Quantitative approaches to political discourse: corpus linguistics and text statistics

Damon Mayaffre, Céline Poudat

Quantitative approaches to political discourse

Corpus linguistics and text statistics

Damon Mayaffre & Céline Poudat
CNRS – Université de Nice-Sophia Antipolis / Université de Paris 13

The present chapter proposes to build bridges between political discourse analysis and corpus linguistics. We intend to bring to light methodological benefits arising from the synergy of (political) discourse analysis and corpus linguistics, pointing to fruitful contribution from French text statistics. Taking the discourses of Nicolas Sarkozy as an example, we show how political discourse analysis can benefit from a reflection on corpora (their constitution, their role in the research process); on linguistic analysis and processing methods (particularly the computer-assisted methods of text statistics); and finally on the interpretative paths at a time of establishment of a numerical hermeneutics.

1. Political corpora: Multidisciplinary perspectives

Political discourse has played a key role in the emergence of Discourse Analysis, starting with (Harris 1952) and subsequently in its development as a multidisciplinary meeting point in France during the period 1970–1980 (French school of discourse analysis: Althusser, Foucault, Pêcheux, etc.) and throughout the world during the period 1990–2000 (Critical Discourse Analysis: Fairclough 1995, 2001; Van Dijk 2009; Wodak 2007, etc.).

Since the turn of the century, corpora have become an important paradigm in the Humanities and Social Sciences (HSS). As naturally occurring data, corpora are platforms for scientific observation in empirical disciplines such as History, Political Science, Sociology, and Textual Linguistics. This central role of corpora, which is now acknowledged, has been enhanced by the development of Corpus Linguistics, which started in the post-war period with Sinclair 1991 in the 60's and in the USA with the development of the Brown Corpus, and which has taken a quantum leap in recent years.

The notion of *political corpus* is therefore very specific within the current scientific landscape, most likely due to its obvious multidisciplinary dimension: indeed, while political corpora are made up of linguistic material (words,

sentences, documents, discourses…), their scope is basically social (*e.g.* understanding the political organisation of men in society, where language is of prime importance, analysing institutional functioning, in which texts are determinant, understanding social or ideological balances, exploring the balance of strength and power throughout speeches).

We will recall here the two most significant aspects of current scientific research on (political) corpora, *i.e.* the technical conditions for the development of corpora studies, i.e. the *digital revolution* (1.1.), and the theoretical progress that has been made in corpora definition and conceptualization (1.2.). These two aspects are pivotal for understanding the methodological proposals that will be put forward with regards to corpus processing in Section 2, where we will explore Europe in French presidential discourse.

1.1 Digital corpora: A revolution in Discourse Analysis, Linguistic Studies, Corpus Linguistics, Political Discourse

As indicated by its etymology, the concept of *corpus* is not new. However, it seems to us that the magnitude of this concept has changed over the last quarter of a century with the advancement of the digital revolution, which is shaking the world every moment, and our scientific practices every day.

First impact: the most obvious impact of the digital revolution is, firstly, the leverage that researchers have for building study corpora in view of the available textual (and non textual) resources, and secondly, the accelerated circulation of these resources, which facilitates data collection work.

Basically, all political institutions (international, European or national) are making thousands of documents available every day: political speeches and declarations, directives, legislation, administrative forms. Further, most national, regional and municipal archives are being computerised. In France, all parliamentary debates occurring at the National Assembly and the Senate are, for example, available through the Official Gazette (*www.journal-officiel.gouv.fr/*). Also, the Elysée's website (*www.elysee.fr/)* constantly updates all speeches by the President of the Republic. Finally, the official website Vie-Publique (*www.vie-publique.fr/)* electronically publishes over 100,000 public declarations of the main leaders of the majority or of the opposition.

Furthermore, a large part of human culture will be in digital format in a few years. In December 2010, Google Books (*books.google.com/)* announced that it had already scanned 4% of the books ever published in the world, i.e. 500 billion words (*Science Express*, 16 December 2010). What is extraordinary about this situation

is not so much the immeasurable volume of documents as their immediate availability: with no other restrictions than money in some cases, these billions of documents are directly and instantaneously accessible to the 3 billion Internet users worldwide in 2011, from their laboratories, their homes or anywhere on the planet thanks to mobile phones or digital tablets.

Second impact: the most unexpected impact of the digital revolution regarding corpora is not virtualisation, but on the contrary, a form of materialization of the data: the corpus had been up to now an *idea* or an *ideal* (and could potentially include all data likely to be of interest to the researcher on a given question). It is now, however, a *fact* or an *input*, *i.e. material* (corpora are indeed text inputs). While corpus data had been very difficult to handle, limited in size, and filed in libraries or archive groups, they can now be shared, modified and classified. As a matter of fact, the political corpora we deal with are getting larger and larger and are becoming progressively easier to handle and analyze.

Third impact: finally, the most fundamental impact of the digital revolution is that our relationship with texts and reading and with archives or corpora, is now more virtual: the digital revolution is basically as significant as the Gutenberg revolution was from the 15th century onward. While printed texts laid the foundation of modernity, as stated by J. Goody (2007), digital hypertext is now setting up the foundation of our hypermodernity.

What in fact is a (political) text in the digital era? What is textuality on our computer screens? How do we read the web, an archive, or a corpus today? Does digital writing create "effects of meaning" or does it add value to meaning, a value that was so far unknown to the reader?

The present contribution stems notably from this simple conviction: digital corpora must relate to digital reading, and we will examine this crucial point. Corpus processing methods must basically take into account the digital nature of the data and create a reading protocol which complements traditional reading with computer-based reading. Texts have become *hypertexts*: reading must also become *hyper-reading*.

In concrete terms, thinkers about texts have shown that digital technology goes beyond traditional text linearity (Rastier 2001; Adam 2008). (Digital) texts, long considered as *series* or *progressions*, must also be considered as networks that we understand through *links* or *echoes*. More specifically, we will show how word-for-word, linear or syntagmatic reading can today be complemented by *tabular reading* (i.e. production of tables, indexes, list of words) and *reticular reading* (i.e. system of references and cooccurrence balance); how qualitative or natural

reading can be complemented by quantitative reading (frequency lists and dictionaries, specific vocabularies and statistical indices); and, finally, to what extent can textual reading be complemented by hypertextual reading (browsing, references, links, or bookmarks). This can be summarized as follows: "(Modern readers will read) horizontally, vertically or diagonally, depending on the directions opened by electronic links" (Darnton 2009, French translation 2011: 180).

In any case, the digital revolution has played and is still playing a vital role in the development of new approaches to (political) corpora. In France, the Saint Cloud school and its political Lexicology laboratory[1] first drew attention to the rising digital reality of analysed corpora at the beginning of the 1980's. At that time, researchers built a methodological approach which is still quite relevant today (see Part 2). At the same time, and looking beyond the French context, the development of Corpus Linguistics in the world is also a direct tributary of the digital revolution. Empirical work on corpus data now seems to counterbalance the introspective approach of Chomskyan linguistics, on the one hand because corpora are larger and more representative than they were before, and on the other because corpora studies are now more numerous and relevant.

Finally, work on evidence, which had long been questionable on account of the paucity of data, can now use overlapping digital corpora which are becoming increasingly evidential.

1.2 Corpora as artefacts

The increase in available sources has led the scientific community to pursue a theoretical reflection on corpora and to define various protocols for assessing data consistency and representativeness – for instance concerning genres and text types (Poudat 2006).

In this context, it is important to recall that corpora are artefacts, *i.e.* constructions. The vast amount of available resources might seem exhaustive, but this should not overshadow the fact that collecting texts always results from a choice and from working hypothesis that clearly impact the whole process. A corpus is neither a natural nor an autonomous object: corpora are artificial objects that are designed for the purposes of a specific research project.

1. re. the first issues of *Mots*: *www.persee.fr/web/revues/home/prescript/revue/mots* accessed 31/08/12.

As the questions of corpus representativeness and relevance are partly settled now, since most researchers work on corpora, the remaining question concerns the use of corpora or of their status in the research process. The most decisive theoretical progress in this regard probably comes from Corpus Linguistics, which makes a distinction between corpus-based and corpus-driven approaches (Tognini-Bonelli 2001). Corpora can indeed be conceived as example databases or as testing grounds. Corpus-based researchers create a corpus – or resort to an existing one – to test a working hypothesis, or to draw an example likely to support an argument or a theory. On the other hand, corpus-driven approaches are undoubtedly more widespread as far as political corpora are concerned. Corpora are indeed considered to produce meaning and interpretation: they are not a resource for the analyst but the very source of the analysis. Far from considering the corpus as a collection of stable data, or as a repository of meaning that already exists, we consider the corpus as a meaning matrix (or as the 'producer of meaning'). Provided it is relevantly constructed, the corpus indeed produces meaning itself and is the very basis of the study. In particular, since we postulate that meaning arises through differences, corpora need to be contrastive in order to be productive (political contrast, *i.e.* comparison of several political speakers; chronological contrast, *i.e.* comparison of several periods; generic contrasts, *i.e.* comparison of several genres or types of discourse production, etc.). Further, we have also demonstrated how corpora can gain by being "reflexive" (Mayaffre 2002; see also mayaffre 2012), i.e. by bringing together texts which are mutually enlightening. Each text conditions the interpretation of the next one, and in that respect reflexive corpora internalize their interpretative resources. The basic objective is to take the hermeneutic turn, which is today relevant to all HSS: the corpus (as with all language) must become the condition for its own interpretation. The approach should remain as endogenous to the corpus as possible, and the latter should be, also as far as possible, self-sufficient. Calling upon exogenous or external resources would weaken the scientificity of the approach: indeed, a corpus must be carefully built and analyzed according to strict criteria. In that respect, texts selected according to other criteria are in a way "alien" to the corpus, and should not be part of the scientific process.

Let us conclude this first section with the further observation that corpora, as constructed objects, are being more clearly defined. Long claimed to be natural because they were composed of "raw" texts validated by philology, corpora are now taking on new forms thanks notably to annotation(s): corpora can now be structured, morphosyntactically labelled and semantically enriched. The following analysis will consider different levels of linguistic granularity (from the word form to its part of speech).

2. Tools and paths to explore Europe

Europe is a key concept in French political discourse and numerous methodological paths may be taken to explore its meaning, its impact and its variations. The following analyses are based on two main corpora:

1. *Sarkozy TV* (130 153 tokens),[2] which encompasses French President Sarkozy's TV appearances (interviews and addresses) from May 2007 (corpus in progress). *Sarkozy TV* will enable us to examine the discourse on Europe that is currently delivered to French citizens in order to capture how Europe is represented in current French political discourse, and;
2. *Corpus V* (2 148 907 tokens), which collects the discourses of Vth Republic French Presidents, from de Gaulle to Sarkozy (1958–2010). It gathers 800 general public speeches, including about a hundred mandate speeches. *Corpus V* is diachronically structured, and it will enable us to assess the evolution of the discourse on Europe and to compare presidential speeches in this perspective.

Europe will be examined using two sets of methods developed by text statistics (2.1.) and corpus linguistics (2.2.).

2.1 Text statistics

Text statistics is particularly well developed in France, under the name *Analyse de Données Textuelles*, or *ADT*. This movement, which we mentioned in Part 1, originated in Saint Cloud, France in the 1980s. As it was mainly devoted to the study of political discourse, this approach was different than that proposed by Dubois and his French School of Discourse Analysis, which was developed in parallel under the inspiration of Althusser and Foucault. *Political lexicometry* was based on textual materialism ("deal first with the formal marks of the text before considering socio-political interpretation") and on systematic quantification. From the beginning, it was a corpus-based approach, which built and contrasted corpora using statistical measures such as relative frequencies. In this framework, the norm is endogenous to the corpus, as linguistic units do not have frequencies in Language (Lafon 1980). Building a corpus is thus a far from easy task and, as seen before, many parameters must be considered to guarantee the representativeness and the *reflexivity* of the sample. This has to be seriously considered, as statistics provides results regardless of the data set, and humans have a propensity

2. *Tokens* are segmented units, including punctuations.

to interpret and generalize figures and visual representations without too much hesitation (see Svensson & Stenvoll this volume for a discussion on overinterpretation). Frequencies and statistical measures can only be interpreted within or between corpora and, on top of that, relevant bases for comparison are required since the approach is, as said before, basically contrastive.

To illustrate this, let us start exploring our guiding light *Europe* in French presidential discourse under the 5th Republic. To what extent is Europe an important topic in our corpora? Is *Europe* a common word in French Presidential discourse? Frequency data will first help us to answer these questions. *Europe* occurs 166 times in *Sarkozy TV* and 3,294 times in *Corpus V*. These figures naturally need some comparison within their respective corpora to be interpreted. For instance, the frequency of *Europe* in Sarkozy needs to be considered relative to the other words or lemmas used in Sarkozy's vocabulary. Let us then examine the ranking of *Europe* in the corpus frequency list. As grammatical words such as *de* or *le* are well-known to be the most frequent, *Europe* is far better compared to words belonging to the same grammatical class, *i.e.* other *nouns*. In this case, it becomes clear that this theme is salient in Sarkozy's TV discourses, as it is the 7th most frequent noun in the corpus, after *France* (396 occ.), *Français* (259 occ.), *politique* (253), *pays* (217), *travail* (190) and *monde*[3] (178).

A question that may also arise at this stage is whether Europe is significantly used by President Sarkozy. In other words, is *Europe* distinctive to Sarkozy? This can be assessed using *Corpus 5*, which is representative of presidential discourses under the 5th Republic. We resorted to hypergeometric distributions,[4] which have been frequently used in Saint-Cloud for political lexicometry studies to explore and compare words with their probabilities of occurrence from one corpus division (*e.g.* periods, authors, works) to another. Figure 1 shows how specific *Europe* is from one presidential term to another. Note that specificity scores can be positive or negative, according to the over- or under-use of the given linguistic unit, and scores around zero are considered inconclusive, or neutral. Moreover, the top items are not associated with a score but are labeled +/– *infinity*. Indeed, they exceed the hypergeometric threshold, which is not even given in this case, and are thus *highly specific* (or *highly non-specific,* depending on whether the item is over-, or under-used).

3. Note that idioms such as *tout le monde* (everybody) are included in these counts.

4. Hypergeometric distribution, or Fisher's exact test, is a discrete probability distribution. See (Lafon 1980).

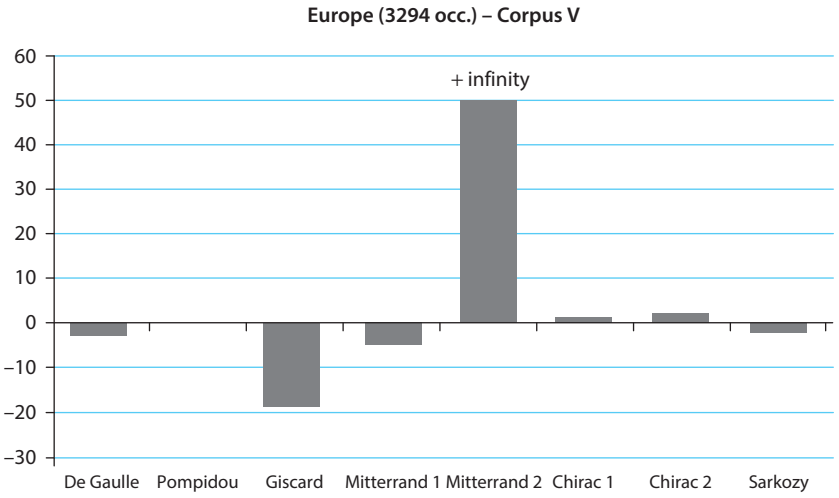


Figure 1. Europe in French Presidential discourse under the 5th Republic

Compared to the other 5th Republic Presidents, Sarkozy infrequently resorts to *Europe* (significant negative use of –2). *Corpus V* indeed reveals another pattern, in which *Europe* is significantly used by Mitterrand; this may be explained by the Maastricht context of the period, notably with the referendum held for the ratification of the Treaty. In addition, Mitterrand had a change of emphasis during his second term and became visibly more involved in international and European affairs – probably to leave a trail in history. Although all of the presidents have – to some extent – contributed to the building of Europe since 1958, the others significantly underuse *Europe* in comparison. Even if the European subject has always been present, it was never fundamental in TV appearances addressed to the French public. In that respect, Giscard d'Estaing's discourse is noteworthy, as he was intimately engaged in *Europe* – he would later participate in the writing of the European Constitution. Yet he took great care to massively underuse *Europe* in his discourse (negative use of –18) and rather used *France* or *pays* ('country') in order to appear closer to the concerns of the French population in everyday domestic affairs.

Figure 2 provides a complementary view of the use of *Europe* among the presidential terms, giving the first factor map of a *Correspondence Factor Analysis* (CFA) that was computed by (Mayaffre 2004: 91) on a subset of Corpus V – note that Sarkozy is absent from the set, as well as the second term of President Chirac:

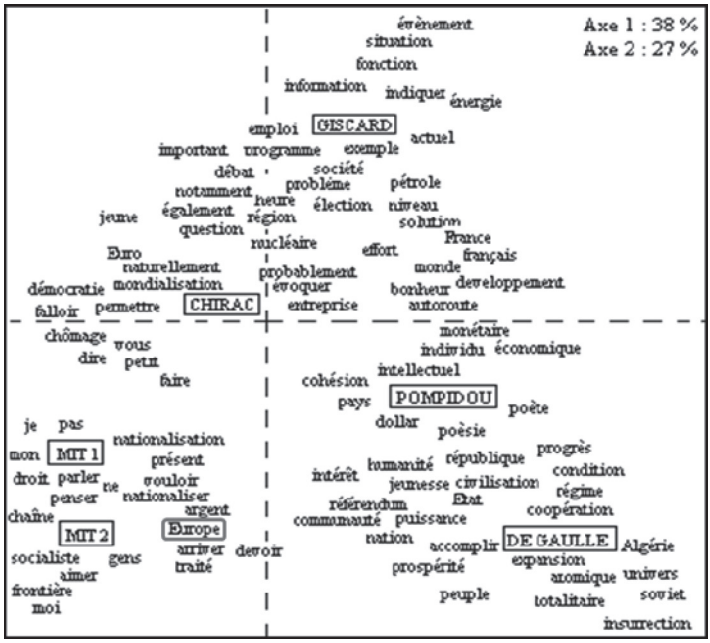


Figure 2. Correspondence Factor Analysis – 120 first lemmas – Corpus V, minus Sarkozy

CFA, which was developed by Benzécri (1973), is a classic multivariate technique in text statistics. Based on a contingency table, the method detects associations and oppositions between individuals (*e.g.* texts, genres, historical periods, politicians, etc.) and observations (*e.g.* words, lemmas, morphosyntactic categories, etc.), which can be visualized separately or simultaneously onto two-dimensional factor maps. This enables researchers to assess the distances between individuals and observations. Figure 2 offers a simultaneous view of both individuals (the French Presidents' terms) and observations (120 first lemmas) that interestingly confirms and refines the preceding analysis: each President has his own priorities, and *Europe* clearly orbits around Mitterrand. On the contrary, the lemma is significantly opposed to Giscard: the distance is indeed the greatest when we consider the positions of the six presidential terms.

As we can see, the methods developed by French political lexicometry at that time were forward-looking and innovative, and are still much used and useful for exploring large corpora. Note that in recent decades, *lexicometry* has interestingly evolved toward *text statistics*, moving beyond an outdated conception of texts as "bags of words". Text structure and sequentiality are now taken into account to contrast and characterize corpora, allowing us to assess for instance whether

Europe is significantly employed at the beginning or at the end of texts, or to observe the distribution of a given form or pattern within a corpus. At the same time, the impressive advances that have been made in Natural Language Processing (NLP), and notably in tagging and parsing, now make it possible to handle annotated or even multi-annotated corpora. Different linguistic levels can be encoded (*i.e.* lemmas, morphosyntax, syntax…) and used – independently or combined – at various stages of the analysis process: the tools we have developed enable us to find and assess patterns using simple or complex concordances or to extract the morphosyntactic specificities of a given speaker. Despite their relevance for corpus processing, these cutting-edge methods are unfortunately not very well known outside of the French-speaking scientific community; this may be partly because most of the programs are in French, with no translation into English, as are some of the key articles that helped shape the field of text statistics. This linguistic restriction has certainly been detrimental to the international visibility of the community.

However, times are changing and a recent project[5] involving the major software designers (Hyperbase,[6] Lexico,[7] Weblex)[8] has led to the development of a stand-alone open source platform which aims at offering the major tools and methods that have been developed to date in text statistics. Even in its current beta state, TXM is a very promising corpus tool: it includes numerous statistical modules, enabling users to perform powerful concordances, to search for cooccurrences, to compute correspondence factor analyses, and to determine which words are *specific* to a given corpus or sub-corpus relative to another. Moreover, the software is in French and in English.

The analyses we conducted in the present article were computed either by TXM or by Hyperbase, which is one of the major tools used by text statistics. Hyperbase was originally developed by Etienne Brunet (UMR BCL, University of Nice) in 1989, and it provides a very wide range of functions, from the calculation of lexical richness to the computation of generalized cooccurrences.

2.2 Corpus linguistics

Text statistics has witnessed the impressive development of Corpus linguistics, which started in the eighties and was made popular with the advent of personal

5. http://textometrie.ens-lyon.fr/, accessed 31/08/12.

6. http://www.unice.fr/bcl/spip.php?rubrique38, accessed 31/08/12.

7. http://www.tal.univ-paris3.fr/lexico/lexico3.htm, accessed 31/08/12.

8. http://weblex.ens-lsh.fr/doc/weblex/, accessed 31/08/12.

computers in the nineties. Concerned with the collection of naturally occurring data to investigate language use, Corpus linguistics is thriving around in the world and has become a leading trend in contemporary linguistics. We will not discuss here whether or not corpus linguistics is a theory or a methodology, because enough ink has already been spilled over that question (see for example Mc Enery & Wilson 1996 or Tognini-Bonelli 2001). Corpus linguistics follows a continuum between *corpus-based* and *corpus-driven* approaches, as explained in Part 1. More inductive, corpus-driven approaches consider, like text statistics, the norm to be external to the corpus, and the studies focus on the distance between intuition and use. On the contrary, corpus-driven studies are more inductive. Like text statistics, they consider the norm as being endogenous to the corpus, which is the empirical material to be described. However, text statistics and corpus-driven approaches differ in terms of their descriptive goals, as the former is clearly *textual* and aims at exploring text dimensions and structuring, and not only collocations (Firth 1957). Moreover, we have no objections to (morphosyntactic, syntactic or even semantic) annotation, contrary to corpus-driven analysts who would rather induce categories from corpora. In spite of these differences, it would be very productive to initiate a discussion, notably on the methodological front. Nevertheless, we will instead concentrate on the similarities and complementarities between the methodologies that are used and that have been designed within the two disciplines.

Corpus linguistics and text statistics both carry out analyses on the patterns used in natural texts. For that purpose, they resort to corpus tools, and notably to *concordance*. Concordance programs enable scientists to search corpora for linguistic items or combinations of linguistic items. An exhaustive list of all occurrences and contexts is then provided, allowing the user to determine the meaning of an item based on relevant interpretative paths (Rastier 2001) or to carry out *qualitative analysis*. Concordance programs are thus essential, and are essentially the main tools corpus linguists resort to. When Baker (2010) lists the most popular corpus tools, he in fact lists the major concordance programs (*Wordsmith tools, Antconc, MonoConc Pro, Xaira, Sketch Engine, Cobuild concordance sampler* and *View*).

In addition to advanced statistical functions, TXM offers powerful concordance searches that would be very useful to corpus linguists, as the above-quoted tools do not have such functionalities. TXM concordance allows users to perform multi-level searches (words, lemmas and morphosyntactic categories, thanks to a Corpus Query Processor[9] that also enables users to use regular expressions) and to

9. http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/, accessed 31/08/12. IMS Corpus Workbench (Christ et al. 1999).

parameter left and right contexts, as well as the global contexts of the occurrences (speaker, genre, text in Figure 3). Advanced sorting options are also available, as is a "return to the text". In that respect, the platform is very useful for both text statisticians and corpus linguists, and it positively benefits from the thorough work that Pincemin (2006) carried out on concordances.

For instance, the concordance in Figure 3 displays the occurrences of *Europe* followed by an item that is not punctuation, allowing us to examine the predicates most often used by Sarkozy when he speaks of Europe on TV.

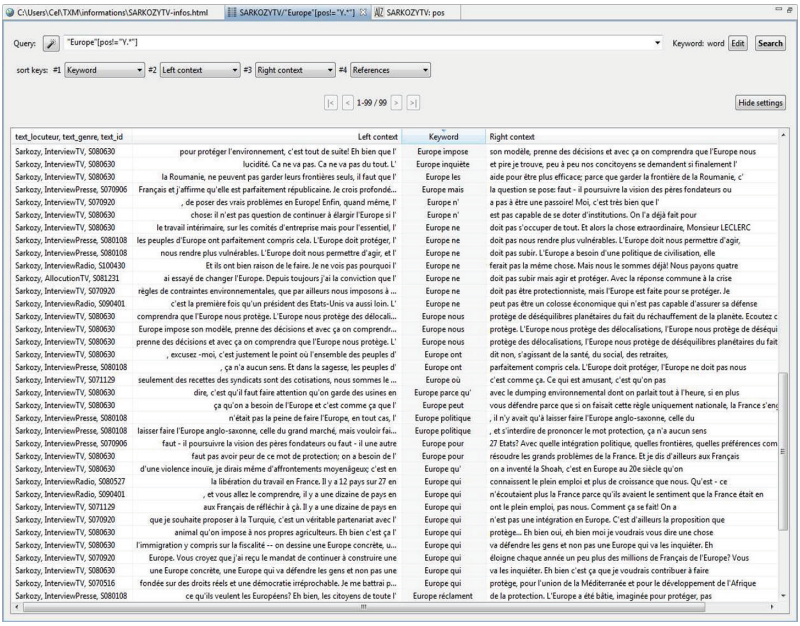


Figure 3. Concordance of Europe sorted on the left and right contexts (Sarkozy TV discourses, TXM)

Depending on the size of the corpus, examining concordances may be quite a heavy task. Corpus linguists have already proposed different ways to get around this problem, such as limiting the observations to 100 concordance lines to examine general linguistic patterns and to 30 lines for detailed patterns (Hunston 2002), or selecting concordance lines randomly. One solution is to use collocation statistics to obtain the frequencies of a node item with its collocates (see WConcord[10] for

10. http://www.linglit.tu-darmstadt.de/index.php?id=linguistics

instance). Table 1 shows the first *Europe* + *verb* patterns we obtained – with one or two words in between (CQP query: "Europe"[]{0,2}[pos = "V.*"]).

Table 1. Searching Europe right collocates – frequencies – TXM index function

Pattern	Frequency
Europe a	6
Europe doit	6
Europe ne doit	5
Europe nous protège	3
Europe, c'est	3
Europe est	2
Europe qui va	2
Europe qui protège	2
Europe ait	2

Europe is clearly a matter of worries in Sarkozy's TV appearances, and the President wants to be reassuring, using *Europe* in contexts where/protection/and/danger/are in constant opposition:

(1) L'Europe a été bâtie, imaginée pour **protéger**, pas pour **inquiéter**. (08/01/2008)
'Europe has been built and designed to **protect**, not to **worry**'

(2) L'Europe doit protéger, l'Europe ne doit pas nous rendre plus **vulnérables**. (08/01/2008)
'Europe must protect, Europe should not make us more **vulnerable**'

(3) On dessine une Europe concrète, une Europe qui va **défendre** les gens et non pas une Europe qui va les **inquiéter**. (30/06/2008)
'We design a concrete Europe, an Europe that will **defend** people, not an Europe that will make people **worry**'

(4) (…) avec ça, on comprendra que l'Europe nous **protège**. L'Europe nous **protège** des délocalisations, l'Europe nous **protège** de déséquilibres planétaires du fait du réchauffement de la planète. (30/06/2008)
'(…) thanks to that, we will understand that Europe **protects** us. Europe **protects** us from relocations, Europe **protects** us from planetary imbalances due to global warming'

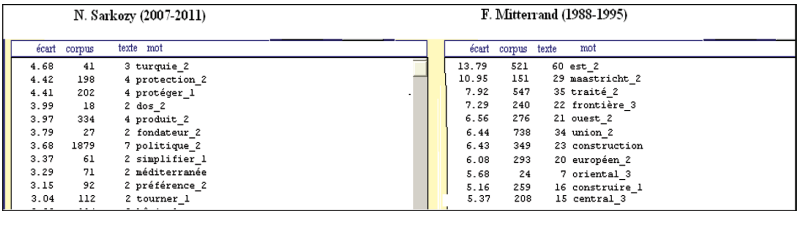
(5) C'est le rôle de l'Europe de **protéger** la sécurité alimentaire de ses citoyens. (30/04/2010)
'The role of Europe is to **protect** food security of the citizens'

In that respect, Sarkozy rejects the point of view according to which Europe may be a danger, using negations in (1), (2) and (3) marking this implicit point of view (for a description of the ScaPoLine framework, see Didriksen & Gjesdal this volume and Gjerstad this volume). Two voices are basically opposed and this opposition can be explained in two ways. One is the political roots of Nicolas Sarkozy, who comes from an old political family that only recently converted to the European idea. The French Right, which he claims to belong to, was at first clearly souvereigntist, with the titulary figure that de Gaulle incarnated. Until Maastricht, it remained eurosceptic, even during Chirac's presidency between 1995 and 2007. On the other hand, French public opinion in 2010 is still affected by the failure of the European Constitution. A large majority of the French, including Sarkozy's voters, indeed expressed a certain degree of concern over, or even rejected, this vision of Europe, which appears more destructive than protective.

Although collocation frequencies are interesting for examining co-occurrences, other statistical measures may be used to refine the analyses. We resorted to co-occurrence statistics to compute the words significantly associated with *Europe*. The software we used takes a node word or expression (in this case, *Europe*) and counts the words occurring within a particular span according to their significant use, computed with the hypergeometric distribution. We generally choose the paragraph span, as paragraphs are well-known to be relevant semantic units.

Since Firth (1957), such methods are essential in text linguistics, as they allow us to objectivize thematic constructions that define discourse, to examine text cohesion and, more generally, to define the discourse meaning of the words. We indeed argue that the meaning of a word depends on its immediate co(n)text; in that respect, we follow Guiraud's old idea (1960) according to which the meaning of a word is not specified in a dictionary but is derived from the sum of its uses in discourse. In this perspective, statistical co-occurrences are the minimal computable forms of the context of a word that participate in its semantization: asserting that A co-occurs with B, C, D, amounts to saying that A is contextualized or semantized by B, C, D.

The results are spectacular and indisputable. Let us for instance examine the statistical universes of *Europe* in Sarkozy's and Mitterrand's discourses (Figure 4).



N. Sarkozy (2007-2011)

écart	corpus	texte	mot
4.68	41	3	turquie_2
4.42	198	4	protection_2
4.41	202	4	protéger_1
3.99	18	2	dos_2
3.97	334	4	produit_2
3.79	27	2	fondateur_2
3.68	1879	7	politique_2
3.37	61	2	simplifier_1
3.29	71	2	méditerranée
3.15	92	2	préférence_2
3.04	112	2	tourner_1

F. Mitterrand (1988-1995)

écart	corpus	texte	mot
13.79	521	60	est_2
10.95	151	29	maastricht_2
7.92	547	35	traité_2
7.29	240	22	frontière_3
6.56	276	21	ouest_2
6.44	738	34	union_2
6.43	349	23	construction
6.08	293	20	européen_2
5.68	24	7	oriental_3
5.16	259	16	construire_1
5.37	208	15	central_3

Figure 4. Cooccurrences of Europe in Sarkozy's and Mitterrand's discourses

In both Presidents' discourses, *Europe* is negatively determined and reflects an opposition shaped by the idea of 'otherness'. Yet the Other changes between 1990 and 2010, and so does European positive identity. In Sarkozy's discourse, Europe is determined together with Turkey (*Turquie*, which is Europe's first co-occurrence), and we know that French President Sarkozy is fiercely opposed to the accession of Turkey to the European Community. In the same way, the European issue is often related to the Mediterranean (*Méditerranée*). Sarkozy sees Europe both as a strict geographical area (Europe vs Asia Minor vs the Mediterranean) and a civilizational reality (Europe vs Arab-Ottoman world, Europe vs Islamic world) – it does not mean for all that that he espouses a view of Europe in a "clash of civilizations".

In the 1990's, the political question was quite different: Mitterrand had to deal with the aftermath of the Cold War and he thought in terms of *East/West* political division, or in terms that are now arcane to us: *central* or *oriental* vs *occidental* Europe.

Co-occurrences especially show that the two presidents have different views on European dynamics: as said before, *Europe* is statistically associated with the noun *protection* and the verb *protéger* ('to protect') in Sarkozy's discourse. Indeed, *Europe*, which de facto exists in 2010, needs to show to what extent it can be efficient and protective for the French citizen. The French governments' liberal policy seems to be rhetorically structured into two contradictory elements: the welfare state needs to be curtailed in the national framework, and curiously, this has to be done within the European context: Europe should indeed insure social, fiscal and health protection. Let us finally add that the protection theme is often met with skepticism, as Sarkozy frequently asserts that Europe disregards (tourner le *dos*, lit. 'to cold-shoulder') its duties and promises.

(6) **En tournant le dos à la Méditerranée**, l'Europe a cru laisser son passé derrière elle mais en réalité **elle a tourné le dos à son avenir**. (06/09/2007)
'Turning its back on the Mediterranean, Europe has believed it was leaving its past behind, but in reality, it has turned its back on the future'

Mitterrand was involved, for his part, in another dynamic, as everything had to be built almost from scratch in the 1990s. The main co-occurrences are indeed the noun *construction* and the verb *construire* ('to build'). Time for recriminations and complaints is not yet at hand, and Europe was still to be figured out, notably with the Maastricht Treaty (*Traité de Maastricht*).

(7) C'est l'Europe que nous sommes en train de **construire**, l'Europe de la Communauté et j'espère l'Europe de la Confédération, l'Europe tout entière. (25/03/1990)
'That is the Europe we are now building, the one of Community, and Confederation, I hope, a whole Europe'

8. Le rayonnement de la France est grand dans le monde, dans cette Europe qu'il faut **construire**, dans cet immense tiers-monde qui a confiance en nous. (31/12/1982)
'The influence of France is great in the world, in this Europe we have to build and in this vast Third world who believes in us'

3. Conclusion

The present paper has explored the concept of *Europe* in French Presidential discourse with a focus on President Sarkozy's speeches. In this perspective, we used and presented a set of methods developed by two trends sharing common interests and intersecting goals: corpus linguistics and text statistics, both of which are concerned with corpora.

Corpora are indeed at the heart of the digital revolution, which has deeply transformed the scientific landscape of discourse analysis – including political discourse analysis. With the increasing number of available digital resources, scientific practices and linguistic methodologies have notably evolved: researchers are now able to build and investigate large corpora, and currently available techniques and software enable us to expand the reading experience by providing tables, figures, graphs, and factor analyses. For instance, the specificities of the Presidents' discourses, which are computed with hypergeometric distributions (as described, supra), can be highlighted in Hyperbase, and this significantly impacts the reading process (Figure 5).

More generally, corpora and digital technology are transforming our relationship to the empirical dimension of language, as the French theoretician François Rastier (2011) underlines in his most recent essay. Rastier examines the experiments corpus linguists have carried out over the last several decades and proposes a reflection on *corpus semantics*, at a time when language data are in the form of corpora and where corpora are in the form of digital data. Using relevant tools, new observables can indeed be retrieved, as corpora are becoming larger and larger, more and more developed, and richly annotated.

"La constitution et l'analyse de corpus sont en passe de modifier les pratiques voire les théories en lettres et sciences sociales. Toutes les disciplines ont maintenant affaire à des documents numériques et cela engage pour elles un nouveau rapport à l'empirique." (*ibidem*, p. 12).

[Corpus constitution and analysis are about to transform the practices, or even the theories in the Humanities and Social Sciences. All disciplines are now confronted with digital documents and this creates a new relationship to the Empirical.]

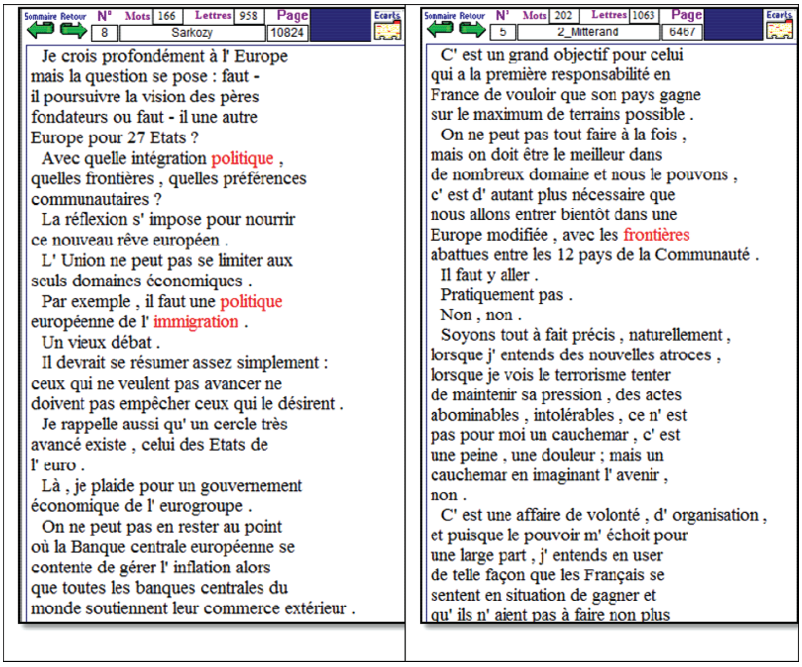


Figure 5. Reading view in Hyperbase – specificities highlighted

In this agenda, text statistics and corpus linguistics enable researchers to combine quantitative and qualitative approaches, and this is what we have tried to show through our analyses. The overall structure of the corpus can be examined quantitatively, whereas the local contexts of the observed linguistic units are determined using a more qualitative approach. Interpretative paths are finally built thanks to a constant back and forth between global and local units, and this leads us towards more regulation – and more relevance – in discourse interpretation.

In French Presidential discourse (1958–2010), the European theme remains ambiguous. Indeed, *Europe* is very present without being necessarily central, except for François Mitterrand during his second term (1988–1995) – and let us recall that the Maastricht context was particular, and even turned out to be historically decisive. On the whole, *Europe* seems to remain a communication issue within a Franco-French perspective. It is always represented as distant, abstract, or external, sometimes as a foil, which threatens the borders of the countries and weakens national identity, sometimes as a protection that may be invoked alongside Nation to challenge globalization. This vision of Europe significantly differs from the findings of both Didriksen and Gjesdal (this volume) and Gjerstad (this volume). It would be particularly interesting to focus

on the contexts where Europe is used with argumentative connectives to refine the results we obtained. On the other hand, Gjerstad, when analyzing Sarkozy's speech at the European Parliament, shows that Sarkozy's argumentative strategy is clearly different there as he conciliates and balances European and national identities, establishing contact and cooperation with quite a heterogeneous audience. Finally, this conciliation strategy may also be run in France as the development of a close, tangible, and intimate Europe will certainly be a central issue in the near future.

References

Adam, J.-M. 2008. *La linguistique textuelle*. Paris: Colin.

Baker, P. 2010. *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Benzécri, J.-P. 1973. Pratique de l'Analyse des Données, Vol. 1: Analyse des Correspondances. Paris: Dunod.

Christ, O. et al. 1999. *The IMS Corpus Workbench: Corpus Query Processor (CQP). User's Mannual*. Stuttgart: University of Stuttgart, Institute for Natural Language Processing.

Darnton, R. 2009. The case for books. Past, present, and future. New York: PlublicAffairs. French translation Apologie du livre (French translation 2011). Paris: Gallimard.

Fairclough, N. 1995. *Critical Discourse Analysis*. Boston: Addison Wesley.

Fairclough, N. 2001. 2nd edition. *Language and Power*. London: Longman.

Firth, J. (1951). 1957. Modes of meaning. In *Papers in Linguistics 1934–1951*. 190–215. Oxford: Oxford U Press.

Firth, J. 1957. "A Synopsis of Linguistic Theory 1930–1955". In *Studies in Linguistic Analysis*, 1–32.

Goody, J. 2007. *Pouvoir et savoirs de l'écrit*. Paris: La dispute.

Guiraud, P. 1960. Problèmes et méthodes de la statistique linguistique. Paris: Larousse.

Harris, Z.S. 1952. "Discourse Analysis". *Language* 28: 1.1–30. (Repr. in The *Structure of Language: Readings in the philosophy of language* ed. by Jerry A[lan] Fodor & Jerrold J[acob] Katz, pp. 355–383. Englewood Cliffs, N.J.: Prentice-Hall, 1964, and also in Harris 1970a, pp. 313–348 as well as in 1981, 107–142.) French translation "Analyse du discours". *Langages* (1969). 13: 8–45. German translation by Peter Eisenberg, "Textanalyse". *Beschreibungsmethoden des amerikanischen Strakturalismus* ed. by Elisabeth Bense, Peter Eisenberg & Hartmut Haberland, 261–298. München: Max Hueber.

Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Lafon, P. 1980. "Sur la variabilité de la fréquence des formes dans un corpus" *in* Mots, Les langages du politique, 1 (1): 127–165.

Mayaffre, D. 2002. Les corpus *réflexifs*: entre architextualité et hypertextualité. *Corpus* 2: 51–69 (http://corpus.revues.org/index11.html).

Mayaffre, D. 2004. *Paroles de président*. Paris: Champion.

Mayaffre, D. 2012. *Mesure et démesure du discours. Nicolas Sarkozy (2007–2012)*. Paris: Presses de Sciences po.

Mcenery, T. and Wilson, A. 1996. *Corpus linguistics*. Edinburgh: Edinburgh University Press.

Pincemin, B. 2006. "Concordances et concordanciers. De l'art du bon KWAC". In *Corpus en Lettres et Sciences Sociales: des documents numériques à l'interprétation*, Actes du XVIIe Colloque d'Albi *Langages et Signification*, Albi, 10–14 juillet 2006, Carine Duteil-Mougel & Baptiste Foulquié (éds), 33–42, and Texto! [web: http://www.revue-texto.net/ ISSN 1773–0120], juin 2006, 2 (2).

Poudat, Céline 2006. "Étude contrastive de l'article scientifique de revue linguistique dans une perspective d'analyse des genres" *in Texto!* [online], septembre-décembre 2006, XI (3–4). Available on http://www.revue-texto.net/1996–2007/Corpus/Corpus.html, accessed 31/08/12.

Rastier, F. 2011. La mesure et le grain. Sémantique de corpus. Paris: Champion.

Rastier, F. 2001. *Arts et Sciences du texte*. Paris: PUF.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Van Dijk, T. 2009. "Texte, Contexte et Connaissance". In *Semen*, 27, 127–155.

Wodak, R. 2007. "Pragmatics and Critical Discourse Analysis: A cross-discipline Inquiry". In *Pragmatic & Cognition*, 15 (1), 203–225.