangle and occupy the top of the figure (in red) around the *town*. The rest is confined to the central area, less violently contrasted: three districts nevertheless emerge: the *political* and social sphere (on the right, light blue), biography and history that hold registers of *deaths* and *births* (green, below the origin) and finally (in black, above the origin) a concentration of human activity gathering thought, research, science and industry. Science and technology are not separated. Except for this detail, we find the same main lines Wikipedia includes in its subtitle: "Art - Geography - History - Science - Society - Sports - Technology". Like Encarta, Wikipedia highlights the cinema and all modern arts that are based on the diffusion of image and sound. It adds the sporting field whose spectacular promotion is linked to this diffusion.

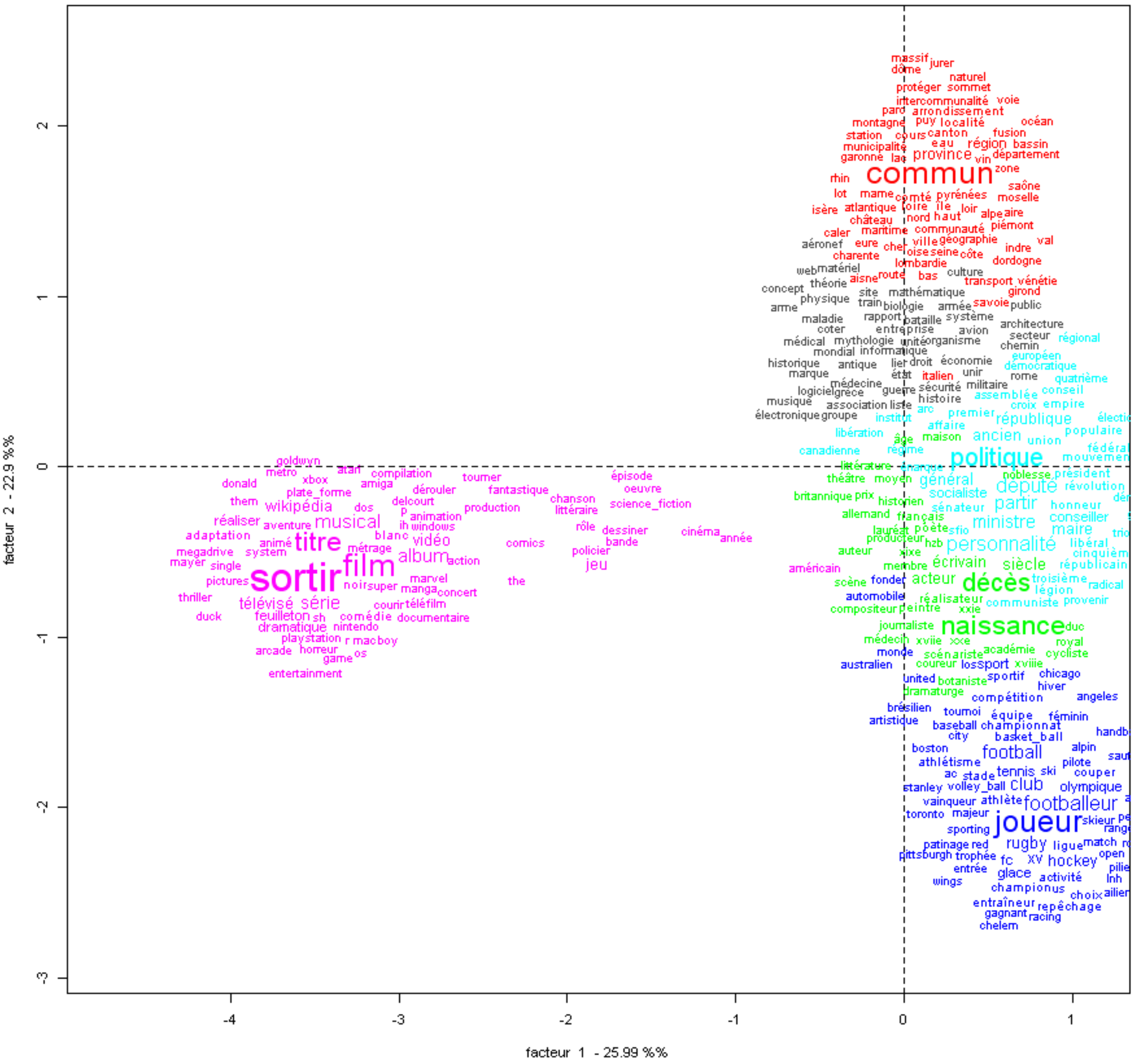


Figure 7. Analysis of the key-words of Wikipedia

Wikipedia is a collective enterprise, linked to individual initiative and free of binding elements of centralization. The underlying ontology that emerges from more than 600,000 articles is based on an improvised provisional architecture[18], 1555 gates, themselves grouped in 11 categories: Arts, Geography, History, Hobby, Medicine, Politics, Religion, Sciences, Society, Sport, and Technology. This ontology is deduced from the keywords ("category" or "portal" in Wikipedia terminology), examined in Figure 7, that accompany each article.

One can wonder wether a classification of the actual texts of the article will produce the same categories. To limit the volume of data to be processed,

[18] It is not forbidden to add others, provided one first checks that the proposal has no articles preceding them. Control is a posteriori. Instead of offering a predefined framework to be completed cell by cell, decision-makers merely register the proposed portals without banning judgement: within1555 portals, the site admits frankly that only 29 are "good quality" and 50 "good gates".

one proceeds by sampling, retaining only one article out of ten, and isolating the class of nouns. In a corpus reduced to four million words, the *Iramuteq* algorithm sees a ten-class structure, which only imperfectly reflects the official nomenclature. Of the eleven groups displayed in the organization chart, more than half are certainly reflected in the results, namely politics, science, technology, history, society and sports. But neither religion nor medicine nor hobbies appears independently. As for art, only sound and the image are taken into account, not the written text that is entitled to an independent constituency, covering literature, the press and scientific publishing. Similarly geography comes in two classes, depending on whether town or country is concerned.

Where does the distortion, the difference between summary and content, come from? Any encyclopedia aims to be a dictionary of knowledge, as well as of places and people. Now as the Wikipedia writing mode is based on unsolicited and unpaid collaborations, voluntary contributions are not immune to interest mingled with people. Let us observe the significant list of Figure 8: if the names of places dominate in the geographical classes (4 and 5) and abstract concepts in the technical or administrative classes (6, 7 and 10), elsewhere the names of persons take center stage: the writer, the teacher, the philosopher in Class 1; the president, the minister, the deputy, the candidate in Class 2; the actor, director, screenwriter in Class 3; and the player, champion, coach, winner in Class 9. As for Class 8 dedicated to history, it is entirely made up of kinship terms, titles of nobility and ecclesiastical dignitaries. Biographical elements occupy such a large place that Wikipedia is becoming a kind of Who's Who where everyone would like to see his or her picture and his or her medals.

Classe 1 : (écrit) littérature ouvrage édition écrivain livre revue professeur lauréat publication poésie roman université école philosophe presse science journal essai critique …

Classe 2 : (politique) parti élection président politique ministre député parlement gouvernement assemblée candidat suffrage constitution république…

Classe 3 : (image et son) film cinéma acteur réalisateur scénariste télévision internet scénario série comédie métrage réalisation feuilleton…

Classe 4 : (ville) pont bâtiment construction quartier architecture architecte pierre édifice métro ville rue hayteur station boulevard façade mètre route béton …

Classe 5 : (campagne) parc montagne réserve zone superficie altitude sud île ouest faune rivière lac forêt nord région massif vallée eau flore…

Classe 6 : (sciences)exemple forme biologie cas molécule propriété type température acide protéine quantité surface chimie particule phénomène effet équation cellule…

Classe 7 : (technologie) logiciel moteur informatique entreprise système fichier utilisateur processeut ordinateur gamme bit type véhicule technologie vitesse gestion…

Classe 8: (histoire) fils mort évêque fille empereur père prince duc royaume pape comte prêtre bataille dieu trône époux archevêque…

Classe 9 (sport) joueur palmarès équipe club championnat football match sport champion entraîneur cyclisme carrière vainqueur classement sélection hockey rugby…

Classe 10 (société) maire population commune identité évolution période mandat monument géographie compte district personnalité municipalité administration statistique village habitant…

Figure 8. Analysis of the text of Wikipedia

III. Monograph-based corpora

Using monographs opens up a fruitful avenue of research. First, the unifying principle of the corpus must be chosen (a language, a theme, an event, a story, an investigation, an author, a review or a newspaper, a time or genre); then texts may be added to the corpus, usually following a chronological axis.

If the texts are big enough, they may be used as sub-corpora for statistical comparisons. If on the contrary, the textual units are numerous and small and one cannot divide the corpus into sub-corpora, we resort to the previous case (the encyclopedias): the entire unstructured corpus must be treated as a single piece, and only co-occurrences phenomena in small contexts may be observed, using the Alceste algorithm (or its Iramuteq implementation).

Such a situation is frequent in the treatment of sociological surveys. For instance, Pascal Marchand and Pierre Rastinaud have conducted a thematic survey about "national identity", based on 18,240 contributions available on the official website of the Immigration ministry (which had opened a forum in 2009). The forums, the social networks or the personal data collection from traffic analysis provide an inexhaustible reservoir for such investigations, some of which may reach gigantic size as soon as industrial, political or commercial interests are involved.

In France, the "classic" methodology, first established by Guiraud and Muller, and then applied in the 'Lexiometry' laboratory, use the corpus as a norm (a reference frequency list) to which its various subcorpora may be compared. There is no external index of the frequencies.

In the political or historical field, data is often public and free, and easier to collect the data. For instance, Damon Mayaffre has analyzed the discourses of several French former presidents using a corpus of three million words. These analyses show that the various former presidents may be distinguished from each other not only according to the subject of their discourses, but also according to their style. Even if he is facing very different situations, the speech of a president remains recognizable. Figure 9 shows that the former president Mitterrand, during his very long period of power, is always characterized by the use of verbs (such as Sarkozy) whereas other former presidents prefer the nominal categories, apart from Chirac who remained undecided throughout two exercises of power.

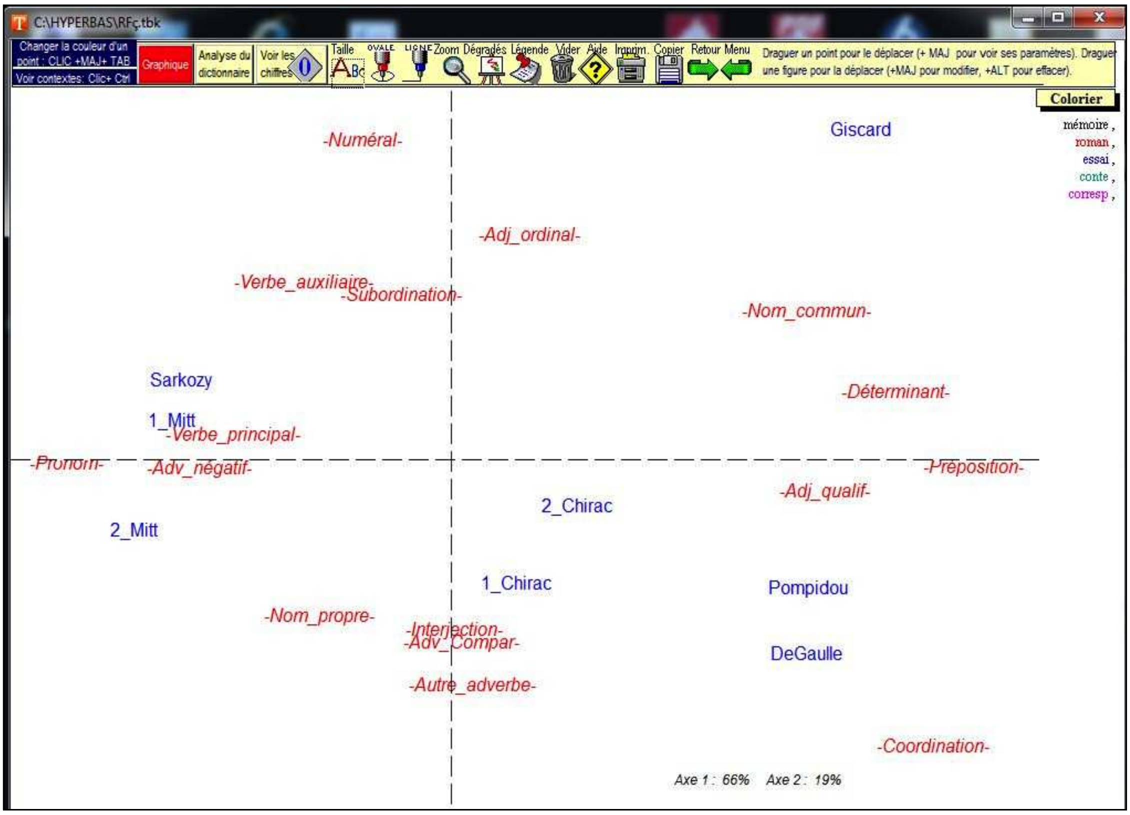


Figure 9. Factorial analysis of parts of speech in former French presidents' discourses.

In the field of literary studies, the typical case is that of studies based on corpora made of all the monographs by a given author. Early works were involved from the very beginning in the study of great texts, like the Bible, St. Thomas Aquinas or Shakespeare. The sizes of the corpora are not expected to defeat those of these pioneering works, due to the lack of prolific writers known to exceed Saint Thomas and Shakespeare... the Nancy ‘treasury’ (Frantext corpus) included works by many writers but the strategy adopted was to build a corpus balanced according to the works, and this choice prevented the inclusion of full texts in the corpus. The most important "complete works" corpora that could be built (La recherche du temps perdu by Proust, the Rougon-Macquart by Zola, Les Miserables by Hugo) have rarely more than one million words. Professor Kiriu, from Japan, devoted years to scan and correct the complete works of Balzac's Comedie humaine. Other passionate contributions have resulted in the scan of the complete works of Voltaire (Y. and R.D. Boudin), Maupassant (Thierry Selva), Jules Verne (Ali Hefied). Today, using the good sources (e.g. Wikisource or the Gutenberg project) we can almost reconstruct the complete work of a writer, even a very productive writer such as Sand or Dumas, provided that there is no copyright. The size of the corpus can then approach 10 million words. Nothing prevents us from going beyond, if one aggregates the works of writers, and compares them inside a textual genre or an historical era. And from there, one can go further and compare various genres or various eras.

However, today the size of the literary corpora remain below that of Frantext (5000 texts and 300 million words). Outside the literary field, several impressive corpora are emerging thanks to the use of newspapers, magazines and electronic documents of all kinds that are produced every day by the administration, industry, research and the media. One year of a regional newspaper such as the “Est Républicain” is 100 million words that are now offered for download and analysis. Most newspapers now have followed the example of the newspaper Le Monde and are open, in digital form, to retrospective research in their archive. We are able to compare in the same corpus the writings of different newspapers during a given period of time.

However, there is a hindrance to the exponential growth of corpora: the inadequacy of conventional software for operating on such large masses. For instance, I have had to deal with all the issues of the magazine *Europe* published between 1923 and 2000. My software, Hyperbase, was not able to deal with a corpus of 28,000 articles and 58 million words. It is difficult to make it fit in the memory of a personal computer a textual corporus whose size is close to one gigabyte. At this level servers and specialized hardware are needed, that are able to handle the long work of entering, correcting, enriching and indexing data; and to distribute this data using index, pointers and references. But such institutional corpora are like huge tanks that distribute their content, word by word, as would a dictionary. The consultation can be only punctual. They do not allow any overview, no overall analysis, as can be seen from three gigantic corpora of the French language built respectively in Germany, in UK and in the USA.

IV. Corpora of the French language made outside of France

1. Wortschatz

The first of these three corpora was build at the University of Leipzig (with collaborators from the University of Neuchâtel). It is a corpus of the French language with 700 million words, 36 million sentences from newspapers (19 million), web (11 million) and Wikipedia (6 million). One can base a query on the entire database or on any of its three components. The querying of the corpus may be done through keywords: absolute and relative frequencies are given for the requested keyword, together with some examples (with their addresses in the corpus) and especially its environment, specified in several ways:
- A list of words that co-occur preferentially with the keyword in the sentence;
- Preferential immediate co-occurrents to the left and right of the keyword
- A graph summarizing the most significant co-occurrences

One example suffices to illustrate the results that can be expected from this base (Figure 10)

Mot-clef: Sarkozy

Nombre d'occurrences: 106536

classe de fréquence: 8 (i.e., has got about 2^8 the number of occurences than the selected word.)

exemple(s): "Lundi depuis Pékin, M. **Sarkozy** avait lancé un appel à "l'apaisement" resté sans effet. (source: *http://fr.biz.yahoo.com/27112007/202/sarkozy-absent-fillon-en-premiere-ligne-face-aux-violences-de.html*) "Il est le cousin de l'ex première Dame de France, Cécilia **Sarkozy**. (source: *http://fr.sports.yahoo.com/29122006/29/aime-jacquet-tire-sa-reverence-avec-un-palmares-unique.html*)
Sarkozy se prévaut d'une baisse de 9,44% des crimes et délits et d'un taux d'élucidation en progression de 8 à plus de 34%. (source: *http://www.rtbf.be/info/international/ARTICLE_080090*)
exemples supplémentaires

cooccurrences significatives de Sarkozy:

Nicolas (838584), président (51762.4), " (48587.6), UMP (36403), M (31913.1), a (30256.3), Royal (25565.4), Ségolène (25195.9), Elysée (22596.7), ministre (22117.5), Intérieur (20576.1), présidentielle (18656), Cécilia (15904.3), François (14851.6), Bayrou (14453.7), Villepin (14450.1), candidat (13928.3), l (13536.4), ' (12823.7), , (12757.2), français (11182.9), Chirac (10215.3), élection (10159.6), visite (9974.35), Fillon (9635.87), Dominique (8574.88), campagne (8432.2), PARIS (7793.58), Carla (7591.07), France (7580.98), avait (7430.25), politique (7166.54), discours (6749.46), son (6746.22), République (6676.53), Bruni (6566.09), etc

voisins de gauche significatifs de Sarkozy:
Nicolas (1326360), Cécilia (17948.7), président (9428.68), . (5228.97), Jean (3338.48), candidat (2430.55), Monsieur (1104.34), Président (774), loi (664.77), couple (595.96), Carla (564.47), Mr (527.76), Guillaume (423.36), Cecilia (386.61), circulaire (365.12), sauf (334.23), Nicoals (301.77), Bruni (295.11), voter (264.16), présidence (260.31), Nicolas (259.26), Royal-Nicolas (253.44), monsieur (246.57), battre (234.84), etc
voisins de droite significatifs de Sarkozy:
a (53559.9), , (28722.9), . (14068.3), avait (11441), s (6645.14), et (5227.08), n (3751.04), veut (3193.77), " (2796), est (2542.41), lors (1744.72), ? (1641.62), ne (1631.29), doit (1446.9), devrait (1353.74), souhaite (1325.86), était (1163.67), etc

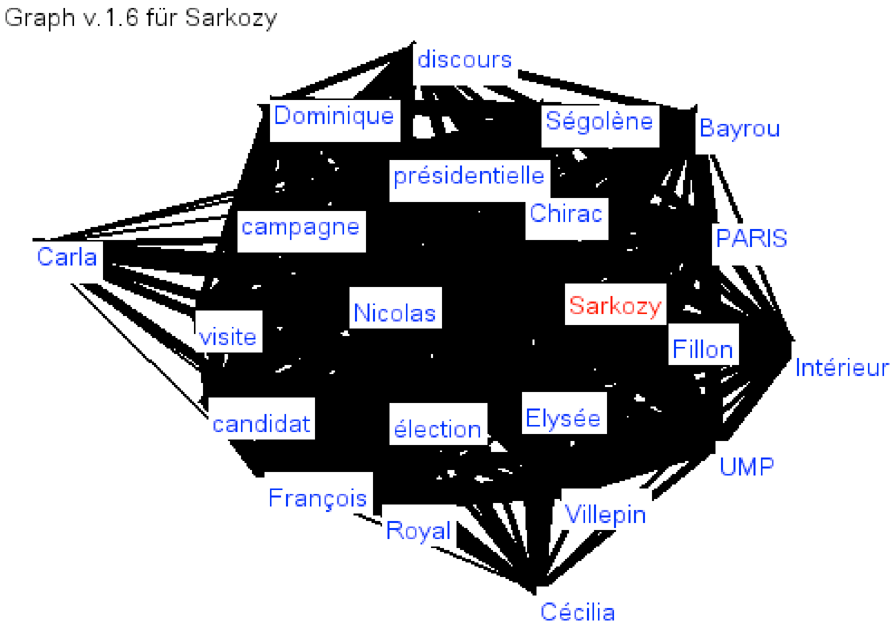


Figure 10. *Sarkozy* according to Wortschatz corpus

2. Sketchengine

Sketchengine is an English website which offers (together with corpora of other languages) a corpus of the French language. This corpus is over ten times larger than the Wortschatz corpus. The range of tools is also much wider. Sketchengine has several points in common with the Frantext corpus: the user needs to subscribe – for reasons of profitability and not, as for Frantext, of copyright –, the freedom to download at least large extracts; and the possibility to manipulate complex objects: lemmas, codes, structures. However, there are also differences. Frantext has its own data. Sketchengine harvests the web. The former focuses on books and full texts, the latter on short contexts. The former is diachronic, the latter is synchronic.

Like many web-based corpora, Sketchengine is harvesting the web in order to build a large representative corpus of a language rather than to build corpora targeted at analyzing lexical innovations. The starting point is a list of a few hundred words of medium frequency, which is the seed of the harvesting. To reap the harvest, thousands of requests on Google, Bing or Yahoo are made in search for pages that contain at least three words from the list. The pages are collected in a cumulative corpus with the associated metadata (at least the address and the title of the site as well as the date of the request). Next, the duplicate pages are eliminated (thanks to the "onion" software) as well as extratextual content (thanks to the "justext" software). Various filters are then applied: the document must meet several conditions: be of sufficient length (at least 500 words), contain a minimum proportion of grammatical words. This automatic control based on simple criteria is helpful for removing many unsuited pages: the relationship between what is retained and what is tested is between 1/10 and 1/1000. The corpus is balanced between various sites in order to increase the corpus diversity. Such a process can harvest up to 1 billion words per day. The collected data receives linguistic processing to ensure lemmatisation (TreeTagger is used for Western languages) and a host of statistical operations to enable a sophisticated consultation.

One of the simplest requests of users is often for a concordance. The concordance tool provided by Sketchengine gives the context (line or sentence) for several kinds of queries: word-form, lemma or complex query with various filters. It can also analyse the distribution of the keyword and rank co-occurring words according to the kind of grammatical relation they have with it, and according to the strength of the statistical attraction for the keyword. For instance, an analysis of the word *Samedi* shows that the co-occurring words are most often *dimanche, dernier, prochain, pluvieux, ensoleillé*. It reflects the major role played by the weekend for people. If one considers the profiles of other days of the week, one gets a sociological typology of the days of the week. If one considers the profiles of the months of the year, one gets a sociological profile of the season. Enquiries into sociological representations are available with keywords such as freedom, justice, equality, community, or deadly sins.

For example, the reputation of French politicians on the internet can easily be observed. It is somewhat reflected in Figure 11. This factor map is the result of a factor analysis of the contexts where there is a mention of one of the major politicians of the fifth republic in France. This corpus is extracted from the FrTenTen12 corpus of Sketchengine and has been built for representing 37 French politicians (Presidents, Ministers or party leaders) through 5,000 randomly selected occurrences of each of them. The 279 most common nouns in the corpus are then selected (they include by definition the names of politicians involved, each with at least 5,000 occurrences). The contingency table cross-tabulates these words (at the intersection of row i and column j there is the number of co-occurrences between the words i and j). For a survey of men with various historical statuses, it is not surprising that the first factor reflects the timeline. The timeline, evident in proper names, is also observed among common names. Those found on the left belong to political events of the 2000s (electoral campaigns and especially the 2007 and 2012 presidential campaign). It shows the competition between candidates (*sondages, campagnes, votes, débat, déclaration, programme, émission, media, opinion, parti, soutien, militant, candidature, primaire, présidentielle, tour, résultat, victoire*). The confrontation is less harsh in the opposite half of the factor map, on the right. The politicians, there, have left the political scene. We see their work rather than their ambition and the history rather than the current events.

The second (vertical) factor does not separate the left and right political tendencies. It might have be the case if the discourses of politicians have been included in the corpus. But the corpus is not about what they say, but about what is said about them. And in the words about them, the right and left tendencies can coexist. *Mitterrand* is close to *De Gaulle*. Public opinion tends to classify people according to their rank. The presidents of the republic occupy the upper part of the figure. Prime ministers are relegated to the lower half, where *Balladur* is close to *Rocard*, *Jospin, Villepin*, *Mauroy*, and *Juppé*. While presidents are characterized with the lexemes referring to the general objectives of politics (*peuple, famille, homme, femme, pays, société, valeur, liberté, justice, loi, démocratie, politique, guerre, mort*), prime ministers are concerned with administrative management of current affairs (*ministère, cabinet, comité, conseil, commission, directeur, secrétaire, conseiller, assemblée, groupe, chef, membre, député, maire, poste, finance, université, presse, fonction, réforme, emploi, etc.*).

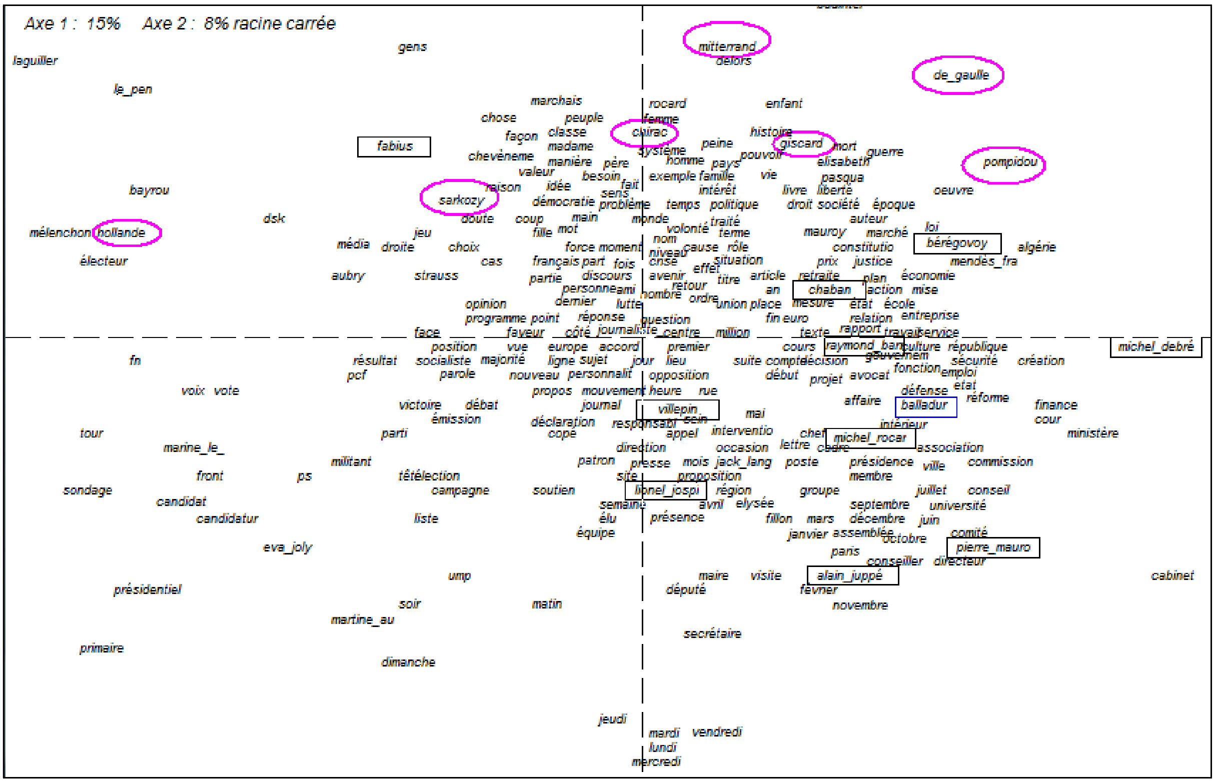


Figure 11. Factorial analysis of co-occurrences in a corpus about French politicians (axes 1 and 2).

3. GOOGLE BOOKS

Such analyses of word distribution are not possible on the Culturomics website we met earlier (Figures 2 and 3). If the contexts are readable in Google Books, they are no more readable in Culturomics, where only indirect pieces of information are available: n-grams, or text sections whose lengths do not exceed five words. However, Google offers significant advantages in quality and quantity. By its size, it is the biggest corpora of the French language, with a size ten times greater than that of Sketchengine (almost 100 billion words in 2012)[19]. While the diachronic dimension is absent in Sketchengine, it extends over centuries in the Culturomics corpus, opening fruitful avenues of research on the history of words and realities of which those words bear witness. The quality of sources is also a strong point in the Google corpus. The internet contains a mix of all kind of discourses and varieties. The methods used by Sketchengine are unable, despite

[19] Between 2009 and 2012, the size of the French corpora has doubled, as did the corpora of the other languages. The current figures at the time of writing are 89 billion words for the French language, 349 the English language (with several dialects), 53 for German, 67 for Spanish, and 33 for Italian, the last corpus built. These figures correspond to the data that can be downloaded. They are higher in the first table of the article published in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, p.170. Three other corpora are available (Russian, Chinese and Hebrew).

all the filtering, to be immune against the barbarisms that are frequent on the social networks. As for Google Books, since it includes only books, such as the BNF and Frantext, it gives access thereby to a certain level of language and culture that Facebook can not guarantee. A simple survey gives the measure: the ratio between the incorrect *fesait* and the correct *faisait* is 2.6% in Sketchengine while it drops to 0.6% in Culturomics. Google Books is still far from the focus on literature found in Frantext, since it accepts all published works, especially in technical news, and social or media domains. But the barrier of printing protects it against insignificant verbal diarrhea that is spreading on blogs and social networks.

Jean Véronis, who has just left us, did not hide his enthusiasm for the birth of *Culturomics* at Christmas 2010. He had also greeted the 2012 version that corrects some defects of the 2009 version and multiplies its power and flexibility. The queries are no longer restricted to word forms or phrases. It is now possible to ask for lemmas (e.g. faire_INF to ask for the details of word forms of the verb *faire*), and for the bare part of speech (_DET_ for determiner) or to use wildcards (such as *), and to select the corpora (symbol ":"). A handicap yet was still preventing the use of *Culturomics*: *Culturomics* was delivering curves only, instead of the underlying numbers, and it was not possible to make further analyses using the raw numbers. The authors of Culturomics have therefore released an API that for a given word gives the 201 frequency counts observed along the timeline from 1800 to 2000. Better still: the raw data used to make tables and curves were delivered for free download, which we used to form a base offering the analysis of unigrams (or individual words) of the French domain.

As an illustration, Figure 12 summarizes the syntactic evolution of the sentence in French. The verb and its acolytes (pronouns, adverbs and conjunctions) lose ground to the benefit of classes related to the name: nouns, adjectives and prepositions. This trend is not unique to French: it is found for the same period in other Western languages. This trend however may be a little suspect. There is the suspicion that this change reflects not so much a change in the use of French, as a change in the composition of the corpus. Recent modern texts are the most numerous and are frequently about technical issues. In those text genres, there is an impersonal style, and information is passed via the nominal categories. On the other hand, the older times are represented by literature, more than by science and technology. The verb is more present in literary discourse, and dialogues are more "personal". The variation of the textual genres between periods may have created a heterogeneous corpus, giving the illusion of an evolution.

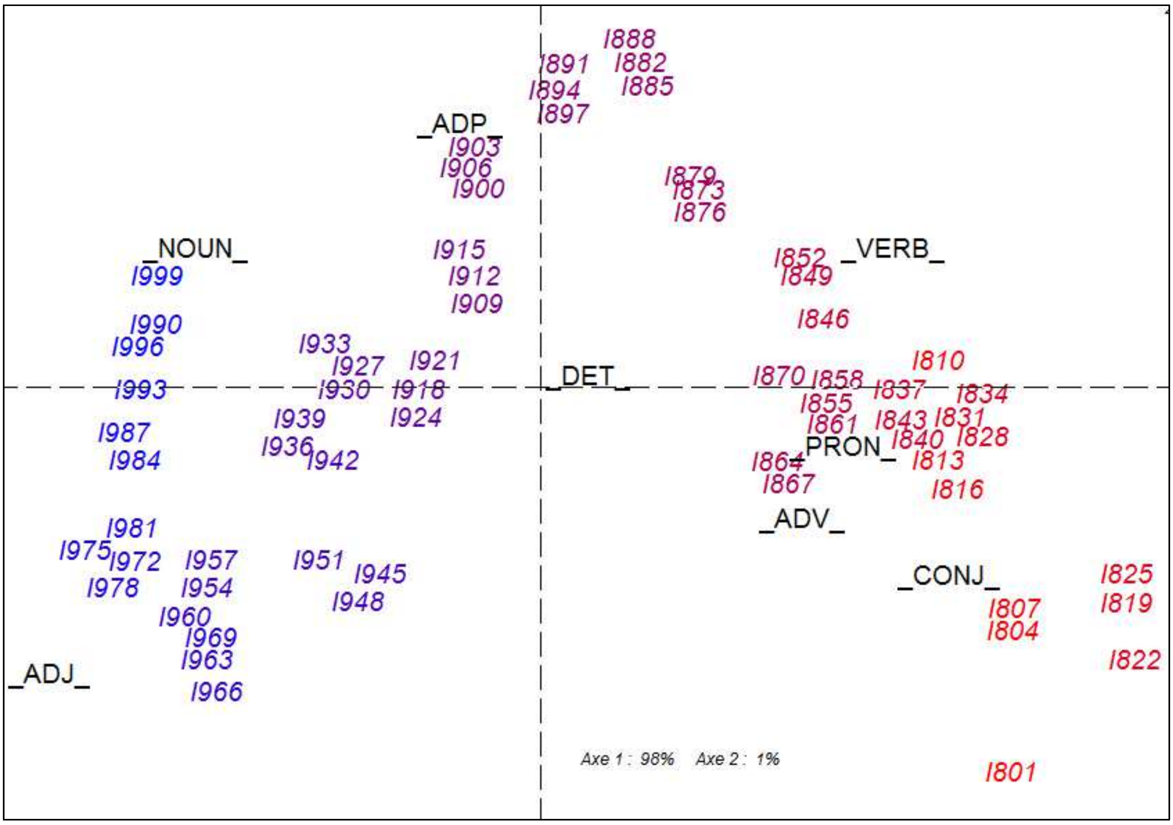


Figure 12. Le dosage des catégories
(24 milliards de substantifs, 10 milliards de verbes)

As we can see, one can be enthusiastic given the huge size of the corpora. But the doubt remains as to the validity of the statistical results. The doubt grows especially as the compositions of the corpora are still “black boxes”. As we saw, even a graph based on large corpora may still be sharply criticised. If the choices underlying the building of the corpus under scrutiny are unknown, the size of the data does not prevent the result from being very difficult to interpret. In such situations, one can talk of “insecurity” in large corpora, as did the reviewers of my book “Vocabulaire français” – which was however based on a corpus a thousand times smaller[20].

[20] Annie Geffroy, Pierre Lafon (1982). *L’insécurité dans les grands ensembles. Mots 5, 129-141.*

Zipf-Mandelbrot's Law Recoded with Finite Memory

Ronan Le Roux

ESPE, Université Paris-Est Créteil
ronan.le.roux@gmail.com

0. Introduction

Although the so-called "Zipf-Mandelbrot law" of word frequencies ("ZML" henceforth) gained immediate fame in the 1950s, its symbolic aura may vary depending on the academic or practical coordinates from which it is considered. Also, the interest it raised has fluctuated throughout the decades. Therefore, writing the history of this scientific law (following, for instance, Petruszewycz 1973) should include a careful mapping and clarification of the various frames of reference involved, since these are likely to shape the perception and historicization of such an object with disciplinary values and expectations. In other words: from their respective vantage points, a linguist, a mathematician, a computer scientist, a communication engineer, or a philosopher, will probably not see ZML with the same "eyes"; and so on for linguists (or mathematicians, etc.) of different schools and periods.

Among the accounts of Benoît Mandelbrot's contribution to this law that have been or could be written, it is worth noticing that the mathematician himself repeatedly wrote about it. Besides his notorious taste for actively constructing his public figure and tendency to downplay some of his influences (Hayes 2013; Bru 2012), it is fair to expect that, in general, autobiographical accounts can both bring valuable contextual information and introduce specific biases in this information.

This paper aims at contributing to the history of ZML in two respects: in the first sections, a comparison of Mandelbrot's early and late statements about the meaning of his work shows a shift in perspective: whereas, after the emergence of fractal geometry, the mathematician presents ZML in retrospect as a precursor application of fractals, early 1950s papers and talks unambiguously claim another unifying framework, that of cybernetics (that Mandelbrot did not reduce to "information theory", but rather wanted to reconstruct as a general science of behaviour). The last section of the paper broadens to the context of postwar France to better assess the originality of Mandelbrot's scientific gesture among the interest of mathematicians for language at the time.

1. ZML in retrospect: a precursor application of fractal geometry?

From time to time, Mandelbrot discussed the meaning of ZML for his broader work, in key publications as well as in his memoirs. In all these late writings, the point is the same: the mathematician retrospectively depicts ZML as a precursor application of fractal geometry. It is enough to pick up a few quotations that speak for themselves:

> "...since the beginning of my scientific career, I concluded that many real-world shapes are so irregular or broken, that the complexity of Nature overwhelms everything that is admitted by Euclidean geometry, not *quantitatively*, but *qualitatively*. […] The existence of such objects addressed geometry with the challenge of describing what was considered as "amorphous". Having decided to take the challenge, I developed and conceived a new geometry of nature, and used it in various domains. I did not say I "applied" it, since it would develop only through its diverse applications" (Mandelbrot 1997b, p. 32).
>
> "...the scaling distribution rules many [...] important phenomena. [...] It turns out that its most elementary example, that is, the most fundamental, is the Zipf law of distribution of words in human discourse" (Mandelbrot 1997b, p. 192).
>
> "Thanks to fractal geometry, those bits of knowledge became understood, acquired a clear identity, and ceased to be "homeless" by becoming part of a new field" (Mandelbrot 1997a, p. 8).
>
> "Zipf's law proved interesting in probabilistic terms and [...] somehow started me on a path that led, first, to finance and economics, and eventually to fractals" (Mandelbrot 1997a, p. 205).
>
> "...I found myself in the position of that child in a story who noticed a bit of string and […] pulled on it to discover that it was just the tip of a very long and increasingly thick thing […]. Oddly but almost ineluctably, that string […] ended up directing me to some of the main themes of my scientific life: [the concept] of fractality" (Mandelbrot 2012, p. 150).

ZML is accordingly featured in Mandelbrot's magnum opus *The Fractal Geometry of Nature* (Mandelbrot 1977, pp. 344 ff.), since its exponent parameter, called "discourse temperature" (or "informational temperature"), generally has a value between 0 and 1 (thus of non-integer dimension, what Mandelbrot considered as an approximate definition of 'fractal').

All the accounts just mentioned converge to construct a coherent picture of Mandelbrot's scientific accomplishments, according to which it was only in the end that ZML found a link to a broader theoretical frame. Before that, the mathematician says, ZML was "homeless". Even the anecdote told by Mandelbrot, of how he came to hear about Zipf's law and start thinking about it[2],

[2] "At the end of a day spent near the Sorbonne, it was not much of a detour – before taking the metro home – to stop at Szolem's flat [i.e., his uncle Szolem Mandelbrojt, a towering mathematician then professor at the Collège de France]. [As a] response to my routine request for reading material for the long ride home […] he pulled out

suggests that ZML came out of the blue, out of solving a fortuitous problem on a blank mental blackboard.

2. Early influences

Was there really nothing at the beginning, as is meant by this picture? When Mandelbrot received the challenge *via* Walsh *via* his uncle Szolem, he *did* have a background already, which played a role in his subsequent development and generalization of Zipf's law.

> "All my scientific work fell under the influence of the branch of physics called thermodynamics, and of other independent traditions ranging from deep to very shallow. I came to scaling and renormalization by cross-fertilizing the influences of probability theory (Levy) and the social sciences (Pareto, Zipf and the economists idea of aggregation)" (Mandelbrot 1997a, p. 105).

Besides thermodynamics (of which Mandelbrot considered himself an adept), the "other traditions" are not detailed here, and barely are they in the 2010 memoirs. However, the influences contemporary of Mandelbrot's 1950s work with Zipf's law can be traced back.

> "To have witnessed the birth of a field [i.e., molecular biology] from close by was an experience I never forgot. It provided exhilarating proof that someone with my bent might have a chance after all. […]
> The timing was ideal because several new developments that had been "bottled up" by war conditions were being revealed in a kind of fireworks I saw on no other occasion. My restless curiosity led me to read works that were widely discussed when they appeared: *Mathematical Theory of Communication* by Claude Shannon, *Cybernetics or Control and Communication in the Animal and the Machine* by Norbert Wiener, and *Theory of Games and Economic Behavior* by John von Neumann and Oskar Morgenstern.
> […] I was beginning to think that the examples of Wiener and von Neumann might guide me to an idea big enough to make me, in some way, the Delbrück of a new field[3].

of his wastebasket a reprint he had recently received from the Harvard mathematician Joseph L. Walsh (1895-1973), president of the American Mathematical Society. […] I became hooked: first deeply mystified, next totally incredulous, and then hopelessly smitten... to this day. […] the metro ride was long and I had nothing else to do. By its end, I had derived a more general version I could explain and was dying to confront it with data. I soon decided to pursue this strange avenue, all the way to a PhD. It is known today as the [ZML]." (Mandelbrot 2012, p. 151)

[3] Max Delbrück was a leading founder of molecular biology, whom Mandelbrot met during his stay at Caltech. Formerly a physicist, Delbrück thus gave Mandelbrot an example that crossing disciplinary boundaries from hard to softer sciences could be

> This is precisely what I set off to do" (Mandelbrot 2012, p. 126).

These early influences date from Mandelbrot's stay at the California Institute of Technology in 1948-49. What the mathematician does not say is that these "hot topics" of the time are linked to his imminent work on Zipf's law, not just by the inspiring figures of their patrons Wiener and von Neumann, but also by the scientific perimeter they set up. Only in a 1985 interview did he clearly mention that link[4], while the memoir is more elusive.

If we turn to Joseph Walsh's account of Zipf's 1949 book, published the same year in *Scientific American*, we can read Mandelbrot's vocation in plain letters. Walsh draws a parallel between, on the one hand, the recent work of von Neumann, Wiener and others, and, on the other hand, the history of celestial mechanics.

> "Certainly it is reasonable to expect that laws in social science, subject to revision and obsolescence, may similarly be established. […] Zipf is at his best, and indeed without a peer, in the statistical study of language. […] In a sense, this work so far serves mainly to suggest further unsolved problems. […] Why, for instance, is the distribution linear and the slope minus one? Opportunity is ripe for new Tycho Brahes, Keplers and Newtons!" (Walsh 1949, p. 56, 58)

This would be Mandelbrot's literal inspiration (for which he credits Walsh in his memoirs), suggesting a continuity between the references he discovered in Caltech, and Zipf's work that needed mathematical clarification (and competition for fame). Not only did Mandelbrot have a background, but he got involved in this corpus, and presented ZML as an application of cybernetics, as evidenced by several sources, some missing from biographical and autobiographical accounts.

3. Mandelbrot and Cybernetics

Several aspects of Mandelbrot's involvement with cybernetics have been documented by Segal (2003). This section expands on the meaning cybernetics had for his ambitions at the time: a general theory of behaviour, besides a theory of information. At the beginning of the 1950s, Mandelbrot attended an informal

fruitful (although it is not the same modality of boundary crossing, as will be seen in the second part of the paper).

[4] "The title [of my dissertation] was *Games of Communication*, largely because for several years before and after my Ph.D. I was very much influenced by the examples of John von Neumann and Norbert Wiener. Indeed, Wiener's book *Cybernetics* and von Neumann & Morgenstern's book *Theory of Games and Economic Behavior* had come out during this time, and they were precisely what I wished to emulate one day. Each seemed a bold attempt to put together and develop a mathematical approach to a set of very old and very concrete problems that overlapped several disciplines" (Mandelbrot 1985, p. 218)

seminar held at the Sorbonne Institute for the History of Science. This small seminar was called the *Cercle d'études cybernétiques*. Created by mathematician Robert Vallée, it gathered about twenty scholars from various disciplines (mathematicians, engineers, biologists...) and featured fourteen sessions between November 1951 and November 1953. The speaker for the third session was Benoît Mandelbrot, on April 5th, 1952. The title of his talk (as translated into English) was: "Relationship between the statistical structure of language and the presumed properties of the brain"[5]. Mandelbrot's name is included in the members' list of the *Cercle*, but it is not clear how he and Vallée got in touch. Both came from the École Polytechnique, but did not graduate the same year. We just know that, somehow, Vallée heard about Mandelbrot's work in progress, and that it was not in the context of Polytechnique. The *Cercle* is not mentioned in Mandelbrot's memoirs. "Indeed I was a member of that *Cercle d'Études* and keep a fond recollection of its meetings. But I stopped when I moved to MIT and don't remember the topics discussed by the other speakers", he simply wrote in 2008[6].

Mandelbrot's PhD took place between 1950 and the end of 1952. His dissertation combines his topics of the moment: thermodynamics, Zipf's law, and the "fireworks" he discovered in 1948-49 at Caltech: Shannon, von Neumann, and Wiener's theories. Wiener and cybernetics are present at a different level than information and game theories, though. The dissertation deals with "games of communications", in the sense that the sender and the receiver form a coalition to play a game against Nature. The strategy of Nature is to produce noise against the transmission of a given message, while the strategy of the communicating coalition is to maximize the transmission for a minimal cost. Feedback mechanisms are not discussed.

The first reference to Wiener is about considering a message as a random sequence (Mandelbrot 1953, p. 5), which is not a cybernetic trademark. But right after, Mandelbrot states that his work is closely related to cybernetics:

> "Communication games [apart from being very special and asymmetrical] can differ a lot from one another, depending on the nature of the messages involved. The existence of a *unique* communication theory thus rests on a special postulate, which unifies, on a functional plane, physical, biological and human phenomena:
> *We assume that the functional conditions of communication (and control) processes belong to an abstract study* that has the rigorous norms of Physics, is quantitative, conceptually homogeneous, *independent from their physical realizations*, hence applicable to living organisms as well as to artificial mechanisms.
> Mr. Wiener, to whom the most striking presentation of this principle is due, proposed to coin 'Cybernetics' the science dealing with its applications" (ibid., p. 5-6).

However, it is not clear, from this statement, whether the "abstract theory

[5] It was most likely preliminary material for Mandelbrot (1953b).

[6] B. Mandelbrot, Personal communication, E-mail, Feb. 27th, 2008.

of communications" and "cybernetics" are one and the same thing, or whether cybernetics only deals with *applications* in machines or organisms. At any rate, according to Mandelbrot, as an application (of the study of communication games in human language), the analysis of Zipf's law belongs to cybernetics. Another influence of Wiener is probably the extensive reference to Maxwell demons, which are formally analyzed in a chapter of the dissertation. Although Mandelbrot claims to have grasped the concept from Physics, its extension to other applications was most notoriously suggested by Wiener.

Thus Wiener, and cybernetics, are involved in the PhD dissertation in a theoretical, or even philosophical, rather than technical, way. That does not mean they are less important in Mandelbrot's vision. Even though he could have presented his dissertation without any reference to cybernetics, he stuck to it as a broader and unifying frame which he then declared to be his actual scientific horizon. He explains that the purpose of his dissertation is not to involve mathematics of the highest possible technical level, but to contribute to a transversal theory. He goes to historical arguments to justify a subtle position: whereas physics and the social sciences have had different mathematics (i.e., "classical" for physics, and game theory for social sciences), cybernetics has provided a bridge between them with information theory. Thus, for Mandelbrot, game theory is (mathematically) more general than cybernetics, but cybernetics is more unifying. The ultimate purpose of his work, though, is to bring cybernetics to a more general form – or, maybe it is better to say that it is to contribute to the construction of a transversal theory within which cybernetics is just a particular case:

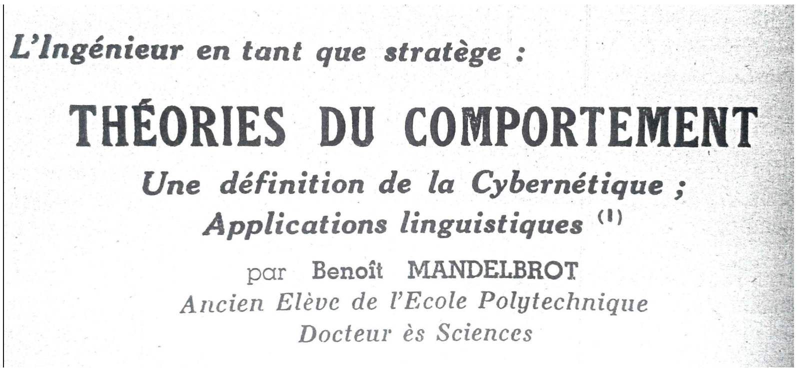
> [The present work] "soon lead us to conclude that Cybernetics' information is only a particular functional linked to certain strategies, within problems of inductive behavior, and part of a very large class of other information concepts" [ibid., p. 9].

Eventually, Mandelbrot gives a somewhat final touch to the theoretical landscape of his work, under the sober title of "terminological remarks": since the "theory of inductive behavior" (referred to Neyman and Wald) and cybernetics overlap to a certain extent, a choice has to be made whether they should be considered one and the same, or exclude one another: "We refuse to take side in this alternative, and avoid the term 'Cybernetics' until further notice", he simply writes, considering that cybernetics remains too confuse (ibid., p. 10). Besides the terminological dilemma, it is perfectly clear that ZML is far from "homeless": it is an application of a theory of communication or inductive behaviour, the generality of which it is even supposed to support and illustrate:

> [The examples of thermodynamics and of the statistical structure of language] "were chosen partly for their valuable support to intuition in the attempt to make the fundamental concepts evolve, and make them more suited to the attack of real problems" [ibid., p. 7].

In 1953, a few months after he defended this PhD thesis, Mandelbrot was invited for a post-doc at MIT, where he wanted to work with Wiener. Unfortunately, Wiener was very depressed at the time, so that Mandelbrot could not expect anything, and stayed only one month (he then went to the Institute of Advanced Studies in Princeton to work with his other role model, John von Neumann – Wiener, though, will durably praise Mandelbrot's work, including ZML in the 1954 revised edition of *The Human Use of Human Beings*[7]). Wiener had split with McCulloch and other scholars (Conway & Siegelman, 2005), which probably somehow undermined cybernetics. Also, that year, Wiener's work in cybernetics was quite off the records. Under such circumstances, it would have been timely for Mandelbrot to give up referring to cybernetics. Yet, he did not.

Mandelbrot came back to France for one year, in 1954 and 1955. In June 1955, he gave a plenary talk at the "Conférences interscientifiques" of the Institut Henri Poincaré, a key place in the French mathematical landscape. This talk was published as a non-technical paper under the title (translated): "The Engineer as a Strategy-Maker: Theories of Behavior. A definition of Cybernetics; Applications to Linguistics". With respect to the 1953 PhD dissertation, this paper brings out significant elements. It stresses Mandelbrot's interest for, and even commitment to, cybernetics, as he wants to "define what [he] would have liked this discipline to become" (Mandelbrot 1955, p. 278). He claims he wants to rehabilitate the term, which implies, he writes, to overcome Wiener's version. He gives his own definition of cybernetics, which is somewhat original, but too long to discuss here. Then, he introduces an "example of application of cybernetics": "In order to show the unity of Cybernetics, we will take an example in communication theory: a problem of message reception […]" (ibid., p. 287). Mandelbrot explains that the construction of an optimal receiver requires searching for and constructing a decision function that corresponds to a certain strategy.



L'Ingénieur en tant que stratège :

THÉORIES DU COMPORTEMENT

Une définition de la Cybernétique ;
Applications linguistiques (1)

par **Benoît MANDELBROT**
Ancien Elève de l'Ecole Polytechnique
Docteur ès Sciences

After presenting "minimax" and "Bayesian" reception strategies, Mandelbrot ends his paper with a discussion of a "model of the structure of language". ZML is featured as a "cybernetic model" of language.

> "As an example of application of cybernetics, let us mention, for a change, a model that allows a description and "explanation" of empirical linguistic laws, thanks to concepts and tools created to deal with normative problems of communication" (ibid., p. 292).

[7] Wiener 1954, pp. 92, 187.

Mandelbrot simply explains that, in a natural text made of statistically independent words, the law of distribution of frequencies corresponds to a minimal Bayesian-decoding cost. (Then he discusses objective and subjective interpretations of this correspondence, and considers that in the second case the structure of language can be considered "intentional"). Hence, natural human (written) language appears as a specific possible communication strategy, beside others (namely, a "minimin" reception strategy, beside "minimax" and "Bayesian" receptors).

This paper is significant, as it suggests that "cybernetic" was not just synonymous with "information-theoretical". Neither was it just a reference Mandelbrot would use to label his early work with the name of an exciting new field. He was sincerely committed to making something out of it. In the references at the end of the paper, two publications labelled "forthcoming" are featured: a joint paper with Schützenberger, under the title "Decision and Informations", and, probably a book by Mandelbrot, simply entitled "Cybernetics". None will ever see the light. Anyway, this paper makes clear that ZML was included in a broader frame that was a theory in the making, a theory of communication and decision for which Mandelbrot temporarily chose to adopt the name "Cybernetics".

Between 1955 and 1957, Jean Piaget hired Mandelbrot at his *Centre international d'épistémologie génétique* in Geneva, a hotspot of cybernetics in Europe (among other general topics). Piaget had probably heard about ZML, and this kind of work, as well as Mandelbrot's profile, was exactly what he was looking for to develop and strengthen the *Centre*. Among the numerous publications produced there, Mandelbrot wrote two papers. The first one deals with the concept of equilibrium (Mandelbrot 1957a). The second one, "Linguistique macroscopique" (Mandelbrot 1957b), is the first part of a volume entitled *Logique, langage et théorie de l'information*, which other parts are handled by logician Leo Apostel and psychologist Albert Morf. In this long paper, Mandelbrot expands in a non-technical fashion about several of the implications of ZML. But this time, the background discussed is the analogy with thermodynamics, at the expense of cybernetics. (A key question for Mandelbrot is how scales relate to one another, i.e., how grammar rules at the "micro" scale of the individual speaker can generate a "macro" structure such as the ZML – he considers that the parallel with the theory of gases is very relevant). Nevertheless, in March 1956, he presents a talk "Macrolinguistique cybernétique" at a two-days conference about cybernetics organized by the *Société d'Études philosophiques*, that brought together several of the main names associated with cybernetics in France.[8]

In 1958, after Mandelbrot went back to Paris, he was a speaker at the 21st *Semaine de synthèse*, a high-level interdisciplinary conference taking place almost every year since 1925. Only an abstract of his talk was published, that reads as follows:

> "Cybernetics. Automata theory; Information theory and its applications in Physics and the Social sciences.
>
> It is notorious that the most splendid syntheses are often unstable. After a decade, it has to be acknowledged that, while the grand ideas of N.

[8] *Revue Philosophique de Louvain*, Chronique générale: Congrès et sociétés savantes, p. 557.

> Wiener's cybernetics have become mainstream, they could not unify the diverse activities between which they had revealed such a deep link. It is such a pain to "define" such a thing as a Cybernetics that one can even wonder whether it does actually exist, despite the importance of its parts.
>
> What can be said, nowadays, of Cybernetics in general? I will present it briefly, using the language of another contemporary theory (which bears obvious links to Cybernetics), that of statistical decision theory as a strategy game.
>
> [After reviewing feedback theory and automata theory] I will expand more thoroughly on another aspect of Cybernetics I am more concerned with: information theory. Its relationships with the foundations of thermodynamics have been noticed since the very beginning. I will illustrate recent advances by drawing on the theory of statistical estimation: first example of a new methodological convergence between Physics and the Social sciences. Next, I will show how information theory shed light on the theory of the structure of certain natural systems of signs, that is, on linguistics. The segment of this science, thus subject to mathematical formalization, remains of course very narrow with respect to the infinite variety of known linguistic facts. But the "macroscopic" linguistic theories feature a striking parallelism with the macroscopic theories of matter, sketching maybe a future interdisciplinary macroscopic science. At any rate, this parallelism opens wide perspectives, whose seed was perhaps in Wiener" (Mandelbrot, 1962, p. 10).

By merely confronting the sources, it appears that cybernetics (whatever it meant for Mandelbrot) played the role of a unifying reference, an early paradigm vis-a-vis which the mathematician was trying to make sense of ZML and to position his work in the scientific field, and that this choice lasted during the entire 1950 decade.

4. Abandon in the 1960s

In the following years, Mandelbrot gave up that reference to cybernetics. At the beginning of the 1960s, he does not work on language anymore. Speaking at the 1962 international conference on "The concept of information in contemporary science", one of the grand interdisciplinary conferences held at the Royaumont abbey, Mandelbrot gives a rather critical talk, with the title "Is information theory *still* useful?" (Mandelbrot's italic), about how the notion of 'information' lost its unifying power as several specialized sub-fields gave it precise, non-equivalent and non-overlapping technical definitions. During his presentation and the discussion that follows, he explains that he still values problems common to different scientific disciplines – nowhere does he talk of cybernetics, although one recognizes the same idea –, but does not believe that the idea of information still constitutes an interesting meeting zone for such problems. He considers information theories as integrated into the normal course of science. This moment is interesting, because he clearly states that he does not see what could replace information theory (and, we could add, cybernetics) in this role of a fruitful cross-disciplinary reference. Later, it will be fractal geometry; but in the 1960s, ZML, and other new works Mandelbrot mentioned in this talk (especially in economics), are orphaned from any such unifying paradigm. This, of course, will make it easier reinterpreting them in retrospect by and after the turn to fractals. As to ZML in particular, or language more generally, we notice that it is not discussed anymore. For Mandelbrot, the subject seems closed. His acrimonious debate with Herbert Simon, between 1959 and 1961, already gave that impression[9]. The only change in the terms in which Mandelbrot will discuss it will be the introduction of the word "fractal" circa 1975. Thus, ZML is in a kind of dormant state in Mandelbrot's memory for at least a decade, during which his former cybernetic framework decays.

From all that precedes, we can suspect that it was tempting for him to reinterpret the meaning of ZML once he coined the word "fractal geometry", and erase his former cybernetic tracks. Certain expressions he uses to narrate or comment on his early work proceed to such a recoding. For instance, when he describes his PhD thesis in his memoirs, he characterizes it as "messy", "flawed", and "ahead of its time". Yet, what precedes suggests that it did take place in a bigger picture, and that it was very timely because of that. Another example is when Mandelbrot mentions the choice of chairmen for his defense committee: Nobel physicist Louis de Broglie was selected for the official reason that "he publicly praised interdisciplinary work"[10]. These seemingly trivial words ring no bell for the unaware reader. In the 1950es, De Broglie was a promoter of cybernetics (though, as in the case of several other French mathematicians and engineers, it was generally synonymous with information theory): the proceedings of his 1951 seminar were published with the title *La Cybernétique : théorie du signal et de l'information* (one of the speakers there, Robert Fortet, a specialist on random functions, was the third chairman of Mandelbrot's defense); De Broglie accepted the symbolic patronage of the *Cercle d'études cybernétiques* (of which, remember, Mandelbrot was a member!); he introduced the first two notes labeled "Cybernétique" for the C*omptes-rendus de l'Académie des Sciences*, and

[9] Herbert Simon proposed an alternative model, in which Zipf's law is explained not by an optimal communication game as in Mandelbrot's, but by a "Matthew effect" instead, a self-reinforcing process by which the words that are most used tend to be the most used. If we consider the titles of the papers by both protagonists, it is amusing to notice how early in the debate Mandelbrot claimed to put a "final" point to it: Simon: "On a class of skew distribution functions " (1955).

Mandelbrot: "A note on a class of skew distribution functions " (1959)

Simon: "Some further notes on a class of skew distribution functions " (1960)

Mandelbrot: "Final note on a class of skew distribution functions: Analysis and critique of a model due to H. A. Simon" (1961)

Simon: "Reply to "Final Note" by Benoit Manbelbrot" (1961)

Mandelbrot: "Post scriptum to "final note"" (1961)

Simon: "Reply to Dr. Mandelbrot's post scriptum" (1961)

[10] Mandelbrot 2012, p. 143. The non-official reason, according to Mandelbrot, being that his supervisor Georges Darmois invited De Broglie, who was a perpetual secretary of the *Académie des Sciences,* because he was campaigning to be elected there.

he wrote a short paper "Philosophical meaning and practical range of cybernetics" in the main popular science journal of the time, *Atomes*. Thus, De Broglie notoriously linked his figure and Nobel Prize aura to the name of cybernetics, simply recoded as "interdisciplinary work" in Mandelbrot's memoirs.

In his 1985 interview, the mathematician gives clues about his abandonment of cybernetics. It is interesting to notice that he presents it as a consequence of a general failure, without mentioning his own commitment to the domain:

> "Unfortunately, cybernetics never really took off, and game theory became yet another very special topic. Colossal claims had been made when there was little to support them, and they were not immediately shrugged off only because of the authors' renown, based on earlier and very different work. It soon became good manners in academia to laugh when someone mentioned "interdisciplinary research." To my bitter disappointment, I had to agree that there was good reason for laughter. I wondered whether things would have been better if von Neumann and Wiener had had the desire and the ability to take an active interest in their progeny" (Mandelbrot 1985, p. 218).

5. Mathematicians and Language in Postwar France

In his memoirs and elsewhere, Mandelbrot, who enjoys calling himself a "maverick" (not less than 99 times in his memoirs!), emphasizes the originality of his work (and more generally of his person and trajectory): "I have never done anything like others" (Taleb 2010). This is true in many respects, in particular in choosing Polytechnique instead of Normale when he ranked first at both schools' entrance exams, and in many other examples not worth mentioning here. If we focus on ZML, Mandelbrot also pictures his choice of working on a linguistic topic in the early 1950es as original. He stresses that he started his PhD without a supervisor. However, Mandelbrot was, so to speak, not alone in the no man's land between mathematics and the study of language (and, more generally, the social sciences).

With the heyday of linguistics as a "pilot science", and figures such as Roman Jakobson who were interested in possible contributions of engineering sciences to the study of human communication, the circumstances were favourable to the development of formal methodologies, and thus possibly appealing for mathematicians. In France, love was in the air too, but to a certain extent only. It is certainly true that the general context was not helping. France benefited much less than its allies from the scientific and technological outcomes of World War 2. Interdisciplinary connections due to the exceptional circumstances of the conflict (such as, typically, operations research groups) were sometimes experienced by French engineers and scientists; but they were embedded into the reconstruction of French research only in a marginal or exceptional way. Strong, enduring traditional disciplinary boundaries remained unchanged and still shaped the habitus of French normal science. More than the War, Auguste Comte, Bourbaki and the "two cultures" gap were names for the implicit or explicit obstacles to the emergence of modelling practices in the life and social sciences.

Thus, interdisciplinary boundary crossing largely rested on individual initiative. Institutional support, whenever it existed, was not to become usual, and rather came from international channels. For instance, the Rockefeller Foundation funded a series of major CNRS conferences in the 1950es and 1960es. Another significant occurrence is the support of UNESCO: besides the international congresses of cybernetics, which took place in Belgium, an interesting example is that of the seminar that took place in 1953 and 1954 in Paris on the use of mathematics in the social sciences. This seminar was organized by Claude Lévi-Strauss with financial support from MIT[11], and obtained the patronage of UNESCO's *Conseil International des Sciences Sociales*, of which Lévi-Straus had been appointed secretary. This seminar was one of the main interdisciplinary "trading zones" of postwar France, bringing together figures of the social sciences (understood at large) such as linguist Emile Benveniste, psychologist and sociologist Paul-Henri Maucorps, psychoanalyst Jacques Lacan, and mathematicians such as Marcel-Paul Schützenberger and Jacques Riguet, specialized in what was soon to be called the theory of automata. Benoît Mandelbrot was there too. (In his memoirs, he only mentions a collaboration with Lévi-Strauss, without further details[12]). Of course, language was a central topic, even from the point of view of disciplines other than linguistics. This short-lived seminar was an emblematic scansion in the emergence of various forms of "structuralism" in those disciplines.[13]

But "language", broadly understood, was not a topic brought to the attention of mathematicians from the outside, as a mere external demand from disciplines eager to harden their methodologies. If we could reconstitute a list, as exhaustive as possible, of all postwar French mathematicians interested in the study of some aspect of language (among which figures such as Jean-Pierre Benzécri and René Moreau would appear), we would notice networks. Mandelbrot was a friend of Schützenberger, who did towering work on algebraic automata, including a collaboration with Chomsky. Beside Schützenberger are two other interdisciplinary mathematicians: Jacques Riguet and Paul Braffort. Riguet, who was interested in the formal analysis of the structure of machines and computer programs, became the mathematical counsellor of Lacan, introducing automata diagrams to represent the constraining necessity of the "symbolic order". Braffort was something of a pioneer in automatic classification and later in artificial intelligence, and a member of the literary think-tank *Oulipo* with other mathematically-minded fellows such as Raymond Queneau and François

[11] The support came from Max F. Millikan, director of the Center for International Studies (Le Roux, forthcoming).

[12] Mandelbrot 2012, p. 164: "Claude Lévi-Strauss, the illustrious anthropologist I had worked with in Paris, had recommended me to his close friend Roman Jakobson".

[13] Another outcome is the famous 1954 issue of the *International Social Science Bulletin* devoted to "Mathematics and the Social Sciences". Surprisingly, Mandelbrot is not mentioned in the volume, although Lévi-Strauss quotes Yule's 1945 study on the statistics of vocabulary (Lévi-Strauss 1954), while Colin Cherry's paper "On the Mathematics of Social Communications" (Cherry 1954) refers to Zipf's 1949 book. Both authors thus seem to skirt Mandelbrot's work, which they were aware of.

Le Lionnais (Braffort 2002, Motte 1986). Braffort (1994) mentions that Schützenberger, Riguet, Mandelbrot and himself, when students in mathematics in the Quartier Latin, formed a group to study the "anatomy and physiology of mathematics", with an original simultaneous interest in Bourbaki's abstract structures and applied topics. Mandelbrot's involvement may not have been to the same extent as others' though[14]. At any rate, in brief, Mandelbrot was linked to mathematicians whose interest in language was *intrinsic*, *permanent*, and *notorious* (even without mainstream institutional support), and he was totally identified and recognized among them for his work on ZML (see, for instance, Moscovici 1959). And while Bourbaki paradoxically inspired many French mathematicians interested in language (Aubin 1997) as well as the purism that would discourage and marginalize such cross-disciplinary mathematical modeling, it did not inspire Mandelbrot at all. Structuralism held no monopoly on the postwar (mathematical) understanding of human language: in a bigger picture that rehabilitates the place of statistical approaches, often bypassed, ZML is emblematic rather than marginal.

Next to this historical perspective on the relative originality of ZML, focusing on specific aspects of the French postwar context, I suggest another perspective, a sociological one, focusing on general aspects by replacing Mandelbrot's gesture in a typology of practices. In the present case, it is a typology of mathematization practices. If we compare Mandelbrot's intellectual trajectory with, for instance, that of the members of *Oulipo*, a major difference is striking: whereas most Oulipian mathematicians are interested in language *per se* and take it as their main and lasting topic, Mandelbrot's study of language was mostly limited to ZML[15], after which he moved on to other domains: economics, finance, signal transmission whether in telephone or the nervous system, galaxy clusters, and all the objects eventually characterized as fractal. By developing a special kind of mathematics which he tried to adapt to a variety of phenomena from different disciplines, he would cross boundaries back and forth, each time solving local problems and enriching his general theory in return. This *modus operandi* corresponds to a perfectly identified ideal-type of practice, known in the sociology of science as the "transversal regime of knowledge production" (Shinn 2007). This regime of activity is that of interdisciplinary instrument makers, who construct material or intellectual generic instruments and adapt them to a series of specific niches of problems arising in different disciplines, while, in turn, their generic knowledge benefits from learning and increased robustness. This is a good sum up of Mandelbrot's trajectory, publishing his papers as an outcome of visiting respective specialized communities, while slowly developing his general theory in parallel. ZML is thus the first chronological outcome of a definite activity of development and circulation of modelling tools.

[14] "I barely knew Riguet and Braffort. I knew Schützenberger very well - or so I thought. You tell me about interests of his of which I was not at all aware. Therefore, they did not affect me" (Mandelbrot, Personal communication, E-mail, Feb. 27th, 2008).

[15] But he seemed to take linguistics quite seriously, that is, not as mere statistical material.

From there, two remarks should be made. First, regarding the cybernetic debuts of Mandelbrot, it is worth noticing that Norbert Wiener was another remarkable figure of transversal scientific modelling. It means that, as a matrix of generic instruments typical of the transversal regime, Mandelbrot's first, cybernetic paradigm, to which he related ZML, has a status similar to fractal geometry. The second remark is that, when comparing the various specific fields targeted by Mandelbrot's transversal trajectory, linguistics appears both as its archetypal experience, and yet did not benefit from the same refinements as other fields. Mandelbrot's role is that of devising a tool for each community, who will then take it over and push further the adaptation to the specific aspects of its problems. "When a fractal theory really starts moving by itself I tend to become technically under-equipped to continue to participate, and it becomes wise to move on" (Mandelbrot 1985). However, from that point of view of the technical and social dynamics of the transversal regime, ZML remained somewhat orphaned: "In linguistics, fractals will not revive. My early work [i.e., ZML] was important to me but peripheral to the field" (ibid.).

Thus, regarding both perspectives – an historical perspective on the French postwar context, and a sociological perspective on modelling practices –, it is possible to better assess and relativize the originality of Mandelbrot's involvement with ZML, that is, to characterize its kind of originality with respect to a context of emerging practices performed by many original individuals: a paradoxical "community of mavericks", of which Mandelbrot was both an emblematic figure, and a "lonesome cowboy". As a scientific gesture, ZML was consistent both with Mandelbrot's early influence (that of Wiener and von Neumann) and his fractal paradigm of his maturity. It was part of a typical mode of scientific activity (the transversal regime of scientific modelling) which provided continuity and coherence to Mandelbrot's trajectory, despite the variety of domains he attacked as he encountered them, and although he had to wait for the mid-seventies to feel unity and achievement.

References

Aubin, David (1997). The Withering Immortality of Nicolas Bourbaki: A Cultural Connector at the Confluence of Mathematics, Structuralism, and the Oulipo in France. *Science in Context* 10, 297-342.

Braffort, Paul (1994). Les digitales du mont analogue. http://www.paulbraffort.net/science_et_tech/atomistique_et_automatique/calc_analog_et_auto/mont_analogue.html

Braffort, Paul (2002). Prolégomènes à une occultation. http://www.paulbraffort.net/litterature/pataphysique/prolegomenes.html

Bru, Bernard (2012). Marc Barbut historien. *Mathématiques et Sciences humaines / Mathematics and Social Sciences 200(4),* 9-25.

Cherry, Colin (1954). On the mathematics of social communications. *International Social Science Bulletin* VI(4), 609-622.

Chomsky, Noam (1957). *Syntactic Structures,* The Hague: Mouton.

Conway, Flo & Siegelman, Jim (2005). *Dark Hero of the Information Age: In Search of Norbert Wiener The Father of Cybernetics,* New York: Basic Books.

Hayes, Brian (2013). Father of Fractals, *American Scientist* 101(1). http://www.americanscientist.org/bookshelf/pub/father-of-fractals

Le Roux, Ronan, *Une histoire de la cybernétique en France, 1948-1975*, Paris: Classiques Garnier (forthcoming).

Lévi-Strauss, Claude (1954). The Mathematics of Man. *International Social Science Bulletin* VI(4), 581-590.

Mandelbrot, Benoît (1953). *Contribution à la théorie des jeux de communication.* PhD dissertation, faculté des sciences de Paris, 1953.

Mandelbrot, Benoît (1953b). An Informational Theory of the Statistical Structure of Language. In: W. Jackson (ed.), *Communication Theory,* London: Butterworths, 486-502.

Mandelbrot Benoît (1955). L'Ingénieur en tant que stratège: Théories du comportement. Une définition de la cybernétique; Applications linguistiques. *Revue des sciences pures et appliquées 62(9-10),* 278-294.

Mandelbrot, Benoît (1957a). Sur la définition abstraite de quelques degrés de l'équilibre. In: Piaget J., Apostel L. & Mandelbrot B. (eds.), *Logique et équilibre. Études d'Épistémologie Génétique vol. II,* Paris : Presses Universitaires de France, 1-26.

Mandelbrot, Benoît (1957b). Linguistique statistique macroscopique. In: Apostel, L., Mandelbrot, B., & Morf, A. (eds.), *Logique, langage et théorie de l'information, Études d'Épistémologie Génétique vol. III,* Paris : Presses Universitaires de France, 1-78.

Mandelbrot, Benoît (1962). La cybernétique. Théorie des automates; Théorie de l'information et ses applications en physique et dans les sciences sociales", abstract for the 21st "Semaine de synthèse", *Revue de synthèse* 83(1), 10.

Mandelbrot, Benoît (1977/1983). *The Fractal Geometry of Nature.* New York: Freeman & Company.

Mandelbrot, Benoît (1997a). *Fractals and Scaling in Finance. Discontinuity, Concentration, Risk. Selecta vol. E,* New York: Springer.

Mandelbrot, Benoît (1997b). *Fractales, hasard et finances*. Paris: Champs Flammarion.

Mandelbrot, Benoît (1985). Interview by Anthony Barcellos. In: D.J. Albers and G.L. Alexanderson (eds), *Mathematical People,* Boston: Birkhäuser, 205-225.

Mandelbrot, Benoît (2012). The Fractalist. Memoir of a Scientific Maverick, New York: Pantheon Books, 2012.

Moscovici, Serge (1959). Review of G.A. Miller "Langage et communication." *Revue d'histoire des sciences et de leurs applications* 12(2), 191-192.

Motte, Walter (ed.) (1986). *Oulipo: A Primer of Potential Literature,* Illinois: Dalkey Archive Press.

Petruszewycz, Micheline (1973). L'histoire de la loi d'Estoup-Zipf: documents. *Mathématiques et Sciences humaines / Mathematics and Social Sciences*

44(1), 41-56.

Segal, Jérôme (2003). *Le Zéro et le Un. Histoire de la notion scientifique d'information au XX^{e} siècle.* Paris: Syllepse.

Shinn, Terry (2007). *Research-Technology and Cultural Change. Instrumentation, Genericity, Transversality*. Oxford: The Bardwell Press.

Taleb, Nassim (2010). "Benoît Mandelbrot", *Time Magazine*, Nov. 1st, 2010, http://content.time.com/time/magazine/article/0,9171,2026995,00.html

Walsh, Joseph (1949). Another contribution to the rapidly growing literature of mathematics and human behavior. *Scientific American* 181(2), 56-58.

Wiener, Norbert (1954). *The Human Use of Human Beings.* Second Edition. Boston: Da Capo Press.

Zipf, George Kingsley (1949)., *Human Behaviour and the Principle of Least-Effort*, Cambridge: Addison-Wesley.

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). Nowadays 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:

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

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.

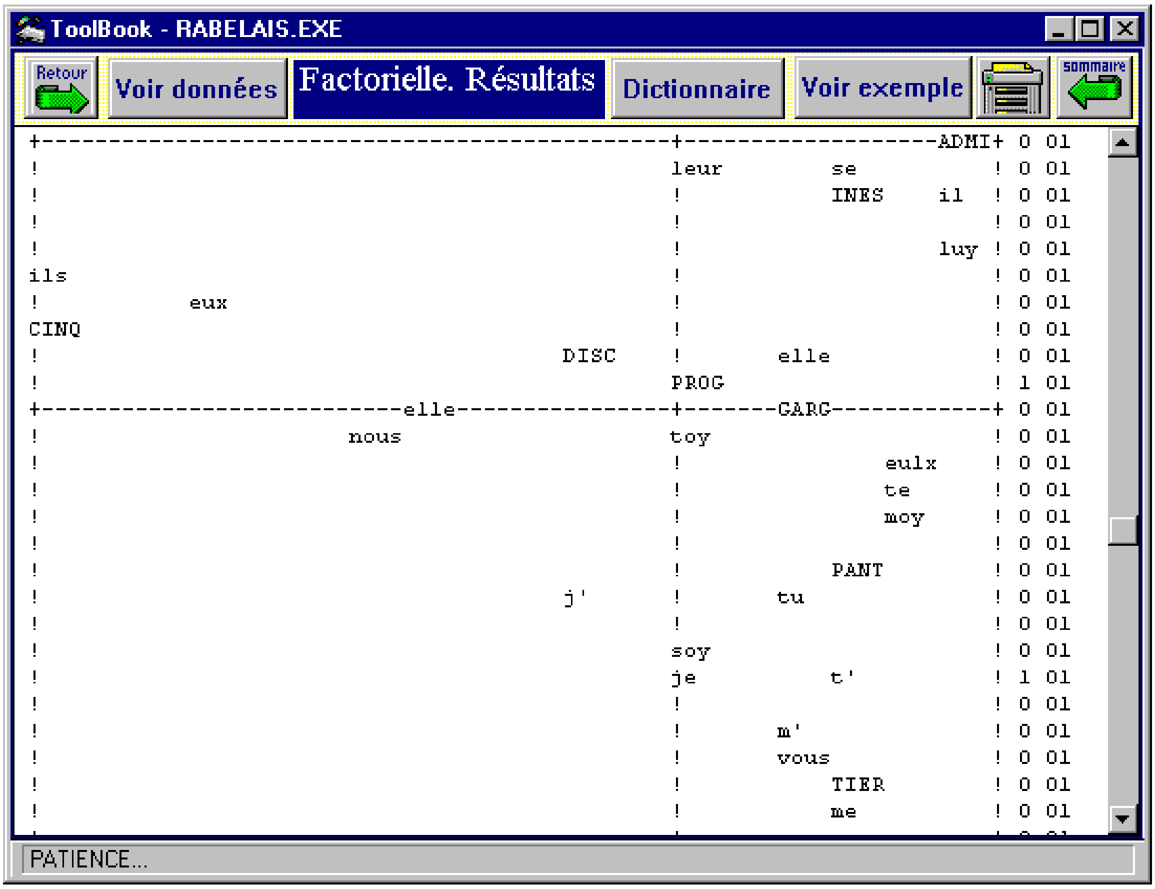
Crédoc (*Centre de recherche pour l'observation des conditions de vie*) was for a long time a powerhouse in the field of textual statistics. Ludovic Lebart worked there for many years (1971-1988), setting up and directing the survey *Aspiration et Conditions de vie des Français*. With André Morineau, 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. Attachments 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.



```
+-----------------------------------------------+-------------------ADMI+ 0 01
!                                               leur        se          ! 0 01
!                                               !           INES    il  ! 0 01
!                                               !                       ! 0 01
!                                               !                   luy ! 0 01
ils                                             !                       ! 0 01
!           eux                                 !                       ! 0 01
CINQ                                            !                       ! 0 01
!                                       DISC    !       elle            ! 0 01
!                                               PROG                    ! 1 01
+---------------------------elle----------------+-------GARG------------+ 0 01
!                       nous                    toy                     ! 0 01
!                                               !               eulx    ! 0 01
!                                               !               te      ! 0 01
!                                               !               moy     ! 0 01
!                                               !                       ! 0 01
!                                               !           PANT        ! 0 01
!                                       j'      !       tu              ! 0 01
!                                               !                       ! 0 01
!                                               soy                     ! 0 01
!                                               je          t'          ! 1 01
!                                               !                       ! 0 01
!                                               !       m'              ! 0 01
!                                               !       vous            ! 0 01
!                                               !           TIER        ! 0 01
!                                               !           me          ! 0 01
```

PATIENCE...

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).

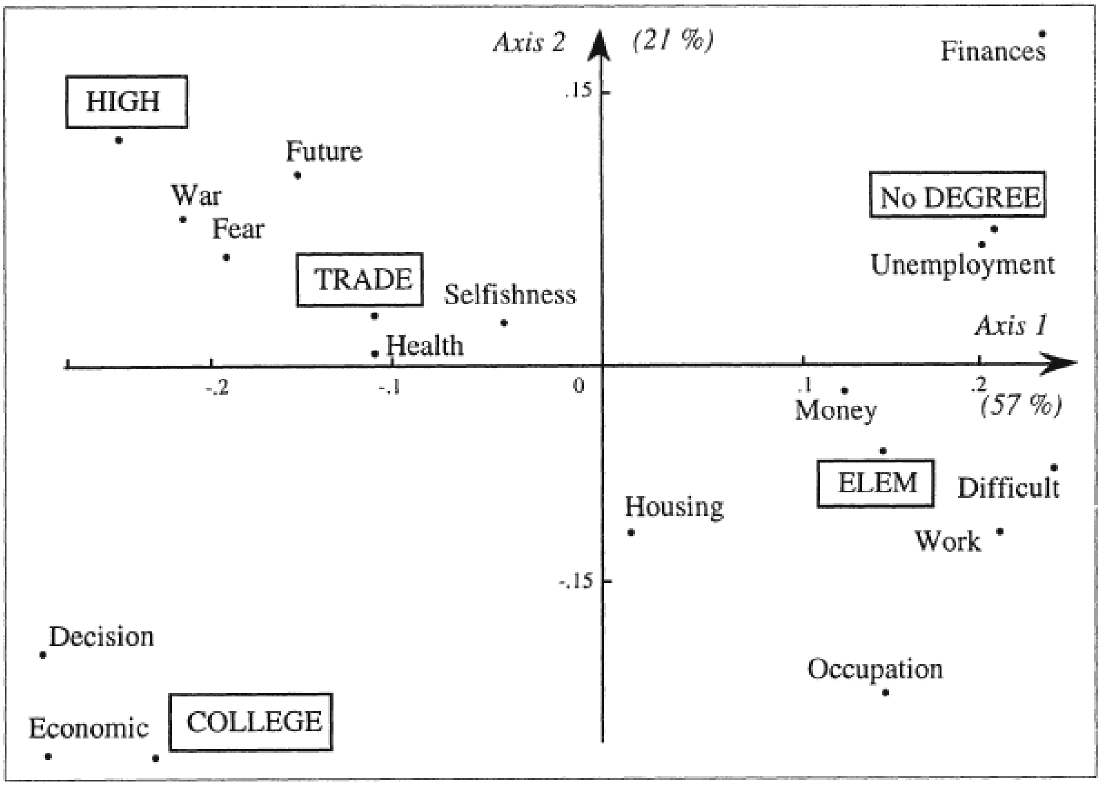


Figure 2. Proximities among words and among educational level (Lebart et al., 1998, p. 52)

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”.

> 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

[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.

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 hierarchical 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.

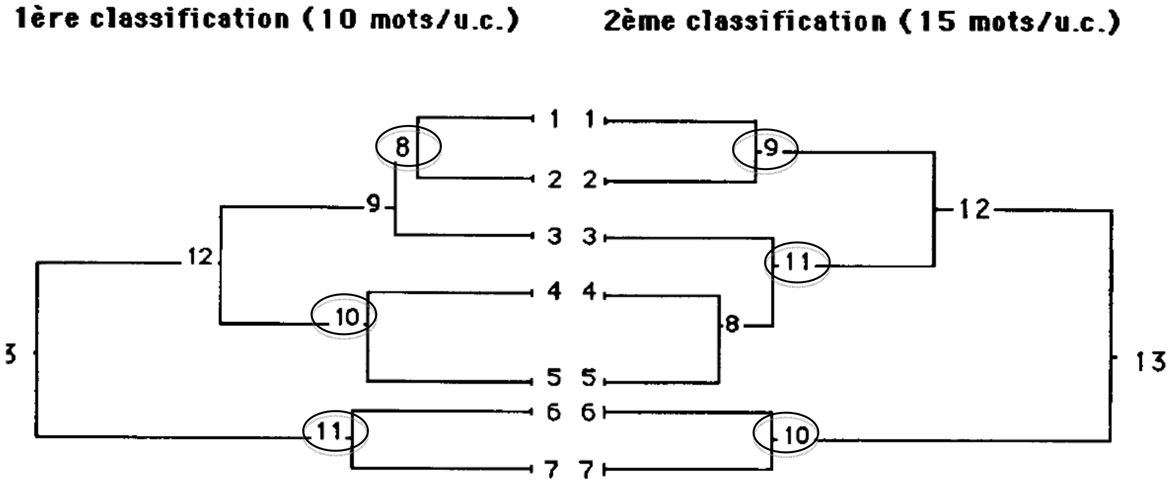

1ère classification (10 mots/u.c.)
2ème classification (15 mots/u.c.)
1
2
3
4
5
6
7
8
9
10
11
12
3
9
11
8
12
10
13

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^{o} 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.

Interview with Jean Petitot

Jacqueline Léon
Histoire des Théories Linguistiques (UMR 7597, CNRS / Université Paris Diderot)
Jacqueline.leon@univ-paris-diderot.fr

Sylvain Loiseau
Sedyl Laboratory (UMR 8202 CNRS/Inalco), sylvain.loiseau@wanadoo.fr

Jean Petitot
CAMS, École des Hautes Études en Sciences Sociales, petitot@ehess.fr

Abstract. From the 1960s onwards, the mathematician René Thom (1923-2002) carried out important contributions to the mathematical modelling of morphogenesis (analysis of forms). The proposed concepts (singularity, structural stability, catastrophe, bifurcation) were re-used in several social sciences, particularly in linguistics. They allowed a Gestalt-like approach, in opposition to the then dominant logico-combinatorial ones, and met some cognitivist trends and connectionist models. Jean Petitot was the first to show interest in the application of Thom's work to linguistics and he developed many studies accordingly.

This article is based on an interview between J. Petitot, J. Léon and S. Loiseau held on the 27 of September, 2014. While preserving the oral style of free conversation, it also includes references, developments and mathematical statements added by the authors.

1. Jean Petitot's biography

After preparatory classes at Louis-le-Grand high school, I entered the École Polytechnique in 1965 where I graduated in 1968. Impassioned by research, I joined the new Centre of mathematics my professor Laurent Schwartz had just created, and I learned algebraic geometry (with Grothendieck's disciple Jean Giraud) and differential geometry. As I investigated singularity theory[1], I met René Thom, one of the leading specialists of the field, who had restructured it entirely since the middle of the 1950s[2].

Besides, I was much interested in structuralism, in particular Claude Lévi-Strauss, whose lectures at the Collège de France I already attended when I was very young, around 18-19 years old. It is through Lévi-Strauss that I discovered

[1] cf. *infra*.

[2] René Thom (1923-2002) was a mathematician and a former student of the Ecole Normale Supérieure. He had lectured in Grenoble, Strasbourg and, near Paris, at the Institut des Hautes Etudes Scientifiques, until 1990. He received the Fields medal in 1958 for his work on differential topology. He developed mathematical models of morphogenesis popularized under the name of "Catastrophe theory" (cf. *Stabilité structurelle et morphogenèse*, 1972, InterÉditions, Paris).

Roman Jakobson. At that time, I did not see any link between structuralism and mathematics.

At the end of the 1960s, René Thom started to circulate the manuscript of *Stabilité structurelle et morphogenèse* (published in 1972). This book, which was focused in biology, also explained why the scope of the morphogenetic approach[3] went far beyond biology and could apply to structuralism in general. Thom had much discussed with Conrad Hal Waddington (1905-1975), an eminent specialist of embryogenesis, and he had benefited from Jakobson's strong support. His book, proposing a mathematical structural approach of biology, became very controversial. Its media audience owed much to Christopher Zeeman (University of Warwick), who coined the term "catastrophe theory" and turned it into a very general methodology whereas Thom's objectives were more focused[4].

In 1969 I discussed with René Thom his applications of singularity theory to structuralism. According to what he told me, I was the first young mathematician of my generation to do it. These discussions filled me with enthusiasm. After having hesitated between pure mathematics (I then had a position at CNRS, Centre National de la Recherche Scientifique) and modelling, I accepted in 1971 a position at the Center of Analysis and Social Mathematics (CAMS) of the 6th section of the EPHE (École Pratique des Hautes Études), which would become later the EHESS - Ecole des Hautes Etudes en Sciences Sociales). I was recruited at the EPHE thanks to the support of Lévi-Strauss, Fernand Braudel, the director of the 6th section, who believed in the role of mathematics in the social sciences, and Charles Morazé, my former professor at the Ecole Polytechnique.

Having joined the EPHE, I naturally got in touch with some structuralists, including A.J. Greimas. Greimas made an announcement in his seminar and some young colleagues got interested. I thus met Jean-François Bordron, Frédéric Nef, Paolo Fabbri, Jean-Claude Coquet, Per Aage Brandt, François Rastier, Claude Chabrol, and later Jacques Fontanille, Ivan Darrault, Jean-Jacques Vincensini and several other semioticians. Greimas did not have a very strong institutional position and his disciples had rather difficult careers, but he compensated for this fragility by his exceptional dimension.

Thanks to Paolo Fabbri who put me in touch with Umberto Eco, I spent one year in Bologna where I wrote a part of my Habiltation thesis ("thèse d'état"). I spread Thom's work on structuralism in the international semiotics community, in Bologna of course, then within Per Aage Brandt's group in Aarhus in Denmark

[3] Morphogenesis studies the formation processes of complex forms, in particular those of life.

[4] Christopher Zeeman (born in 1925) founded the Department of Mathematics and the Research Centre in Mathematics of the University of Warwick in 1964. In 1969-1970, during a sabbatical year in Paris, he discovered René Thom's Catastroph theory. Next, he largely contributed to the notoriety of this theory by providing it with many applications in various fields, in particular in social and behavioural sciences (cf. Isnard C. A. & Zeeman E. C. (1976). Some models from catastrophe theory in the social sciences". In: Collins L. (ed.) *Use of Models in the Social Sciences*, Tavistock, London, pp. 44-100).

and in Toronto. I very early discussed with Jean-Pierre Desclés who worked at the time with Antoine Culioli. Thus, I became involved in a social sciences community where I could use my double competence in mathematics and semiotics.

Later on, I remained primarily at the CAMS. I defended my "thèse d'état" in 1982[5], and, until 1985, I remained focused on applications of Thom's morphodynamical models (*i*) to phonetics (*Les catastrophes de la parole. De Roman Jakobson à René Thom.* Maloine, Paris, 1985), (*ii*) to elementary structures in semiotics, (*iii*) and to theories of actantial syntax, in particular case grammars (*Morphogenèse du sens. Pour un schematisme de la structure.* PUF, Paris, 1985).

Afterwards, I became more and more interested in cognitive neurosciences. In 1986, I joined a team of cognitive sciences which had just been created by Daniel Andler in a lab of the Ecole Polytechnique, the CREA (Centre de Recherche en Epistémologie Appliquée) founded in 1982 and directed by Jean-Pierre Dupuy. A little later, Michel Imbert, a specialist in neurosciences, created the first DEA (Diplôme d'Etudes Avancées - a post-graduate diploma) of cognitive sciences, and I became actively involved there in this new context, which led me establishing footbridges with American, in particular Californian, cognitivism.

2. René Thom's contributions

Q.: You said you were initially interested in Thom's work in mathematics, in particular in his work on the theory of singularities you were working on. Then, you were interested in his work on structuralism in linguistics. Can you tell us about Thom's contributions in these fields? To start with, what is his theory of singularities?

Singularity theory aims to study, analyze and classify geometrical structures of a specific type, which are called "singular" because they are not "regular". One considers a class of objects for which (*i*) the opposition between local and global properties is meaningful, and (*ii*) there are standard "simple" objects whose structure is "trivial". Then, one calls "regular" the objects that are everywhere locally simple (or "locally trivial"), even if they can be globally very complex and not trivial at all. There exist singularities when, locally, the considered object is not regular.

For example, let us consider surfaces and define a regular point as a point where the surface admits a tangent plane. Let us take a cone: apart from the vertex there is a tangent plane at every point and the cone is thus locally regular. But the vertex does not have a tangent plane and is thus a singular point. And as

[5] *Pour un Schématisme de la Structure: de quelques implications sémiotiques de la théorie des catastrophes,* thèse d'État defended in 1982 at the École des Hautes Études en Sciences Sociales, Paris.

it is the only singular point in a small neighbourhood, it is said to be an isolated singularity.

One is immediately confronted with a theoretical problem: how to classify the singularities? One can observe for example that there exist points that are more or less singular. Consider for example a roof: apart from the ridge, points are regular. The points of the ridge are singular but are not isolated singularities since the whole ridge consists of singular points. The cone apex is more singular, it has a larger “degree” of singularity than the ridge of a roof.

One can thus establish a hierarchy of singularities and it is necessary to build a battery of theoretical concepts to analyze all the possibilities.

The interest of singularities — it was one of Thom’s great ideas — is that if all the local singularities of an object are known, then the object can be globally known qualitatively. The singularities concentrate the qualitative information on the objects. I give an example, introduced by Moebius in the 19th century, but known for centuries by sculptors. A good way of understanding a three-dimensional form is to cut it out in two-dimensional slices and to consider these successive slices. Take a torus (considered vertically) and cut it out in horizontal slices of increasing height (see fig. 1). If the cutting plane is too low, you do not meet the torus. At a certain time you meet the torus at a first singular point (a minimum). When you still go up, you get circles. You still go up and you meet another singular point (a saddle): the section has the form of an 8. The following level lines contain two circles, then again one 8, then only circle, and finally you reach the point at the apex of the torus (a maximum).

Conversely, if by cutting out a surface in slices you meet four singular points of this type (a minimum, two saddles, a maximum), then the surface is topologically a torus. Topologically these four singular points (with their type) characterize a closed surface with a central hole. The fact that the list of the singular points with their type define the object topologically is called Morse theorem.

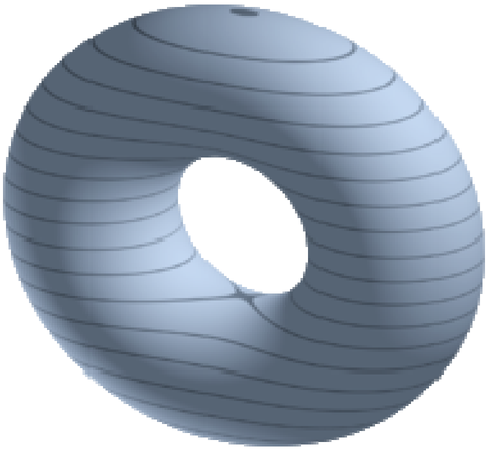


Figure 1. A torus with its level lines. Morse theorem
(source: http://fr.wikipedia.org/wiki/Théorie_de_Morse)

The classes of structures you can analyze with such methods are of a great diversity. You can look for example at what are called differentiable manifolds which generalize the concept of surface; you can also look at maps between spaces. In all these cases, you consider classes of objects and you want to study their possible non-trivial local properties.

The levels of structure can be very different from one another. You can consider topological objects (a very low level of structure but where the concept of continuity has a meaning nevertheless); or objects having more rigid properties. For example, if you take an orange peel and you try to crush it on a plane, as it is not elastic, it tears. This is a metric property: at the metric level, the sphere has curvature whereas the plane does not have any. Thom focused on the level of structure known as “differentiable”, which is intermediate between topological and metric levels and means that you can take as many derivatives as you want of the functions describing the objects.

Q: What are Thom’s contributions in linguistics?

To understand Thom’ contribution in the fields of semiotics and linguistics, it is necessary to come back to the specific notion of structure in linguistics, which concerns the mereological problem: how totalities can be organized with constituents, relations and transformation rules between constituents, and show an organization which is more than the sum of their constituents.

There are many fields where mereological structures can be encountered: grammatical rules and syntax (whatever the theories), but also, in psychology, visual perception spatial objects linked by spatial relations; in biology, the constituent structure of organs; in chemistry, the molecules where atoms are linked by their valence electrons, etc. These structures have been pinpointed for a long time, but, until recently, the adequate mathematics to model them in biology and in linguistics was completely missing (in chemistry it is only with quantum mechanics that they could be modelled).

This is why the issue concerning structures refers to formalization and modelling. In linguistics, the mathematical models used are multiple but generally rest on formal tools, that is algebraic, combinatorial and logical tools (Chomsky, Shaumyan, Montague, etc). In other fields, like visual perception and biological morphogenesis, structures are interpreted in a much more geometrical, topological and dynamical manner, as organized forms and Gestalts. The concept of structure is no longer algebraic and logico-combinatorial but morphological and dynamical, “morphodynamical” as I like to say.

The problem of a topological and morphodynamical mathematical theory for forms, primarily in biology and perception, is fundamental and extremely old. If we look back in history, there was a rather good theory in Aristote (cf. homeomeres and anhomeomeres in *The Parts of the Animals*) but it was completely eliminated by modern Galileo-Newtonian physics.

As André Robinet brilliantly showed, Leibniz was obsessed all his life by the antinomy thus created: one needs neo-Aristotelian concepts to work out a theory of form, but those seem to be incompatible with mechanist physics[6]. To overcome that antinomy, Kant had to write his third Critique, *The Critique of Judgment.* After him, many philosophers and scientists raised these issues. But these remained open until the 1960-70s when one suddenly saw flowering several radically new theoretical proposals: Thom and Zeeman with catastrophe theory, Ilya Prigogine with dissipative structures[7], Hermann Haken with synergetics[8], Henri Atlan with self-organization[9].

It is Thom who introduced the deepest mathematical tools. The only precedent had been, about fifteen years before, that of Turing who, just before his death, had been interested in morphogenesis and had introduced the first models explaining the emergence of forms and patterns in biochemical substrates using reaction-diffusion equations[10].

In the late 1960s, one thus started to have an idea of how the old problem of a theory of forms could be apprehended. Thom then introduced, in a very radical (and very controversial) way, the assumption that these morphodynamical tools could be transferred from biological morphogenesis to structural linguistics and semiotics. As a result, he found himself at the very heart of a linguistic debate which had a rich history: that of the opposition between gestaltic views (Guillaume, Tesnière, etc.) and formal views (Chomsky, etc.). Some linguists, like Hansjakob Seiler and Bernard Pottier, were enthusiastic. Others, like the Chomskyans, were more careful, even hostile.

I had the privilege to take part in the historical (and polemic) meeting between Jean Piaget and Noam Chomsky organized in 1975 at the Center of Royaumont by Massimo Piattelli-Palmarini, where I presented the principal differences between Chomsky and Thom[11].

[6] See André Robinet (1986) *Architectonique disjonctive, Automates systémiques et Idéalité transcendantale dans l'oeuvre de G. W. Leibniz,* Paris, Vrin. See also J. Petitot (1999). "Le troisième labyrinthe: dynamique des formes et architectonique disjonctive", *L'actualité de Leibniz: les deux labyrinthes* (D. Berlioz, F. Nef eds), *Studia Leibnitiana Supplementa*, 34, 617-632, Stuttgart, Franz Steiner.

[7] See for example Ilya Prigogine, Isabelle Stengers (1979) *La Nouvelle Alliance. Métamorphose de la science*, Paris, Gallimard.

[8] H. Haken (1981) *The Science of Structure: Synergetics* (Van Nostrand Reinhold).

[9] H. Atlan (1972/1992) *L'Organisation biologique et la Théorie de l'information*, Hermann,

[10] A. Turing (1952), “The chemical basis of morphogenesis”, *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, Vol. 237, No. 641, pp. 37-72. See also J. Petitot (2013) “Complexity and self-organization in Turing”, *The Legacy of A.M. Turing*, (E. Agazzi, ed.), Franco Angeli, Milano, 149-182. ArXiv: http://arxiv.org/abs/1502.05328v1. A. Lesne, P. Bourgine (eds.) (2006). *Morphogenèse. L’origine des formes*. Belin, Paris. Murray J.D. (2005) *Mathematical Biology*, Springer, New York.

[11] Petitot, J. (1979) “Hypothèse Localiste et Théorie des Catastrophes. Note sur le Débat”, *Théories du Langage, Théories de l'Apprentissage, le Débat Chomsky / Piaget*, (M. Piattelli-Palmarini, ed.), 516-524, Le Seuil, Paris.

Q. How did Thom become interested in linguistics?

Thom much admired Lucien Tesnière. He deeply regretted that he did not meet this great linguist in Strasbourg when he was a young researcher working there with Henri Cartan between 1947 and 1951. His interest focused on the way Tesnière conceived the dependence relations between the constituents of a sentence (still the mereological problem!) and developed an almost narratologic idea of the sentence as a "scene" making actants interact: "The verbal node [...] expresses a small drama by itself. "[12]

Thom was philosophically a realist in linguistics. He estimated that, below the great variability and complexity of the morphosyntaxic surface structures, the universals of language result from evolution and are rooted in the cognitive abilities of primates, in particular in perception and action. Consequently, he tackled the linguistic problems from the point of view of the biological evolution of cognitive structures.

Q. Which other researchers besides you were interested in Thom's work on linguistics?

Among the linguists and semioticians who very early were deeply interested in Thom, one can quote, besides masters like Jakobson, Seiler and Pottier[13], two researchers of my generation: Wolfgang Wildgen[14] of the University of Bremen in Germany and Per Aage Brandt[15] of the University of Aarhus in Denmark. Their work developed, like mine, in the 1970-80s.

Then, in a completely independent way, without any reference to the European debate, something relatively similar happened in the United States in the 1980-90s with the emergence of West Coast cognitive linguistics: Charles Fillmore and George Lakoff at Berkeley, Len Talmy at Berkeley then at Buffalo, Ron Langacker at San Diego (where Gilles Fauconnier was, too). These linguists developed approaches which, on the one hand, were very structural (although with few references to European structuralism) and, on the other hand, explicitly supported the same theses on the evolutionary origin of language in relation to perception and action. At the time of the conference on Tesnière organized in 1992 in Rouen by Françoise Madray-Lesigne (Tesnière was born in 1893 in

[12] L. Tesnière (1959), *Éléments de syntaxe structurale*, Klincksieck, Paris, 1959 (2ème éd. 1988), 48, 1. Voir aussi J. Petitot (1985) *Morphogenèse du Sens. Pour un Schématisme de la Structure*, PUF, Paris.

[13] See for example B. Pottier (2000) *Représentations mentales et catégorisations linguistiques*, Paris, Louvain, Peeters.

[14] cf. W. Wildgen (1982) *Catastrophe Theoretic Semantics. An Elaboration and Application of René Thom's Theory*, Benjamins, Amsterdam, and (1999) *De la grammaire au discours. Une approche morphodynamique*, Peter Lang, Bern.

[15] cf. Per Aage Brandt (1994) *Dynamiques du sens*, Aarhus University Press, Aarhus, and (1995) *Morphologies of Meaning*, Aarhus University Press, Aarhus.

Mount-Saint-Aignan close to Rouen), Langacker made an emphatic praise of Tesnière by regarding him as one of his precursors[16] .

In addition, connectionist models of neural networks, opposed to formal models, were developing powerfully and rapidly. One of the linguists more in sight in that field was Paul Smolensky whose models raised a violent debate with Jerry Fodor and Zenon Pylyshyn. I took part in that debate[17].

Wildgen, Brandt and myself made contact with these new trends. We organized meetings, for example in San Marino, at Umberto Eco and Patrizia Violi's "International Center for Semiotic and Cognitive Studies", a conference with Len Talmy, and also two important conferences at the CREA on the issue of constituent structures in connectionist models. These conferences were relayed by another one, organized this time at Bloomington by Tim van Gelder and Bob Port, entitled "Mind as Motion"[18].

Q. How were you informed of the works existing in the United States?

I was much interested in Fillmore — Case linguistics, with Tesnière's structural syntax, is the closest to Thom's views —, Langacker, Jackendoff, Lakoff. But the one that made contact with the most interesting linguist for us, namely Len Talmy, was Per Aage Brandt. Len achieved a splendid work on linguistic Gestalt while showing empirically, on a large corpus of data, the existence of very close connections between syntax (more precisely deep "syncategorematic" structures) perception and action. He went much further than case markers like prepositions[19].

Q. Which modelling for linguistics?

Once the empirical regularities are described, one asks how to model them I regarded the works by Fillmore, Langacker, Talmy, etc. as well supported results; I trusted them, and what interested me was to see how the syntactico-semantic structures they had identified could be modelled.

[16] cf. Langacker, R.W. (1995) "Structural Syntax: The View from Cognitive Grammar ", *Lucien Tesnière aujourd'hui,* (F. Madray-Lesigne and J. Richard-Zappella, eds.), Actes du Colloque international CNRS, Université de Rouen 16-18 novembre 1992, Louvain/Paris, 13-39.

[17] See for example P. Smolensky (1988). "On the Proper Treatment of Connectionism", *The Behavioral and Brain Sciences*, 11, (1988), 1-74. J. Fodor, Z. Pylyshyn (1988) "Connectionism and cognitive architecture: a critical analysis", *Cognition*, 28, 1-2 (1988) 3-71. J. Petitot (2011) *Cognitive Morphodynamics. Dynamical Morphological Models of Constituency in Perception and Syntax* (with R. Doursat), Peter Lang, Bern.

[18] T. van Gelder, R. Port (eds.) 1995, *Mind as Motion*, Cambridge, MIT Press.

[19] Len Talmy (2000). *Toward a Cognitive Semantics*, Vol. I: *Concept Structuring Systems*, Vol. II: *Typology and Process in Concept Structuring*, Cambridge, MA, MIT Press, 2000.

For this purpose, I applied a general methodological principle. It is not because the structures under scrutiny are of a linguistic nature that the good tool, *a priori*, is formal languages. In sciences, one should not make any assumption on the fact that the mathematical tools should be of the same order as the objects. I am not anti-formalist *a priori*. I am ready to admit that formal models can prove to be the best in some cases. But I do not see any *a priori* reason that formal languages should constitute good tools to understand syntactic structures of natural languages, no more than to understand perceptive, biological or molecular mereological structures. My methodological principle is: the structures of natural languages are natural phenomena (I underline “natural”) and, as in other sciences, it is necessary to invent (I underline “to invent”), starting from appropriate bases, the suitable mathematical tools to model them.

These appropriate bases are two-fold. On the one hand, they come from properly linguistic studies and on the other hand from other disciplines like cognitive sciences and neurosciences. Of course, no need to make brain imagery to explain the conditional mood in French, but one must use neurocognitive results on universal sensorimotor schemes of interaction between actants to understand verbal valence.

III. Phonetics, phonology and catastrophe theory

Q. You were interested in phonetics.

In certain cases, non-formal mathematical models of a topological-geometrical-dynamical type have proved to be rather good. The first example which convinced me of the validity of morphodynamical models in the field of language was phonetics. As you know, structuralism comes mainly from the phonological work of the Moscow and Prague Circles, and, when I began to work with Thom, I already knew structural phonology a little and the remarkable results that Jakobson transfered to general structuralism, in particular in collaboration with Claude Lévi-Strauss. I thus tried to test Thom’s models on it and I discovered (I consider that it was my first scientific “discovery”) that they were completely adapted to phonology.

It happens that at the EHESS there was (and there still is) a very good laboratory of cognitive psychology[20] some members of which worked in phonetics. Following their advice, I read many things in this field and I noted that Thom’s models were not only relevant but that they were quantitatively and qualitatively exact.

In phonetic perception, several levels should be distinguished: the acoustic level, the peripheral level of sensory transduction, the perceptive level and the linguistic level. At one end of the chain, one can make a lot of acoustico-physical experiments and at the other end one has at one's disposal a very important linguistic corpus of thousands of languages.

[20] The LSCP (Laboratoire de Sciences Cognitives et Psycholinguistique).

One characteristic of phonetic perception is what is called its “categorical” character. What does that means? When one makes a sonogram one can identify the “formants” of the sounds produced by the vocal tract. Sounds produced by vocal cords are very rich in harmonics, and the articulatory controls control the shape of the resonators of the vocal tract. Each of these resonators (mouth, nose …) amplify or damp specific harmonics. In other words, the amplitudes of the harmonics are modulated by a continuous curve having strongly marked peaks. These resonance peaks select frequency bands which are called formants. Vowels are stationary sounds having characteristic formants, and consonants are transient sounds carrying out transitions between formants and possibly introducing turbulence (plosives, fricatives).

When you look at the equations, you observe that the formants correspond to the maxima of what is called “the transfer function” (the output/input ratio) of the vocal tract. In fact this function H is the inverse of a function G and the maxima of H correspond to the minima of G.

One can simplify the problem by preserving only a few resonators, for example three: the front cavity (mouth), the back cavity (pharynx) and the nasal cavity. Each cavity is described by a tube (with length and diameter) and constrictions of the vocal tract are described by small intermediate tubes. One knows how to explicitly compute the way in which the formants depend on these articulatory parameters. Personnally, I used as a guide the classic *Preliminaries to Speech Analysis* by Jakobson, Fant and Halle[21]. From this audio-acoustic base, structuralist works, in particular Jakobson’s, show how phonological (linguistically relevant) distinctive features can be recovered using a qualitative description of the formant configurations. For example if one considers the universal vocalic triangle /*a*/, /*i*/, /*u*/ in simple models with two formants:

/*a*/ corresponds to close formants of medium frequency (feature “compact”),

/*i*/ corresponds to well separated formants (feature “diffuse”) with predominance of the “acute” formant (high frequencies),

/*u*/ corresponds to well separated formants (feature “diffuse”) with predominance of the “bass” formant (low frequencies).

If more detailed models are used, one can still qualitatively describe phonological distinctive features in this way by using not the true formants quantitatively defined, but “formantial masses” as Ludmilla Chistovich proposed a long time ago.

Then, I looked at the explicit formulas connecting the formants to articulatory controls and I discovered that, for the models with tubes, the function G exactly is an unfolding of singularity in Thom’s sense and that the formants and their configurations are consequently describable in terms of catastrophes:

[21] Jakobson, R., Fant, G. & Halle, M. (1952) *Preliminaries to Speech Analysis. The distinctive features and their correlates,* MIT Technical Report.

there are abrupt — discontinuous — changes in formantial masses according to continuous changes in articulatory controls[22].

Let us be a little more technical. The transfer function is a function *H(s)* of a complex variable $s = \sigma + i\omega$ where $\omega/2\pi$ is the frequency and σ a damping factor. The restriction of $H(s)$ to the imaginary axis ω gives the modulation of the harmonics frequencies. Let us consider the model with one resonator (i.e. two tubes and one formant) of figure 2.

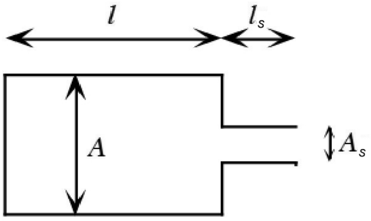


Figure 2. Model with two tubes

One obtains (for conveniently chosen values of l, l_s and A_s) the already complex formula:

$$H(s) = \frac{1}{LC\,s^2 + (RC + GL)\,s + GR + 1} = \frac{{\omega_0}^2}{(s - s_1)\,(s - s_2)}$$

with

$$L = {}^{\rho l_s}\!/_{A}$$
$$C = {}^{lA}\!/_{\rho c^2}$$
$$R = \frac{S}{A^2}\sqrt{\frac{\omega\rho\mu}{2}}(l + l_s)$$
$$G = S\frac{\eta - 1}{\rho c^2}\sqrt{\frac{\lambda\omega}{2c_p\rho}}(l + l_s)$$
,

where A = section of the open tube, S = circumference of diameter A, ρ = density of the air, c = speed of sound, μ = coefficient of viscosity, λ = coefficient of conduction of the heat, η = adiabatic constant, c_p = specific heat of air under constant pressure. The poles of *H(s)* are

[22] See J. Petitot (1985) *Les Catastrophes de la Parole. De Roman Jakobson à René Thom*, Maloine, Paris; and (1997) “Modèles morphodynamiques de catégorisations phonétiques”, *The Roman Jakobson Centennial Symposium* (P.A. Brandt, F. Gregersen eds), *Acta Linguistica Hafniensia*,29,239-269. http://jeanpetitot.com/ArticlesPDF/Petitot_Jakobson.pdf

$$s_1 = -\frac{1}{2}\left(\frac{R}{L}+\frac{G}{C}\right)+i\sqrt{\frac{GR+1}{LC}-\frac{1}{4}\left(\frac{R}{L}+\frac{G}{C}\right)^2}$$
$$s_1 = \sigma_1 + i\omega_1$$
$$s_2 = \bar{s}_1 = \sigma_1 - i\omega_1$$

$$\omega_1 = \sqrt{{\omega_{01}}^2 - {\sigma_1}^2}\ ,\quad {\omega_{01}}^2 = \frac{GR+1}{LC}$$

Only s_1 is relevant because its imaginary part is > 0 and frequency must be > 0.

For a model with two resonators (four tubes and two formants), one obtains the formula:

$$H(s) = \frac{\omega_1\,\omega_2}{s^4 + as^3 + bs^2 + cs + d}$$

with

$$\omega_1 = \frac{1}{\sqrt{L_1\,C_1}}\ ,\ \omega_2 = \frac{1}{\sqrt{L_2\,C_2}}$$

$$a = \frac{R_1}{L_1}+\frac{R_2}{L_2}+\frac{G_1}{C_1}+\frac{G_2}{C_2}$$

$$b = \frac{1+R_1G_1}{L_1\,C_1}+\frac{1+R_2G_2}{L_2\,C_2}+\frac{R_1G_2}{L_1\,C_2}+\frac{R_2G_1}{L_2\,C_1}+\frac{G_1G_2}{C_1C_2}+\frac{R_1R_2}{L_1\,L_2}+\frac{1}{L_2\,C_1}$$

$$c = \frac{\left(R_1(1+R_2G_2)\right)}{L_1L_2\,C_2}+\frac{\left(G_2(1+R_1G_1)\right)}{L_1C_1\,C_2}+\frac{R_1+R_2+R_1R_2G_1}{L_1L_2\,C_1}+\frac{G_1+G_2+G_1G_2R_2}{L_2C_1\,C_2}$$

$$d = \frac{\left(1+R_1G_1+R_2G_2+R_1G_2+R_1R_2G_1G_2\right)}{L_1C_1L_2\,C_2}$$

Figure 3 shows the graph of the logarithm of the module of *H*(*s*) and the damping of the formants.

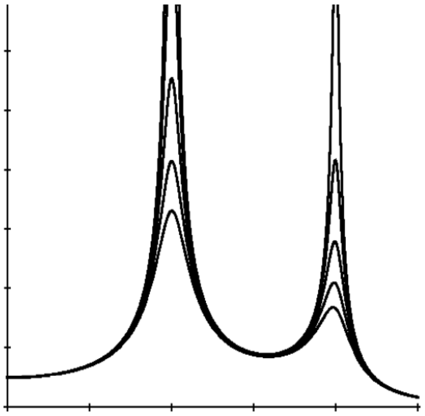


Figure 3. Damping of the two formants

The key point is that, for *n* formants, (*i*) the denominator of the transfer function *H*(*s*) is the universal unfolding of the singularity known as A^{2n}

$$A^{2n}(s) = s^{2n} + a_{2n-1}\, s^{2n-1} + \quad + a_1\, s + a_0$$

and (*ii*) the coefficients of this unfolding are complicated functions of the 2*n* articulatory controls.

To move from formants to formantial masses, and thus from quantitative to quailtative, one introduces an "auditive transformation" which merges the sufficiently close formants. Then one really obtains models in the sense of Thom.

In short, one can, thanks to Thom's models, explicitly move from the acoustic level (physical) to the auditive level (sensorial) then to the phonological level (linguistic), the key being the interpretation of formantial masses in terms of unfoldings of singularities parameterized by articulatory controls.

Let us address now the issue of categorical perception[23]. In addition to articulatory controls determining the shape of the vocal tract, there exist other acoustic cues that one can vary in a continuous manner, e.g. voicing (VOT: voice onset time) that measures the moment of excitation of the fundamental harmonic. In short, while one can vary many parameters continuously, perception does not vary continuously. It is the fundamental reason why some sounds can be the substrate of a phonological code. For example, you can vary voicing in order to move from [b] (voiced labial) to [p] (not-voiced labial). But at the perceptive level, on the other hand, you perceive only allophones of /*b*/ or /*p*/ and no intermediate state.

To explain this remarkable phenomenon, psychologists distinguish two fundamental mechanisms. On the one hand, discrimination: can I discriminate two close [b]; and on the other hand, identification: do I identify a /*b*/ or a /*p*/, i.e. a sound as an allophone of a phoneme or of another one.

For colours, discrimination corresponds to shades and identification corresponds to colour names. The perception of colours is "continuous" in the sense that shade discrimination depends very little on colour identification: one perceives gradual shades independently of the existence of the categories of colours. In categorical perception, the situation is quite different: discrimination degenerates inside the categories; as one says, it is subordinated to identification: one discriminates two close sounds only if they are identified as different. One is unable to discriminate two close [b] sounds identified as allophones of /*b*/, but on the other hand one is able to discriminate two close intermediate sounds if one is identified as an allophone of /*b*/ and the other as an allophone of /*p*/.

It is a little as in geography: there are areas delimited by boundaries (the domains of the parameter space corresponding to a single phoneme), inside an

[23] Among a rich bibliography, three references were important for me: Liberman, A. M., Cooper, F. S., Shankweiler, D. P., Studdert-Kennedy, M., (1967) Perception of the Speech Code, *Psychological Review*, 74, 6, 431-461. Stevens, K., (1972) The Quantal Nature of Speech, *Human Communication, a Unified View* (P. B. Denes, E. E. David Jr. eds.). Malmberg, B., (1974) *Manuel de phonétique générale,* Paris, Picard.

area the various positions (allophones) have an equivalent type (they are tokens of the same phoneme: no intra categorial discrimination), but at the boundary crossing, the type (the corresponding phoneme) changes abruptly. In categorical perception, there exist thresholds between categories which are induced by perception itself. That is due to the fact that percepts vary in an extremely non-linear way compared to their audio-acoustic and articulatory controls. Kenneth Stevens well studied this phenomenon in his article quoted above "On the quantal nature of speech".

Categorical perception is a fundamental property of phonetic perception and, I repeat, explains why and how some sounds can become the substrate of a code.

Catastrophe theory is particularly well adapted to the modelling of categorical perception because its general model rests on the concept of "bifurcation". A bifurcation occurs in a system when a small change of a continuous control produces a qualitative jump of the internal state of the system, in other words when a small variation of causes involves great differences on effects. It is precisely what occurs with categorical perception when, for example, a small articulatory change qualitatively moves the configureation of formantial masses from a single formantial mass to two formantial masses ("compact/ diffuse" opposition).

I thus showed that the catastrophes related to audio-acoustic equations match the phonological structures observed in languages. For phonetics, Thom's models are thus valid models. Jean-Luc Schwartz and the Grenoble group, among them Christian Abry and Louis-Jean Boë, went more thoroughly into them[24]. In particular they identified the "auditive transform" as a mechanism of "large scale spectral integration".

To summarize, catastrophe models help understanding in a detailed way the link between audio-acoustics, psycho-physics, perception and structural phonology.

Q. What exactly is a catastrophe model?

A catastrophe model starts with a system which has internal states. For instance, in the case of phonetic perception, there are neuronal states corresponding to percepts. In the case of a chimical element such as water, the thermodynamical states are called "phases": solid, liquid, gas. These internal states are attractors of the internal dynamics of the system and the transient states, induced by the inputs of the system, are rapidly stabilizing toward them: for instance, an acoustic input turns into a perceptive state (after having gone through the external ear, the cochlea, the auditory cortex). Moreover, the system is controlled by external parameters (articulatory parameters, acoustic cues, temperature, pressure…). When these controls change, the inner states change in turn, and there is two

[24] See for example Abry, C., Boë, L.-J., Schwartz, J.-L. (1989). Plateaux, catastrophes and the structuring of vowel systems. *Journal of Phonetics* 17, 47-54. Schwartz, J.-L., Boë, L.-J., Vallée, N., Abry C. (1997). The dispersion-focalization theory of vowel systems. *Journal of Phonetics* 25, 255-286.

possible outcomes. Either a small change in the controls is without consequences and does not change the inner state qualitatively (e.g. the sound is still perceived as a /*b*/, the water temperature shifts from 50°C to 51°C), or a small change in the controls changes the qualitative type of the inner state (the sound is now perceived as a /*p*/, the water temperature shifts from 99°C to 100°C and the water starts boiling). Such qualitative changes are called "bifurcations". In thermodynamics they are called "phase transitions".

Q. Can a shift from /*b*/ to /*p*/ be predicted, in the sense that it followssome constraints?

There are strong differences across languages. But one can assume a universal innate "initial state" for new born humans. During language acquisition, some thresholds move, others split, others disappear. For instance, in the case of the Japanese language, the threshold between [r] and [l] disappears and [r] and [l] become two allophones of a single phoneme. Young Japanese are able to discriminate between [r] and [l] but, while learning the language, these discriminations disappear due to the categorical property of perception.

Thus, maybe there is an initial phonological "geography" that evolves with the learning of a specific phonological system. All the phonological systems categorize the same space of sounds defined by anatomically possible articulatory controls and harmonics. The question is to know whether the categorysations of actual languages are efficient and whether they reach an optimum of the quantity of information conveyed by the phonological code.

Q. Is there a definition of what is an optimum vocalic system for communication?

This is a fascinating question. We know of a great number of phonological systems and they may be grouped into classes. Numerous models have been proposed. The problem is the following: within the space of the possible sounds, defined by universal anatomic constraints, we have to find the best categorisation into sub-regions (the phonemes). There are several strategies in order to solve the problem of the optimisation of the categorisation and there are several studies showing how these strategies are related to each other. All the phonological systems are based on the universal vocalic triangle /*a*/, /*i*/, /*u*/, and can be described as a progressive refinement complexifying that triangle, leading eventually to the most complicated vocalic systems, such as that of French.

Q. We mentioned "basins of attraction". Could you explain that notion more thoroughly?

When you have a dynamics defined on a given space *M*, all the points *x* belonging to *M* have a trajectory $\gamma(x)$ and you can consider the asymptotic behaviour of that trajectory. Generally, $\gamma(x)$ is attracted by an attractor *A* which is a sub-set topologically closed and dynamically invariant, minimal for these two properties, and which attracts all the trajectories coming from points in its neighbourhood. All the points *x* whose trajectory $\gamma(x)$ are asymptotically attracted by *A* constitute the basin of attraction *B*(*A*) of *A*. Thus the dynamics decomposes *M* in several basins of attraction separated by boundaries (some of them can be very complicated).

In the case of the internal dynamics of a system, every input puts the system into an initial state *x* and, generally, *x* is not on the boundary but inside a basin of attraction *B*(*A*) and is therefore attracted by the attractor *A*. This means that the initial transient state *x* will be attracted by the internal state *A*. This projection of the input on an attractor models the process of "identification". In this type of models for categorisation, the basins of attraction are the categories and the attractors are the prototypes. An input induces an initial state that is associated with a prototype. In the phonetic domain, a sound is recognized and identified as an allophone of a given phoneme.

Some boundaries between basins of attraction are more complex than others. When an initial state is on a boundary separating two basins as a sort of ridge-line, it is possible to fall into one or the other of the basins. And, last but not least, the control space allows the modification of the basins of attraction and their boundaries.

That is why Thom proposed to distinguish two kinds of bifurcation: (*i*) internal bifurcations, when the system moves from a basin of attraction to another in the internal space *M* and (*ii*) external bifurcations, when the system is coerced into another attractor by the effects of the controls, for instance due to the fact that an attractor disappeared or that two attractors have merged. In practice, external bifurcations matter most: the systems are in general of the "slow/fast" kind, which means that the internal dynamic is fast, while the variation of the controls is slow and then it is possible to do as if the system were always in an internal stable state (on an attractor). It is called an "adiabaticity hypothesis". What chiefly matters then is the bifurcation of the attractors and not the fast internal transient trajectories.

Q. From a mathematical point of view, which kinds of mathematics are involved?

When building the mathematics for his models, Thom chose – for the spaces, the functions and the maps between spaces he needed – the level called differentiable, that is the level where the objects have locally almost everywhere well-defined derivatives, except sometimes in some singular points. This level is more constrained than the basic continuous level (he does not allow objects such as fractals). However, it is far less constrained than the algebraic or metric level. The differentiable objects are very "flexible".

Thom introduced two kinds of models: the elementary models, and the extended ones. In the first ones, the internal dynamics is the steepest descent of an energy potential function *f*(*x*) defined in the inner space *M*: the system optimizes its state by minimizing its internal energy. The attractors (the internal states) are in that case the minima of $f(x)$: an initial state goes (according to the specificities of the system) either to the closest minimum or to the absolute minimum.

The control parameters allow the variation of the potential functions, and therefore the change of the minima and their height. A minimum may then disappear and the system will have to go into another minimum: these are bifurcations.

One of Thom great achievements has been the (difficult) proof of the classification theorem for elementary catastrophes. The main idea is the following: if you consider a potential function f where several minima, maxima or saddles (called "critical points") are merged in a single point x, f has an unstable singularity at x (according to a natural notion of stability). If you deform such a singularity through external small parameters w embedding f into families $f_w(x)$ defined in a small neighbourhood of x with $f_0(x) = f(x)$, it is possible to stabilize the singularity in many ways, partially or totally, through the dissociation of the critical points that have been merged. The key result is that, given such a singularity, there exists a universal deformation, called "universal unfolding", that gathers optimally all the possible stabilisations.

Figure 4 shows the catastrophe named "cusp" that plays the key role in modelling the universal vocalic triangle. The unstable singularity x^4 merges two simple minima (non-degenerated minima, i.e. that are not composed by merging simpler critical points) and one simple maximum. The external space W of the universal unfolding is two dimensional. It is partitioned into three regions by a catastrophe set K, containing the two branches K_b of a cusp curve and the median half-line K_c. Along the branches K_b one simple minimum remains simple while the other simple minimum and the simple maximum merge into an inflection point: the K_b are lines of catastrophes of bifurcation. Along K_c, the two simple minima and the simple maximum remain simple but the two minima have now the same height: K_c is a line of catastrophes of conflict. Apart from K, $f_w(x)$ have either a simple minimum, either two simple minima separated by the simple maximum with one of the minima that dominates the other.

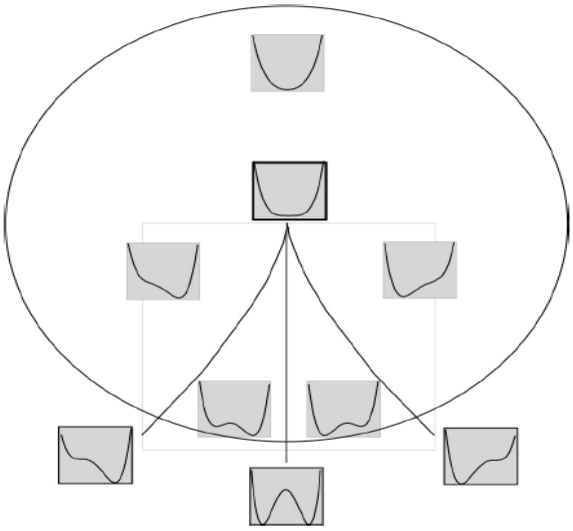


Figure 4. The universal unfolding of the "cusp" catastrophe

The classification theorem says that, whatever the system under scrutiny, if there are one or two internal dimensions and no more than four external controls *w*, and if the potential function *f*(*x*) of the system has unstable singularities, and if its unfolding is structurally stable (that is the process stabilizing internal unstabilities is itself stable), then these singularities belong to a finite list: “cusp”, “swallowtail”, “butterfly” in the one dimensional case, elliptic, hyperbolic or parabolic “umbilic” in the two dimensional case.

From a methodological point of view, this result is very important because it exhausts the field of possibilities. It is as important as the theorem of classification for the platonic solids (the finite sub-groups of the group of the rotations) [25].

III. Language and perception

Q. What are the consequences for linguistics?

As I said Thom was interested in a realist approach to language. For him, language had an evolutionary origin. The ability to describe perceptive scenes of the outer world communicate them to those that do not see them was, for him, a fundamental requisite constraining natural languages. It appears that a large part of these perceptive scenes are interactions between “actants”[26] (being either agents or objects), and the transformations of their spatial relations can be described by verbs (to go into a place, to seize something, to attack, to run away, etc.).

The basic assumption is that the structure of the sentences describing a perceptive scene with verbs and actants situated in space and time is a result of the evolution pressure and that there is an analogy between the constituent structure (mereology) of actantial syntax and the constituent structure of the perceptive scenes.

This assumption of a foundation of the actantial structures in the structures of perception and action has a long history. One of its components is the “localist hypothesis”, which has been supported under various guises by linguists such as Anderson, Langacker or Talmy, and according to which the basic syntactic structures of elementary sentences categorize the generic interactions in space and time. Structural syntaxes like Tesnière’s and case grammars like Fillmore’s belong to the same paradigm. All these theories rely on an actantial theory using semantic roles defined by spatio-temporal schemas similar to schemas of perception and action and therefore rooted in cognitive evolution.

In this context, Thom's theorem is of primary importance: it is possible to classify the actantial spatio-temporal interactions thanks to the classification of

[25] Readers interested in the theory of mathematics may wish to read (in French): Chenciner, A., (1980). “Singularités des fonctions différentiables”, *Encyclopædia Universalis*, Paris; as well as the compilation I made in 1982 (in French): *Eléments de théorie des singularités,*
http://jeanpetitot.com/ArticlesPDF/Petitot_Sing.pdf

[26] We use the term “actant” analogue to what are called “semantic roles” in case grammars. It is a deeper concept than that of “actor” or “character”.

elementary catastrophes. This theorem proves the existence of case universals. It is obviously a fundamental result.

These hypotheses have been controversial. Numerous linguists objected that language is independent from perception and that it is a cognitive faculty *sui generis*. Chomsky for instance argues for the notion of autonomy of syntax. Other linguists acknowledge the existence of links with perception but argue that the linguistic categories of perception cannot be extracted from the perception itself. And, in any case, these hypotheses are but of low interest for linguists describing actual natural languages since they pertain to a deep "proto-linguistic" level, far below from the morphosyntactic diversity of natural languages. However, these hypotheses have a great theoretical significance for building bridges between linguistics and cognitive neurosciences. They have also a great technological relevance, in order to build robots able to convert natural language instructions in terms of perceptual structures and motor programs.

Q. Could you elaborate upon the relation between localist hypothesis and catastrophe theory?

If one try to schematize perceptual scenes and actantial relations (schematization is a strong simplification focusing only on the essential forms), one encounters again structures that are derived from elementary catastrophe.

Let's take objects distributed in space, that is to say static configurations of spatial actants. Let us add a temporal evolution that changes this configuration dynamically. These temporal evolutions generally lead the actants to interact. It is then possible to represent the actants through minima of a potential function (here is the schematization) so that to turn the interactions into bifurcations and so that it is possible to apply the models of elementary catastrophes. I explain this in details in *Cognitive Morphodynamics*[27]. A schema such as "to take an object" is the fact that there is an actant and an object which are initially disjoint and which, later, are conjoint. The verbal node lexicalized by the verb "to take" describes this interaction, which is a bifurcation derivable from the cusp catastrophe. As soon as early 1970s, Thom made the list of the "archetypal actantial graphs" that are derivable from elementary catastrophes[28].

Later on, Thom's archetypes have turned out to be great precursors of several cognitive models of language: Fillmore's frames, Langacker, Talmy, Lakoff's image-schemas, Haiman's "iconicity in syntax", Desclés' "cognitive archetypes", Shank and Abelson's scripts, etc. (for more details see *Cognitive Morphodynamics*[29]).

[27] Peter Lang, Bern, 2011.

[28] See for instance "Topologie et linguistique", *Essays on topology and related topics*, A. Haefliger and R. Narasimhan (eds), Springer, 1970, 226-248. Reprinted in: *Modèles mathématiques de la morphogenèse*, Paris, 10-18 UGE, 1974.

[29] See for instance Fillmore, C., (1976) "Frame semantics and the nature of language", In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*. Volume 280, 20-32. Haiman, J., (ed.) (1985) *Iconicity in Syntax*, Amsterdam, J. Benjamins. J.-P. Desclés (1990) *Langages*

Using the fact that elementary verbal nodes grammaticalize bifurcations of actantial relations, you can build a theory of verbal valency. It was one of the results that mattered the most for Thom. All the linguists that have been interested in verbal valency know that there is a limit of 4 actants (the few controversial cases with 5 actants use indeed a double actant). Where does this limit come from? Why could not we be able to create new semantic roles allowing for an increase of the valency? According to Thom, it is one of the strongest evidences of the rooting of actantial syntax into perception and action. Indeed, perception and action take place in a 4-dimensional space-time, and archetypal actantial graphs derive from elementary catastrophes whose external space have 4 dimensions at most. This closed list of catastrophes, drawn from the classification theorem, puts a drastic limit on the complexity of the bifurcations and, then, on the verbal valency. We then observe that, in all archetypes, the valency has a limit of four[30]. According to Thom, this constraint comes from our outer world.

Of course, several linguists objected that, in most of the verbs denoting action, there is an agentivity, and that agents are generally intentional agents. However, numerous remarkable experiments have shown how strongly agentivity itself is deeply rooted in perception and action. As early as the 1940s, F. Heider and M. Simmel have shown that movements, even complex ones (such as movements including accelerations, decelerations, changes of direction, etc.) of simple forms (triangles, circles, rectangles of various sizes) were spontaneously described by way of intentional action verbs (“come in”, “come out”, “give”, but also “hide”, “escape”, “hunt”, “attack”, “force”, etc. [31]). Since these pioneering experiments, numerous works were devoted to such phenomena. Let us mention for instance J. Scholl and P. D. Tremoulet on the perception of causality and the animacy of objects; D. Premack on the perception of intentional movements by children; S. J. Blakemore and J. Decety on the comprehension of intentions; M. E. Zibetti on the fact that we interpret as if the movements we perceived were caused by intentional agents [32]. All these works try to unveil the evolutionary and cognitive roots of the tendency we have to interpret purely cinematical and dynamical motions as if they resulted from an intentional agentivity. All these authors showed that this tendency is automatic, non-conceptual, “hardwired” and “rooted in automatic visual processing”. Their works offered a general confirmation of Thom’s theses.

applicatifs, Langues naturelles et Cognition, Paris, Hermès. Schank, R., Abelson, R.P. (1977) *Scripts, Plans, Goals and Understanding*, Hillsdale, Lawrence Erlbaum.

[30] With 4 as the number of dimensions d of space-time and with 4 as the maximum valency $v(d)$, we have $v(4) = 4$. But it is not the case that $v(d) = d$ generally.

[31] Heider, F., Simmel, M. (1944) “An experimental study of apparent behavior”, *American Journal of Psychology,* 57 (1944) 243-259.

[32] See for example Scholl, B.J., Tremoulet, P.D. (2000) “Perceptual causality and animacy”, *Trends in Cognitive Science* 4(8), (2000), 299-309. Blakemore, S. J., Decety, J. (2001) “From the perception of the action to the understanding of intention”, *Nature Reviews Neuroscience* 2, (2001), 561-567. Zibetti, E., Tijus, C. (2003) “Perceiving Action from Static Images: the Role of Spatial Context”, *CONTEXT* (2003) 397-410.

Q. How to build a model of the perceptive scene?

This is an old and interesting issue. In order to move from perception to language, it is necessary to describe the perceptive scenes very schematically in order to specify the relevant information that should be translated into elementary proto-sentences. But how to simplify the perceptive scenes in a bottom-up and data-driven way, and how to define the linguistically relevant information that lays inside?

For instance, let us consider a preposition as "across" (this example is drawn from my book *Cognitive Morphodynamics*). This preposition can be applied to a huge diversity of perceptive scenes. "Across" can be applied to the crossing of a street, a lake, a field, a country, to walking haphazardly into a forest, etc. How would it be possible to define the geometric and topological invariant content of "across"? Obviously, that invariant pertains to the notion of "transversality", but how could we design algorithms extracting such a schema from a real and complex scene? The problem is to find good tools for simplifying the scene, tools strong enough for skeletonizing it, but also able to preserve the relevant information ("transversality"). In *Cognitive morphodynamics*, I showed, together with René Doursat, that some morphological algorithms implemented in neural networks can do the job[33].

Figure 5 schematizes the clause "zigzagging across the woods": (*a*) is the input image; it contains two objects: the path and the wood. (*b*) and (*c*) result from a first preprocessing using morphological algorithms of dilation and skeletonization. (*d*) and (*e*) result from a second preprocessing; in (*e*) the intersection of skeletons *sk*(*A*) and *sk*(*B*) allows one to extract the invariant schema of "transversality".

[33] These algorithms are discrete variants of dynamic processes of diffusion and skeletonization often used in morphodymics. They are computationally very efficient. They have been developed by G. Matheron and J. Serra and their colleagues. See for instance: Serra, J. (1982). *Image Analysis and Mathematical Morphology*, New York, Academic Press, 1982.

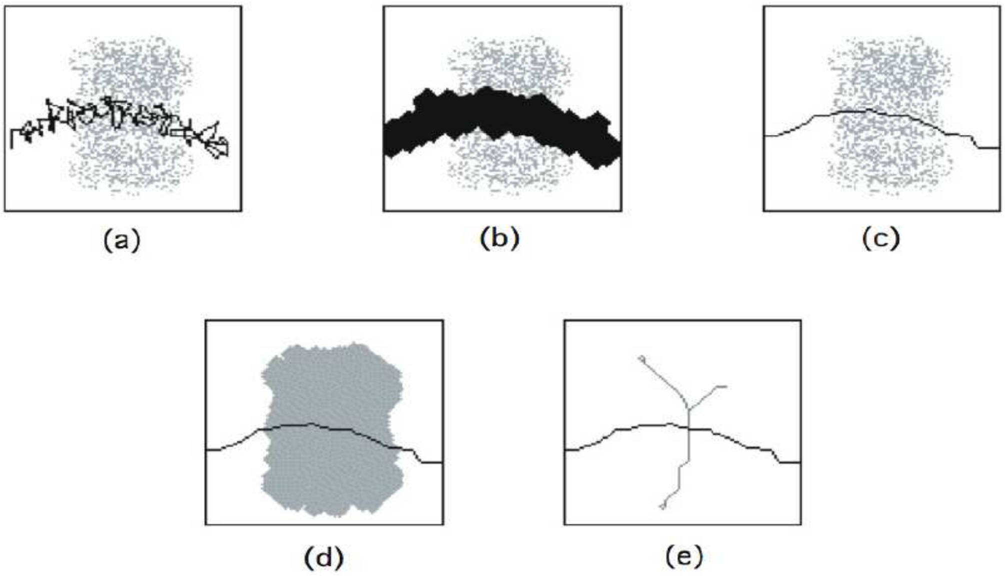


Figure 5. The extraction of the invariant content “transversality” in a pixelized image described using the preposition “across”.

I stress the fact that these morphological analyses of images are not obvious at all and that our capacity to easily and correctly apply prepositions to visual scenes is far from being understood. This issue is far most complex than that, already quite difficult, of hand writing recognition.

IV. Modelling and simulation

Q. What is the relation between modelling and simulation?

Models are mathematical, but do not belong to reality. I am not a realist concerning mathematics. Let us consider classical physics. Newton’s equation is perfect, but planets do not do differential calculus. They move, but they do not solve equations. However, Newton’s equation allows for computing (either explicitly or only numerically) some solutions that simulate perfectly the observed motions. The same holds for all models. We start with collecting a great corpus of empirical data and then we try to find good models able to generate a virtual reality that simulates the empirical reality.

If the morphodynamical models based on the hypothesis of the rooting of language in perception are correct, then it would be right to try to understand their neuronal implementation. It is not easy at all. Stephen Kosslyn, a well known neurophysiologist of vision, has studied with current methods of brain imaging the neural activity during the use of prepositions. He showed that there exist two systems for the processing of spatial relations: one is a continuous quantitative processing (*A* is more or less above *B*), and the other is a categorical discontinuous processing (*A* is above or beside *B*). Moreover, the two neural processings are lateralized: the continuous one takes place in the right hemisphere, while the categorical one takes place in the left hemisphere[34]. Hence, if one wants to know exactly how the brain deals with prepositions, he has to go deeply into the analysis and modelling of the link between perceptive structures and linguistic categoryization.

Interested readers will find more information in *Cognitive Morphodynamics* as well as in my 2008 book *Neurogéométrie de la vision. Modèles mathématiques et physiques des architectures fonctionnelles*[35] in which I deal with the neural implementation of basic properties of perception (which are already very difficult to understand even though they remain very far from the complexity of language).

Q. Is simulation a form of explanation?

It depends on the structure of the models on which the simulation is based. “It works!” is not by itself an explanation since it can pertain to the mere fine-tuning of ad hoc parameters. Models are explanatory when they arise from general and strong hypotheses while being able at the same time to generate good simulations. It is the case with Newton’s equation, which results from general physical principles; it is the case with the elementary catastrophes which result from general principles of structural stability and from the dimensions of space-time.

Q. Numerous linguistics phenomena are quantitatively characterized by a Zipfian distribution. Is there any relation between this characteristic and the modelling proposed by catastrophe theory?

I have never worked in the field of statistical linguistics. Regarding Zipf’s law in particular, I haven’t worked on this subject, although the CAMS did work a lot on it[36].

However, one can’t ignore that statistics are a good way for approaching regularities and that, during acquisition, children are learning rules in a statistical way: they extract linguistic rules by generalizing over a finite set of examples.

[34] Kosslyn, S.M. (2006). “You can play 20 questions with nature and win: Categorical versus coordinate spatial relations as a case study”, *Neuropsychologia*, 44 (2006) 1519-1523. See also Kemmerer, D. (2007) “A Neuroscientific Perspective on the Linguistic Encoding of Categorical Spatial Relations”, *Language, Cognition and Space*, (V. Evans et P. Chilton eds), *Advances in Cognitive Linguistics*, London, Equinox Publishing Co.

[35] Les Editions de l'Ecole Polytechnique, Distribution Ellipses, Paris.

[36] Cf. Micheline Petruszewycz, “L'histoire de la loi d'Estoup-Zipf: documents”, *Mathématiques et sciences humaines*, 44 (1973) 41-56.

There are very interesting connectionist models (those by Jeffrey Elman seem to me to be the most interesting) that model how the syntagmatic statistical regularities induce semantic paradigms.[37]. You consider a small corpus containing various classes of nouns (animate agent, non animate object, etc.) and various classes of verbs (*to eat*, *to read*…). Then you do supervised learning with a neural network: you give a word as input to the network, you ask it to add one more word, and you correct it if it outputs an incoherent sentence. At the beginning, the network produces outputs that haven't any coherence, neither syntactically nor semantically. The corrections you pointed-out allow it to change its internal structure (i.e. to change the weight of its hidden layers) by retro-propagating the errors. When learning is done, the network does not make errors anymore. Then, you look at its hidden layer, and you see that it has built paradigms (animate agents, inanimate objects, state verbs, transitive and intransitive action verbs, etc.). "Paradigm" here means that the words are grouped into clusters. In other words, in order to produce correct syntagmatic sentences, the network has built some semantic rules.

Q. In the field of complex systems, what are the differences today between dynamical models and connectionnist models?

The difference between (morpho-)dynamical models and connectionist models is the following: connectionist models do have internal dynamics and, hence, attractors. They are made of atomic units (the formal neurons) linked by inhibitory/excitatory connections having synaptic weight. Each unit influences the units to which it is connected, which produces a global internal dynamics of the network.

The main interest of these connectionist models is to make explicit the underlying differential equations, whereas they remained implicit in Thom's and Zeeman's works. These equations (introduced by Jack Cowan, Hugh Wilson and John Hopfield[38]) have very interesting properties and look very similar to those found in statistical physics in the theory of spin glasses. This has allowed, during the 1980s, a massive transfer of a large bulk of results from statistical physics to connectionist models.

But the fundamental limit of connectionist models is that they do not model the bifurcations of attractors that can result from an external dynamics modifying the attractors. They do use external dynamics but chiefly for modelling learning processes. The consequence is that they cannot afford models for constituent structures needed by all syntactic theories. A sharp debate took place at the end of

[37] Cf. Elman, J. (1989). Representation and Structure in Connectionist Models, *Cognitive Models of Speech Processing*, (G. T. M. Altmann, ed.), Cambridge, MA, MIT Press, 1989, 345-382.

[38] H.R. Wilson and J.D. Cowan (1972). Excitatory and inhibitory interactions in localized populations of model neurons, *Biophys. J.*, 12 (1972) 1–24. J. J. Hopfield (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the USA*, 79, 8 (1982) 2554–2558.

the 1980s between classic cognitivism (Jerry Fodor and Zenon Pylyshyn) and connectionist cognitivism (Paul Smolensky). Fodor and Pylyshyn's thesis was that if one models the components of a sentence by attractors of a neural network, then it is not possible to model constituency. They were right. In order to model syntax, a model needs to be able to model constituency, which is impossible with attractors only.

However, as I wrote[39], Thom's actantial models provided an answer at the beginning of the 1970's, to this key issue of the late 1980's! Indeed, thanks to their built-in bifurcations, these models allow for what I call an "attractor syntax". If one models constituents (for instance, actants) with the attractors of some network, then it is not possible to model the relations between these constituents (for instance, actantial relations in a verbal node) through the attractors of the same network. One needs interactions between attractors, that is bifurcations. Attractors' bifurcations allow for the dynamical modelling of verbal nodes and constituent structures. It was the central idea of Thom's actantial graphs we have already discussed.

Q. Have these models proved seminal? What are the actual research results that are based on your work in cognitive morphodynamics?

We already talked about phonetics. In actantial syntax, the most important works are those by my friends Wolfgang Wildgen and Per Aage Brandt. In the teams of Aarhus and Copenhagen Peer Bundgaard[40], Svend Østergaard and Frederik Stjernfelt have used morphodynamic models. In Paris, David Piotrowski, a structuralist in the line of Hjelmslev has elaborated upon my propositions and plans to use neuroimaging (EEG). He claims that good neuroimaging experiments may help decide between linguistic theories since acceptability may be tested with neural waves, in particular N400[41].

Again in the field of linguistics, there are works by Bernard Victorri on synonymy that use dynamic models in an innovative way[42]. About prepositions, there are many works that still need to be modelled, in particular those by Claude

[39] (1991) Why Connectionism is such a Good Thing. A Criticism of Fodor's and Pylyshyn's Criticism of Smolensky, *Philosophica*, 47, 1 (1991) 49-79. (1994) Attractor Syntax: Morphodynamics and Cognitive Grammars, *Continuity in Linguistic Semantics*, (C. Fuchs et B. Victorri eds), Amsterdam, John Benjamins, 1994, 167-187. (1995) Morphodynamics and Attractor Syntax. Dynamical and morphological models for constituency in visual perception and cognitive grammar, *Mind as Motion*, (R. Port and T. van Gelder eds.), Cambridge, MA, MIT Press, 1995, 227-281. Articles summarized in *Cognitive Morphodynamics*.

[40] Cf. for instance P. Bundgaard and J. Petitot (eds), (2010) *Aesthetic Cognition*, Special Issue of *Cognitive Semiotics*, 5, 2010. F. Stjernfelt and P. Bundgaard (eds) (2011) *Semiotics. Critical Concepts in Language Studies*, New York, Routledge.

[41] Piotrowski, D. (2009) *Phénoménalité* et *objectivité linguistiques*, Champion, Paris.

[42] See Victorri B., Fuchs C. (1996), *La polysémie. Construction dynamique du sens*. Paris, Hermès.

Vandeloise[43]. In the volume edited in tribute to Vandeloise, there is a very interesting paper by Langacker[44].

In the field of perception, perceptive bifurcations have been studied extensively. There are models that follow Thom explicitly, others that follow Prigogine, and others synergetics; however, all these models are based on bifurcations. There is a large amount of empirical data. For instance, the Necker cube (figure 6), with its well-known double perspective. The same bi-dimensional stimulus can be interpreted as a tri-dimensional object in two different ways, and these two ways are bifurcating one in the other in a spontaneous and alternating manner along temporal series that have been studied in depth. The inversion of perspective is easy to understand. In bi-dimensional images, there are two points particularly salient and informative (the two edges in the centre of the figure); and according to the way you focus on one or the other of these two points, the cube is seen under one or the other perspective. There is also the example of the Rubin's face (figure 7) [45].

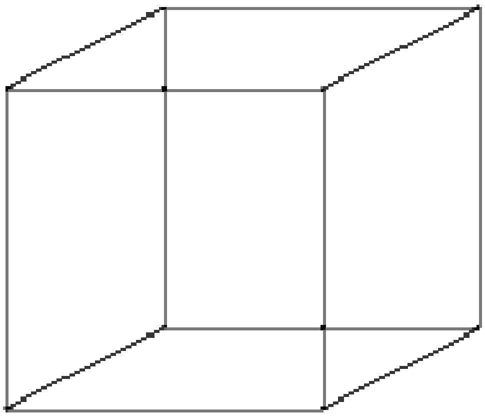


Figure 6. Necker's cube
(source: http://fr.wikipedia.org/wiki/Cube_de_Necker)

[43] Vandeloise, C. (1986) *L'Espace en Français: Sémantique des prépositions spatiales*, Paris, Editions du Seuil. (2009) “The genesis of spatial terms”, *Language, Cognition and Space: the State of the Art and New Directions*, (V. Evans, P. Chilton eds), London, Equinox (*Advances in Cognitive Linguistics*), 157-178.

[44] Langacker, R. (2010) Reflections on the Functional Characterization of Spatial Prepositions, *Espace, Préposition, Cognition. Hommage à Claude Vandeloise*, (G. Col, C. Collin, eds), Corela.

[45] The vase-face by Edgar Rubin (Rubin, 1921) shows the importance of the figure–ground contrast in perception. According to whether one looks at the white area as the ground or the form, one sees two faces or a vase.

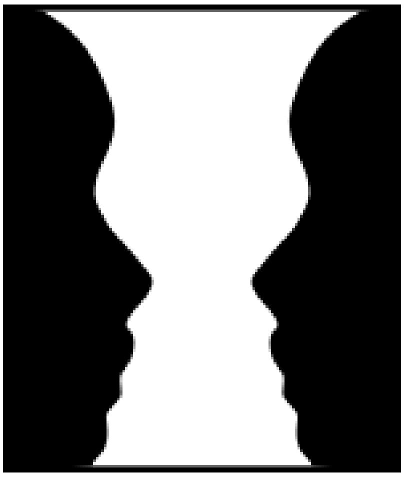


Figure 7. Rubin’s face
(source: http://fr.wikipedia.org/wiki/Perception_figure-fond)

Thus, in many domains where mereological concepts of morphology and of structure mean something, one of the major issues is understanding how categories can emerge in continuous substrates. For this, one needs models where, in one way or another, there exist processes that produce discontinuities. One cannot escape this necessity and this explains the relevance of morphodynamical models.

Lexical Semantics and Topological Models

Sabine Ploux
Institut des Sciences Cognitives, UMR 5304 (CNRS, Université Lyon 1)
sploux@isc.cnrs.fr

This article presents some elements and milestones that led to the emergence and development of the topological paradigm in lexical semantics in France, and more specifically, to the computational realizations that ensued. The concepts that laid the groundwork for this paradigm arose at the interface between mathematics, linguistics, and computer science, conveyed by a scientific community that has inherited both a modelling tradition and a set of linguistic theories whose foundation includes mathematical terminology. The paradigm rests on topological notions, data-analysis methods, and a booming technological context made possible by the development of computer science and the possibilities it offers: large calculations, databases, and digital corpora.

There are many ways, of course, to grasp and interpret *a posteriori* the history of ideas and scientific realizations. In computational lexical semantics, authors often reduce the presentations of their work to goals initiated in natural language processing (NLP), such as the problem of sense disambiguation in context. I propose instead to trace the link between abstract mathematical concepts and the sometimes-oblique influence they have had in this domain. This link has grounded the representation of lexical semantics in a process of unfolding potentialities, unlike sense disambiguation which is more like a process of reducing the set of possible meanings.

I will begin by attempting to present the linguistic and mathematical inheritances, which, by way of conceptual links, laid the foundation for the discipline. I will then examine the various solutions proposed (dynamic systems, vector spaces, geometrical space, etc.) and their implementation (neural nets, data analysis, etc.), while trying to highlight the fit between the semantic phenomena to be described and the abstract ways of doing so. Unlike generative grammar, whose formal framework is essentially grafted onto the concept of inference, the topological approach to semantics, which looks at meaning variability in context, is built on the notions of "space" and "dynamics". I will pay particular attention to the different levels of organization proposed to support the explanation of variability: the observable level of lexical units, and the underlying level of theoretical units (features, semes, cliques).

1. Scientific Inheritance

1.1 Linguistic Theory and Topology

Antoine Culioli's theory of enunciative operations has certainly had a great influence in France and has enriched the link between continuous mathematics and linguistics. The terminology used therein is borrowed directly from topology.

Accordingly, a notional domain that structures the class of occurrences of a given notion (which itself refers to a system of cognitive representations, e.g., the notion denoted /cat/) is composed of an *interior*[1] endowed with an *organizing centre* (which is a type occurrence), an *exterior*, and a *boundary* (Culioli 1999). The theory also includes the concepts of stability and deformation, which we find in the theory of dynamic systems. "Without *stability*, there would be no regulated adjustment, no communication [...] but stability cannot be confused with rigidity or immutability. Linguistic phenomena form *dynamic systems* that are regular, but with a margin of *variation* due to a wide variety of factors: these phenomena are both stable and malleable. [...] *Deformation* is a transformation that modifies a configuration in such a way that certain properties do not vary under the transformation, while others do. [...] For deformability to exist, we have to have a schematic form (such that we can see both modification and invariance in it), there must be some deformation factors and some leeway, i.e., an adjustment space endowed with *topological* properties." (Culioli 1991; 1986 T1:129-130). This quoted remark places enunciation theory in a paradigm very much like the one proposed by René Thom (1977,1980), in which the main task is to account for possible deformations of a concept or category in a way that allows it to retain its structural stability.

Other linguists like L. Gosselin founded their research on a framework involving topological representations. In Gosselin (1996), this author proposes a computation-based model of French temporality and aspect, grounded in a representation of linguistic processes based on temporal intervals (in the sense of continuous bounded time frames).

References to these mathematical notions have often been grouped under the heading "continuist models", with continuity being a concomitant notion to that of spatiality. The concept of continuity has also been used to delineate an approach opposing the discrete one, wherein linguistic units are finite elements that create meaning by combining with each other. At a conference entitled "Continuity in Linguistic Semantics" headed by C. Fuchs and B. Victorri (1994), various types of arguments were advanced to explain and support this trend of ideas:

- The extralinguistic parameters that convey language — such as perception, movement, space and time, and pragmatic context — are continuous in nature, and it is necessary to account for how one goes from these infra-linguistic levels to the linguistic level.
- Polysemy and the intrinsic variability of word meaning is better represented in a continuous space.
- Difficulty using a purely symbolic framework to account for the results of psycholinguistic experiments on categorization calls for a paradigm that includes either weighting (using values taken from real numbers) or gradients in a continuous space.

[1] Terms common to the quoted passage and topological mathematics are shown in italics.

- Continuity allows one to synthesize a large set of data. (This is a methodological argument that does not assume that continuity is used to adapt to the intrinsic nature of objects.)

Apart from the work by L. Gosselin, who used temporal representations to build a computational system, a continuous topological framework is called upon the most because it allows one not only to analyze (perception- and context-based analysis, meaning analysis, comparative processing of concepts, data processing, etc.) but also to explain category-identification processes. This approach differs from generative theories in linguistics, wherein the concept of inference is implemented within a discrete paradigm compatible with a logical framework. Here, the objective is to synthesize linguistic mechanisms by proposing rules and then calculating and producing the set of well-formed utterances.

Thus, the features of the models are in keeping with the formal framework chosen. The characteristic feature of the topological framework is the presence of different, incommensurable levels of units: the units of the studied objects, and the units of a substrate level where the objects' units delineate shapes or are defined in terms of their deformation potential or their dynamics. This framework is thus more suitable for a semantic analysis "within" lexical units. The discrete approach[2] enables one to compose units in order to generate new ones of the same nature, which is why it was chosen as the generative framework for combining linguistic units when the goal is more to understand the operating rules interlinking them, than to uncover their internal structure.

1.2 Space, Topology, Dynamics, and Modelling

The idea that topology is a necessary paradigm — first for modelling perception, and later, lexical semantics because of the latter's link to the perceptual modalities — is anchored in a tradition of work in mathematics and philosophy initiated by Poincaré (1921), who "strove ... to analyze the psychological origin of the notion of space." In *La Science et l'hypothèse* (Science and the Hypothesis, 1968)[3], Poincaré analyzes the constitution of the physical continuum and the space containing our representations, based on our experiences and immediate sensory data. R. Thom's work, which aligns with Poincaré's, goes beyond the perception of shapes and space by adding dynamic modelling of morphogenesis

[2] Although the concepts "movement" and "displacement" exist in generative linguistics, as does the concept "dynamic" in formal semantics, they have not, to my knowledge, received any mathematical underpinning in the form of a mathematics of dynamic systems, and they are still described in terms of constraints attached either to the sentence's tree structure or to mathematical logic.

[3] Poincaré's work influenced philosophers such as Jean Nicod, who published a study on the construction of a geometry of space from motion sensations. Nicod was able to take advantage of several mathematical frameworks, each according to its specific characteristics: Russell's mathematical logic for induction, and Poincaré's work on topology for constituting a proper space.

phenomena like those found in language and semantics. Accordingly, Thom describes verbs as specific processes that unfold singularities linking the subject of a process to its object and any other existing constituents. J. Petitot (2011) showed how the theory of dynamic systems can account for categorical perception in phonology and also in syntax, where he stresses the fundamental links between vision and syntactic structures.

However, while the dynamic topological paradigm offers an explanatory framework for semantic analysis - in the sense that it allows one to derive the set of all possible variations, and only those variations, as a function of the internal parameters associated with the organizing centre or type schema - its theoretical proposals are confined solely to case studies. And there is no systematic computation-based realization following directly from it that is able to assign a semantic space to each lexical unit, one capable of modelling its constitutive dynamics and the organization of the different semantic values it takes on in context.

As a result, the massive realizations (massive in the sense that they apply to the entire set of words in the language) have not been derived from a topological model of the spaces proper to each lexical unit, but rather from a systemic approach using linear algebra, i.e., data analysis. The data-analysis approach encompasses work done by several researchers worldwide in view of organizing large sets of data coming from different domains of science and the humanities. The main idea is to extract components or axes that allow one to determine the principal variations that structure the data.

In the 1960's in France, Benzécri's work and his correspondence analysis (hereafter called CA) were well received and utilized by the scientific community in many fields, particularly sociology (Bourdieu 1996). When applied to textual data, a correspondence analysis calculates the coordinates of words in a space, using tables filled in by extracting information from text corpora and questionnaires, etc. The coordinates are then used to compute semantic neighbourhoods and semantic proximities between words.

In sum, major differences separate structural topological approaches like those proposed by J. Petitot or B. Victorri and Y.M. Visetti, from approaches based on correspondence analysis. The former look at the topological and dynamic structure of lexical units; the latter (at least the first realizations) attempt to represent the structure of part or all of the lexicon. The topological approach defines a dynamic configuration within which it directly relates the different types of utterances containing a given lexical unit, to the unit's variations in semantic content. The statistical approach uses semantic proximities to interlink the lexical units in the set of words under study. Lastly, while the goal of the former is to dynamically study the construction of meaning, the latter are static models that represent meaning.

1.3 Computers, Corpora, Graphs, and Calculations

The advent of computers constituted a major technological advancement towards the implementation of models and the validation of their linguistic output.

Computers gave us the ability to compile large text corpora, develop computerized dictionaries, and perform enormous calculations.

Compiling large corpora. For the English language, the compilation of large corpora began in the 1960's (Brown Corpus, 1967; BNC, 1991, etc.). France followed in the 1980's and 1990's (Frantext 1988; the Inalf corpus of synonym dictionaries). These sources led to many corpus-based studies in linguistics (Habert, Nazarenko, Salem 1997) and in natural language processing.

Software and implementation of mathematical methods. The arrival of computers had another consequence, namely, the implementation of connectionist models and statistical methods for analyzing text, the latter of which gave rise to many studies in laboratories like the Saint-Cloud Lexicometry Lab in France. This work went beyond the study of lexical semantics, covering domains such as sociolinguistics, literary analysis, and discourse analysis. The development of computer systems like Alceste, Semantic Atlases, Lexico, Hyperbase, Prospero, TreeCloud, etc., which are still in use today, is currently a very active area of research. These software packages enable the study of word frequencies in large reference corpora, co-occurrences of words in texts, and techniques for visualizing semantic proximities using various methods such as correspondence analysis and hierarchical classification. The ability to see output in spatial format has had a great impact on society, with graphic presentations of news analyses and political discourse (Jean Véronis, http://blog.veronis.fr/), and on the world of design, with the use of word clouds that have invaded our daily lives (publicity, banners, etc.) and inspired artistic works like Boris Nordmann's Semographe (http://www.borisnordmann.com/semographe/).

In the next section, I will begin by giving some prototypical examples of realizations of connectionist networks that construct meaning in context, while relating each one to its formal source framework. Then I will present some spatial models devoted to the representation of meaning (including some of the so-called vector space models).

2. Models and Realizations

2.1 Dynamic Models of Meaning Construction in Context and Connectionist Realizations

In order to implement dynamic models, researchers have used connectionist networks. This type of network was chosen because of a shared objective: using data or initial states to determine a system's convergence states. In this vein, researchers in linguistics attempted to find a model that relied on dynamic convergence to determine the possible semantic value or values of a word, as a function of its context of usage (considered to be its initial state in a connectionist network) and as a function of its initial weights, the network's architecture, and the learning rule. After several iterations of the system, the values of the weights stabilize — this corresponds to the network's response. In Victorri, Fuchs (1996), we find a description and a precise justification of the fit between dynamic models and connectionist networks in semantics. The theory of dynamic systems serves as a framework for synthesizing and understanding the set of semantic variations of a lexical item in the various utterances that contain it. For example, the French word *encore*, which is presented as a paradigmatic example of this approach, is endowed with a core meaning ("the part of its meaning that remains invariant" over and above the different modifications of the co-text) or abstract operator, described in a diagram like the ones used by A. Culioli. "Let D be any domain (temporal, spatial, notional, etc.). Let P be a proposition whose domain of definition is D and whose domain of validity, denoted D(P), belongs to D. Now let T be a trajectory in D and t_0, a privileged point on that trajectory. Then the various senses of *encore* have in common the fact of indicating that the boundary between D(P) and D(non-P) crosses T at a point t_1 beyond t_0, while it would be conceivable or even predictable that it would be before it." The authors show how different utterances containing the word *encore* induce a dynamic process described by the diagram. This operator constitutes the invariable part of the word's semantics. The variable part (such as the fact that *encore* can refer to the repetition or sustainment of a process) is described via different dimensions: "The first has to do with domain D ... The second concerns a way of travelling along the trajectory and corresponds to the distinction, fundamental for *encore* as well as for other grammatical markers, between discrete and continuous. Lastly, a third dimension, which pertains more to the enunciator's point of view, emphasizes that such and such an aspect of the operation described by *encore* proves necessary in the end to differentiate between the typical values." Another dynamic system is then used to model the process that determines semantic values as a function of the co-text's characteristics. The co-text is described as follows: Each lexical unit in the utterance is described using linguistic features: singular or plural for nouns; stative, spatial, notional for verbs; etc. These features are encoded in the form of vectors composed of 0's and 1's. The semantic values output by the system are associated with various reformulations of *encore — à nouveau* (again), *une fois de plus* (once again), *davantage* (more), *un peu plus* (a little more) — and constitute the attractors of the dynamic system.

Two levels of units are distinguished here. The first level contains semantic values paraphrasable by locutions and represented by the system's attraction and convergence basins. The second level contains features composed of elements of linguistic description, represented by values between 0 and 1 in a multidimensional space whose dimension is equal to the number of features. Despite the intrinsic merits of this type of modelling for obtaining a concise understanding of a lexical unit's semantic variations in context, the infatuation for connectionist networks and their use in lexical semantics, or in cognition in general, dwindled in the late 1990's.[4] I would be tempted to analyze this disappearance in terms of the intrinsic limitations of having to choose features and primitives by hand and the pitfalls involved in encoding them as real number values. Clearly — and this is not a characteristic of connectionist models alone, since we find it in any

[4] A few new projects (see Christopher D. Manning's works for example) are using this paradigm to reproduce and improve vector space models derived from data analysis, while also taking the order of words in a sentence into account, something that data analysis cannot do.

proposal defined in a metalanguage, such as a list of attributes (argumental, evenemential, or qualia-related (Pustejovsky 1998) or semes (Rastier 1987) assumed to be finite in number — recourse to a metalanguage with a finite vocabulary raises unresolved metatheoretical questions, such as whether a complete and potentially minimal set of primitives exists (where a set of primitives is complete if it can describe and simulate any construction process, in this case semantic) and whether we are capable of selecting that set. The problem of choosing primitives also goes back to the problems of how to encode them (in the above example, by a vector composed of 0's and 1's, where 1 stands for the presence of the feature, 0 otherwise) and whether we can determine the correspondence in a systematic way.

The replacement of connectionist models by data-analysis methods — which while not fully responding to the goal of modelling the meaning-construction process but nevertheless offering the advantages of systematicity and automaticity — undoubtedly also precipitated their disappearance.

2.2 Spatial Representation Models

Unlike generative (Pustejovsky 1998) or connectionist models, which simulate the processes that compute meaning, models that use space as a paradigm do not give access to processes; instead, they provide a representation of word meaning. Now called vector space models, since most use a vector in a multidimensional space to represent a word (as explained below), these models rely on data analyses, especially text data (Benzécri 1980; Lebart, Piron, Steiner 2003). Among the wide variety of studies using methods derived from data analysis, two major types can be distinguished: those whose units are linguistic entities only, and those that construct a level of units other than directly observable ones, such as cliques or small worlds. This distinction was chosen because it also corresponds to certain objectives and particularities. The former type of method essentially focuses on the organization of all or part of the lexicon, taken in its globality; the latter also permits a representation of the internal structure of a lexical unit's semantics.

Vector space models. Vector space models are generated via the automatic extraction of co-occurrence links from *corpora*. The initial units are the word, the sentence, the paragraph, and the text, all delineated by separators such as blanks, punctuation, a carriage return, etc. Note that the process of segmenting into word units, even very basic ones, involves making some critical choices. In this approach, for example, the French term for potato *pomme de terre* (literally, apple of the earth), is made up of three words, not one, whereas the words *maison* and *maisons* (the singular and plural of the word *house*) are two distinct lexical units. The vector space model automatically performs a factor analysis between the words on the one hand, and the texts or paragraphs that contain them on the other. To do so, it builds a table with the documents, sentences, and paragraphs in the rows, and all words in the corpus in the columns. The cells of the table are filled with the number of occurrences of the word (the column header) in that portion of the text (the row header), sometimes reduced to 1 if the word is present and 0 otherwise. Then the chosen data-analysis method is applied to this table. The output associates a vector in a multidimensional space to each word, paragraph, or text, and defines a system of neighbourhoods pertaining both to the terms and the texts or paragraphs. The framework chosen for representing the semantic neighbourhoods is thus a Euclidean space. In some cases, semantic similarity between two words is not measured by the Euclidean distance between their associated vectors but by the cosine of the angle formed by those vectors.

Below, I outline some of the features of these models.

Vector space models do not assume any kind of *a priori* organization for the semantics of lexical units, nor any substrate level of organization (there are no semantic features, nor an organization level finer than the word). Only the paragraph, the sentence, and the text, taken as "bags" of words, constitute another level (whose own organization is not studied directly).

In addition to being used in NLP, these models have been taken up in psycholinguistics, where they provide an answer to a criticism addressed to the classic approach to concepts (also called Aristotelian) originating in an alternative trend to the one that developed following E. Rosch's work (Collins, Loftus 1975; Rosch, Mervis 1975). Researchers in this trend question the feasibility of defining concepts and word meaning (Kintsch 2001), and, via experimental paradigms, have pointed out contradictions that follow from the presupposed existence of definitions based on a fixed list of necessary and sufficient conditions or properties. Rather than a system hierarchically organized into a list of properties, they prefer an organization founded on the notion of similarity, with spatial models offering a form of realization. "*We seek a mechanism by which the experienced and functional similarity of concepts ... are created from an interaction of experience with the logical (or mathematical or neural) machinery of mind*" (Landauer, Dumais 1997).

However, from the standpoint of lexical semantics, spatial-representation models have some intrinsic limitations. Indeed, the representation associated with a word is atomic, in the sense that the different components of the vector are not interpretable in terms of semantic characteristics. For this reason, the semantic values of a word cannot be represented in and of themselves. The only thing given is a list of neighbouring words, which, while each is associated with one or more of the word's values,[5] does not allow one to separate them. The "logic" of the word's meaning (i.e., the distinctions and interconnections between the different semantic values) is not a direct output of the model; the sole calculation is a measure of the distance between words. In sum, the organization of the different senses of a word into classes or a tree structure (as in dictionaries or the WordNet database (Fellbaum 1998) has no counterpart in a vector space model.

Representation of polysemy. The initial plan to model meaning in context and characterize the semantic space proper to a given word is not directly achievable using data analysis applied to text corpora, unless one introduces an

[5] Note, however, that Schütze's (1998) work addresses this issue by proposing an automatic classification of the various values of a term, obtained by determining the centres of gravity of the vectors representing the word's different contexts.

organization level other than a directly observable one like the word, the sentence, the paragraph, or the text. Several proposals for infra-linguistic units — infra-linguistic in the sense that they are not directly interpretable as units of the language — have been made. They do not use a metalanguage formed of semantic features or semes, but rather units calculated from the properties of word graphs (cliques in Ploux 1997 or small worlds in Gaume 2004), with each word being associated not with a vector but with a domain in a multidimensional space, in such a way that within that domain, one can discern various areas likely to represent the different meanings of the word.

Cliques. To construct the domain associated with a lexical unit, it is necessary to build or define its constituent elements. The idea of recourse to the notion of clique arose from the hypothesis that attaching to a word a list of words that are semantically related to it will constrain its meaning and thereby result in the fragmentation of its semantics. For the word *good*, for instance, the following lists (Ploux, Hyngsuk Ji 2003) will further specify the meaning of *good* by dividing up its values into ones linked either to an aptitude or to a moral quality:

- 6: able, adequate, capable, competent, effective, good
- 7: able, adroit, clever, dexterous, expert, good, skilful
- 8: able, capable, clever, expert, good, skilful
- 111: friendly, gentle, good, kind, kindly, nice, sweet
- 112: friendly, good, gracious, kind, kindly, nice, sweet
- 113: friendly, good, helpful, kind

Cliques offer a way of compiling such lists. To generate them, one proceeds as follows. The lexicon is seen as a graph whose vertices are words and whose sides are the semantic links (by synonymy or association) between those words. On the graph, a clique is a maximal, complete, and connected subgraph. If the link is synonymy, for example, then a clique will be composed of words all of which are synonyms of each other.

From the spatial standpoint, one can regard a clique as the intersection of the set of areas associated to the list of words the clique contains. An implication of the property of maximality is that there exists no other term in the language that can divide up the intersection of the areas associated to the clique's list of terms. For this reason, a clique represents a minimal unit of meaning, a "grain" of meaning. Each clique will be represented by a point or vector.

The topology that underlies the set of all cliques can be discerned in a list of cliques. Below is a path in which each clique shares at least one term with the preceding clique. We can see that the meaning moves from a taste-related value to a value that might refer to a person:

- 80: delectable, delicious, good, lovely, savoury, scrumptious, tasty
- 78: delectable, delicious, excellent, exquisite, good, lovely, scrumptious
- 77: delectable, delicious, enjoyable, good, pleasant
- 79: delectable, delicious, good, lovely, pleasant

- 82: delicious, good, lovely, nice, pleasant
- 114: friendly, good, kind, kindly, nice, pleasant, sweet
- 111: friendly, gentle, good, kind, kindly, nice, sweet

The construction of the shape associated with the initial word (here, *good*) via a CA is used to summarize the set of proximity links. Then a hierarchical classification on the clique's coordinates outputs a map of the organization of the different semantic values. At the centre of the map is the generic value, if there is one, and the various other values are situated around it. The gradient coming from the map's centre is a measure of the specificity of the semantic values (which is comparable to an organization into prototypes). Overlapping meanings or relative proximities among values are depicted by their distances from each other.

Small worlds. Small worlds are properties of graphs utilized in Bruno Gaume's lexical graphs. They act as a sort of generalization of cliques, not by taking into account complete connected subgraphs, but rather by using the more flexible notion of strongly connected subgraphs. In strongly connected subgraphs, all vertices (words) can be joined by a path that is shorter than a given length. In Gaume (2004), the author shows how using such subgraphs enables one to represent not only lexical polysemy but also the structure of graphs extending beyond the word's immediate neighbourhood (i.e., all words directly related to it). The method itself takes the graph's adjacency matrix G and calculates profiles (as in CA) representing the weighted probability of there being a link between two vertices, and then raises this matrix to the power *n*, thereby tolerating paths between vertices that are shorter than a fixed value *n*. The structure of the resulting graph can be viewed by applying a principal component analysis (PCA) to the output matrix. The coordinates of each word are then calculated from the first three dimensions of the PCA in order to obtain a representation in space.

Status of infra-linguistic units. In the two approaches described above, two levels of organization, different in nature, mutually determine each other: the lexical-unit level and the level of the units extracted from the lexical graph (cliques or small worlds). In short, if these approaches are able to represent the polysemy of a word, it is firstly because they assume that a geometric shape in a multidimensional space must be associated with the word, and secondly because they propose a means of defining an infra-linguistic level of units whose granularity is finer than, and even incommensurate with (in the sense that a point is incommensurate with any portion of space whose dimension is greater than 1), the level of the words themselves. The term *"infra-linguistic"* is employed here to describe those units whose individual semantic specificity cannot be characterized, each one often being too close, semantically, to several others to allow one to distinguish them by their meaning (as is possible in the SynSets of WordNet, which, moreover, are less numerous).

The addition of an abstract organization level incommensurate with the object level in a to-be-modelled domain seems to be essential, for the absence of this level inevitably generates homogeneity conflicts. Indeed, objects of the same nature in a linguistic domain must be associated with mathematical entities that are also of the same nature, and conversely, differences in nature in the to-be-modelled domain must be accompanied by the same differences within the model. Failure to satisfy this principle generates pitfalls or contradictions. Thus, representing each lexical unit by a vector does not enable one to assign it its own semantic organization, since a vector is an atomic unit. On the other hand, using vectors to represent objects that are linguistic in nature (features, semes) forces the composing process to obtain only units that are alike in nature, so it seems difficult to end up with realizations whose output would correspond to the semantics of a lexical unit.

3. Conclusion and Cognitive Perspectives

This article does not claim to be exhaustive but simply attempts to trace the history of the topological paradigm in lexical semantics in France. This history points out a deviation from the initial goal of representing the semantics of words in terms of their own individual dynamics. However, the topological notions that grounded the paradigm (neighbourhoods, interior, boundary, etc.) have spread to another paradigm, this time based on data analysis, that has resulted in the creation of models with broad lexical coverage. To develop such models, new units were generated from the theory of graphs. The new units (e.g., cliques) are not interpretable as linguistic units, which raises the interesting question of their cognitive relevance. To answer this question, a promising route would be to search for equivalents of these units in the dynamics of neural activeation during word comprehension in context. By providing a structural characterization of these dynamics from the standpoint of their neural temporality and spatiality, such an approach would bring us back to the initial goal.

References

Benzécri, J-P. (1980). *L'analyse des données: l'analyse des correspondances*. Paris: Bordas.

Bourdieu, P. (1996). *Raisons pratiques*. Seuil: Points.

Collins, A., Loftus, E. (1975). A spreading activation theory of semantic memory. *Psychological Review 82,407-428.*

Culioli, A. (1991). *Pour une linguistique de l'énonciation : Opérations et représentations,* volume 1. Paris: Ophrys.

Culioli, A. (1999). *Pour une linguistique de l'énonciation. Formalisation et opérations de repérage,* volume 2 of *L'homme dans la langue*. Paris: Ophrys.

Fellbaum, C. (Ed.) (1998). *Wordnet, An Electronic Lexical Database*. Cambridge, Massachusetts: MIT Press.

Fuchs, C., Victorri, B. (Eds.) (1994) *Continuity in Linguistic Semantics*. Amsterdam: Benjamins.

Gaume, B. (2004). Balades aléatoires dans les petits mondes lexicaux. *I3: Information Interaction Intelligence 4(2), 39-96.*

Gosselin, L. (1996). *Sémantique de la temporalité en français. Un modèle calculatoire et cognitif du temps et de l'aspect. Recherches*. Louvain-la-Neuve: Duculot.

Habert, B., Nazarenko, A., Salem, A. (1997). *Les Linguistiques de Corpus*. Paris: Armand Colin.

Kintsch, W. (2001). Predication. *Cognitive Science 25, 173-202.*

Landauer, T.K., Dumais, S.T. (1997). A Solution to Plato's Problem: the latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review* 104, 211-240.

Lebart, L., Piron, M., Steiner, J-F. (2003). *La sémiométrie - Essai de statistique structurale*. Paris: Dunod.

Ploux, S., Ji H. (2003). A model for matching semantic maps between languages (French/English, English/French). *Computational Linguistics* 29(2), 155-178.

Ploux, S. (1997). Modélisation et traitement informatique de la synonymie. *Linguisticae Investigationes* 21(1), 1-28.

Poincaré, H. (1921). Analyse des travaux scientifiques. *Acta mathematica* 38(1), 3-135.

Poincaré, H. (1968). *La Science et l'hypothèse* Paris: Flammarion.

Pustejovsky, J. (1998). *The Generative Lexicon*. Cambridge, Massachusetts: MIT Press.

Rastier, F. (1987). *Sémantique interprétative*. Paris: PUF.

Rosch, E., Mervis, C. (1975). Family resemblances: studies in the internal structure of categories. *Cognitive Psychology 7, 573-605.*

Schütze, H. (1998). Automatic sense discrimination. *Computational Linguistics 24(1), 97-123.*

Thom, R. (1977). *Stabilité structurelle et morphogénèse*. Paris: InterEditions,

Thom, R. (1980). *Modèles mathématiques de la morphogénèse*. Paris: Christian Bourgeois Editeur.

Victorri, B., Fuchs, C. (1996). *Polysémie et construction dynamique du sens*. Paris: Hermès.

Petitot, J. (with **Doursat, R.**) (2011). *Cognitive Morphodynamics. Dynamical Morphological Models of Constituency in Perception and Syntax.* Berne: Peter Lang.

The RAM-Verlag Publishing House edits since 2001 also the journal ***Glottometrics* –** up to now 35 issues – containing articles treating similar themes. The abstracts can be found in http://www.ram-verlag.eu/journals-e-journals/glottometrics/.

The contents of the last issue (35, 2016) is as follows:

Herausgeber – Editors of Glottometrics