Quantitative Linguistics and Political History

Damon Mayaffre

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, structuralism, 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).

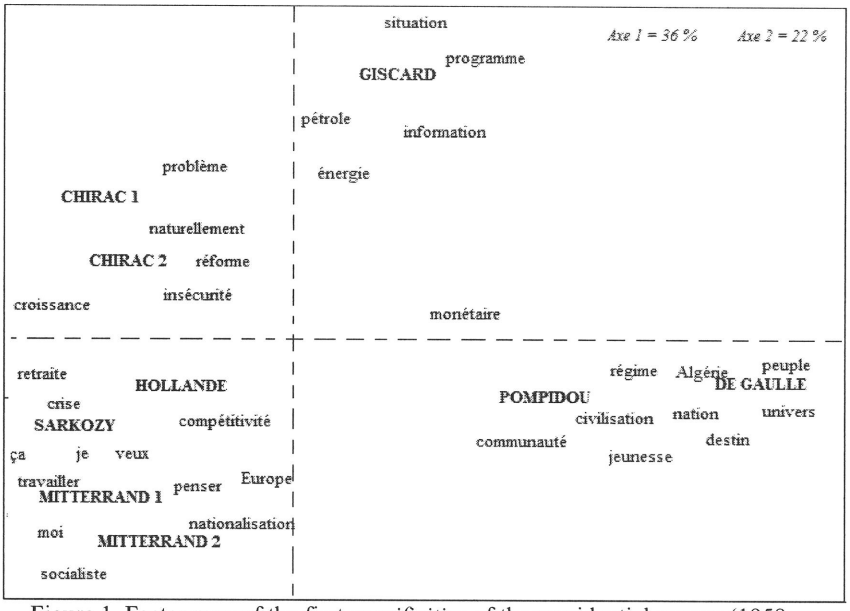

Figure 1. Factor map of the first *specificities* of the presidential corpus (1958-2014)

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).

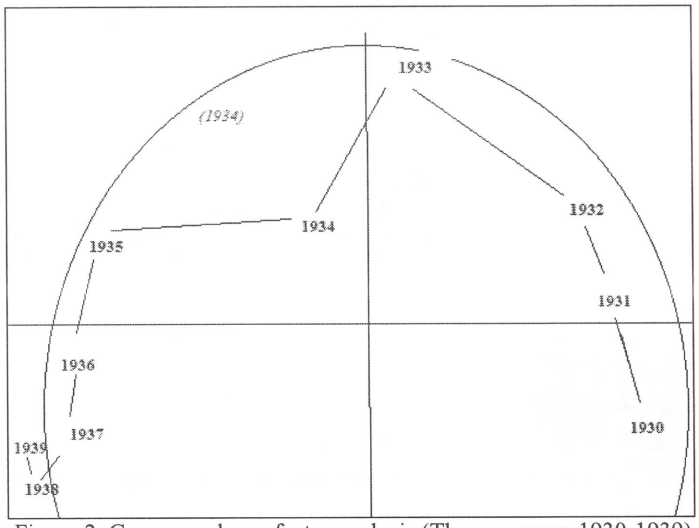

Figure 2. Correspondence factor analysis (Thorez corpus 1930-1939)

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)

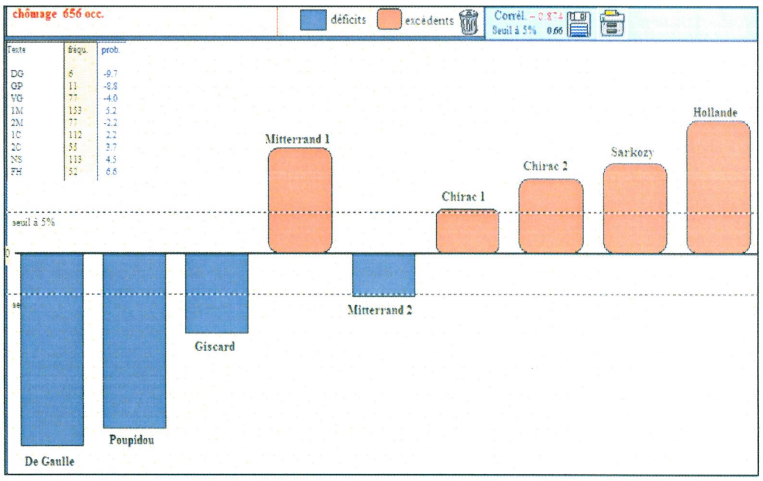

Figure 3. Chronological distribution of "unemployment" (chômage) in the presidential corpus (1958-2014)

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).

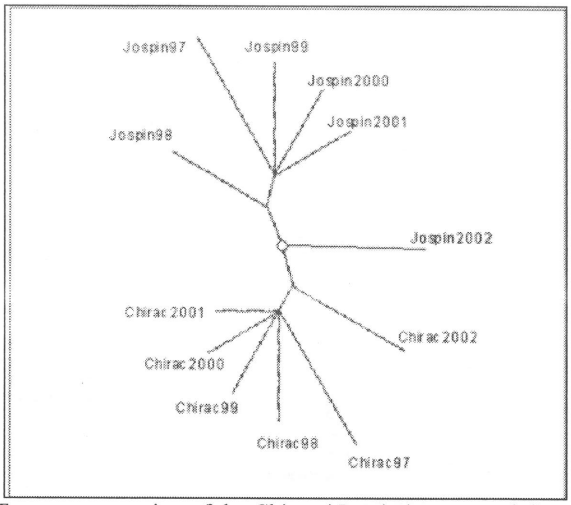

Figure 4. Tree representation of the Chirac / Jospin intertextual distance (1997-2002)

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.).

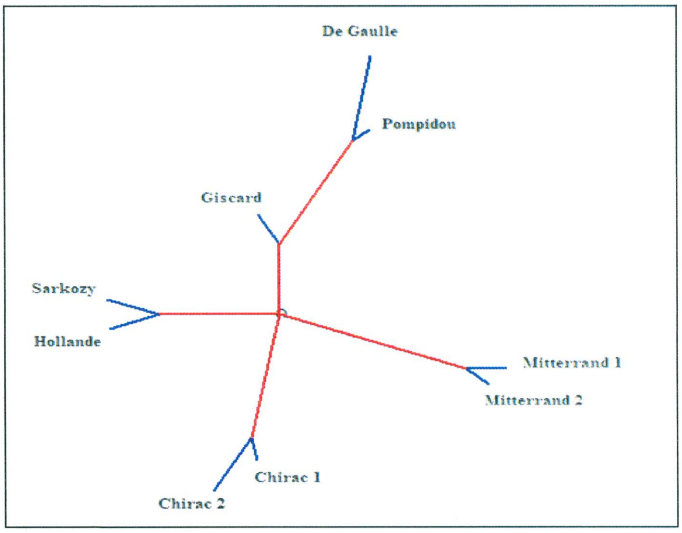

Figure 5. Tree representation of the intertextual distance in the presidential corpus (1958-2014)

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)

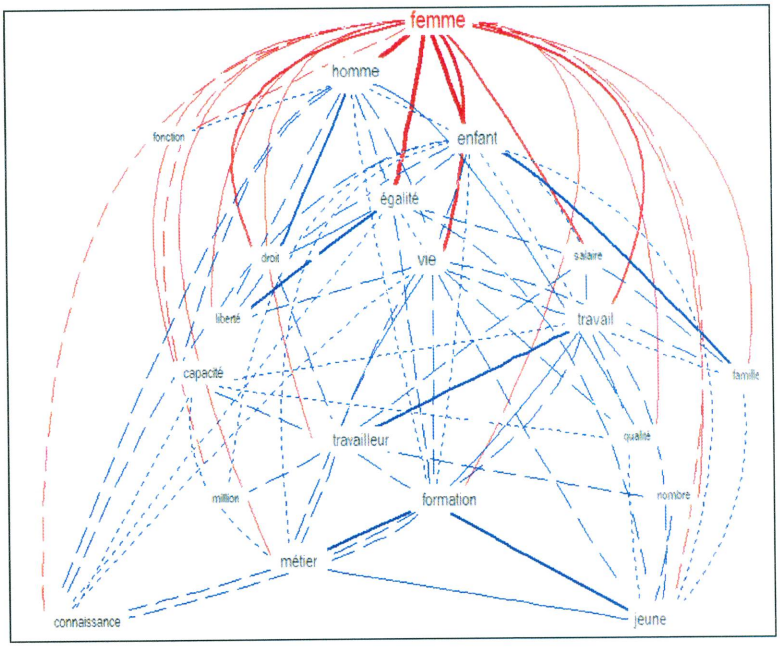

Figure 6. Graph showing multiple co-occurrences from the word "woman" in the presidential corpus (1958-2014) (Hyperbase 10.0 - 2015)

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).

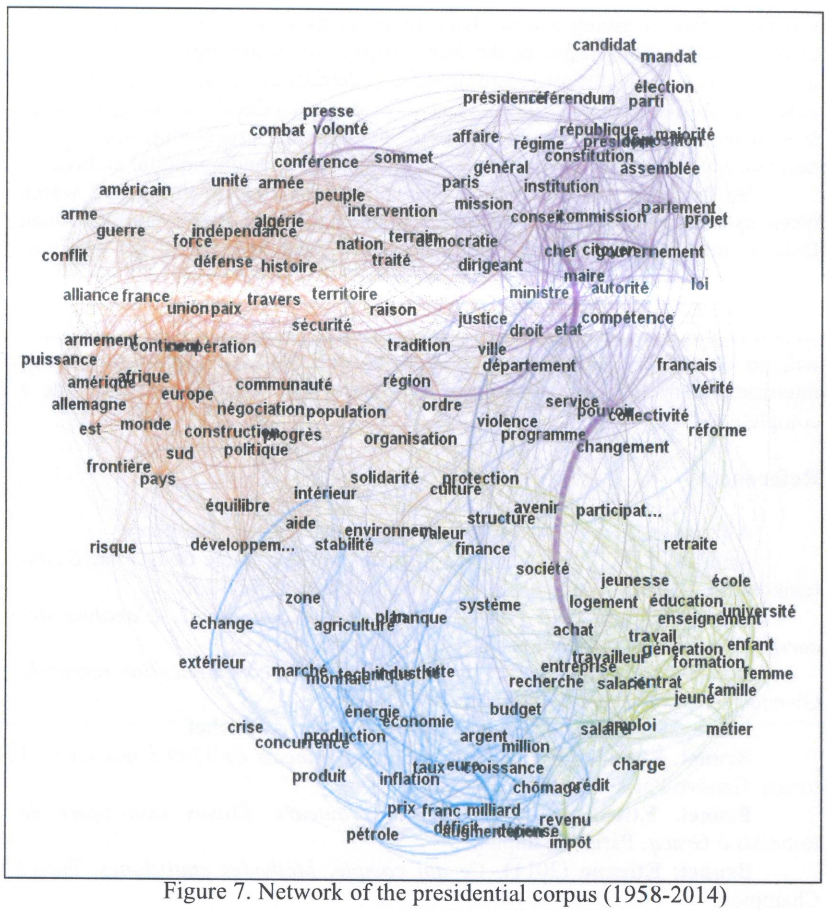

Figure 7. Network of the presidential corpus (1958-2014)

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. Néé, 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.