Automatic Segmentation of Texts and Corpora

Cyril Labbé, Dominique Labbé, Pierre Hubert

Automatic Segmentation of Texts and Corpora

Cyril Labbé
Université Grenoble I
Cyril.labbe@imag.fr

Dominique Labbé
PACTE (CNRS – Institut d'Etudes Politiques de Grenoble)
dominique.labbe@iep-grenoble.fr

Pierre Hubert
Université de Paris VI
pjy.hubert@free.fr

Abstract
Segmentation of large textual corpora is one of the major questions asked of literary studies. We present a combination of two relevant methods. First, vocabulary growth analysis highlights the main discontinuities in a work. Second, these results are supplemented with the analysis of variations in vocabulary diversity within corpora. A segmentation algorithm, associated with a test of validity, indicates the optimal succession in distinct stages. This method is applied to Racine's works and those of various other works in French.

Résumé
Le découpage des grands corpus de textes est l'une des questions cruciales posées aux études littéraires. Il est proposé une double méthode. L'analyse de la croissance du vocabulaire (type-token ratio) met en lumière les principaux changements de rythme. Ces résultats sont complétés par l'étude de la diversité du vocabulaire. Un algorithme de segmentation, associé à un test de validité, indique le découpage optimal. La méthode est appliquée aux œuvres de Racine, Corneille et aux discours du Général de Gaulle.

Key words: statistics for linguistics ; segmentation ; corpora ; vocabulary growth ; vocabulary diversity ; stylistics

Draft of the paper published in:
Journal of Quantitative Linguistics. December 2004, vol. 11, n° 3, p. 193-213.

INTRODUCTION

How to isolate generally homogeneous parts in a work or in a corpus? It is one of the major questions which confront critics and scholars in literary studies. This question becomes increasingly important as software programmes are used with an ever-growing body of electronic texts available to researchers.

Usually, one uses the major events of an author's life in connection with his works, — or various natural textual caesurae like divisions by chapters, books, etc. We propose here a set of procedures which operate more objectively and which can help critics or scholars in their studies. These procedures involve two major stylistic indices: vocabulary growth and vocabulary diversity which can be included within a more general topic, the relation between the text length and the vocabulary size or Type-Token-Relation (Müller, 2002).

The calculations are related to the work of Jean Racine (1639-1699), a well-known French author of the 17th century (titles and dates of the plays in the Appendix, see also Bernet, 1983).

Preliminary statement

Texts are first normalised and tagged. The "part-of-speech" tagging is necessary because in any text written in French, an average of more than one-third of the words are "homographs" (one spelling, several dictionary meanings). Hence standardisation of spelling and word tagging are first steps for any high level research in quantitative linguistics with French texts (norms and software are described in Labbé, 1990). All the calculations presented in this paper utilise these lemmas.

Moreover, tagging, by grouping tokens under the categories of fewer types, has many additional advantages, in particular: a major reduction in the number of different units to be counted.

This operation is comparable with the calibration of sensors in any experimental science.

VOCABULARY GROWTH

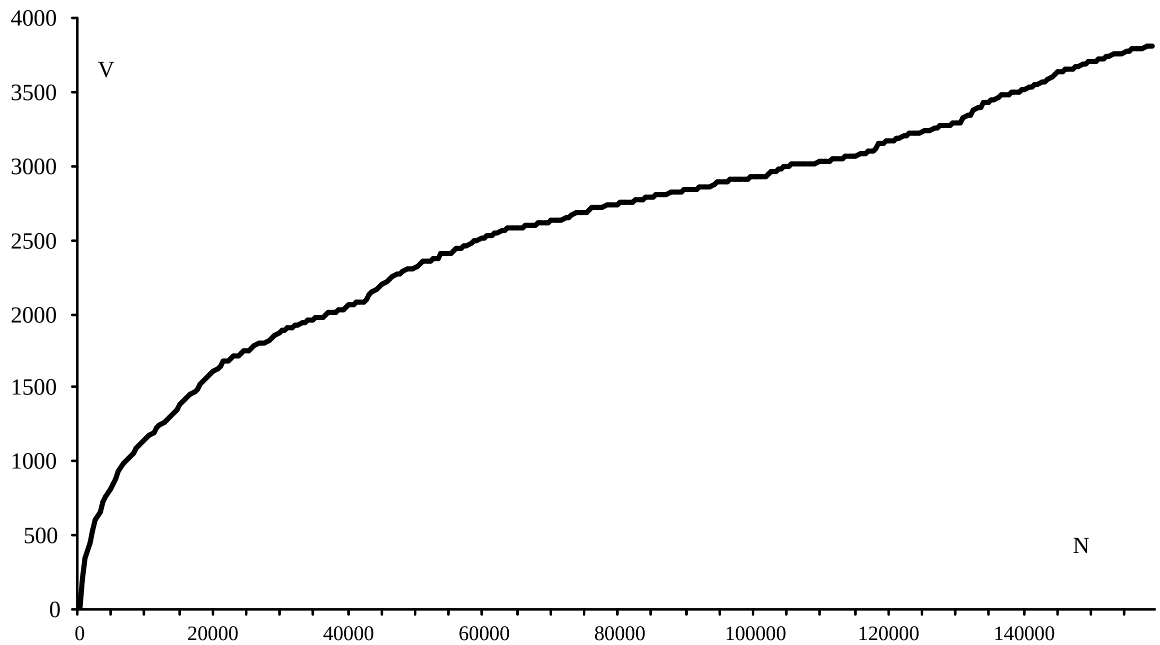
Vocabulary growth is a well known topic in quantitative linguistics (Wimmer & Altmann, 1999). In any natural text, the rate at which new types appear is very high at the beginning and decreases slowly, while remaining positive even in extremely long works (Hubert-Labbé, 1988b, Hubert-Labbé, 1994).

Let N equal the length of a text (in tokens). For example, with Racine's tragedies, N equals 158,585 tokens.

Let V equal the number of different types used in this text. With Racine's tragedies, V equals 3,814 types.

First, the observed vocabulary grows more or less exponentially (with an exponent of less than 1). In a second phase, this growth seems to become roughly linear (Müller, 2002). Figure 1 shows this curve in Racine's work (following a step interval of 500 tokens along N and measuring the number of different types (V) from the beginning of the work).

Fig. 1. Chart of vocabulary growth in the tragedies of Jean Racine
(chronological order, 500 token intervals)



This chart is irregular and suggests the existence of certain "disconformities" or "break-points" in the work. To locate accurately the main caesurae, we propose a two-part procedure: first, a mathematical curve is fitted and then, oscillations about the trend are highlighted.

Adjustment of the chart of vocabulary growth

First, this chart is adjusted by calculating V' — the number of different types expected in an excerpt of N' tokens — according to the following procedure (Hubert & Labbé, 1988a).

The V types, in the whole work, are graded in order of frequency into n frequency bins. Define V_i as the number of types which occur i times; V' is approximated by this formula:

$$(1)\ V'(u) = p.u.V + (1 - p)\left[V - \sum_{i=1}^{i=n} V_i Q_i(u)\right]$$

with :

$u = \frac{N'}{N}$ (in this experiment, N' varies from 500 to N)

$Q_i(u) = (1 - u)^i$

p is the "coefficient of vocabulary partition".

The coefficient of vocabulary partition measures the relative size of the two sets of vocabulary (Hubert-Labbé, 1988b). The first set contains pV specialised word types which are devoted to a special part of the text, such as nouns of figures, towns and countries or technical vocabularies... The average growth of this first set is a linear function of N' (first part of the formula (1)). The second set contains $(1-p)V$ types which belong to the general vocabulary. This set contains the vocabulary used whatever the topic is: articles, prepositions, auxiliary and modal verbs, etc. The probability of their appearing is constant at any stage of the text and can be estimated as if they belong to a sample of size N' tokens randomly drawn, without replacement, from the N tokens of the whole corpus. The size of this second set is estimated with the help of the Muller's formula (second part of the formula (1)) (Muller, 1977)

In practice, the p coefficient is calculated in this way: at each interval of 500 words, the different types are counted from the beginning of the corpus. For the K milestones — 500, 1,000, …, N — let:

— N'_k be the number of tokens counted since the beginning of the texts until the k_{th} milestone;

— $u_k = \frac{N'_k}{N}$

— $V_*(u_k)$ be the number of different types counted since the beginning of the texts until the k_{th} milestone;

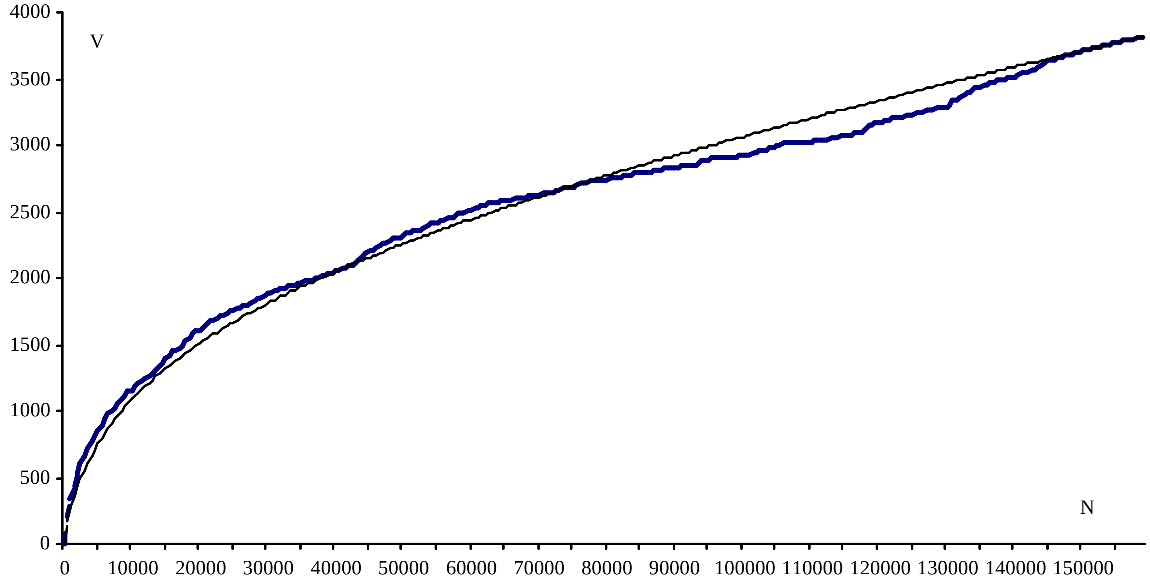
— $V(u_k)$ be the theoretical number of different types, since the beginning of the texts until the k_{th} milestone, calculated with formula 1.

The value of p is that which minimises the sum of the squared deviations between the observed values and the calculated ones. We thus obtain:

$$(2)\quad p = \frac{\sum_{k=1}^{k=K}\left[(u_k - 1)V + \sum_{i=1}^{i=n} V_i Q_i(u_k)\right]\left[V_*^{'}(u_k) - V + \sum_{i=1}^{i=n} V_i Q_i(u_k)\right]}{\sum_{k=1}^{k=K}\left[(u_k - 1)V + \sum_{i=1}^{i=n} V_i Q_i(u_k)\right]^2} \quad with\ u_k = \frac{N'_k}{N}$$

Formulae (1) and (2) are easy to compute. Notice that for the calculation of (2), the intervals are not necessarily equal or proportional: the counts of $V'(u_k)$ can take place anywhere along the corpus. Of course, the accuracy of results depends on the number and quality of these observations. It seems that no less than ten values of $V_*(u_k)$ is necessary, evenly distributed within the texts or corpus. Given this minimum requirement, numerous experiments have proven that p is actually independent of the size and number of the excerpts. Figure 2 presents the results on the same texts by Jean Racine: the theoretical curve (plain line) actually goes through the chart of the observed values (bold line).

Fig. 2. Observed and estimated growth of vocabulary in the 11 tragedies by Jean Racine (chronological order).



Variations about the trend

Figure 2 suggests that some caesurae (or disconformities) occur in this work. To locate them, there is an existing procedure used in economics which can be applied to series that exhibit cyclical variations about a stable trend. Which deviations may be considered to occur by chance and which ones are non-randomly significant? To answer this question:

— theoretical values become the *X*-axis and observed values are centred on theoretical ones: $V'_*(u_k)$ - $V'(u_k)$;

— a theoretical variance is calculated. Given that the general vocabulary is the only "probabilistic" part, this variance can be calculated solely on this part of the whole vocabulary ie on *(1-p)V*. Considering each *k* observation and the *n* classes of frequency in the whole vocabulary, the variance can be estimated by:

$$(3) \quad Var[V'(u_k)] = (1-p) \; . \; \sum_{i=1}^{i=n} V_i Q_i(u_k) \; [1 - Q_i(u_k)]$$

— reduction of the centred values using the standard deviation (square root of the variance). For each of the *k* points, we obtain:

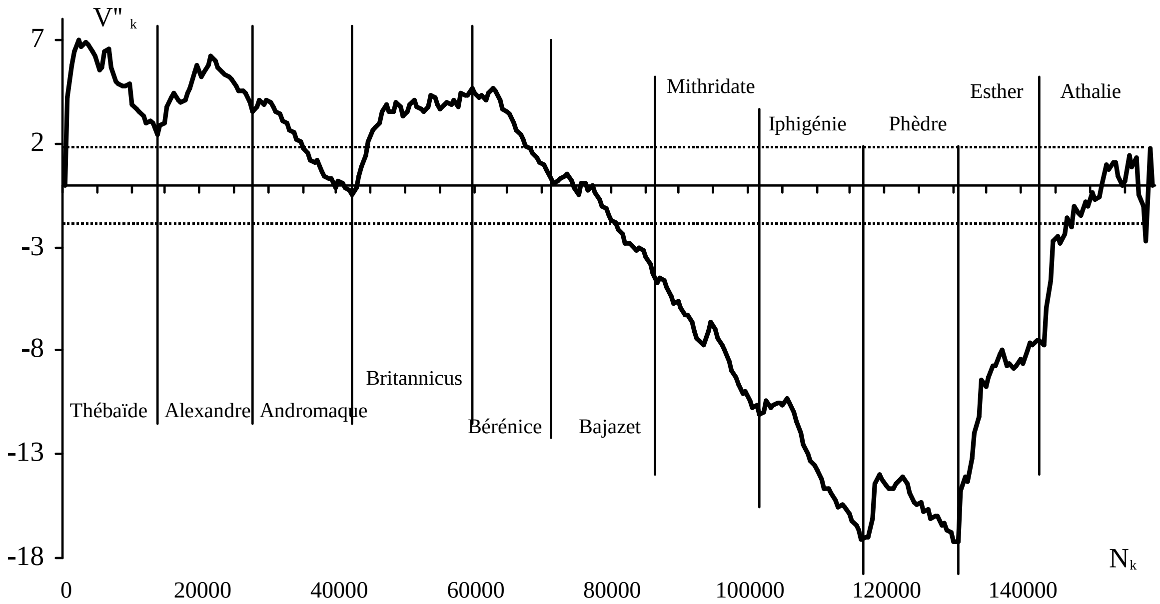
$$(4) \quad V''_*(u_k) = \frac{V'_*(u_k) - V'(u_k)}{\sigma(u_k)}$$

Figure 3 shows the results of this procedure applied to the tragedies of Jean Racine. Compared with Fig. 2, Fig. 3 gives a kind of "zoom effect", highlighting movements about the general trend (now the *X*-axis).

Interpretation of the chart

First, for a given portion of the chart, the slope must be taken into consideration. If this portion or segment of the chart is moving upwards (positive slope), an influx of new vocabulary occurs at this point and new ideas appear in the writing. In Racine's work, such episodes seem generally to occur at the beginning of each play. Almost all the plays present a characteristic wave shape brought about by the emergence of the main characters, or by countries and cities where the action takes place. Sometimes, there is a strong influx of new words as with: Thébaïde, Alexandre, Britannicus, Esther, Athalie. Except for the first, these plays can be considered as disconformities or turning points in Racine's work. By contrast, portions or segments with negative slopes indicate that few fresh word types are present: the author is drawing on his usual themes. The endings of all Racine's plays exemplify this feature with the exception of Britannicus, Esther and Athalie, in which unexpected renewals occur right up to their endings. This suggests that these plays display an unusual pattern.

Fig. 3. Growth of vocabulary in Racine's tragedies (centred and reduced values)

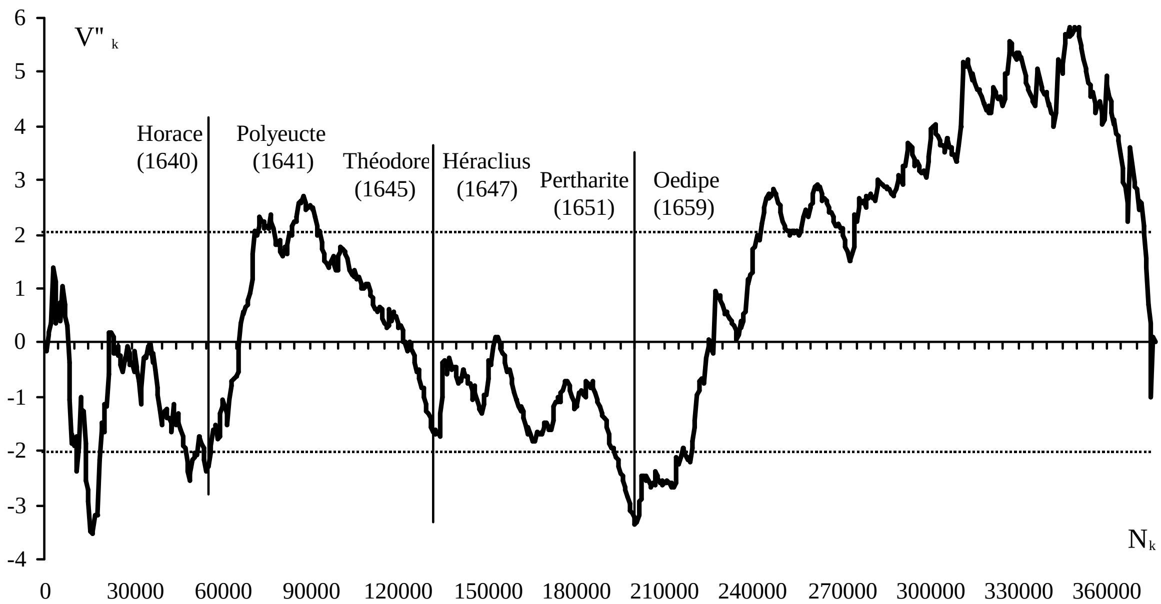


Second, the general slope of the chart. The dotted lines define a confidence interval (operating limits) — two standard deviations above and below the X-axis — by means of which one can form an opinion as to what extent the author conforms to his general trend. Above two standard deviations, the chart signals a period of inventiveness: for Racine, this occurs in the two first plays and in Britannicus. Below two standard deviations, the author reuses his former vocabulary, and we may assume that he repeats old ideas. Jean Racine seems to have done this during the middle part of his career from Bérénice (1670) to Phédre (1677)…

Finally, the chart clearly shows that the major turning point in Racine's work occurred around the time of Phèdre. Does this play belong to the first part of the chart or to the last part, that composed of the two tragédies sacrées ("sacred tragedies"), Esther and Athalie? To answer this question, supplementary methods can be used. For example, intertextual distances, combined with cluster analysis (Labbé & Labbé, 2001), or variation of vocabulary richness as treated below.

In long works like those of Racine, patterns such as the one in Fig. 3 are relatively common: initial stages often reveal a period of creativity and invention, and repetitions become common with the passage of time, even if occasional renewals occur from time to time. But it is not a hard and fast rule, as may be seen in Fig. 4. Pierre Corneille (1606-1684) was the most famous author of the 17th century.

Fig. 4. Growth of vocabulary in Pierre Corneille's tragedies



First, the dispersion of values around the mean in Fig. 4 indicates that vocabulary growth is far more regular in Corneille's than in Racine's work, even though we might expect the opposite in view of the fact that the larger a work and the longer the time taken to write it, the more its vocabulary is likely to change. Corneille's tragedies are nearly 400,000 tokens long, compared with a word length of 160,000 in Racine's; Corneille's composition is spread over 40 as opposed to 27 years for Racine…

Second, Fig. 4 shows that, like J. Racine, P. Corneille experienced a comparable decrease in creativity during his middle period, right up until Pertharite. This play was a failure, and was followed by nine years of silence. The return of Corneille to the theatre in 1659 (with Œdipe) is clearly the major turning point in his work. From that year onwards till his last tragedy (Surena, 1674), each play reveals an influx of new word types that moves the chart up and keeps it above the upper dotted line: fifteen years of continuous inventiveness, not to mention his collaboration in writing the major plays of Molière published at the same time (Labbé & Labbé, 2001).

Furthermore, the calculation offers a solution to the much-debated question of vocabulary richness (Hubert & Labbé, 1994; Labbé, 1998; Wimmer & Altmann, 1999). For a given text, richness of vocabulary is a function of two variables. The first is vocabulary specialisation which reduces vocabulary diversity in the short term, but generally increases the global variety of word in the middle and long term. Racine and Corneille illustrate two contrasting choices. Racine uses a large specialised vocabulary (p = 0.33): one out of three word types derives from a specialised vocabulary and is not reused in other plays. At the other extreme, Corneille uses only a generalised vocabulary with just the few words needed to give his plays local colour (p= 0.02). In other words, the two playwrighters make opposite choices: Jean Racine tells very similar stories using different words, whereas Pierre Corneille writes different stories with nearly the same vocabulary!

Diversity is the second variable connected with the calculation of vocabulary richness (for an examination of diversity measurement and indices, see: Pielou 1982).

VOCABULARY DIVERSITY

Vocabulary diversity measures the author's tendency to vary his vocabulary within a short length of text (as, for example, a few hundred tokens). Every author holds in mind certain themes or ideas, and, when writing about them, can employ a great diversity of words and complex sentences structures as they are available and stored in his memory. By contrast, however, when speaking or writing about things which do not really matter to him, the diversity of the author's vocabulary decreases significantly. Genre must also be taken into account: a person does not talk in private conversation the same way in which he writes for a scientific journal. Therefore, when studying written literary texts like novels or plays, one must expect a greater variety of words than for newspapers or letters, let alone transcriptions of oral speech. For literary studies, we propose considering the number of different types in any extract of 1,000 contiguous tokens.

In the entire work of J. Racine this vocabulary diversity is equal to 360‰. That is to say, an average of 360 different types may be expected in any sample of 1,000 contiguous tokens. Is this relatively high or low? Pierre Corneille, for example, in his total of more than a half-million tokens has an average diversity of 352‰, very close to the value noted with Racine, his young rival. But the ratio is not merely a characteristic of the 17th century, as it can be seen, for example, in the work of J.-M. Le Clézio, the most popular — according to pools — contemporary French novel writer. In his writing, from beginning in 1965 until 1999 (roughly a million words), the average diversity is 363‰.

Even if vocabulary diversity does not clearly characterise an author's style, it may nevertheless prove a useful tool for text or corpora segmentation, in addition to the observation of vocabulary growth.

Main steps

First, calculation of the average diversity (for 1,000 contiguous tokens) with the help of the vocabulary partition model (see formula 1 above). Second, as is the case with any natural phenomenon, vocabulary diversity shows random variation which can be estimated with a theoretical variance calculated by using the partition vocabulary model (see formula 3, above).

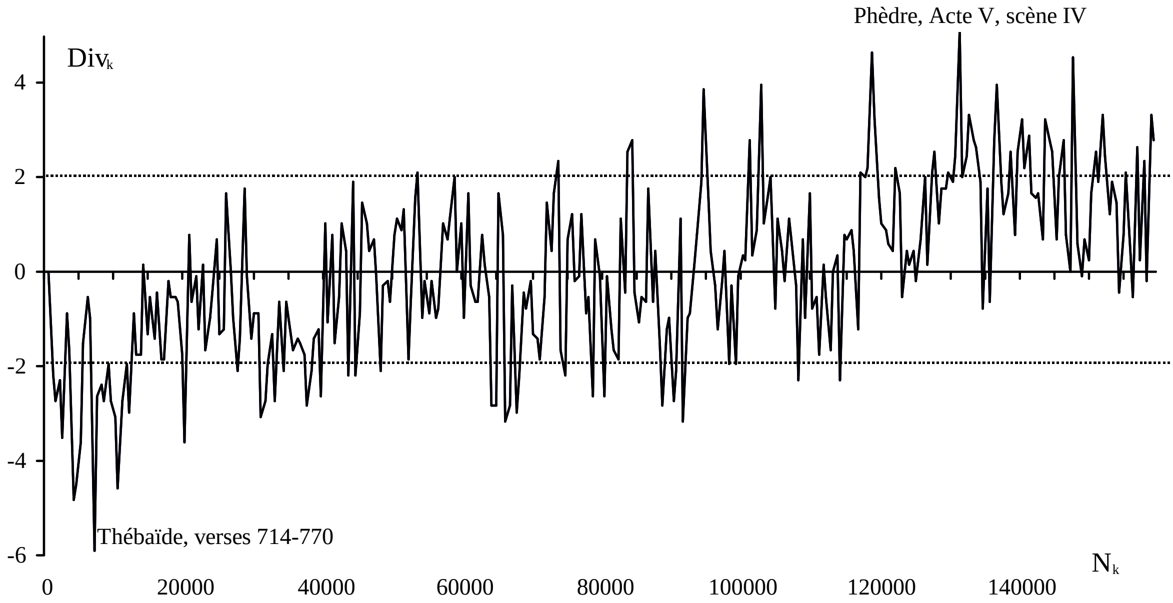
For example, Racine's average diversity is: 360 ± 12.3 types (for 1,000 contiguous tokens). If deviations about the mean are due to random events only, one can expect that the observed values for the whole work will remain within the limits of normal deviation, between 336 and 384 types (mean ± 2 σ). And furthermore, one can expect if some values occur above or below these limits that a significant change occurs in Racine's style at this point.

The algorithm proceeds along the course of the work — depending on the span chosen by the operator, in this case 500 tokens —, and detects the number of different types which are used in the last k^{th} block of the 1,000 tokens just read. These observed values (V'_{*k}) are centred on the mean (V'_{1000}) and are normalised following the technique presented above in relation to vocabulary growth. For the k^{th} segment, the ordinate of the graph will be:

$$Div_k = \frac{V'_{*k} - V'_{(1000)}}{\sqrt{Var(V'_{(1000)})}}$$

Notice that the theoretical variance is slightly overestimated (see Hubert & Labbé, 1988b). The abscissa represents the chronological growth of the work at this k^{th} point (N'_k). Figure 5 gives the result with Racine's work. The interval of confidence (± 2σ) is marked by the dotted lines.

Fig. 5. Evolution of vocabulary diversity in Racine's tragedies (chronological order)



This Figure clearly leads to three observations:

— while the majority of values fall within the confidence limits, deviations about the mean are numerous and are somewhat greater than expected in terms of normal random variation;

— many "accidents" occur and these accidents can be precisely located. The deviations are of short duration and provide an accurate means with which to observe stylistic events. For example, the lowest point of the chart corresponds to a scene of the first play (Thébaïde) and in particular to a dialogue between Créon — Thebes' "Prime minister" — and his main adviser. The fragment with the richest vocabulary is a dialogue in Phèdre which highlights one of the major themes of Racine, the relationship between father and son.

— actual average diversity is not uniform throughout the work. Figure 5 shows a clear rising trend. In other words, Jean Racine was relatively abstemious in his first plays. Then he gradually came to be more assiduous in avoiding short-term repetitions and in diversifying his vocabulary.

Is it possible to proceed further from the stylistic point of view in locating homogeneous periods and turning points or break-points in Racine's work?

Segmentation procedure

The calculation which follows is inspired by a model designed for hydro-climatological studies by Hubert, Carbonnel and Chaouche (1989). This procedure was first applied to rainfall and annual time series for hydrological discharge in West Africa since the beginning of the 20th century. A simple summary of this procedure, adapted for textual data series like vocabulary diversity, is given below.

Let x_i be the number of different types observed in the i_{th} 1000 tokens excerpt and n the number of observations along the span of the entire corpus. Any partition of the n measurements of vocabulary diversity in Racine's work into m segments is a m-order segmentation, with: $0 = i_0 < i_1 < \ldots < i_m = n$.

The sub-series x_i (with $i_{k-1} \leq i \leq i_k$) is a segment. The mean of the k_{th} segment is:

$$\overline{x_k} = \frac{\sum_{i=i_{k-1}+1}^{i=i_k} x_i}{i_k - i_{k-1}}$$

And we will define:

$$d_k = \sum_{i=i_{k-1}+1}^{i=i_k} (x_i - \overline{x_k})^2$$

The sum of all such quantities for the m segment is the quadratic departure (squared difference) between the segment and the original series:

$$D_m = D_{(1,\ldots,m)} = \sum_{k=1}^{k=m} \sum_{i=i_{k-1}+1}^{i=i_k} (x_i - \overline{x_k})^2$$

The “best” m-order segmentation is that which minimises D_m. Considering n (the length of the original series), the number of possible m-order segmentations is equal to:

$$N(n,m) = C_{n-1}^{m-1} = \frac{(n-1)!}{(m-1)!\,(n-m)!}$$

Applied to large corpora, like that of Racine, this involves an extremely large number; therefore an efficient and economical algorithm is necessary to find the optimal solution. This algorithm is fully presented in Hubert, Carbonnel & Chaouche, 1989 and Paéquin, 2003. It can be described as the search for the best path in descending a tree structure with the help of a "branch-and-bound" procedure. This searches for optimal segmentations, beginning with order 1, and successively considering the *1, 2,…, n* last terms of the series. For each segment, the values are placed in an array where they can be reused for calculation at next tree-level. For any given level, the result of the on-going segmentation is compared with the optimal segmentation previously established; if the result is negative, the algorithm can ignore all paths sited below this node.

If a maximum number of segmentations has been chosen *a priori*, the algorithm stops when it reaches this number. One may, however, ask how can one know what is the optimal number of segmentations. In answer, we propose to define the optimal number as*: the maximal segmentation in which the mean of each segment is significantly different from the means of its two neighbours* (of course, the decision on the first and the last segment are made only on the basis of one neighbour). A test, modelled on Sheffé's contrast (Sheffé, 1959 and Dagnelie, 1970) is applied to ensure that: given $\overline{x_k}$ and $\overline{x_{k+1}}$ the mean of the k^{th} and $k+1^{th}$ segments under study, n_k and n_{k+1} their sizes (number of values), the difference between the two segments:

$$C_{(k,\,k+1)} = \overline{x_k} - \overline{x_{k+1}}$$

must confirm, with a probability equal to 1-α, the inequality:

$$(5)\ C_{(k,k+1)} - Ss \leq 0 \leq C_{(k,k+1)} + Ss$$

with σ being the square root of the variance calculated on all the series:

$$\sigma^2 = \frac{D_m}{n-m}$$

S is defined as:

$$(6)\ S^2 = (m-1)F_{m-1,n-m}(a)\left(\frac{1}{n_k} + \frac{1}{n_{k+1}}\right)$$

in which $F_{m-1,n-m}(\alpha)$ is Fisher's variable, with m-1 and n-m degrees of freedom, whose probability is set at α.

If the values calculated with formula (5) fall within the critical range — that is to say, if the quantities $C_{(k,k+1)} - S\sigma$ and $C_{(k,k+1)} + S\sigma$ are respectively positive and negative —, the means of the two segments (k and $k+1$) do not significantly differ; in other words, the sub-series composed by these two segments is "uniform" ie has the "same" mean (their difference can be considered as occurring by chance);

. The segmentation must be interrupted at the level immediately higher, considered to be the optimal level.

On the other hand, if the signs of these two quantities are the same, the contrast between segments k and $k+1$ is significant, and the procedure continues further.

It will be noticed, in formulae (5) and (6), that the variance is calculated on all the m segments of the series. We also may calculate the variance of just the two segments *(k, k+1)* under analysis, as if the series were constituted of only these two segments (granted that these segments are sufficiently large as to permit the calculation of variance). The two different methods occasionally yield different results. As can be seen below (in the test), the second method seems to yield slightly but appreciably more accurate results.

With formulae (5) and (6), the operator needs, *a priori*, to choose a value for α. In the algorithm developed by P. Hubert in 1989, α acts effectively as a threshold: if the dissimilarity between two segments exceeds this threshold value by ε, it causes the operator to reject a significant segmentation unless additional information is present. In order to overcome this threshold-effect and to improve the calculation, we propose having the algorithm search this value on its own, beginning at the highest possible value (0.01) and decreasing by stages of 0.0001 until the test is null: the previous value is associated with the pair of segments under consideration. For an *m-order* segmentation, *n-1* values of α are obtained. Firstly, the highest value of α is to be adopted in order to decide whether this segmentation can be accepted or not. Secondly, for each pair of segments *(k, k+1)*, a *quality index* is calculated, graded out of 100:

$$\lambda_{(k,k+1)} = (1 - \alpha_{(k,k+1)})*100$$

By the help of this quality index and by considering all the values of variance, the operator may choose the best segmentation according to his particular needs, without threshold effects tied to certain critical values.

Similarly, to provide the operator with more information, the algorithm allows him to disregard segments less than a minimal size; in some cases, it may eliminate small erratic changes in the series as apparent in the test results.

Tests and simulations

This segmentation procedure may be considered a test of uniformity. For the whole series, the null hypothesis, "the series is uniform", will be accepted if the algorithm cannot find an acceptable segmentation of order 2 or more. Of course, this decision is subject to the risk of a type 1 error which rejects a null hypothesis that is true. To evaluate this risk, the authors have tested the algorithm on several random series.

With this in mind, a large number of random series were generated with the help of SCILAB software from the Institut National de Recherche en Informatique et Automatisme (INRIA), as follows:

(a) a large number of normally distributed series (varying from 50 to 100). For example, the values are randomly distributed about a mean of 0 with a standard deviation of ±1. These series are uniform and the algorithm should not accept any segmentation.

(b) three or more normal series, appended one to the other. For some of these, the means were deliberately chosen to differ significantly; the algorithm should discover the number of possible segmentations and accurately locate their caesurae.

(c) "explosive" series inserted in the middle of normal series. They are called "explosive" because they show dispersion with very large variation, incorporating extreme values equal in magnitude to several times the mean (see an example below).

A synthesis of the results is shown in Figs. 6 .1, 6.2, and 6.3 below (for further details about these tests, see: Paéquin, 2003).

Fig. 6. Results on random series

6.1 First version of Sheffé's test (calculation of variance on the whole series). No minimal size for a segment.

Type of series	Number of series	α	Number of correct segmentations (Mean %)
(a)	100	0.01	85
(a)	100	0.002	96
(b)	100	0.01	74
(b)	100	0.002	86

6.2 Second version of Sheffé's test (calculation of variances on the *k, k+1* segments). No minimal size for a segment.

Type of series	Number of series	α	Number of correct segmentations (Mean %)
(a)	100	0.01	86
(a)	100	0.002	96
(b)	100	0.01	91
(b)	100	0.002	98

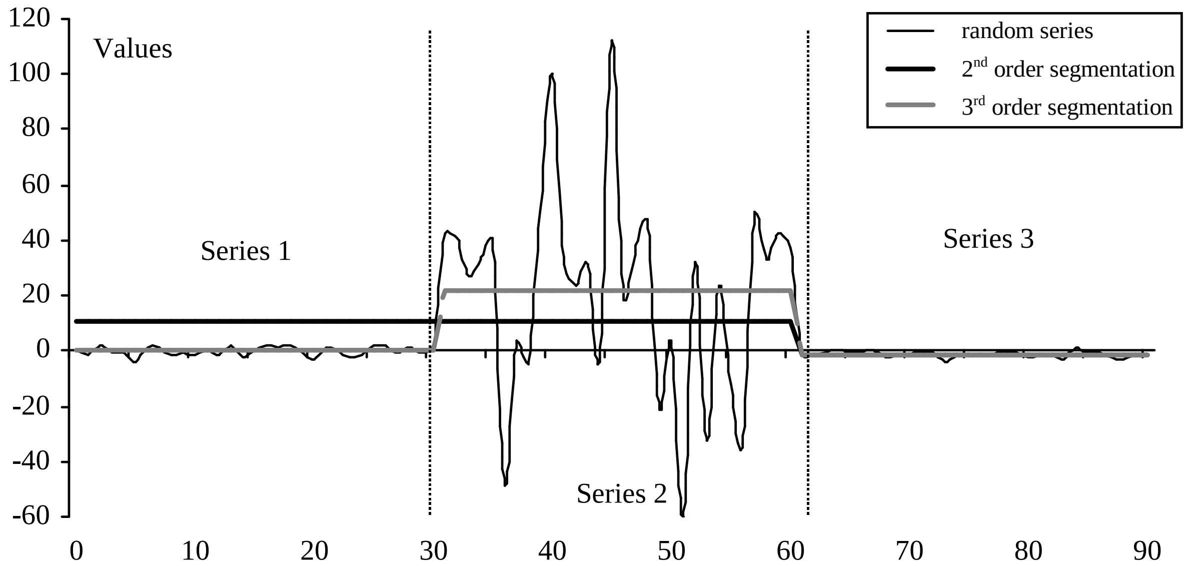
6.3 Second version of Sheffé's test (calculation of variances on the k, k+1 segments). Minimal size for a segment:5 values (each series appended together has 30 values)

Type of series	Number of series	α	Number of correct segmentationsn (Mean %)
(a)	100	0.01	88
(a)	100	0.002	97
(b)	100	0.01	96
(b)	100	0.002	99

It appears that the results are generally favourable. Figure 6.1 confirms the values obtained by Hubert, Carbonnel and Chaouche in 1989. The second variance calculation (6.2) and the choice of a reasonable minimal segment size (6.3) lead to more accurate results.

Tests on "explosive" series clearly show the actual limits of a complete automatic algorithm (see Figure 7 for an example).

Fig. 7. Problem of segmentation with a random "explosive" series inserted between two normal series.



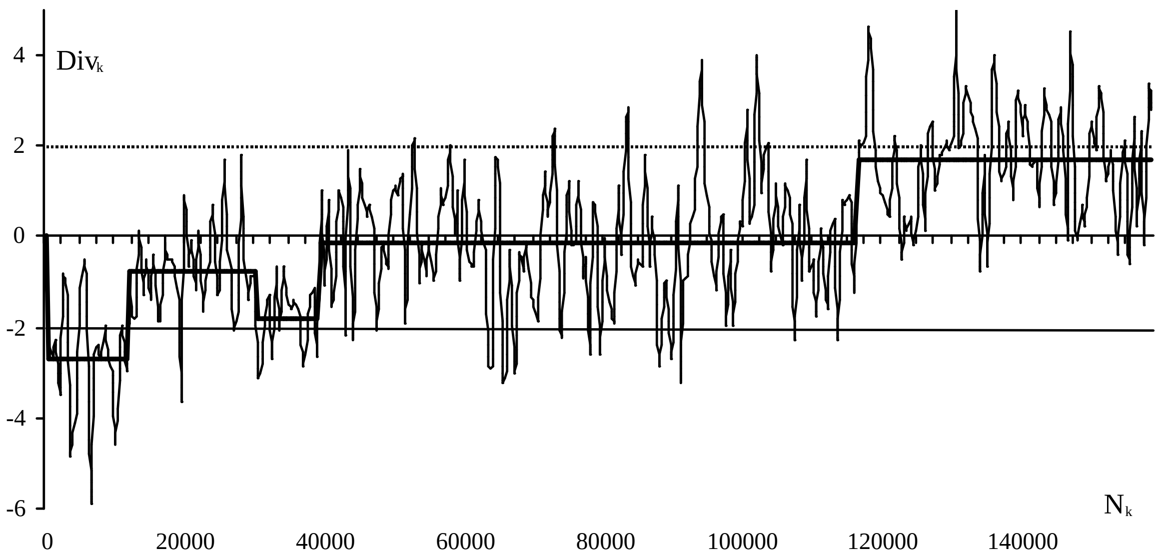
In this example, there are three series appended one to the other (plain line); the values of the middle series are generally spread above and below those of the other two with a very large range of variation. Given the fact that the 2nd order segmentation is not acceptable (black bold line), the automatic algorithm ceases its search and considers the whole series as uniform. If the automatic disablement is deactivated, the algorithm passes over the 2nd order segmentation, discovers that the 3rd order segmentation is highly significant (grey bold line) and indicates that the 4th order is no longer acceptable. Naturally, this kind of series is highly unlikely to occur, especially with natural language. But this experiment demonstrates that, for a very reliable segmentation of large corpora, one must use all the algorithm's potentialities and perform several iterations before coming to a decision.

Applications

Figure 8 illustrates the result of the segmentation procedure when applied to Racine's work. The disconformities from a stylistic point of view can be located with a precision of ± 500 tokens.

Fig. 8. Segmentation of Racine's tragedies according to variation of vocabulary diversity

(α = 0.002)



The ascending structure of the chart clearly confirms the fact that a trend toward diversity increases throughout Racine's work. Furthermore, a comparison with Fig. 3 reveals certain similarities between the two phenomena displayed. Significant stylistic changes are generally linked with the main thematic changes in the playwright's work. For example, the final break-point in diversity occurs at the end of Iphigénie (or at the beginning of Phèdre), at the lowest point of the vocabulary growth chart, when the last major inflexion (upward) occurs after a long and stable decreasing period with negative slope.

— in the first play (Thébaïde), the style is austere, very much like that in Corneille's last plays;

— the second segment corresponds to the second play (Alexandre) and to the first part of the third play(Andromaque);

— the third segment, representing the one significant decline in the series, corresponds with the second part of this third play, which also shows a remarkable decline in vocabulary growth (Fig. 3);

— the five succeeding plays are, stylistically speaking, just below the mean and relatively homogeneous; they are Racine's most successful plays. On this part of the chart, corresponding to the middle half of the opus, most of the values fall within the confidence limits; this sub-series is clearly uniform.

— a new level is attained in the final scenes of Iphigénie and characterises Phèdre and the two last Racine's plays (written a long time after Phèdre).

The position of the discontinuities should be noted: most of them occur inside a play rather than between two plays as might be expected. In the case of the nine first plays, this is not very surprising because the writing of each successive play took place immediately on completion of the previous one. The nine plays may thus be considered as the result of a continuous stream of creation. However, twelve years elapsed between Phèdre and Esther and, during this time, Racine seems to have seriously changed his mind about the theatre and religion. It appears that, from the stylistic point of view (Fig. 8), these changes had few repercussions and that the style of Esther may be regarded as a continuation of Phèdre's.

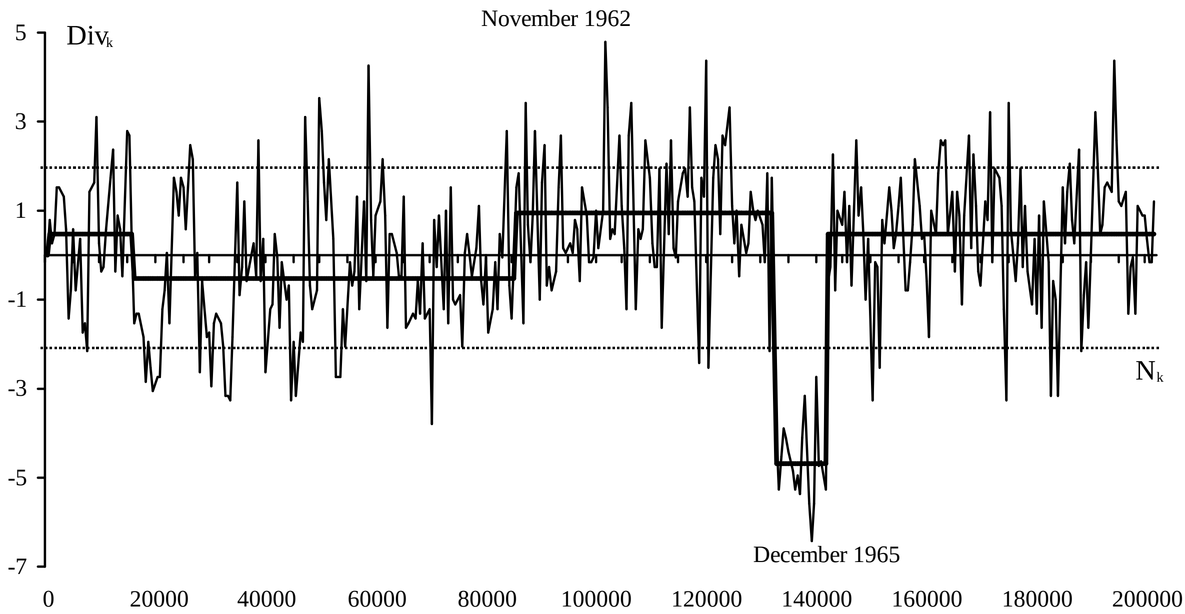
It should also be noted that:

— only the first segment in Fig. 8 exceeds the limits of random variation (dotted lines), while the last segment is just below the upper limit of this confidence interval: our measures permit an analysis more accurate than the classic tests based on variance;

— the best possible segmentation is the last one for which all the contrasts between each segment have a difference of null (for α varying between 0.01 and 0.001).

This procedure has been applied to a large number of corpora (for an example on Spanish texts: Alvarez & Al, 2000). The results are always productive and occasionally surprising. For example, this technique can distinguish written from oral speeches as seen in Fig. 9.

Fig. 9. Evolution of vocabulary diversity in General de Gaulle's broadcast speeches (June 1958 - April 1969).



During his eleven years as head of state, General de Gaulle always learnt his speeches by heart and played them out like theatre roles, even during his press conferences (the journalists had to submit their questions before the interview). His mean vocabulary diversity is very high (390‰): these Gaullist texts are literary and quite unusual in French politics (see Labbé & Monière, 2002 & 2003). But such vocabulary diversity was not the case on one occasion in December 1965 when de Gaulle was not re-elected in the presidential election on the first ballot and thus had to improvise a two-hour interview in order to campaign for the second ballot. One can make out a deep notch in the chart: with respect to this 1965 interview, diversity falls by nearly twenty per cent (changes are also clear in the vocabulary and sentence structures).

The other break-points in General de Gaulle's vocabulary diversity are closely linked with major political events, especially in autumn 1962 — independence of Algeria and the advent of a Gaullist majority in parliament. After this point, the diversity reaches a very high level and remains at that level until December 1965.

CONCLUSIONS

These segmentation procedures appear to be efficient tools for literary analysis. They enable quick and simple explorations of large series such as literary corpora. They also provide new information, as, for instance, an author's propensity to specialise his vocabulary or to diversify his words in the short term. As for the second procedure, the semi-automatic method — which allows comparison between results obtained using several values for α and different minimum segment sizes, combined with a quality index — this second procedure provides a large amount of information which permits an operator accurately to locate homogenous sub-corpora in a clear and precise fashion.

When homogeneous parts in a corpus are located within spatial limits, their vocabularies can be described with the help of other tools such as the calculation of their "specific" types, associations of words, or sentences (Labbé & Labbé, 1993). In the field of quantitative linguistics, other stylistic indices can be brought to bear on the problem, like regularity (or irregularity) of occurrence of selected function words, or selected grammatical categories.

Because it is futile to accurately measure phenomena the observations of which are made without precision, all calculations, as a necessity, need strict standardisation of word spelling and — for French (with its many inflections and homographs) — tagging ("lemmatisation") of each token in the texts.

These tools can be used in many fields of research, such as sociology, econometrics, climatology — everywhere that large series of data are to be analysed.

ACKNOWLEDGMENTS

The authors are grateful to Gaétan Paéquin (Polytech' Grenoble) — who wrote the software and carried out the experiments presented in this paper under our supervision during the summer of 2003 — to Charles Bernet (Institut National de la Langue Française) who provided the texts of Racine and to Tom Merriam for his accurate reading of our first translation and for his most helpful comments and advice.

REFERENCES

Alvarez R., Becue M. & Lanero J.-J. (2000). Vocabulary Diversity and its Variability: A Tool for the Analysis of Discoursive Strategies. Application to the Investiture Speeches of the Spanish Democracy. In Rajman M. & Chappelier J.-C. (eds). *Actes des 5e journées internationales d'analyse des données textuelles*. Lausanne : Ecole polytechnique fédérale, vol 1, 111-118.

Bernet C. (1983). *Le vocabulaire des tragédies de Racine.* Genève-Paris: Slatkine-Champion.

Dagnelie P. (1970). *Théories et méthodes statistiques*. Tome 2. Gembloux: Duculot.

Hubert P. & Labbé D. (1988a). Note sur l'approximation de la loi hypergéométrique par la formule de Muller. In Labbé D., Serant D. & Thoiron P. *Etudes sur la richesse et la structure lexicales*. Paris-Genève: Slatkine-Champion, 77-91.

Hubert P. & Labbé D. (1988b). Un modèle de partition du vocabulaire. In Labbé D., Serant D. & Thoiron P. *Etudes sur la richesse et la structure lexicales*. Paris-Genève: Slatkine-Champion, 93-114.

Hubert P., Carbonnel J.-P. & Chaouche A. (1989). Segmentation des séries hydrométéorologiques - Application à des séries de précipitations et de débits de l'Afrique de l'Ouest. *Journal of hydrology*, 110, 349-367.

Hubert P. & Labbé D. (1994). La richesse du vocabulaire. *Communication au congrès de l'ALLC-ACH*, Paris: La Sorbonne. Reproduced in *Lexicometrica*, 0, 1997 (http://www.cavi.univ-paris3.fr/lexicometrica/).

Labbé C. & Labbé D. (1994). Que mesure la spécificité du vocabulaire?. Grenoble: CERAT. Reproduced in *Lexicometrica*, 3, 2001 (http://www.cavi.univ-paris3.fr/lexicometrica/).

Labbé C. & Labbé D. (2001). Inter-textual Distance and Authorship Attribution Corneille and Molière. *Journal of Quantitative Linguistics*, 8, 212-231.

Labbé D. (1990). *Normes de saisie et de dépouillement des textes politiques*. Grenoble: Cahier du CERAT.

Labbé D. (1998). La richesse du vocabulaire politique : de Gaulle et Mitterrand. In Mellet S. & Vuillaume M. (eds). *Mots chiffrés et déchiffrés: mélanges offerts à Étienne Brunet*. Paris: Champion, 173-186.

Labbé D. & Monière D. (2003). *Le discours gouvernemental*. Paris: Champion.

Monière D. & Labbé D. (2002). Essai de stylistique quantitative. In Morin A. & Sébillot P. *VIe Journées d'Analyse des Données Textuelles*. Rennes: IRISA - INRIA, 561-570.

Muller C. (1977). *Principes et méthodes de statistique lexicale*. Paris: Hachette.

Müller D. (2002). Computing the Type Token Relation from the *A Priori* Distribution of Types. *Journal of Quantitative Linguistics*, 9-3, 193-214.

Péaquin G. (2003). *Segmentation automatique des corpus Rapport de stage*. Grenoble: Polytech'Grenoble & Institut d'Etudes Politiques.

Pielou E.C. (1982). "Diversity Indices". Johnson & Kotz (eds). *Encyclopedia of Statistical Sciences*. Vol 2. New York: Wiley.

Scheffé M. (1959). *The Analysis of Variance*. New York: Wiley.

Wimmer G. & Altmann G. (1999). Review Article: On Vocabulary Richness. *Journal of Quantitative Linguistics*, 6-1, 1-9.

Appendix Racine's work

N°	Titre	Genre	Date	Length (tokens)	Types
1	Thébaïde	Tragedy	1664	13,813	1,313
2	Alexandre	Tragedy	1665	13,864	1,372
3	Andromaque	Tragedy	1667	15,076	1,392
4	Plaideurs	Comedy	1668	8,041	1,312
5	Britannicus	Tragedy	1669	15,387	1,637
6	Bérénice	Tragedy	1670	13,242	1,346
7	Bajazet	Tragedy	1672	15,297	1,507
8	Mithridate	Tragedy	1673	15,091	1,550
9	Iphigénie	Tragedy	1674	15,782	1,604
10	Phèdre	Tragedy	1677	14,394	1,775
11	Esther	Tragedy	1989	11,147	1,656
12	Athalie	Tragedy	1691	15,492	1,656
	Entire work			166,626	4,322