Vocabulary diversity and its variability: A tool for the analysis of discoursive strategies. Application to the investiture speeches of the Spanish democracy

Ramon Álvarez[1], Mónica Bécue[2], Juan José Lanero[3]

[1]Área de Estadística e Investigación Operativa. Universidad de León. Campus de Vegazana, s.n. 24071 León. España. e-mail : dderae@unileon.es

[2]Departament Estadística i Investigació Operativa. Universitat Politècnica de Catalunya. c/ Pau Gargallo, 5 - 08028 Barcelona. España. e-mail : monica@eio.upc.es

[3] Departamento Filología Moderna. Universidad de León. Campus de Vegazana, s.n. 24071 León. España. e-mail : dfmjlf@unileon.es

Abstract

Our target is the analysis of discoursive strategy through style-statistical tools and statistical multidimensional descriptive methods which yield answer elements and allow the characterization of both structure and construction of speeches from a lexicometric approach; to these tools, we suggest to add up the study of the variation of the vocabulary diversity, presenting the working-out process and the results obtained over the corpus of the Spanish democracy's Heads of government's investiture speeches.

Key-words : vocabulary diversity, discourse analysis

1. Introduction

Within the study of political speeches, the discoursive strategy is an outstanding element which has to be analyzed. Statistic tools can contribute to it. In this work, we will centre ourselves on the tools that enable to segment the corpus into topics. We will use the Hubert-Labbé vocabulary growth model (Thoiron et al., 1988). We also propose a new tool that consists in the study of the vocabulary diversity variation; it reveals itself as a useful complementary tool to the vocabulary growth study. The results will be presented using the corpus of the Spanish democracy's Heads of government's investiture speeches.

2. Spanish democracy's investiture speeches

The corpus studied is made up of the speeches between 1979 and 1996 (67197 occurrences and 7374 different words), which correspond to the last Spanish democratic period.

The only Head of Government who has delivered more than one speech is Felipe Gonzalez; in a first study (Álvarez et al., 2000), it has been verified that Gonzalez's four speeches do not have a vocabulary particularly common; they do not present either similar style-statistical indexes either. The 82 speech is possibly the one which presents bigger differences with the rest, which corresponds to the first time the socialists came to power and means the end of the transition in Spain.

Table 1 reproduces some quantitative results of the first study of this corpus.

	Suarez	Calvo-Sotelo	Gonzalez				Aznar	Total
	79	81	82	86	89	93	96	
Length	12140	8281	9426	11354	7591	8142	10263	67197
Diversity mean. Sections of 300 words	171	168	175	154	166	167	170	167
% own words	8,7%	7,3%	8,6%	5,3%	5,8%	6,5%	6,3%	7,0%
Originality Index of own words	1,2	1,1	1,3	0,70	0,82	0,97	0,90	1
Originality Index of total hapax	1,2	1,1	1,4	0,66	0,80	0,99	0,90	1
Originality Index of partial hapax	0,99	1,2	1,2	0,69	0,99	1,1	0,96	1

Table 1. Some quantitative results of the speeches

3.Vocabulary growth

3.1 Vocabulary growth models

The notation established by Muller (1977) is used. The vocabulary growth study consists in analysing the rhythm with which the V vocables (different words) appear in the corpus made up with N words (occurrences). In a given corpus, when the corpus has achieved the length N', V' words are counted. The observed growth curve is the bivariate plot (N',V') for N' varying between 1 and N.

Muller (1977) and Hubert and Labbé (1994) suggest to fit the (N',V') points by a smooth curve which will be considered as the regular growth for this speaker in the present communicating situation. The first author only considers one vocabulary class, the second ones take into account that there is a general vocabulary and a specialised one, that allow to obtain a better fit.

We use the last fit. It allows to work out, in any point of the corpus, the expected number of words that would have been used from the beginning under the hypothesis of a regular growth. The marginal growth will be the difference between the number of words observed and the number of words expected. A topic break is denoted by a new word flow.

The vocabulary growth model can be applied to the whole corpus when it exists a chronological order and when they are due to the same textual source. It can also be applied text by text, that is to say, successively to each of the seven speeches. In other words, the model can be used as a global tool or as a local tool. The different results must be interpreted within the context used.

3.2 Vocabulary growth of the seven speeches

In the whole corpus, the seven texts are made up by four different candidates. Despite of this fact, we could consider the possibility of studying them as a whole because they correspond to the same textual source. Figure 1 shows the differences between the vocabulary growth observed and expected.

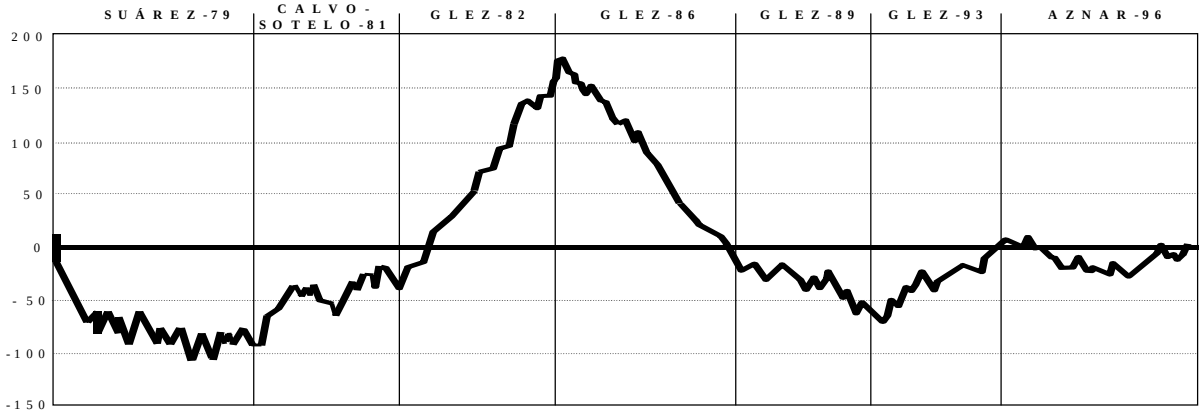

Figure 1: Differences between observed and expected vocabulary growth

One could see that the first two speeches provide relatively few new words, and that Gonzalez-82 breaks this trend and introduces a brand new vocabulary. Surprisingly enough, when the right came back to power, with Aznar's government, his speech does not correspond to a rush of new words.

In order to interpret the big vocabulary flow within González-82 speech, it is useful to work out the originality. Originality indexes can be calculated using the total hapax (a word is a total hapax if it is pronounced only once in the whole corpus), the partial hapax (a word is a partial hapax if it is only used once in a text) and the one for the own words (a word is an owned word for a speech if it is only used in this speech). It can be seen (table 1) that the three indexes give very similar values. A value superior to 1 indicates an originality over the mean, a value inferior to 1 indicates an originality under the mean.

Gónzalez-82 introduces a brand new vocabulary and is the most original speech of them all what means that many new words will not be fixed in the political vocabulary of the following one.

3.3 Vocabulary growth of Suarez's speech

The following figure shows de vocabulary growth curve across Suarez's speech. One could see a big difference, should we compare it with the first part of figure 1, which gathered all the speeches as a whole.

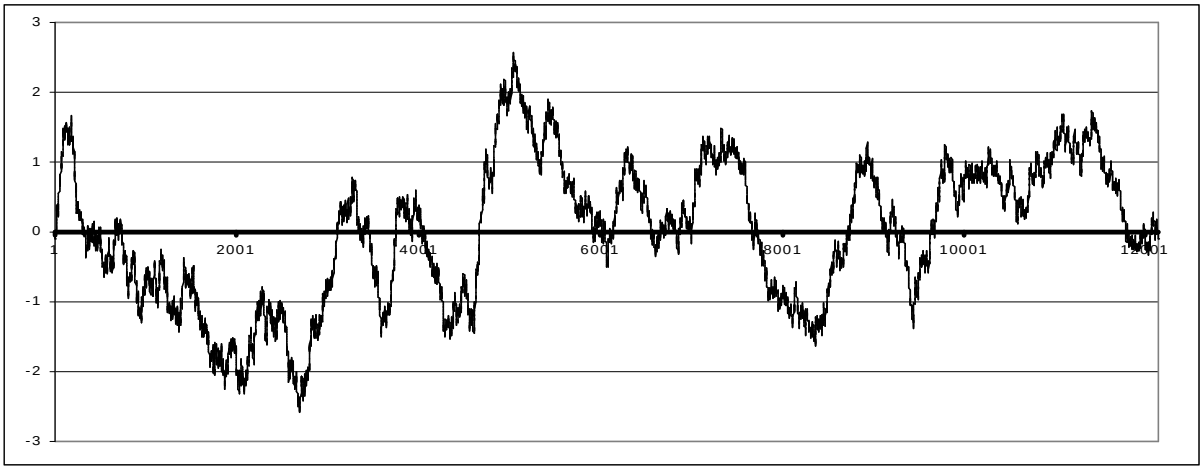

Figure 2: Vocabulary growth of Suarez's speech

4. Diversity

4.1. Diversity index

The vocabulary diversity index (Hubert et Labbé, 1990) of a corpus measures the speaker's capacity to avoid word repetitions, in other words, his capacity to diversify the vocabulary and to know how to use synonyms and nuances which denote a good knowledge and handling of the language.

The diversity index is worked out as the mean of a number of different words mentioned in all the blocks of *n* consecutive occurrences which can be built up starting from the N total occurrences. The obtained diversity index values have no sense unless in comparison.

In our analysis, we have proved values of 300, 500 and 1000 for *n* and kept the results obtained with 300. The problem of the missing information for the *n-1* first words has been solved supposing that the end of a speech could be considered linked to the beginning in order to determine the topic breaks.

A low diversity index could be interpreted as a low vocabulary richness, but also as the result of the wish to send a clear message.

4.2. Diversity variability

As it has been noted in paragraph 3.1, a topic break can be translated by a new words rush. But, generally, it is true only when a new topic appears, not when the speaker comes back to a topic dealt with earlier on. The study of the diversity variability will help solve this problem out.

A high variability or dispersion would indicate important diversity changes, that is, frequent changes in the speech topic.

The diversity growth along one section of a text could be interpreted as the appearance of a new topic. Sometimes, the variations are produced in a very reduced space, generally due to style variations such as repetitions to reinforce an idea (we cannot allow..., we cannot allow ...).

The diversity variability must be interpreted depending on the topic dealt with. When the speaker talks about specific topics which allow the use of a broad vocabulary (for instance: foreign politics or the economic globalization), the diversity growth could reach levels which are significantly high. Nevertheless, the appearance of topics whose contents are "narrow", that is, very specific aspects, it implies a lesser but quicker diversity growth and then the diversity falls considerably.

For this particular reason, in the study of the diversity it is useful to set up the typified values in order to determine the importance of the variations and also to allow the comparisons among different speeches and to eliminate the effect of the difference of vocabulary richness in each of them.

4.3. Contributions of the diversity and the diversity variation

An analysis was made for the seven speeches as a whole and later on each one in an independent way, working out the diversity mean and analyzing its variation, considering sections of 300 words.

Figure 3 shows the compared results of the growth curve and of the diversity variability curve. One could note that as far as Suarez's speech is concerned, the diversity clearly is above the growth curve, whereas for Gonzalez-86 is under it (see results table 1). It is precisely at this point where attention should be drawn. It seems that Gonzalez-82's growth and Gonzalez-86's fall coincide with the values obtained for the originality (studied in the hapaxes and own words) and for the diversity.

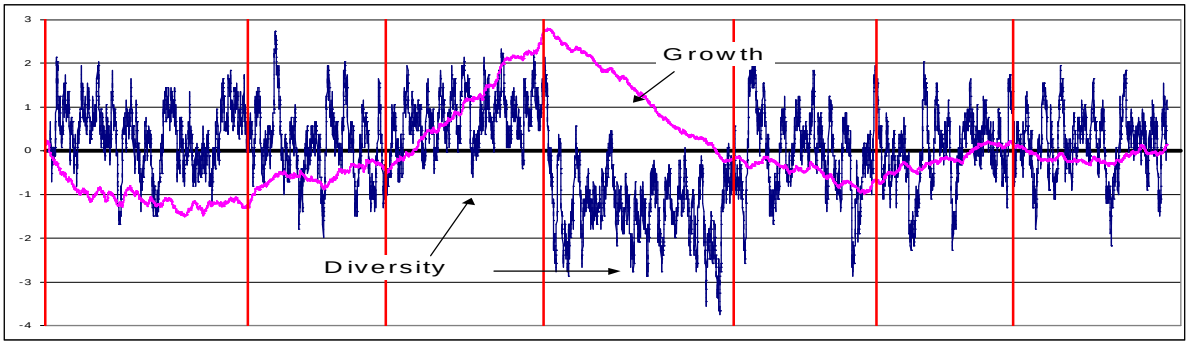

Figure 3: Differences in vocabulary growth and variability diversity

It is difficult to quantify and qualify certain extralinguistic, paralinguistic and metalinguistic factors because, for example, the parliamentary majorities lead the protagonist through different tracks. Let us see one clear example: Gonzalez had an absolute majority in 1982, whereas Aznar in 1996 (and Gonzalez in 1993) were the leaders of the biggest group in the chamber but without the necessary majority; that implies the need to compromise what appears in the results of the correspondence analysis.

5.Individual study of the speeches

Now, despite external circumstances determine that a speech has its own characteristics which distinguish it from others said by other candidates or even from the one he could offer at a different moment, what produces significative effects on the construction of figure 3, we think it is necessary to study the speeches individually in order to be able to compare them later on.

Among the seven speeches, we have chosen the first one (Suarez) to show how useful it is the study of the diversity and its variation. Initially, after a thorough reading, we divided the text subjectively into forty parts in order to being able to confirm whether the topic breaks detected using the diversity and the vocabulary growth correspond to the subjective divisions. The results for the sections of 300 words happened to be more sensitive to the topic variations than the bigger sections and, therefore detected topics already dealt with or mentioned previously. The effect is similar to the choice of the number of periods in the moving average of time series.

5.1 Comparison of the contributions of the vocabulary growth and the diversity variation

In order to compare the vocabulary growth and the diversity as tools to detect the topic breaks, both results were categorized, getting the following figure for a diversity worked out over sections of 300 words:

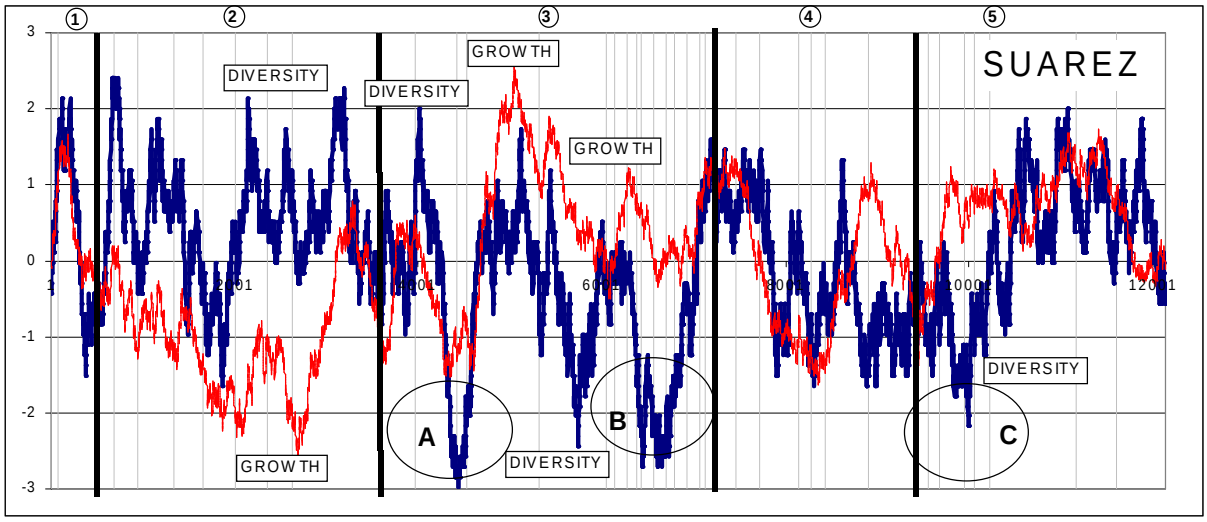

Figure 4. Diversity and growth evolution in Suarez's speech

On the ordinate axis both the standardized diversity and growth are expressed, whereas on the horizontal axis the length of the text is represented. In a first part (1) one could find the introduction and background with similar drawing of the curves. In a second phase (2) the diversity takes high values and is above the growth curve, what shows that generic topics are dealt with, decreasing at half of this period. This descending part makes reference to general aspects of the government manifesto presented by Suarez and the last one refers to foreign politics with a growing part and a decreasing one which would indicate the topic is exhausted.

In the third part (3) one can see that in most of the drawing the diversity is below the growth, what implies the appearance of specific topics with a very "narrow" vocabulary and therefore repetitive. It is necessary to stress the low levels of diversity marked with the A circle (antiterrorist fight), circle B contains very specific topics; between them there is the fifth section corresponding to employment, which grows rapidly and decreases gradually once the topics is exhausted.

The fourth part (4) contains general topics about the future, family (both curves fall), education and culture, only to get an increase with the bill of rights finishing off with unresolved reforms.

In the fifth part (5) very specific topics appear such as the reform of criminal law, the reform and restructuring of the Central, Public and Local Administrations and other institutions (circle C), to finish right off with the autonomous regions and the final conclusions.

5.2 Detailed study of the third part

Now, let us give some considerations over the contents of the third part previously mentioned. The diversity is represented with thick strokes (figure 5); one could note, within the section marked with letter A, that the diversity falls whereas the growth raises, what indicates that the Defense topic has recently been dealt with (first part of A) and it will not be anywhere else. In B the diversity raises as a consequence of dealing with economic aspects of the army; by the same token, C is characterized for indicating the problems of the economic crisis in order to improve the public safety (growing diversity). The first part of D has got decreasing diversity containing the institutions which should guarantee the public safety, whereas the second contains the terrorist problem. Part 14 has a first section E with the economic and world crisis, F with the imbalance of the regions and the emigrations, and G with Spain joining the EEC.

There are two differences between diversity and growth: H with the imbalances initially named in section F, and I with the topics of world unemployment and inflation (the diversity increases but the growth decreases because these topics are deeply dealt with in some other places of the speech).

Section 15 (employment policies) is subdivided in two parts: J with employment problems, inflation and unemployment, and K where the diversity increases due to the appearance of new terms such as "employment insurance", "social security benefits" and "fraud fighting", for instance.

Section 16 corresponds to structural reforms and 17 to taxation. There is a difference marked with letter L due to the fact that taxation was expressed with the structural reforms. The following sections were made *a priori* corresponding to very specific topics, what shows the quick diversity decrease. They are: 18 (labour relations), 19 (public companies), 20 (financial and banking system), 21 (health service), 22 (industrial restructuring), 23 (energy), 24 (agriculture) and 25 (housing). The big distance between diversity and growth which separates most of them, marks the very specific topics.

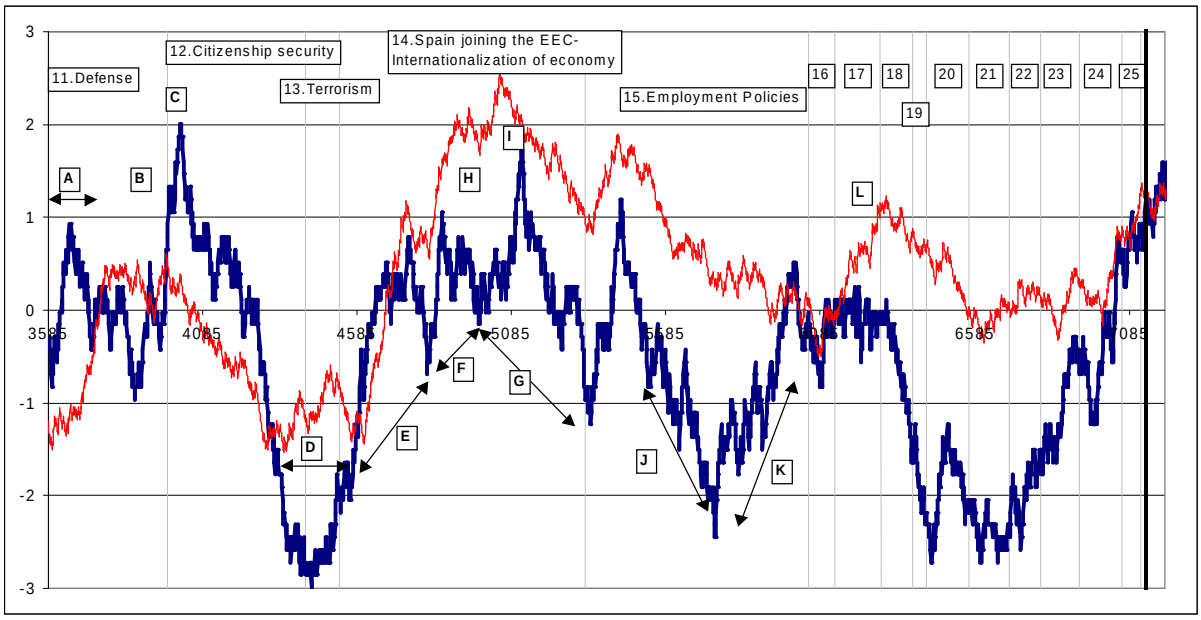

Figure 5. Diversity and growth evolution in the third part of Suarez's speech

The following figure has been obtained as the difference between diversity and growth ; one could note how the general topics (negative differences) are to be found in the first part of the text, whereas the specific topics have positive differences, mainly in the central part.

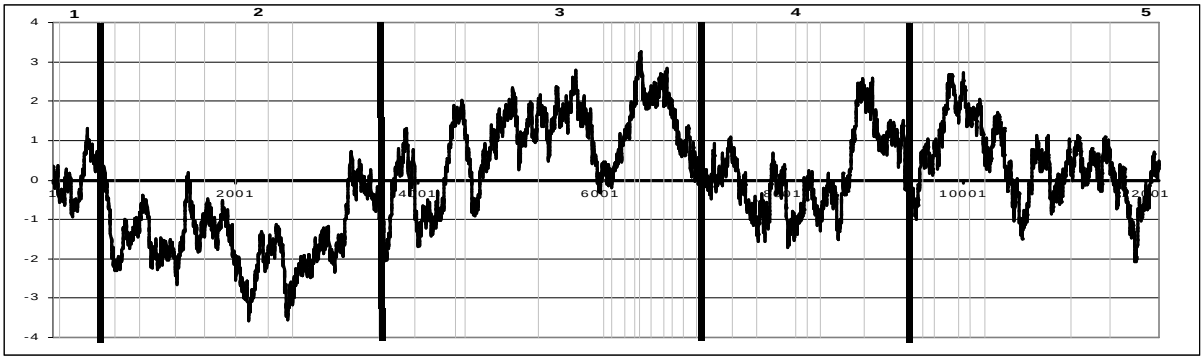

Figure 6. Evolution of the difference growth minus diversity

6. Conclusion

Establishing the parts of a speech is somehow subjective, but it could be supported by the vocabulary growth analysis and the diversity variability based on some other analysis. The results are, in global terms, similar, despite some differences have to be stressed:

- Strong growing slopes of the curves indicate the appearance of new words, corresponding to general topics when the growth curve is below the diversity and to specific topics on the contrary.
- The decreasing slopes indicate the exhaustion of a topic, which can be shown with deep falls or in a staggered way, in the shape of steps of a ladder. In the latter case it is due to the appearance of subtopics.
- When the growth curve increases whereas the diversity remains constant or decreases, it could indicate the appearance of new vocabulary within a very specific subtopic, that is the case when talking about the Public Administration (part of circle C in figure 4), or a topic to be found almost exclusively in that section. This sloping difference tends to disappear when the size of the diversity sections increases.
- If the growth curve drops or remains constant but the diversity increases, this would indicate the appearance of a topic which also appears somewhere else in the text, although not close to it, because in the latter case the diversity would decrease.
- High values of the growth curve and the diversity indicate either generic topics or specific ones which allow a broad vocabulary; in other words what in linguistics is called specialized language.
- Low values of the growth curve and the diversity, indicate very specific topics dealt with a quick way, whose semantic load is so well defined that its appearance is conclusive, avoiding any kind of paraphrase.

References

Álvarez R., Bécue M. and Lanero JJ. (2000). Le vocabulaire gouvernemental espagnol 1976-1996. *Mots*, nº62.

Bécue M., Álvarez R., Lanero JJ. (1999). Etude statistique des discours d'investiture de la démocratie espagnole (1979-1996). Technical Report DR 99/10. Departament d'Estadística i Inv. Operativa, Universitat Politècnica de Catalunya, Barcelona

Hubert P. and Labbé D. (1990). La répartition des mots dans le vocabulaire présidentiel, *Mots*, n°22: 80-92

Hubert P. and Labbé D. (1994), La richesse du vocabulaire. *Communication au Colloque de l'ALLC-ACH*, Paris, pages 19-23 avril 1994.

Labbé,D. (1990). *Le vocabulaire de François Mitterrand*. Paris, Presses de la Fondation Nationale des Sciences Politiques.

Lebart L. and Salem A. (1994). *Statistique Textuelle*. Paris, Dunod.

Muller Ch. (1977), *Principes et méthodes de statistique lexicale.* Paris, Hachette.

Thoiron Ph., Labbé D. and Serant D. (1988), *Etudes sur la richesse et la structure lexicales*. Champion-Slatkine, Paris-Genève.