On Very Large Corpora of French

Etienne Brunet

On Very Large Corpora of French

Etienne Brunet
(Bases, Corpus et Langage, CNRS, Université de Nice)

The first to imagine an automatic (or mechanographical) processing of a large textual corpus is an Italian Jesuit, Roberto Busa (who passed away in 2011; he was nearly one hundred years old). Father Busa liked to tell of a visit he made in 1949 to the headquarters of IBM. In the anteroom leading to the office of Thomas J. Watson, the founder, had provided a sign touting the power and speed of the company: "For emergencies, it's already done. For miracles, it is ongoing." Father Busa brandished the sign under the director's nose and as if he believed in miracles, he got it in the form of a thirty-year sponsorship which lead to the *Thomisticus Index* in 56 volumes, large format, bound in leather.

In France, a few years later, thanks to the support of René Moreau, the director of scientific development of IBM-France, Bernard Quemada and the researchers from Besançon initiated the *Centre for the study of French vocabulary* by extending an earlier undertaking initiated by Wagner and Guiraud as early as 1953 and dedicated to the establishment of the *Vocabulary Index of Classical Theater*.

Using the same method, the Rector Paul Imbs started the lexicographical project that would become the TLF (Trésor de la Langue Française) and where no example was to be found that was not dated and signed. The technique provided only examples, references and statements. At a time when the text input could only be manual, the creation of a large corpus required substantial resources, a long time and much effort, especially as the text input could not be conceived without grammatical correcting and corpus enrichment, related and expensive operations that were self-evident without recourse to the word *lemmatization*.

The word *corpus* itself was then a rare and almost new Latinism to designate the crude product of such undertakings. Looking back fifty years later, one can see in Figure 1 the various fates of some terms associated with the study of language and the remarkable extension of the word "corpus", which participates with a slight delay, in the explosion of linguistics in the 1960s, but without being affected by the decline observed from 1980. Note that this figure comes from the *Google Books corpus* which accounts for 100 billion words pertaining to the French domain and which we will discuss later in this study[1]. The combination of the two words is itself evolving (in the right part of Figure 1): the

[1] As absolute frequencies are very unequal, comparison was facilitated, without distorting, by increasing the lowest by a factor 2 or 3 (or 200 for the rare word *lemmatisation*).

corpus tends to break free from the linguistic tutelage in favour of an association with the text, while conversely linguistics tends to link its fate to the corpus in the expression *Corpus Linguistics* that acquires a sudden favour in 2000.

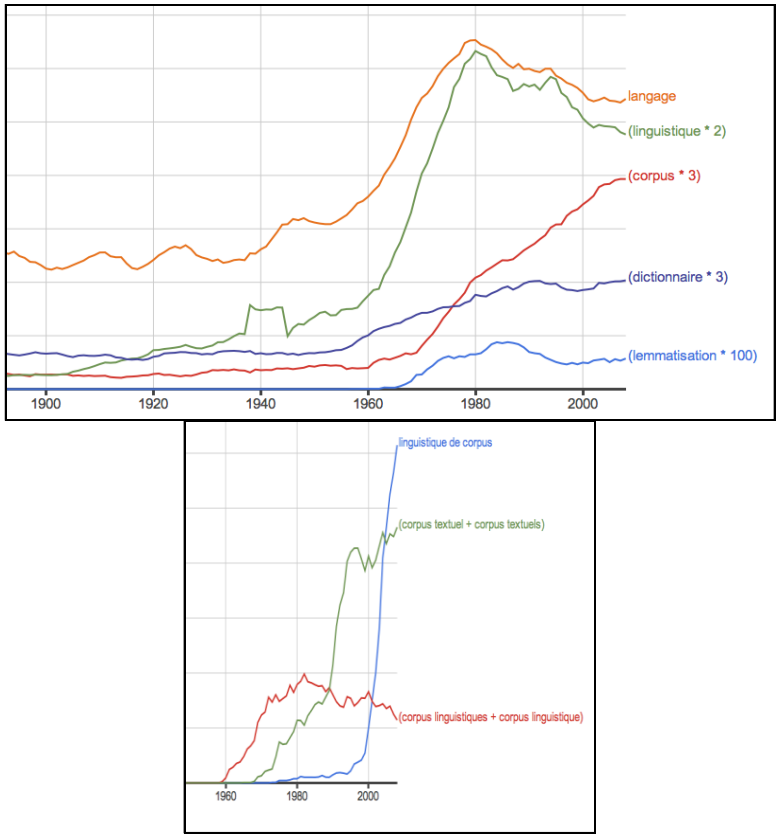

Figure 1. The evolution of some words associated with *corpus* since 1900

If the proper linguistic exploitation of corpora took time before asserting itself, it was probably due to their low availability. To judge the grammaticality of an utterance, the innate or acquired intuition of language seemed a sufficient guarantee and, at a time when the Internet did not exist, it was faster to make a hand-out than to question existing databases.

The TLF data have led, even in the early stages, to some outside services, mainly concordances and indexes, delivered on printed paper[2]. But sales of digital texts were exceptional, especially as the copyright slowed their spread[3].

[2] To prevent ingratitude, we should acknowledge that we benefited without restraint from the Nancy sources which were widely available to us as index or frequency dictionaries. If the text was not transferred directly to magnetic tapes, at least it could be reconstructed from the index and allow monographs, like Giraudoux, Proust, Zola and Hugo established in the 1980s.

[3] This brake still exists even for copyright free texts. CNTRL resource center provides only a sample of 500 texts out of the 4000 *Frantext* ones, half of which escapes the copyright

The gestation of the TLF lasted a long time and during the 1970s, the data processing hardly crossed Nancy borders. And without being shelved, the project remained the prerogative of the TLF editors. Outside, some impatience accompanied this great project advancing slowly, absorbing a considerable part of research funding. In 1978, the corpus was there, almost untouched, and we could write in *Le français moderne*: "The largest world linguistic data base is French. Available. Untapped. And almost unexplored. This vast forest that covers two centuries, 350 authors, 1000 titles, 70 million words, awaits its Livingstone or Stanley[4]. "

I – National projects

1 – The National Library

Concerning French, it would be natural to turn to the French National Library, which is rich in 14 million documents including 11 million books on the Tolbiac site. This would be comparable to *Google Books* offer, if access was similarly electronic. Unfortunately the number of documents accessible on the Internet, mainly in the *Gallica* base, is far from reaching that figure. We certainly have access to the catalog and a sophisticated choice of metadata parameters allows a selection as accurate as you want. But the text itself is often unavailable to the internet user. And when the text is transferred, it is usually readable only in image mode. The transcription in text mode, which is sometimes proposed, is often the raw output of the optical scanner, with mention of the probable error rate. When the rate falls below 99%, that means that a character out of a hundred is questionable or wrong, that is a word out of twenty (the average length of a word being five letters). Naturally the success rate decreases as one moves away in the past, old documents suffering from ravages of time and often offering unusual fonts. These defects are common to every corpus or base founded on automatic reading of documents, but they are more significant in *Gallica* because past centuries are less under-represented. Yet we see in the results only the correct reading of the proposed word, because it does not come to mind to look for erroneous readings. Although *Gallica* is a long-standing base and widely predates *Google Books*, its extension does not have the same scope, thus limiting its statistical interest. The number of usable documents in text mode is limited to 200,000 while its US rival offers millions. And the proper statistical information is minimized, with only the mere mention of the frequency of the desired word. It is difficult with so few elements to establish a curve, let alone a table[5].

(address of CNRTL: http://www.cnrtl.fr/corpus/). It is barely more than the 300 texts which were selected for the cdrom DISCOTEXT twenty years ago.

[4] *Le français moderne*, 46/ n°1, Editions d'Artrey, Paris, 1978, pp 54-66.

[5] *Google Books* does not offer more quantitative information and, similarly, is happy to indicate the number of documents concerned by the query. But that number is of another order

2 - FRANTEXT

In reality, the most reliable texts of *Gallica*, aside from newer ones transmitted by publishers in digital form, are those coming from the Frantext legacy. Those owe nothing to scanning, whose invention in 1974 by Ray Kurzweil is after the initial capturing, carried out by keyboardists on perforated tape. This manual input, duly revised and corrected for fifty years, resisted all changes of systems or supports, passing unhindered from the perforated tape to magnetic tape and disk, and finally to all types of memories available today.

To that reliability of texts, even when they are older editions, Frantext adds many other virtues: a balance between eras, allowing comparisons and providing a solid basis for analysing the evolution of the language; covering a wide chronological span of five centuries of publication; a desired homogeneity of texts whose choice is governed by specific criteria, concerning genre and language level; consistency in the services offered to the scientific community, the same software being kept unchanged for twenty years on the Internet[6]; a moderate increase and a controlled enrichment of data ensuring compatibility with the previous treatment. In brief, in the original draft of the *Treasury of the French Language* as in the derivatives *TLFI* (the digitalized version) and *Frantext*, there is a clear understanding of objectives and a precise definition of the means that have made the French project a model. Now one feature of this model interests us: it is the part played by statistics. From the beginning, TLF reserved for each article a final section where the word's frequency in the whole corpus is noted, but also in the subsets formed by time and genre. Throughout the making of the dictionary, the editors had at their disposal, besides concordances, encrypted information concerning frequencies and co-occurrences[7] and attached to written forms, lemma, parts of speech, expressions and structures. Most documentary and statistical functions that made the success of Frantext were already operational in local mode on the Nancy site or even, in distributed mode, on national networks (Transpac or Minitel) that preceded the Internet. They were also used in the CD *Discotext* produced and distributed in 1984. But it was in 1998, with *Frantext on Internet*, that proper statistical research was greatly facilitated. The use remained primarily documentary and numerical results were quite modest. For, in order not to frighten the literary populations, *Frantext* often merely provides for percentages or relative frequencies. But it is easy to deduce the actual frequencies, calculate variances, and build curves, distribution tables and multivariate analyzes, by opposing texts to each other, or authors or genres

---

of magnitude and proper statistical analysis is performed by a derivative site *Culturomics*, which has no counterpart in the BNF.

6 This software, named *Stella*, was achieved by an exceptional engineer, Jacques Dendien. It regulates Frantext but also TLFI.

7 The name "binary groups" had been given to these co-occurring records, sorted by grammatical categories.

or epochs. The statistical treatment not being fully supported by *Frantext*, the user needs additional and specialized programs.

Figure 5. Statistical analysis of Frantext. *THIEF* database

Our base THIEF (Helping Tools for Interrogation and Exploitation of Frantext) addresses this need by offering the usual array of statistical tools and applying them to Frantext data, whether preloaded or downloaded on demand. In the first case, the data are frozen in the state they were in in 1998, a corpus of 117 million words divided into 12 time slices, from 1600 to 1990. This allows us to see the evolution of the literary language[8] for four centuries and discover lexical and linguistic properties for each period. One works then without connection and without text, on recorded frequencies reachable by the buttons on the top margin of the main menu (Figure 5). Actually the functions spread over the left margin deliver direct access to *Frantext* in its current state. The user is then connected to *Frantext* and can define his working corpus, according to various criteria (title, author, genre, time) and can extract any frequency or textual data he wants. Once saved in a file, the results are taken over by the software THIEF to deliver histograms (by period or author), tables, factor analysis, co-occurrence graphs, etc. However whatever its reputation and merits, Frantext has limited prospects. Fifty years of history overshadow its future. This is due in part to the timidity of its statistical apparatus: simply distributing the text as word lists or number

[8] Technical texts were excluded, to give more coherence to the corpus.

series, at a time when images have invaded the Internet, means depriving oneself of the immediate readability specific to the graphical representation. The most serious handicap is the data: we praised its reliability and homogeneity, but it merely represents a single use of French: high language level, literary and classic. It is the French that is learnt in school textbooks or books you read in libraries. It is not the French one speaks or is used in everyday life, in newspapers and the media. It bears testimony to the culture, not the reflection of current events. The catalogue is now expanding by adding more recent production: it has currently 4000 references and 270 million words. But the BNF weighs ten times more; *Google Books* is a thousand times more and its pace of growth is much faster. Finally, *Frantext* remains constrained by an agreement made with publishers, which limits the size of extracts and forces the user to a prior subscription. This subscription can be justified if it is to communicate a text or a copyrighted extract. But we do not see the legal legitimacy if it concerns quantitative information from the text, whether or not in the public domain. Furthermore, *Frantext* has not fully retained the intermediate solution which would be to export the text, at least the copyright free text, offering it for download so that the users may apply any statistical and computing processing of their choice. This distribution function has been outsourced to a subsidiary organization, the CNRTL, whose current catalogue is too small.

II - Bases and corpora. Encyclopedia

The notion of corpus became extensive and now tends to designate any set of texts likely to be submitted to statistical and computational processing. In principle, one should distinguish structured data, such as a library catalogue, from those that are not, and where the text will scroll continuously. It would be appropriate to call the first “bases” reserving the term “corpus” for the second[9]. Thus *Frantext* is clearly a corpus, while the *TLFI* (or computerized TLF) is a base. The criterion that differentiates them is the presence or absence of a fixed frame having ordered sections that items must fill in one way or another, by a number, a code or text. But the opposition is not absolute: firstly a corpus is usually partitioned into multiple texts and comparing each of which can receive qualifications or metadata: title, author, date of publication, genre, register already constitute an external structure that can continue internally with chapters, acts or scenes, collections or more generally components, disjoint or nested, of textual content[10]. In addition, opinion and market surveys, next to boxes where there appears various coded information (occupation, age, sex, education, income, etc.), often give way in form to a free section where respondents express their opinion without any binding directive. To treat this part of the survey,

[9] This distinction is what prompted the name "Bases, corpus and language" to the laboratory where this research was conducted.

[10] The normalization of textual data is greatly facilitated by the standard TEI guidelines (Text Encoding Initiative) and XML tags

specialized software then uses the same tools used in the processing of corpora. It even happens that structured database information may be processed directly by ignoring or blanking out every structure and tag[11]. Now that most of the historical dictionaries are available on the net or on CD or DVD[12], one could flatten the text and treat it as a corpus. But the statistical significance of such an operation seems low, since there is nothing to define partitions that can be compared. One cannot oppose the words that begin with A to those beginning with B. At most, one might isolate some elements of the structure, such as entries, definitions, examples or quotations, synonyms, areas of application[13].

The enterprise is justified more easily when it concerns an encyclopedia, first because the inclusion of proper names gives the corpus a space-time dimension which a language dictionary is lacking and also because an ontology of knowledge and disciplines takes shape more accurately. We will give two examples borrowed from the editorial news.

1 - **Encarta**

Indeed *Encarta encyclopedia* is no longer relevant since this cultural product, launched in 1993 by Microsoft, ended its existence in 2009, Microsoft having withdrawn it from the market in the face of *Wikipedia*'s dominance[14] in the global network and the *Encyclopaedia Universalis* on the French market. Benefiting from a personal contract with Microsoft in 2000, we had access to the

---

11 This unscrupulous scanning is often practiced by automata that scour the Internet. More or less coarse filters chop up the site pages to draw the best pieces, usually cutting off the head and tail.

12 For example, here is the rich catalogue of Redon editions, to which are added Diderot's *Encyclopedia*, various editions of the *Dictionary of the French Academy* and the *Larousse Universal Dictionary of the Nineteenth Century*.

- Dictionnaire de la Curne de Sainte-Palaye (1876)
- Curiositez françoises d'Antoine Oudin (1640)
- Dictionnaire universel d'Antoine Furetière (1690)
- Dictionnaire de l'Académie française (éd. de 1762)
- Dictionnaire philosophique de Voltaire et compléments (1765)
- Dictionnaire universel des synonymes de Guizot (1822)
- Dictionnaire de la langue française d'Emile Littré (1872 et supp. de 1877)
- Le Thresor de la langve francoyse de Jean Nicot (1606)
- Dictionnaire français contenant les mots et les choses de Pierre Richelet (1680)
- Dictionnaire étymologique de Gilles Ménage (1694)
- Dictionnaire des arts et des sciences de Thomas Corneille (1694)
- Dictionnaire universel françois et latin de Trévoux (1743-1752)
- Dictionnaire [sic] critique de l'Abbé J.F. Féraud
- Dictionnaire grammatical portatif de la langue française de l'Abbé J.F. Féraud

13 Specialized research in proxemy was published by Bruno Gaume from the definitions of French verbs ("For a cognitive ergonomics of electronic dictionaries" in *Document numérique*, 2004/3 (Vol.8), pp.157-181).

14 At that time *Encarta* represented only 1% of Internet queries against 97% for Wikipedia. But it is true that *Encarta* users used the CD more readily. It is in this form that this encyclopedia still continues its career, even though marketing has stopped.

full text of *Encarta*, which Microsoft wanted to submit to our software *Hyperbase*. As was expected, the lack of partitions reduced the interest in this undertaking. However, a function remains that can be applied to any corpus delivered in one piece and based on co-occurrences. This function can be limited to one word, by observing its lexical environment, that is identifying its closest terms, but it can also extend to the whole corpus. A separation technique and progressive refining allows one to decant and isolate lexical themes or constellations that structure the corpus[15]. Applied to the text of *Encarta*, the decomposition process delivers a spectrum of ten colours of which Table 6 details the nuances[16].

1 (science) *vitesse onde rayon énergie surface métal électricité gaz particule atome*, etc.
2 (littérature) *roman œuvre auteur publier écrire poésie poème récit écrivain*, etc.
3 (géographie) *région nord sud département ouest habitant plateau vallée population massif côte*, etc.
4 (arts) *film cinéma carrière cinéaste théâtre scène acteur réalisateur peintre*, etc.
5(guerre) *guerre armée troupe militaire allemand Allemagne britannique allié force accord conflit offensive camp soviétique*, etc.
6 (politique) *président république élection gouvernement ministre politique député parti socialiste républicain*, etc.
7 (pensée) *philosophie dieu philosophe pensée connaissance Christ science esprit idée vérité évangile sociologie pratique âme histoire foi*, etc.
8 (société) *droit loi économique public juge entreprise salarié justice tribunal état privé social*, etc.
9 (histoire) *roi empereur empire fils Charles royaume pape Louis duc trône prince Henri Angleterre dynastie*, etc.
10 (patrimoine) *siècle musée église cathédrale château gothique ancien chapelle édifice Notre-Dame monument*, etc.

Table 6. The disciplinary spectrum of *Encarta*[17]

As can be seen, the editorial board of Encarta hardly deals with geography, circumscribed in section 3, whether physical or human. History is best treated, whether actors of the past, especially kings or monuments that bear witness of the time (themes 9 and 10). But the distinction is made between the old and the contemporary: the theme of war makes clear reference to the world wars (theme 5). And a distinction is made in the arts between the literary tradition (theme 2) and modern performing arts, especially cinema (theme 4). It would be interesting to make a comparison with a similar editorial company and we think of the *Grand Larousse Encyclopédique du XIXe* and the *Encyclopaedia Universalis*. But in both cases, there is no honest way to obtain the full text of these bases,

15 The algorithm used is the *Alceste* software is one.

16 The software (IRAMUTEQ) preventing us from treating all the articles at one time, we merely treated a sample of 2 million word-occurrences, nearly a random tenth of the whole set.

17 The name of the list designated in brackets results from the interpretation of the elements of the list, which are much more numerous than those we supply, for lack of space.

which can be searched word by word but not as a whole. Like most bases available online or in DVD, they answer all questions except those relating to themselves.

2 – Wikipedia

However there is one base which reveals its secrets ingenuously: **Wikipedia**. There is no need to describe this cooperative encyclopedia; everyone uses it daily. What is less known is the capacity to download its content through the site http://dumps.wikimedia.org/, or more easily through the site REDAC offering exploitable resources from Wikipedia (http://redac.univ-tlse2.fr/corpus/wikipedia.html). A first approach, which is indirect, is to identify products and to note to which discipline they relate. While writing Wikipedia articles is free, it is customary to indicate, at the end of the article, to which categories and portals they can be linked, for any aspect of the content. As most articles refer to several descriptors or keywords simultaneously, one can, noting these associations or co-occurrences, map the disciplines represented in Wikipedia.

The software programme *Iramuteq*, that was used for studying *Encarta*, provides, in Figure 7, an unexpected distribution where one can hardly recognize the traditional ontology of knowledge and activities. Two areas are particularly highlighted: on the left (magenta) the performing arts, around the *cinema*, television and music, and at the bottom (dark blue) the circle of *players* who compete for the ball. The divisions of the geographical area complete the triangle and occupy the top of the figure (in red) around the *town*. The rest is confined to the central area, less violently contrasted: three districts nevertheless emerge: the *political* and social sphere (on the right, light blue), biography and history that hold registers of *deaths* and *births* (green, below the origin) and finally (in black, above the origin) a concentration of human activity gathering thought, research, science and industry. Science and technology are not separated. Except for this detail, we find the same main lines Wikipedia includes in its subtitle: "Art - Geography - History - Science - Society - Sports - Technology". Like Encarta, Wikipedia highlights the cinema and all modern arts that are based on the diffusion of image and sound. It adds the sporting field whose spectacular promotion is linked to this diffusion.

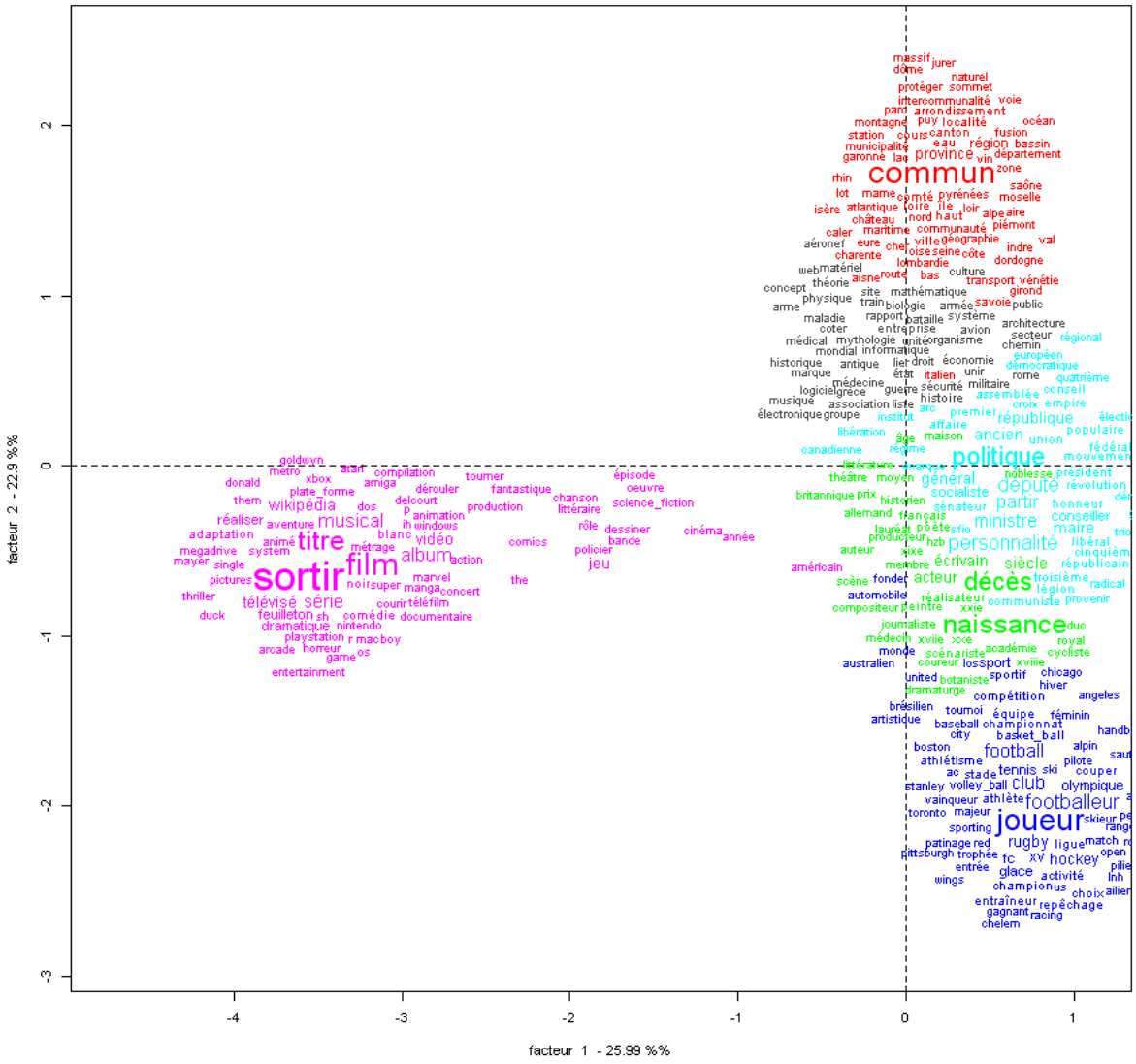

Figure 7. Analysis of the key-words of Wikipedia

Wikipedia is a collective enterprise, linked to individual initiative and free of binding elements of centralization. The underlying ontology that emerges from more than 600,000 articles is based on an improvised provisional architecture[18], 1555 gates, themselves grouped in 11 categories: Arts, Geography, History, Hobby, Medicine, Politics, Religion, Sciences, Society, Sport, and Technology. This ontology is deduced from the keywords ("category" or "portal" in Wikipedia terminology), examined in Figure 7, that accompany each article.

[18] It is not forbidden to add others, provided one first checks that the proposal has no articles preceding them. Control is a posteriori. Instead of offering a predefined framework to be completed cell by cell, decision-makers merely register the proposed portals without banning judgement: within1555 portals, the site admits frankly that only 29 are "good quality" and 50 "good gates".

One can wonder wether a classification of the actual texts of the article will produce the same categories. To limit the volume of data to be processed, one proceeds by sampling, retaining only one article out of ten, and isolating the class of nouns. In a corpus reduced to four million words, the *Iramuteq* algorithm sees a ten-class structure, which only imperfectly reflects the official nomenclature. Of the eleven groups displayed in the organization chart, more than half are certainly reflected in the results, namely politics, science, technology, history, society and sports. But neither religion nor medicine nor hobbies appears independently. As for art, only sound and the image are taken into account, not the written text that is entitled to an independent constituency, covering literature, the press and scientific publishing. Similarly geography comes in two classes, depending on whether town or country is concerned.

Where does the distortion, the difference between summary and content, come from? Any encyclopedia aims to be a dictionary of knowledge, as well as of places and people. Now as the Wikipedia writing mode is based on unsolicited and unpaid collaborations, voluntary contributions are not immune to interest mingled with people. Let us observe the significant list of Table 8: if the names of places dominate in the geographical classes (4 and 5) and abstract concepts in the technical or administrative classes (6, 7 and 10), elsewhere the names of persons take center stage: the writer, the teacher, the philosopher in Class 1; the president, the minister, the deputy, the candidate in Class 2; the actor, director, screenwriter in Class 3; and the player, champion, coach, winner in Class 9. As for Class 8 dedicated to history, it is entirely made up of kinship terms, titles of nobility and ecclesiastical dignitaries. Biographical elements occupy such a large place that Wikipedia is becoming a kind of Who's Who where everyone would like to see his or her picture and his or her medals.

Classe 1 : (écrit) littérature ouvrage édition écrivain livre revue professeur lauréat publication poésie roman université école philosophe presse science journal essai critique …
Classe 2 : (politique) parti élection président politique ministre député parlement gouvernement assemblée candidat suffrage constitution république…
Classe 3 : (image et son) film cinéma acteur réalisateur scénariste télévision internet scénario série comédie métrage réalisation feuilleton…
Classe 4 : (ville) pont bâtiment construction quartier architecture architecte pierre édifice métro ville rue hayteur station boulevard façade mètre route béton …
Classe 5 : (campagne) parc montagne réserve zone superficie altitude sud île ouest faune rivière lac forêt nord région massif vallée eau flore…
Classe 6 : (sciences)exemple forme biologie cas molécule propriété type température acide protéine quantité surface chimie particule phénomène effet équation cellule…
Classe 7 : (technologie) logiciel moteur informatique entreprise système fichier utilisateur processeut ordinateur gamme bit type véhicule technologie vitesse gestion…
Classe 8: (histoire) fils mort évêque fille empereur père prince duc royaume pape comte prêtre bataille dieu trône époux archevêque…
Classe 9 (sport) joueur palmarès équipe club championnat football match sport champion entraîneur cyclisme carrière vainqueur classement sélection hockey rugby…
Classe 10 (société) maire population commune identité évolution période mandat monument géographie compte district personnalité municipalité administration statistique village habitant…

Table 8. Analysis of the text of Wikipedia

III. Monograph-based corpora

Using monographs opens up a fruitful avenue of research. First, the unifying principle of the corpus must be chosen (a language, a theme, an event, a story, an investigation, an author, a review or a newspaper, a time or genre); then texts may be added to the corpus, usually following a chronological axis.

If the texts are big enough, they may be used as sub-corpora for statistical comparisons. If on the contrary, the textual units are numerous and small and one cannot divide the corpus into sub-corpora, we resort to the previous case (the encyclopedias): the entire unstructured corpus must be treated as a single piece, and only co-occurrences phenomena in small contexts may be observed, using the Alceste algorithm (or its Iramuteq implementation).

Such a situation is frequent in the treatment of sociological surveys. For instance, Pascal Marchand and Pierre Rastinaud have conducted a thematic survey about "national identity", based on 18,240 contributions available on the official website of the Immigration ministry (which had opened a forum in 2009). The forums, the social networks or the personal data collection from traffic analysis provide an inexhaustible reservoir for such investigations, some of which may reach gigantic size as soon as industrial, political or commercial interests are involved.

In France, the "classic" methodology, first established by Guiraud and Muller, and then applied in the 'Lexicometry' laboratory, use the corpus as a norm (a reference frequency list) to which its various subcorpora may be compared. There is no external index of the frequencies.

In the political or historical field, data is often public and free, and easier to collect the data. For instance, Damon Mayaffre has analyzed the discourses of several French former presidents using a corpus of three million words. These analyses show that the various former presidents may be distinguished from each other not only according to the subject of their discourses, but also according to their style. Even if he is facing very different situations, the speech of a president remains recognizable. Figure 9 shows that the former president Mitterrand, during his very long period of power, is always characterized by the use of verbs (such as Sarkozy) whereas other former presidents prefer the nominal categories, apart from Chirac who remained undecided throughout two exercises of power.

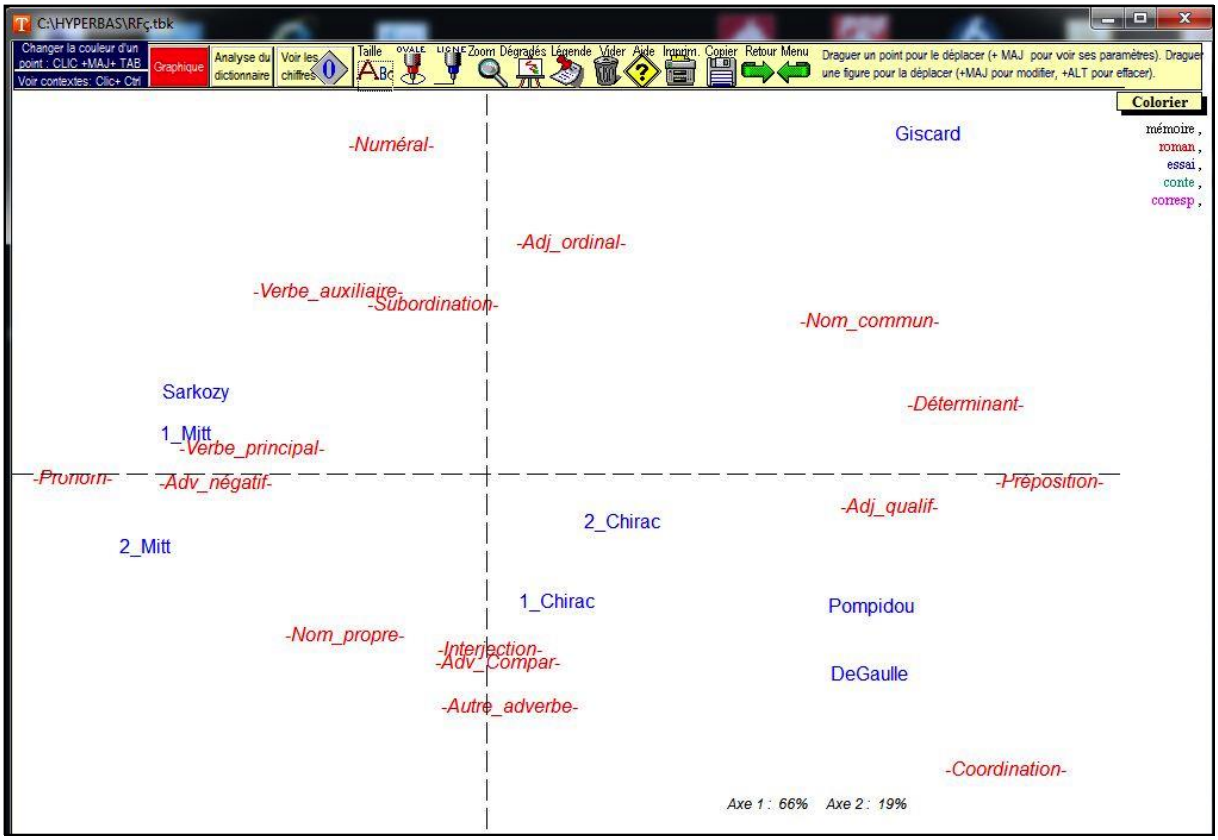

Figure 9. Factorial analysis of parts of speech in former French presidents' discourses.

In the field of literary studies, the typical case is that of studies based on corpora made of all the monographs by a given author. Early works were involved from the very beginning in the study of great texts, like the Bible, St. Thomas Aquinas or Shakespeare. The sizes of the corpora are not expected to defeat those of these pioneering works, due to the lack of prolific writers known to exceed Saint Thomas and Shakespeare... the Nancy 'treasury' (Frantext corpus) included works by many writers but the strategy adopted was to build a corpus balanced according to the works, and this choice prevented the inclusion of full texts in the corpus. The most important "complete works" corpora that could be built (La recherche du temps perdu by Proust, the Rougon-Macquart by Zola, Les Miserables by Hugo) have rarely more than one million words. Professor Kiriu, from Japan, devoted years to scan and correct the complete works of Balzac's Comedie humaine. Other passionate contributions have resulted in the scan of the complete works of Voltaire (Y. and R.D. Boudin), Maupassant (Thierry Selva), Jules Verne (Ali Hefied). Today, using the good sources (e.g. Wikisource or the Gutenberg project) we can almost reconstruct the complete work of a writer, even a very productive writer such as Sand or Dumas, provided that there is no copyright. The size of the corpus can then approach 10 million words. Nothing prevents us from going beyond, if one aggregates the works of writers, and compares them inside a textual genre or an historical era. And from there, one can go further and compare various genres or various eras.

However, today the size of the literary corpora remain below that of Frantext (5000 texts and 300 million words). Outside the literary field, several impressive corpora are emerging thanks to the use of newspapers, magazines and electronic documents of all kinds that are produced every day by the administration, industry, research and the media. One year of a regional newspaper such as the "Est Républicain" is 100 million words that are now offered for download and analysis. Most newspapers now have followed the example of the newspaper Le Monde and are open, in digital form, to retrospective research in their archive. We are able to compare in the same corpus the writings of different newspapers during a given period of time.

However, there is a hindrance to the exponential growth of corpora: the inadequacy of conventional software for operating on such large masses. For instance, I have had to deal with all the issues of the magazine *Europe* published between 1923 and 2000. My software, Hyperbase, was not able to deal with a corpus of 28,000 articles and 58 million words. It is difficult to make it fit in the memory of a personal computer a textual corporus whose size is close to one gigabyte. At this level servers and specialized hardware are needed, that are able to handle the long work of entering, correcting, enriching and indexing data; and to distribute this data using index, pointers and references. But such institutional corpora are like huge tanks that distribute their content, word by word, as would a dictionary. The consultation can be only punctual. They do not allow any overview, no overall analysis, as can be seen from three gigantic corpora of the French language built respectively in Germany, in UK and in the USA.

IV – Corpora of the French language made outside of France

1 – Wortschatz

The first of these three corpora was build at the University of Leipzig (with collaborators from the University of Neuchâtel). It is a corpus of the French language with 700 million words, 36 million sentences from newspapers (19 million), web (11 million) and Wikipedia (6 million). One can base a query on the entire database or on any of its three components. The querying of the corpus may be done through keywords: absolute and relative frequencies are given for the requested keyword, together with some examples (with their addresses in the corpus) and especially its environment, specified in several ways:

- A list of words that co-occur preferentially with the keyword in the sentence;
- Preferential immediate co-occurrents to the left and right of the keyword
- A graph summarizing the most significant co-occurrences

One example suffices to illustrate the results that can be expected from this base (Figure 10)

Mot-clef: Sarkozy

Nombre d'occurrences: 106536

classe de fréquence: 8 (i.e., has got about 2^8 the number of occurences than the selected word.)

exemple(s): Lundi depuis Pékin, M. **Sarkozy** avait lancé un appel à "l'apaisement" resté sans effet. (source: *http://fr.biz.yahoo.com/27112007/202/sarkozy-absent-fillon-en-premiere-ligne-face-aux-violences-de.html*) Il est le cousin de l'ex première Dame de France, Cécilia **Sarkozy**. (source: *http://fr.sports.yahoo.com/29122006/29/aime-jacquet-tire-sa-reverence-avec-un-palmares-unique.html*) **Sarkozy** se prévaut d'une baisse de 9,44% des crimes et délits et d'un taux d'élucidation en progression de 8 à plus de 34%. (source: *http://www.rtbf.be/info/international/ARTICLE_080090*) exemples supplémentaires

cooccurrences significatives de Sarkozy:

Nicolas (838584), président (51762.4), " (48587.6), UMP (36403), M (31913.1), a (30256.3), Royal (25565.4), Ségolène (25195.9), Elysée (22596.7), ministre (22117.5), Intérieur (20576.1), présidentielle (18656), Cécilia (15904.3), François (14851.6), Bayrou (14453.7), Villepin (14450.1), candidat (13928.3), l (13536.4), ' (12823.7), , (12757.2), français (11182.9), Chirac (10215.3), élection (10159.6), visite (9974.35), Fillon (9635.87), Dominique (8574.88), campagne (8432.2), PARIS (7793.58), Carla (7591.07), France (7580.98), avait (7430.25), politique (7166.54), discours (6749.46), son (6746.22), République (6676.53), Bruni (6566.09), etc

voisins de gauche significatifs de Sarkozy:
Nicolas (1326360), Cécilia (17948.7), président (9428.68), . (5228.97), Jean (3338.48), candidat (2430.55), Monsieur (1104.34), Président (774), loi (664.77), couple (595.96), Carla (564.47), Mr (527.76), Guillaume (423.36), Cecilia (386.61), circulaire (365.12), sauf (334.23), Nicoals (301.77), Bruni (295.11), voter (264.16), présidence (260.31), Nicolas (259.26), Royal-Nicolas (253.44), monsieur (246.57), battre (234.84), etc

voisins de droite significatifs de Sarkozy:
a (53559.9), , (28722.9), . (14068.3), avait (11441), s (6645.14), et (5227.08), n (3751.04), veut (3193.77), " (2796), est (2542.41), lors (1744.72), ? (1641.62), ne (1631.29), doit (1446.9), devrait (1353.74), souhaite (1325.86), était (1163.67), etc

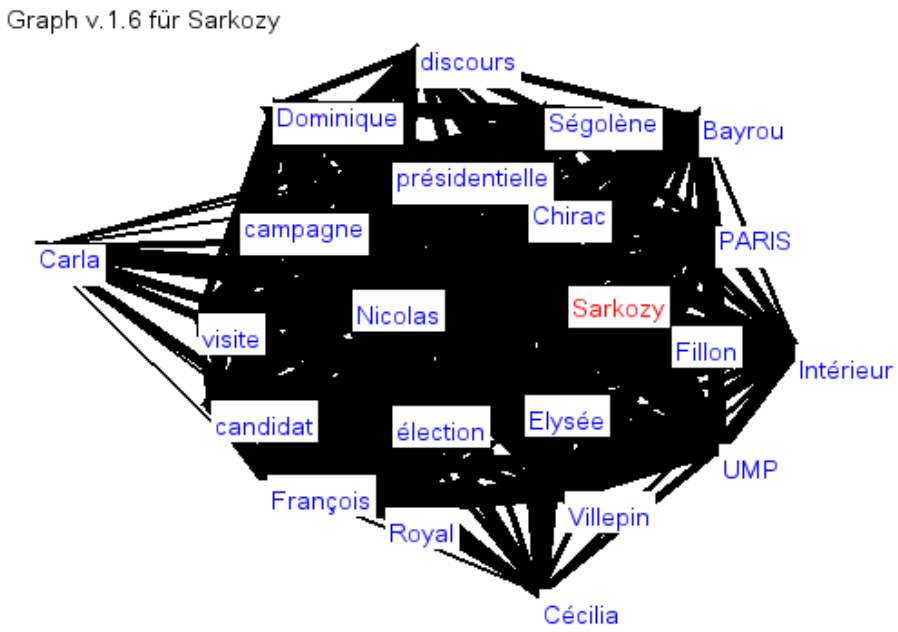

Figure 10. *Sarkozy* according to Wortschatz corpus

2 - Sketchengine

Sketchengine is an English website which offers (together with corpora of other languages) a corpus of the French language. This corpus is over ten times larger than the Wortschatz corpus. The range of tools is also much wider. Sketchengine has several points in common with the Frantext corpus: the user needs to subscribe – for reasons of profitability and not, as for Frantext, of copyright –, the freedom to download at least large extracts; and the possibility to manipulate complex objects: lemmas, codes, structures. However, there are also differences. Frantext has its own data. Sketchengine harvests the web. The former focuses on books and full texts, the latter on short contexts. The former is diachronic, the latter is synchronic.

Like many web-based corpora, Sketchengine is harvesting the web in order to build a large representative corpus of a language rather than to build corpora targeted at analyzing lexical innovations. The starting point is a list of a few hundred words of medium frequency, which is the seed of the harvesting. To reap the harvest, thousands of requests on Google, Bing or Yahoo are made in search for pages that contain at least three words from the list. The pages are collected in a cumulative corpus with the associated metadata (at least the address and the title of the site as well as the date of the request). Next, the duplicate pages are eliminated (thanks to the "onion" software) as well as extra-textual content (thanks to the "justext" software). Various filters are then applied: the document must meet several conditions: be of sufficient length (at least 500 words), contain a minimum proportion of grammatical words. This automatic control based on simple criteria is helpful for removing many unsuited pages: the relationship between what is retained and what is tested is between 1/10 and 1/1000. The corpus is balanced between various sites in order to increase the corpus diversity. Such a process can harvest up to 1 billion words per day. The collected data receives linguistic processing to ensure lemmatisation (TreeTagger is used for Western languages) and a host of statistical operations to enable a sophisticated consultation.

One of the simplest requests of users is often for a concordance. The concordance tool provided by Sketchengine gives the context (line or sentence) for several kinds of queries: word-form, lemma or complex query with various filters. It can also analyse the distribution of the keyword and rank co-occurring words according to the kind of grammatical relation they have with it, and according to the strength of the statistical attraction for the keyword. For instance, an analysis of the word *Samedi* shows that the co-occurring words are most often *dimanche*, *dernier*, *prochain*, *pluvieux*, *ensoleillé*. It reflects the major role played by the weekend for people. If one considers the profiles of other days of the week, one gets a sociological typology of the days of the week. If one considers the profiles of the months of the year, one gets a sociological profile of the season. Enquiries into sociological representations are available with keywords such as freedom, justice, equality, community, or deadly sins.

For example, the reputation of French politicians on the internet can easily be observed. It is somewhat reflected in Figure 11. This factor map is the result of a factor analysis of the contexts where there is a mention of one of the major politicians of the fifth republic in France. This corpus is extracted from the FrTenTen12 corpus of Sketchengine and has been built for representing 37 French politicians (Presidents, Ministers or party leaders) through 5,000 randomly selected occurrences of each of them. The 279 most common nouns in the corpus are then selected (they include by definition the names of politicians involved, each with at least 5,000 occurrences). The contingency table cross-tabulates these words (at the intersection of row i and column j there is the number of co-occurrences between the words i and j). For a survey of men with various historical statuses, it is not surprising that the first factor reflects the timeline. The timeline, evident in proper names, is also observed among common names. Those found on the left belong to political events of the 2000s (electoral campaigns and especially the 2007 and 2012 presidential campaign). It shows the competition between candidates (*sondages, campagnes, votes, débat, déclaration, programme, émission, media, opinion, parti, soutien, militant, candidature, primaire, présidentielle, tour, résultat, victoire*). The confrontation is less harsh in the opposite half of the factor map, on the right. The politicians, there, have left the political scene. We see their work rather than their ambition and the history rather than the current events.

The second (vertical) factor does not separate the left and right political tendencies. It might have be the case if the discourses of politicians have been included in the corpus. But the corpus is not about what they say, but about what is said about them. And in the words about them, the right and left tendencies can coexist. *Mitterrand* is close to *De Gaulle*. Public opinion tends to classify people according to their rank. The presidents of the republic occupy the upper part of the figure. Prime ministers are relegated to the lower half, where *Balladur* is close to *Rocard*, *Jospin*, *Villepin*, *Mauroy*, and *Juppé*. While presidents are characterized with the lexemes referring to the general objectives of politics (*peuple, famille, homme, femme, pays, société, valeur, liberté, justice, loi, démocratie, politique, guerre, mort*), prime ministers are concerned with administrative management of current affairs (*ministère, cabinet, comité, conseil, commission, directeur, secrétaire, conseiller, assemblée, groupe, chef, membre, député, maire, poste, finance, université, presse, fonction, réforme, emploi, etc.*).

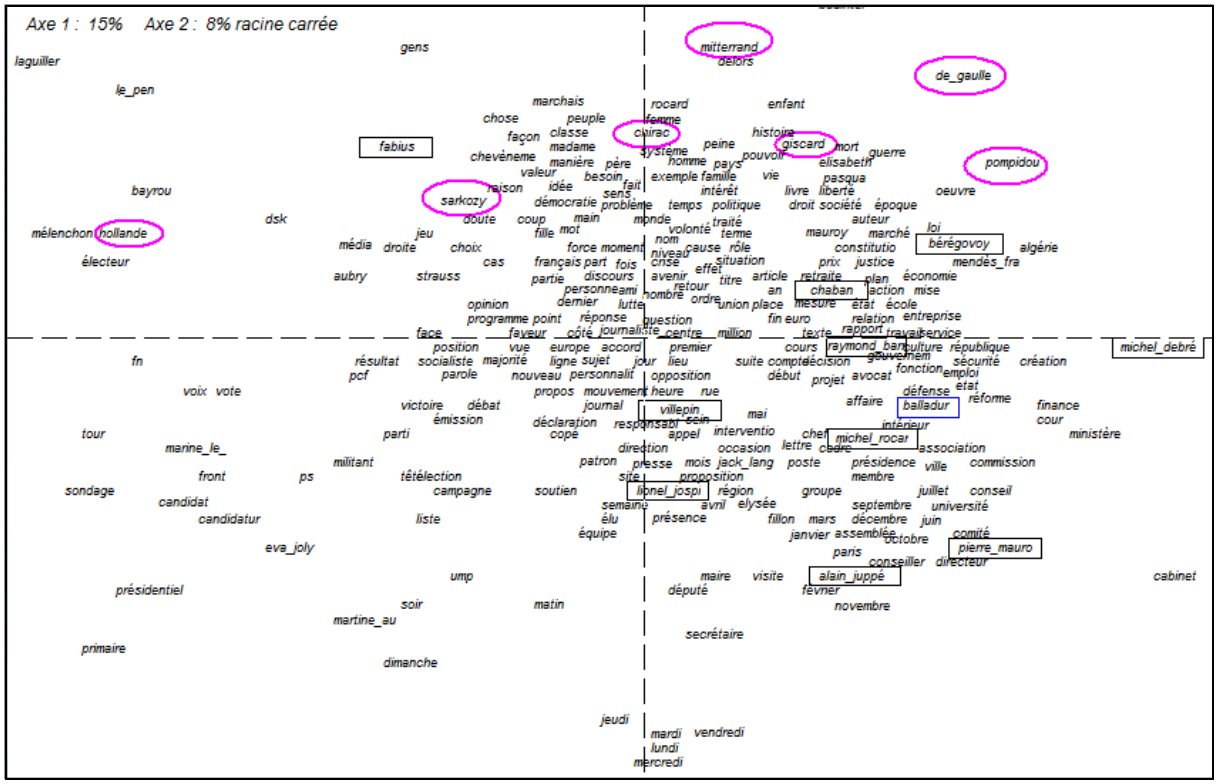

Figure 11. Factorial analysis of co-occurrences in a corpus about French politicians (axes 1 and 2).

3 - GOOGLE BOOKS

Such analyses of word distribution are not possible on the Culturomics website we met earlier (Figures 2 and 3). If the contexts are readable in Google Books, they are no more readable in Culturomics, where only indirect pieces of information are available: ngrams, or text sections whose lengths do not exceed five words. However, Google offers significant advantages in quality and quantity. By its size, it is the biggest corpora of the French language, with a size ten times greater than that of Sketchengine (almost 100 billion words in 2012)[19]. While the diachronic dimension is absent in Sketchengine, it extends over centuries in the Culturomics corpus, opening fruitful avenues of research on the history of words and realities of which those words bear witness. The quality of sources is also a strong point in the Google corpus. The internet contains a mix of all kind of

[19] Between 2009 and 2012, the size of the French corpora has doubled, as did the corpora of the other languages. The current figures at the time of writing are 89 billion words for the French language, 349 the English language (with several dialects), 53 for German, 67 for Spanish, and 33 for Italian, the last corpus built. These figures correspond to the data that can be downloaded. They are higher in the first table of the article published in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, p.170. Three other corpora are availabe (Russian, Chinese and Hebrew).

discourses and varieties. The methods used by Sketchengine are unable, despite all the filtering, to be immune against the barbarisms that are frequent on the social networks. As for Google Books, since it includes only books, such as the BNF and Frantext, it gives access thereby to a certain level of language and culture, that Facebook can not guarantee. A simple survey gives the measure: the ratio between the incorrect *fesait* and the correct *faisait* is 2.6% in Sketchengine while it drops to 0.6% in Culturomics. Google Books is still far from the focus on literature found in Frantext, since it accepts all published works, especially in technical news, and social or media domains. But the barrier of printing protects it against insignificant verbal diarrhea that is spreading on blogs and social networks.

Jean Véronis, who has just left us, did not hide his enthusiasm for the birth of *Culturomics* at Christmas 2010. He had also greeted the 2012 version that corrects some defects of the 2009 version and multiplies its power and flexibility. The queries are no longer restricted to word forms or phrases. It is now possible to ask for lemmas (eg faire_INF to ask for the details of word forms of the verb *faire*), and for the bare part of speech (_DET_ for determiner) or to use wildcards (such as *), and to select the corpora (symbol ":"). A handicap yet was still preventing the use of *Culturomics*: *Culturomics* was delivering curves only, instead of the underlying numbers, and it was not possible to make further analyses using the raw numbers. The authors of Culturomics have therefore released an API that for a given word gives the 201 frequency counts observed along the timeline from 1800 to 2000. Better still: the raw data used to make tables and curves were delivered for free download, which we used to form a base offering the analysis of unigrams (or individual words) of the French domain.

As an illustration, Figure 12 summarizes the syntactic evolution of the sentence in French. The verb and its acolytes (pronouns, adverbs and conjunctions) lose ground to the benefit of classes related to the name: nouns, adjectives and prepositions. This trend is not unique to French: it is found for the same period in other Western languages. This trend however may be a little suspect. There is the suspicion that this change reflects not so much a change in the use of French, as a change in the composition of the corpus. Recent modern texts are the most numerous and are frequently about technical issues. In those text genres, there is an impersonal style, and information is passed via the nominal categories. On the other hand, the older times are represented by literature, more than by science and technology. The verb is more present in literary discourse, and dialogues are more "personal". The variation of the textual genres between periods may have created a heterogeneous corpus, giving the illusion of an evolution.

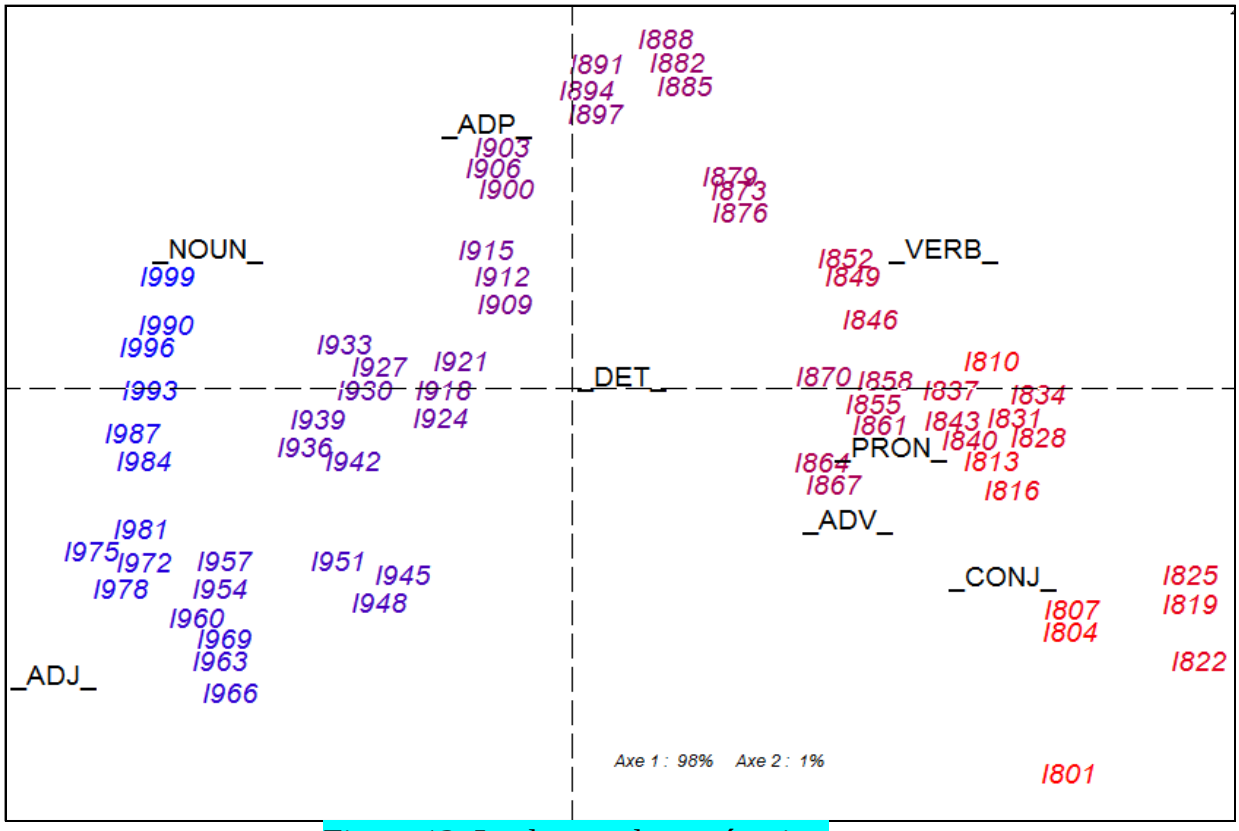

Figure 12. Le dosage des catégories
(24 milliards de substantifs,10 milliards de verbes)

As we can see, one can be enthusiastic given the huge size of the corpora. But the doubt remains as to the validity of the statistical results. The doubt grows especially as the composition of the corpora are still "black boxes". As we saw, even a graph based on large corpora may still be sharply criticised. If the choices underlying the building of the corpus under scrutiny are unknown, the size of the data does not prevent the result from being very difficult to interpret. In such situations, one can talk of "insecurity" in large corpora, as did the reviewers of my book "Vocabulaire français" – which was however based on a corpus a thousand times smaller[20].

[20] Annie Geffroy, Pierre Lafon (1982). *L'insécurité dans les grands ensembles. Mots 5, 129-141.*