A Model of Vocabulary Partition

Pierre Hubert, Dominique Labbé

A Model of Vocabulary Partition

Pierre HUBERT
Ecole des Mines de Paris (France)

D. LABBE
Université de Grenoble II (France)

Abstract
The model proposed here is used to describe the vocabulary of a corpus. It is divided into two groups: general vocabulary which is used whatever the circumstances and several local (or 'specialized') vocabularies, each of which is used in only one part of the corpus. General words may appear everywhere in the text and their increase with corpus length can be estimated with Muller's formula. In this model, a partition parameter measures the relative importance of both types of vocabularies: so the value of this parameter gives an estimation of the lexical 'specialization' in the text.
This model has been applied to Racine's plays and can also be used to measure the increase of vocabulary with corpus length, to locate stylistic changes or to compare several texts from the point of view of their lexical richness.

Résumé
On propose un modèle destiné à décrire le vocabulaire d'un corpus. Il est divisé en deux groupes : le vocabulaire général, utilisé quelles que soient les circonstances, et de plusieurs vocabulaires locaux ou "spécialisés", utilisés uniquement dans une partie du corpus. Les mots appartenant au vocabulaire général apparaissent partout dans le texte et leur rythme d'apparition peut être estimé grâce à la formule de Muller. Un paramètre de partition mesure le poids relatif des deux vocabulaires. Ce paramètre donne donc une estimation de la spécialisation du vocabulaire dans un texte ou un corpus.
Ce modèle est utilisé pour mesurer l'accroissement du vocabulaire avec l'allongement du corpus, pour localiser les ruptures thématiques et stylistiques dans ce corpus et pour comparer différents textes du point de vue de leur richesse lexicale. On présente une application aux pièces de Racine.

Key words
statistics for linguistics; corpora; vocabulary growth; vocabulary specialization; vocabulary richness

Mots clefs
Lexicométrie ; corpus ; croissance du vocabulaire ; specialization du vocabulaire ; richesse du vocabulaire

Draft of the paper published in:
Literary and Linguistic Computing, Vol. 3, No. 4, 1988, pp. 223-225.

More than twenty years ago, C. Muller introduced the classical notion of the ballot box into lexical statistics (Muller, 1964): this views a text of N words as the result of N random excerpts from a vocabulary which contains V types. Define V_i as the number of types which have an absolute frequency *i* in the text. In this model, any excerpt, of N' tokens in length, may be expressed as a sample composed of V types which is obtained as a random drawing of N' words without replacement from the entire text. The expected value of V (i.e. the number of different types in excerpts of this length) is approximated by:

$$V'(u) = V - \sum_{i=1}^{i=n} V_i \; Q_i(u) \;\; avec \; u = \frac{N'}{N}$$

where n is the frequency of the most frequently occurring word. The probability of each type with frequency *i* not appearing in the sample, is:

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

This formula is an excellent approximation to the hypergeometric pattern (Hubert-Labbé, 1988a, p. 83-85). The theoretical variance of V'(u) is:

$$Var[V'(u)] = \sum_{v=1}^{v=V} Q_{i(v)}(u)\left[1 - Q_{i(v)}(u)\right] \quad (1)$$

if V’ is assumed to have been obtained through independent sampling. As D. Serant has demonstrated, this calculation leads to a slight overestimation of the variance which would be obtained by the strict application of the hypergeometric law (Serant, 1988).

If this model for the description of the vocabulary used in a text is accurate, 95% of the values V'_* observed in any excerpt of a text should be located in an interval of + 2σ around V’ (as calculated by Muller's formula). However, close examination of numerous corpora has shown that empirical values are almost always outside this confidence interval and below the calculated values. This has been demonstrated, for example, for the works of Corneille (Muller, 1967) or for those of Racine (Bernet, 1983, p. 111-126; see also below, graph number 2). These two findings leave a doubt hanging over the validity of the random drawing without replacement model: the 'lexical specialization' of excerpts of the text has an influence

on the probability that certain types will appear, which rules out the random drawing hypothesis in its strictest sense (Muller, 1970, p. 302; Muller, 1977, p. 142-144).

Is it possible to devise a model which accommodates and allows the measurement of the phenomenon of lexical specialization, and this enables estimation of the influence of specialization on vocabulary? This is the ambition of the 'partition model' presented here.

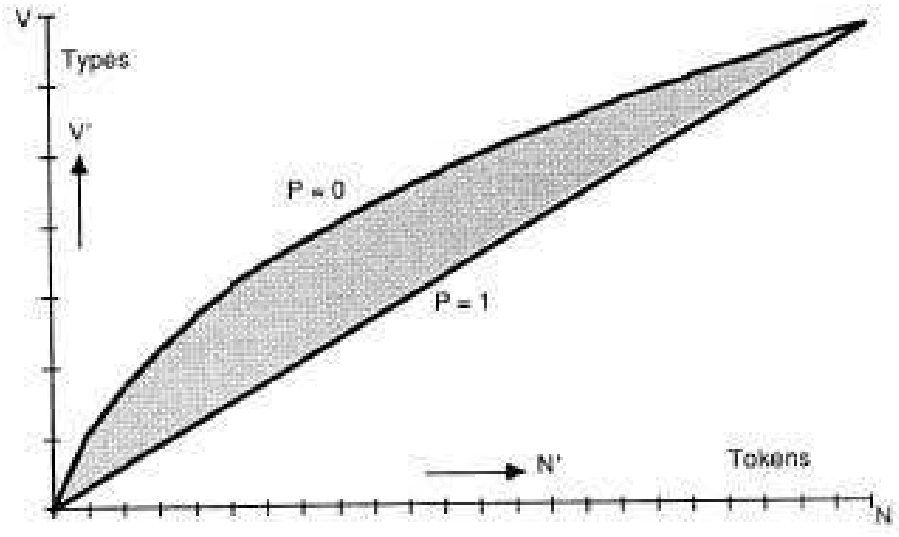
The foundations of the partition model

If we acknowledge the existence of lexical specialization, a corpus is seen to result from a two-fold process (Hubert & Labbé, 1988b). On the one hand, the author uses a general, 'non-specialized' vocabulary which contains, for example, 'words of relation' (articles, prepositions and pronouns) and the most common verbs. On the other hand, the author draws upon several local or specialized vocabularies which encompass the terms used within a single excerpt only. The contribution made by the local vocabularies (V_s) and by the general vocabulary (V_g) to the total vocabulary (V) is measured by a coefficient p:

$$p = \frac{V_s}{V} \text{ and } q = \frac{V_g}{V} = 1 - p \; ; 0 \leq p \leq 1$$

Let us apply this idea to a classic problem of lexical statistics: the prediction of the number of different types (V') as N' increases from 1 to N. The theoretical growth of the variable V' is described in Figure 1.

Fig. 1 The possible values of V in function of p.



All the possible values of V’ are included in the grey shaded area. We can therefore envisage two extreme cases:

- If p is equal to one, all the words of the text have been drawn from local stocks of words. All excerpts are distinct from one another. V’ will be a linear function of N':

$V'_1(u) = u.V$ (with u = N’/N)

- Conversely, if p is equal to zero, all the types of the text come from the general vocabulary: the text presents no lexical specialization, it constitutes of types whose probability of appearing at some stage in the text is constant. The mathematical expectation of the number of different types in an excerpt of N' tokens in length, follows the hypergeometric law that can be estimated with Muller's formula:

$$V'_0(u) = V - \sum_{i=1}^{i=n} V_i Q_i(u)$$

Between these two extreme cases, fall the values V'_p:

$$V'_p(u) = p\,u\,V + q\left[V - \sum_{i=1}^{i=n} V_i Q_i\right] \quad (2)$$

The coefficient p estimates an intrinsic character of the text namely the division between the general vocabulary and the specialized vocabularies which have been drawn on to produce it. We propose to name p: 'coefficient of vocabulary specialization' or better still, 'coefficient of vocabulary partition’.

In order to identify and estimate this coefficient, we generally have at our disposal a series of empirical values observed in the text: V (the total number of different types noted in the totality of the text); V-t (the frequency distribution of the types). Finally, we know, for a number of values of u, the number of different types that have appeared since the beginning of the text, that is:

$$V'_*(u_k) \text{ pour } u_k = \frac{N'_k}{N} \text{ avec } k = 1, 2, \ldots, K$$

We consider that the most appropriate value for p will be that value which minimizes the sum of the quadratic deviations between the observed-values (the $V'_{*}+ (u_k)$) and the calculated values: the $V'_p(u_k)$. We obtain:

$$p = \frac{\sum_{k=1}^{k=K} X(u_k) \,.\, Y(u_k)}{\sum_{k=1}^{k=K} X(u_k)^2} \qquad (3)$$

with:

$$X(u_k) = (u_k - 1)V + \sum_{i=1}^{i=n} V_i Q_i(u_k) \text{ and } Y(u_k) = (u_k - 1) - V + \sum_{i=1}^{i=n} V_i Q_i(u_k)$$

Practically, as with any statistic, the accuracy of p will be influenced by the number and the quality of the observations used to calculate it (it would appear that six to eight values of V' — almost evenly distributed within the range of u — are necessary to obtain a reliable value for p).

Applications of the partition model

Among the possible applications, our model can provide information on the style of an author or of a work. For example, in the whole of Racine's tragedies, p reaches 0.33. In other words, on average for his works, one type in three is specialized; that is to say it is employed exclusively in a single play. This is an indication of the great diversity of vocabulary in the complete works of Racine, such diversity being found even within single play. We should remember in effect that classical French tragedy was constrained by somewhat strict rules, and that the vocabulary of the 17th century was probably less extensive than that of contemporary French.

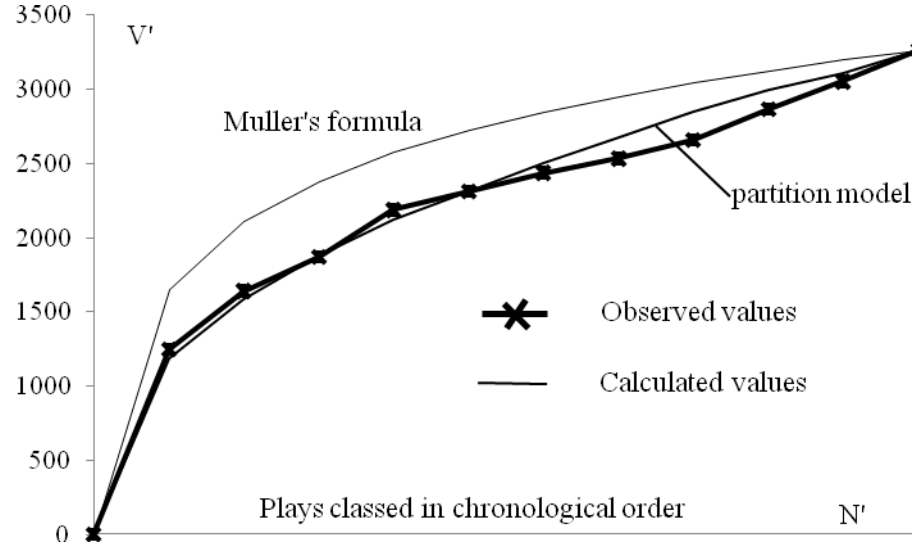
For the whole of a text, the model also allows us to adjust the theoretical values ($V'_p(u)$) to the observed values ($V'_{*}u$)) by applying formulae (2) and (3). These theoretical values can be given confidence intervals that are calculated, using formula (1), on the only 'probabilistic' part of the whole vocabulary (the general types). One can see that the last term of formula (2), which corresponds to the calculation of the number of general types, simulates a drawing: therefore $V'_p(u)$ is partly a random variable. The results of an experiment carried out on

Racine's works are shown below. This experiment involves describing the appearance of new types in the tragedies (classed in chronological order: see Table 1 and Figure 2).

Table 1: The increase of vocabulary in Racine's works. V'_*: number of different types which appeared since the beginning of the first tragedy; V'_p: values calculated using Muller's formula and partition model; and the associated standard deviations (according to the findings of C. Bernet).

Tragedies	Observed values	Theoretical Values		
	V'_*	Muller's formula	Partition Model (V'_p)	Standard deviation (σ)
La Thébaïde	1244	1656.6	1184.3	13.8
Alexandre	1638	2111.6	1583.9	13.4
Andromaque	1868	2382.6	1880.5	12.9
Britannicus	2185	2576.0	2123.7	12.2
Bérénice	2310	2726.7	2306.4	11.5
Bazajet	2435	2850.2	2498.5	10.7
Mithridate	2533	2954.8	2674.2	9.7
Iphigénie	2659	3045.5	2847.1	8.4
Phèdre	2867	3125.6	2997.4	6.9
Esther	3052	3197.2	3110.1	5.3
Athalie	3262	3262	3262	0

Fig. 2 The increase of vocabulary size in Racine's tragédies, using Muller's formula and partition model (data from table 1).

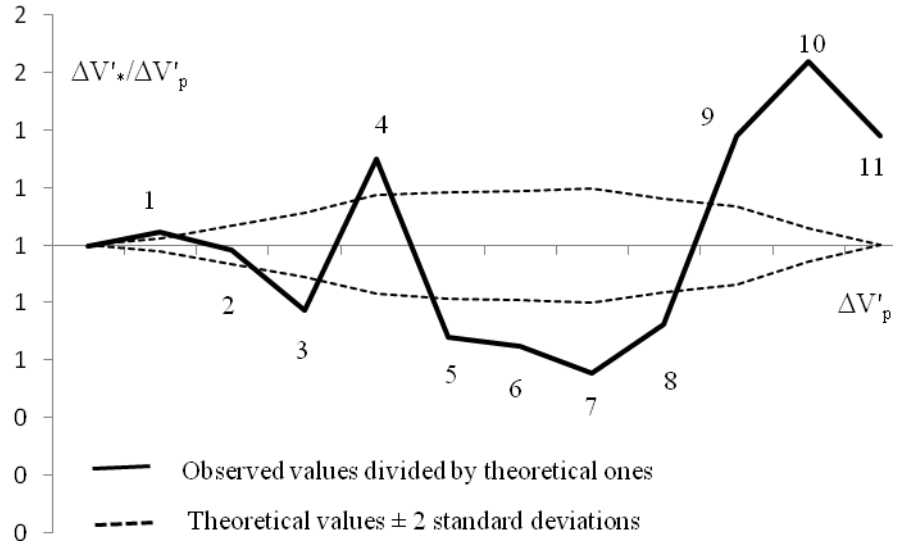


It is worth noting the nature of the adjustment obtained with the partition model: the curve of the calculated values passes through the concentration of points and not above them as with

Muller's formula. The observed values are distributed around both sides of points and not above them as with Muller's formula. The observed values are distributed around both sides of the theoretical values and are located within or near the confidence interval. The values calculated with our model can therefore be considered as an accurate description of Racine's works, had the style and the themes treated remained constant from the first to the last tragedy. When this is not the case, two situations are possible. When the observed values are greater than the calculated values, it can be considered that there is an influx of new terms from a specialized vocabulary, and therefore that the author is tackling a new theme. On the other hand, a value which is lower than the average indicates that less new specialized words are being introduced. Consequently, our model can be used to pin-point the thematic variations within a corpus.

To confirm this hypothesis on Racine's works, we have studied the development of vocabulary from one play to another: Figure 3. The theoretical growth ($\Delta V'_p$) forms the x-axis. It have been given a confidence interval: if the law of progression of vocabulary observed in the totality of the works was in operation throughout every play, all the points of the curve ($\Delta V'_* / \Delta V'_p$) should be found within the dotted lines around the horizontal axis. As it is not the case, we can conclude that there are significant stylistic (or thematic) variations in the corpus.

Fig. 3 The apparition of new types in Racine's tragedies using the model of vocabulary partition (V'_p).



1. La Thébaïde, 2. Alexandre, 3. Andromaque, 4. Britannicus, 5. Bérénice, 6. Bazajet, 7. Mithridate, 8. Iphigénie, 9. Phèdre, 10. Esther, 11. Athalie.

The graph brings to light some facts which have already been noted by commentators on Racine: the special place occupied by Britannicus and the three 'sacred' tragedies: Phèdre, Esther and Athalie (Bernet 1983, p. 118).

As a **conclusion**, we would like to emphasize that the main feature of our vocabulary partition model is that it provides a supplementary parameter (p) in the now classic procedures of lexical statisticians. This calculation offers at least a new information: the precise estimation of lexical specialization, or, in other words, of the diversity of the vocabulary used in the corpus, More generally, taking account of this coefficient improves the accuracy of calculations which are traditionally used in lexical statistics.

Bibliography

Bernet C, 1983, *Le vocabulaire des tragédies de Racine (Analyse statistique)*, Genève, Slatkine.

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. (Eds), *Etudes sur la richesse et la structure lexicales (Vocabulary Structure and Lexical Richness)*, Genève-Paris, Slatkine-Champion, pp. 77-91.

Hubert P., Labbé D., 1988b, 'Un modèle de partition du vocabulaire' in Labbé D., Serant D., Thoiron P. (Eds), *Etudes sur la richesse et la structure lexicales (Vocabulary Structure and Lexical Richness)*, Genève-Paris, Slatkine-Champion, pp. 93-114.

Muller C., 1964, 'Calcul des probabilités et calcul d'un vocabulaire', in *Langue française et linguistique quantitative*, Genève, Slatkine, 1979, pp. 167-176.

Muller C., 1967, *Etude de statistique lexicale, Le vocabulaire du théâtre de Pierre Corneille*, Genève, Slatkine, 1979.

Muller C., 1970, 'Sur la mesure de la richesse lexicale', in *Langue française et linguistique quantitative*, Genève, Slatkine, 1979, pp.281 307.

Muller C., 1977, *Principes et méthodes de statistique lexicale*, Paris, Hachette.

Serant D., 1988, 'A propos des modèle de raccourcissement de textes' in Labbé D., Serant D., Thoiron P., *Etudes sur la richesse et la structure lexicales (Vocabulary Structure and Lexical Richness)*, Genève-Paris, Slatkine-Champion, pp. 77-91.