Un modèle de partition du vocabulaire

Pierre Hubert, Dominique Labbé

Pierre HUBERT
(Ecole des Mines de Paris)

Dominique LABBE
(Université de Grenoble II)

Un modèle de partition du vocabulaire

Manuscrit auteurs de l'article paru dans :
Dominique Labbé, Philippe Thoiron, Daniel Serant (Ed.). *Etudes sur la richesse et la structures lexicales*. Genève-Paris : Slatkine-Champion, 1988, p. 93-114.

Résumé
On propose ici un modèle de description du vocabulaire employé dans un corpus ; il est partagé en deux groupes : un vocabulaire général employé quelles que soient les circonstances et des vocabulaires locaux (ou "spécialisés") dont chacun est mobilisé dans une partie seulement du corpus. Les vocables généraux peuvent apparaître en n'importe quel point du texte et leur accroissement, en fonction de la taille du corpus, peut être estimé grâce à la formule de Muller. Dans le modèle, un paramètre de partition estime le poids relatif des deux vocabulaires : la valeur de ce paramètre donne donc une estimation de la spécialisation lexicale à l'œuvre dans le corpus. Des applications de ce modèle sont conduites sur l'œuvre de Racine et sur des débats télévisés (Giscard-Mitterrand et Chirac-Fabius),
Le modèle de partition peut être également utilisé pour calculer l'accroissement du vocabulaire dans un corpus, pour y localiser des variations stylistiques ou pour comparer plusieurs textes du point de vue de leur "richesse de vocabulaire".

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 'specialised') 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 size 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 'specialisation' in the text. This model has been applied to Racine's plays and TV debates (Giscard *vs* Mitterrand, Chirac *vs* Fabius).
The partition model 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.

Considérons le schéma classique de l'urne tel que le pose la statistique lexicale (cf. dans le même ouvrage, notre article "Note sur l'approximation de la loi hypergéométrique par la formule de Muller" : Hubert & Labbé 1988) : soit un échantillon, d'effectif *N'*, obtenu par tirage exhaustif sur un texte contenant *V* vocables dont V_i de fréquence absolue *i* (*i* variant de 1 à *n*). L'espérance mathématique du nombre de vocables contenus dans un tel échantillon s'exprime comme :

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

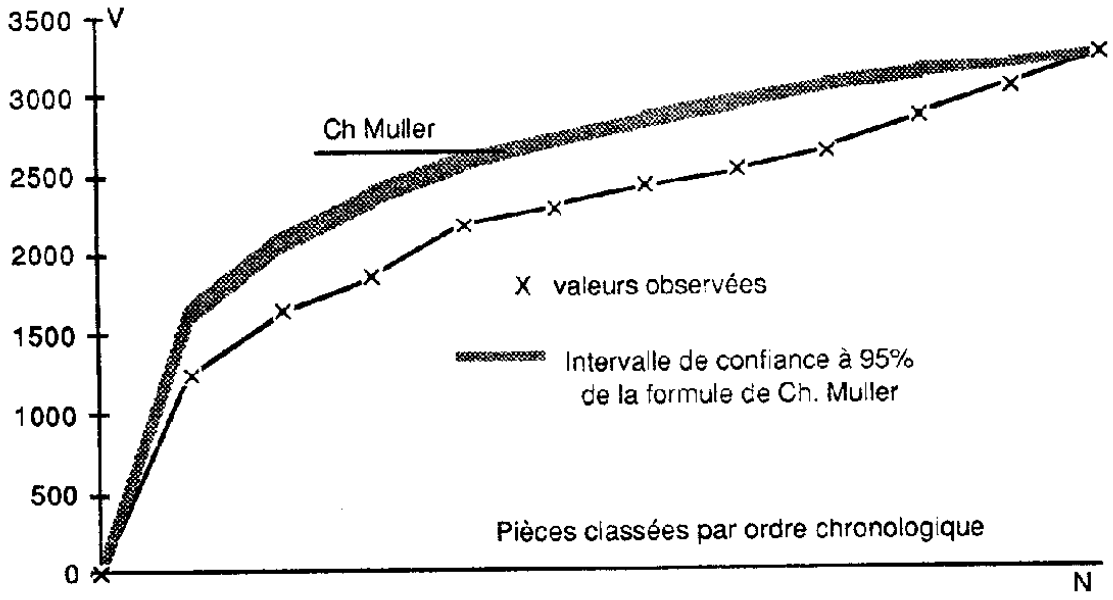
puisque chaque vocable de fréquence *i* (il en existe V_i) a une probabilité $Q_i(u)$ de ne pas figurer dans l'échantillon :

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

Cette expression, rappelons-le, aboutit à une erreur négligeable par rapport au schéma hypergéométrique rigoureux. Or les valeurs empiriques observées par les statisticiens du lexique sont presque toujours hors de l'intervalle de confiance égal à 2 écarts types et, quasi-systématiquement, inférieures aux valeurs calculées. Ainsi, pour l'accroissement du vocabulaire dans l'œuvre de J. Racine, le graphique I ci-contre - établi grâce au dépouillement de C. Bernet - montre que, sauf dans les observations les plus proches de la limite supérieure du corpus, la courbe en fuseau de l'intervalle de confiance, qui est censée s'ajuster aux données connues, se situe toujours au dessus de celles-ci.

Cette caractéristique avait déjà été notée par C. Bernet (1983, pp. 111-126). A l'instar de ce qu'avait fait C. Muller pour P. Corneille, C. Bernet a partiellement corrigé ce biais en sortant les noms propres du calcul de *V*. D'autres solutions ingénieuses ont été proposées comme, par exemple, celle de E. Brunet consistant à effectuer un deuxième ajustement en commençant par la fin du corpus (Brunet : 1978). Aucune de ces tentatives n'a été totalement couronnée de succès. Pour notre part, nous avons travaillé sur l'ensemble des vocables utilisés par J. Racine (noms propres et noms communs).

Graphique I. Nombre de vocables cumulés dans l'œuvre de J. Racine : (valeurs calculées avec la formule de Muller et valeurs observées, dépouillement de C. Bernet)



Cet ajustement défectueux se retrouve dans toutes les études de statistiques lexicales fondées sur l'utilisation de la formule de Muller. Comme nous l'avons démontré, on ne peut incriminer l'erreur commise en substituant cette formule à celle de la loi hypergéométrique (Hubert & Labbé 1988). Cette double constatation laisse alors planer un doute sur la validité même du modèle - la fameuse urne - qui se trouve à la base de ces opérations. Pour le moins, on doit admettre que le schéma probabiliste strict introduit un biais systématique. L'explication de ce biais est d'ailleurs connue ; ainsi que le notait C. Muller à propos de ces calculs :

"Cet écart augmente à mesure que l'étendue des fragments s'éloigne de celle de l'ensemble pris comme référence, et la moyenne des nombres observés est régulièrement inférieure aux nombres calculés. Il faut voir là un effet de la spécialisation lexicale qui, quoique faible, est ainsi rendue sensible".

"On se souviendra que les calculs théoriques sont fondés sur une hypothèse qui rejette la spécialisation lexicale des fragments et qui, postulant une stabilité parfaite du lexique virtuel, admet que chaque vocable a autant de chances d'apparaître dans un fragment que dans un autre, à longueur égale de ceux-ci, or il est évident que la réalité s'écarte d'autant plus de cette hypothèse stylistiquement nulle que des causes diverses, thématiques ou stylistiques modifient cette probabilité" (Muller, 1970 : p. 302 ; également Muller : 1977 pp. 142-144))

Est-il néanmoins possible de concevoir un schéma qui tienne compte de ce phénomène et qui le neutralise, au moins partiellement, ne serait-ce que pour en estimer l'influence réelle sur le vocabulaire ? Telle est l'ambition du modèle que nous présentons ci-dessous.

LES FONDEMENTS DU MODELE

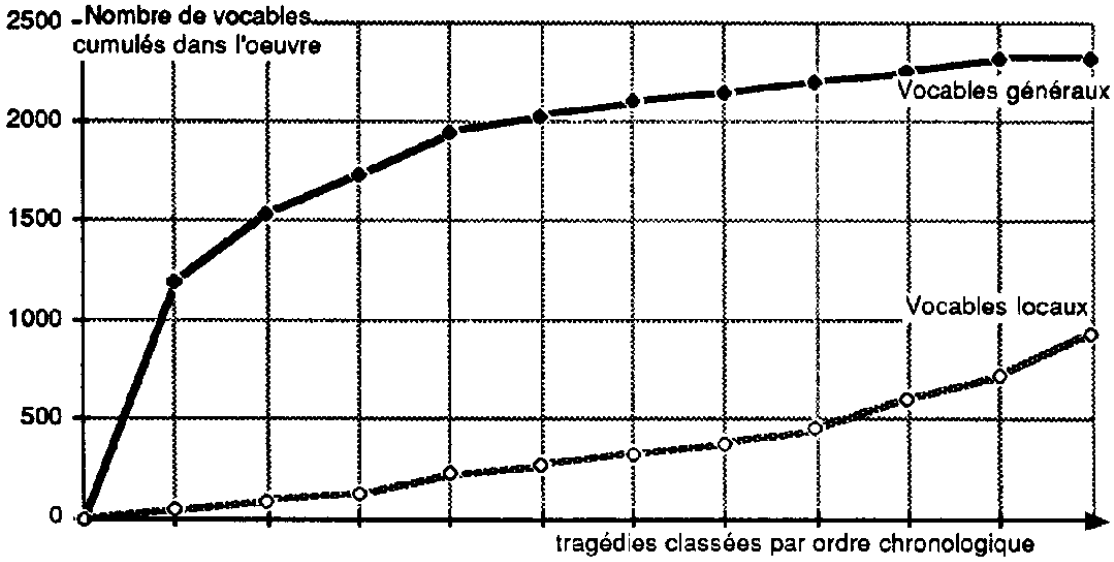
Au point de départ, nous poserons que tout corpus est le résultat d'un double processus.

D'une part, l'auteur mobilise un vocabulaire que l'on peut nommer général, ou "polyvalent" ou encore "non-spécialisé", en ce sens qu'il contient tous les vocables utilisés quelles que soient les circonstances. On trouve, dans cette catégorie, les mots dit "de relation" (articles, prépositions, pronoms), les verbes les plus usuels, les auxiliaires "être" et "avoir", "faire" et les quatre pseudo-auxiliaires- souvent appelés verbes modalisateurs - ("vouloir", "savoir", "pouvoir", "devoir"). Au delà d'une liste courte, et relativement commune à tous les utilisateurs du français, la composition de ce vocabulaire dépendra de chaque locuteur. Ainsi le passionné d'un sport se reconnaît à ce qu'il glisse, dans n'importe quelle conversation, des expressions tirées de sa discipline favorite.

D'autre part, l'auteur mobilise des vocabulaires "locaux" ou "spécialisés". Ces réservoirs sont aussi nombreux que l'exige la situation de communication et que le permet la culture de l'auteur. Par exemple, dans les articles de ce livre, on pourra observer l'utilisation conjointe, ou successive, de notions de statistique et de lexicologie.

Pour attester l'existence de ces réservoirs multiples nous proposons une expérience sur le vocabulaire de J. Racine (en nous appuyant sur les dépouillements de C. Bernet. Nous posons que tout vocable employé dans plus d'une tragédie de J. Racine appartient au vocabulaire "général" de l'auteur et que les autres - ceux qui sont employés dans une seule pièce -sont des vocables "locaux" appartenant à un vocabulaire particulier que l'auteur a utilisé pour la circonstance. Observons l'évolution de ces deux catégories de vocables au long des onze tragédies classées par ordre chronologique : graphique II.

Graphique II : vocables généraux et vocables locaux dans les tragédies de J. Racine (d'après le dépouillement de C. Bernet).



Les vocables généraux - du fait même de leur polyvalence, de leur nature d'outils - apparaissent en grand nombre au début de l'œuvre et, au fur et à mesure que celle-ci s'allonge, la probabilité de rencontre d'un nouveau vocable général diminue rapidement. De par la nature même de son contenu, l'épuisement de ce réservoir est donc rapide ; la courbe se rapproche assez vite de l'horizontale. Ce phénomène n'est pas propre à J. Racine : toutes les expériences comparables, menées sur des textes très divers, donnent une courbe de même profil : la croissance du vocabulaire généraï semble obéir à une fonction puissance de *N* dont l'exposant serait nettement inférieur à 1.

Pour reprendre le schéma probabiliste nous dirons donc que le texte provient partiellement d'une série de tirages aléatoires effectués dans le vocabulaire général. Les vocables qui en sont issus sont des vocables "à tout faire". Ici, "emploi non-spécialisé" ne signifie pas probabilité d'emploi stable mais que nous sommes assurés de les rencontrer dans un texte ayant une longueur minimale et ceci quelles que soient les circonstances d'émission de ce texte.

En revanche, l'apparition des vocables locaux semble à peu près constante au cours du temps (avec, dans les quatre dernières tragédies, une légère accélération sur laquelle nous reviendrons). Dès la moitié de l'œuvre, l'accroissement du vocabulaire provient en majorité de cette seconde catégorie. Il semble donc que nous soyons ici en présence d'une fonction linéaire de *N*. Autrement dit, la croissance de cette partie du vocabulaire ne dépend pas du hasard, elle semble simplement liée à l'allongement du corpus. Toutes les expériences du même genre, tentées sur des textes divers, nous conduisent à la même conclusion : manifestement une partie du vocabulaire obéit à des lois qui ne doivent rien au hasard...

Naturellement, l'expérience ne fait qu'approcher le phénomène. D'un côté, elle surestime le nombre de vocables locaux en y incluant des hapax qui, probablement, ne devraient pas s'y trouver. Mais, d'un autre côté, elle sous-estime ce nombre puisque nous avons respecté la partition en tragédies alors que l'on sait que J. Racine a utilisé le même registre - par exemple l'antiquité gréco-romaine - dans plusieurs de ses pièces : notre partition place donc des éléments tirés de ce registre dans le vocabulaire général - tel que défini ci-dessus - alors qu'ils devraient se trouver dans un réservoir spécifique... Cette remarque permet de comprendre pourquoi l'idée ancienne des réservoirs multiples n'a jamais été poursuivie très loin malgré son intérêt : le problème réside dans l'observation et la mesure du phénomène. Si les mesures ne sont pas indépendantes du découpage opéré par l'observateur, le modèle n'a plus grande utilité ; de plus, tout tri entre les vocables peut être accusé d'arbitraire même s'il repose sur des critères simples comme ceux que nous avons employés ci-dessus.

Pour résoudre cette difficulté, nous poserons que le modèle ne peut dire si tel ou tel vocable particulier provient du réservoir général ou des réservoirs locaux. Il a simplement pour fonction d'indiquer le poids relatif de chacune de ces deux catégories dans la constitution du corpus ; il identifiera l'importance relative des vocables généraux par rapport aux vocables locaux. Cette proportion est attachée au corpus lui-même : on l'observera en n'importe quel point du texte et elle sera indépendante des découpages opérés par l'observateur. Autrement dit, le modèle a pour ambition de réaliser une partition du vocabulaire, utilisé dans un corpus donné, en deux sous-ensembles : nous le baptiserons donc "**modèle de partition du vocabulaire**".

Les autres éléments du modèle s'ordonnent autour de cette idée :

- tout d'abord, le lecteur aura remarqué que nous employons "vocabulaire" et non pas "lexique". L'existence d'un lexique virtuel est aujourd'hui une idée incontestée en statistique lexicale et nous recherchons justement des informations qui permettront d'estimer certaines caractéristiques propres au lexique de l'auteur étudié. Mais l'estimation de l'étendue du lexique se heurte, encore aujourd'hui, à des difficultés et donne lieu à des controverses. C'est pourquoi nous travaillerons sur !e vocabulaire observé : nous poserons que le contenu de tous les réservoirs, définis ci-dessus, est épuisé lorsque la prestation s'achève ; à la fin du texte, sa structure et sa population se trouvent entièrement connus par l'observateur.

- enfin, dernier postulat, tout texte d'une longueur minimale peut être découpé en fragments possédant chacun une unité thématique, c'est-à-dire que chaque fragment provient en partie d'un réservoir mobilisé localement et à cette seule occasion. Par exemple, pour l'œuvre de J. Racine, ces fragments peuvent être les différentes tragédies ou, pour chacune de ces tragédies, les actes, les rôles, etc. Qu'ils soient voulus par l'auteur ou détectés par l'analyse, ces découpages fournissent des valeurs empiriques permettant de chiffrer le modèle.

Ce chiffrage est inscrit dans certaines contraintes. Ainsi la dimension des réservoirs locaux n'est évidement pas connue à l'issue d'un dépouillement "standard". Dès lors, nous poserons comme hypothèse de départ que tous les réservoirs locaux contiennent un nombre de vocables proportionnel à la taille des fragments où ils sont mobilisés. A fortiori, nous ignorons également la loi de distribution des vocables au sein de ces différents vocabulaires : nous devons poser que tous ces réservoirs (général et locaux) ont des structures identiques et homothétiques à la seule que nous connaissions : la distribution des fréquences qu'on observera sur le texte entier.

Naturellement, ces hypothèses ne sont pas vraiment conformes à la réalité. Voici rapidement énumérées les objections qui peuvent être faites ;

- nous savons que les vocables, même les plus usuels, connaissent de grandes fluctuations d'emploi : c'est toute la question de la "répartition" tant débattue en statistique lexicale. Cependant, il n'est pas absurde de considérer que la structure du vocabulaire (la distribution des fréquences) est nettement plus stable que chacun des vocables qui la composent. Or c'est sur cette structure que nous raisonnons ;

- nous savons également que, suivant les sujets qu'il traite, un auteur disposera d'un vocabulaire plus ou moins étendu ;

- la situation, entendue dans un sens restreint, joue également son rôle ; elle interdit ou impose l'utilisation de certains vocables ce qui renforce ou minimise marginalement la spécialisation du vocabulaire utilisé.

Mais ces deux derniers mouvements sont nécessairement localisés. Notre dernière hypothèse se trouvera prise en défaut en un point précis ; il sera aisé de localiser le phénomène puis de le mettre en relation avec le thème, les événements qui ont pu provoquer cet "accident".

Aucune des objections habituelles n'atteint donc vraiment le modèle simple dont nous venons de poser les fondements. Pour illustrer la manière dont il peut être mis en œuvre, nous commencerons par reprendre l'expérience classique de Yule consistant à découper un texte en parties égales et à y observer l'apparition des vocables nouveaux.

LE MODELE DE PARTITION DU VOCABULAIRE

Soit donc un texte de N mots découpé en K tranches égales. Ce texte comporte V vocables dont V_i de fréquence i (i variant de 1 à n) avec :

$$V = \sum_{i=1}^{i=n} V_i$$

$$N = \sum_{i=1}^{i=n} i\, V_i$$

Nous supposons donc que chaque segment de ce texte est issu de tirages exhaustifs réalisés parallèlement dans deux urnes. Si le corpus comprend K tranches distinctes, sa réalisation aura mobilisé K+1 urnes (K urnes locales pour chacune des K segments et une urne contenant les vocables généraux). K de ces ensembles, d'effectifs m égaux, sont associés aux K tranches du texte initial ; ils comportent le même nombre w de vocables dont w_i de fréquence i.

Nous avons donc, pour ces K urnes locales,

$$w = \sum_{i=1}^{i=n} w_i$$

$$m = \sum_{i=1}^{i=n} i\, w_i$$

Le dernier ensemble contient les vocables "généraux" que l'auteur utilise quel que soit le passage, c'est-à-dire sans considération du thème traité. Cet ensemble se trouve associé à la totalité du texte et contient M mots dont W vocables, dont la fréquence i varie de 1 à n, dans chacun des ensembles définis ci-dessus. On a donc :

$$W = \sum_{i=1}^{i=n} W_i$$

$$M = \sum_{i=1}^{i=n} i\, W_i$$

Comme nous l'avons exposé en introduction, nous compléterons ces égalités par deux postulats :

- premièrement, nous poserons que la spécialisation du vocabulaire est "parfaite". Les K+1 ensembles définis ci-dessus sont disjoints : un même vocable ne peut se retrouver dans deux ensembles différents ; ou bien il appartient au vocabulaire général (les W_i) et peut apparaître en n'importe quel point du texte, ou bien il appartient à l'un des K ensembles

propres à une tranche (les w_i et ne peut être employé que dans cette tranche. Ce dernier postulat entraîne les égalités suivantes :

$V = W + kw$

$N = M + Km$

- deuxièmement, nous supposerons que ces K+1 ensembles ont une structure semblable c'est-à-dire que les fréquences i des w_i et des Wi suivent une distribution identique à celle observée sur le texte entier. Comme nous l'avons déjà signalé, ce dernier postulat est imposé par la nature même des renseignements dont on dispose habituellement sur un texte : ainsi, à l'issue d'un dépouillement "standard", connaît-on la distribution des fréquences sur l'ensemble mais non sur les différents réservoirs a priori inconnus.

Notre second postulat conduit donc aux égalités suivantes :

$$\frac{W_1}{w_1} = \frac{W_2}{w_2} = ... = \frac{W_u}{w_u} = \frac{W_1 + W_2 + ... + W_n}{w_1 + w_2 + ... + w_n} = \frac{W}{w} = \lambda$$

Compte tenu de ce même postulat, le coefficient de proportionnalité λ se retrouvera dans le vocabulaire général :

$$M = \sum_{i=1}^{i=n} i\, W_i \;=\; \sum_{i=1}^{i=n} i\, \lambda\, w_i \;=\; \lambda \sum_{i=1}^{i=n} i\, w_i = \lambda\, m$$

On pourra alors écrire :

$$V = W + Kw = W + \frac{K}{\lambda} W = \left(\frac{K + \lambda}{\lambda} \right) W$$

$$N = M + Km = M + \frac{K}{\lambda} M = \left(\frac{K + \lambda}{\lambda} \right) M$$

Le processus d'accroissement du vocabulaire d'un texte ou d'une œuvre peut alors être décrit de la manière suivante. Dans chaque tranche, d'effectif N/K, on rencontrera les m mots de l'ensemble associé à la tranche. Le reste des mots rencontrés proviendra de l'ensemble général par tirage exhaustif. De telle sorte que, au sortir de chaque tranche, l'ensemble associé (le vocabulaire propre à cette tranche) se trouve épuisé et, au sortir du texte, le vocabulaire général à son tour est entièrement connu : il comporte M mots en Wi vocables de fréquence 1 à n.

On peut alors évaluer l'espérance mathématique du nombre de vocables contenus dans les k premières tranches du texte. Celle-ci dépendra de deux processus :

- d'une part, on aura épuisé les k urnes associées aux k tranches, soit un nombre de vocables égal à : kw.

- on aura, d'autre part, réalisé un tirage exhaustif d'un certain nombre de vocables dans l'ensemble général. L'espérance mathématique du nombre de vocables issus de cette opération étant, selon le modèle classique,

$$W - \sum_{i=1}^{i=n} W_i Q_i \left[\frac{k\left(\frac{N}{K} - m\right)}{M} \right]$$

qui peut s'écrire, puisque $N = M + Km$,

$$W - \sum_{i=1}^{i=n} W_i Q_i \left[\frac{k}{K} \right]$$

On aura donc en définitive :

$$V'(k,K) = kw + W - \sum_{i=1}^{i=n} W_i Q_i \left[\frac{k}{K} \right]$$

Tenant compte des relations entre, d'une part, V, W et w et, d'autre part, entre les V_i et les W_i, il vient :

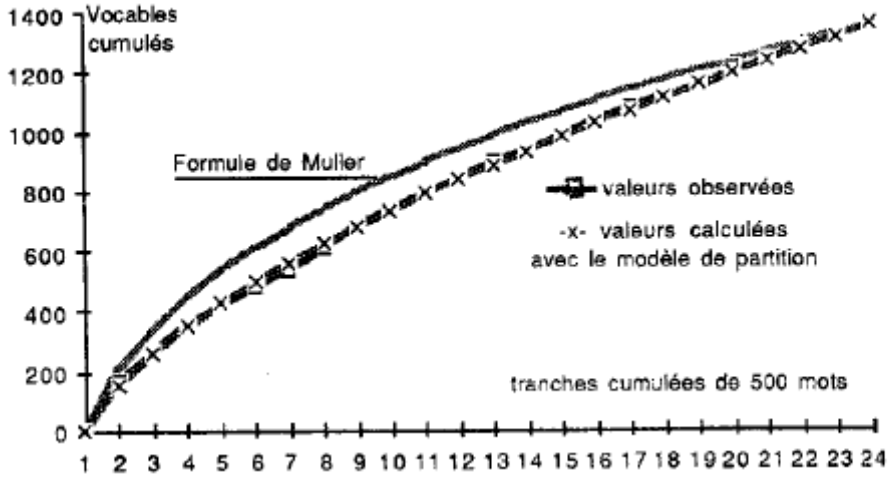
$$V'_{\lambda}(k,K) = \left(\frac{k+\lambda}{K+\lambda} \right) V - \frac{\lambda}{K+\lambda} \sum_{i=1}^{i=n} W_i Q_i \left[\frac{k}{M} \right]$$

Comme nous l'avons dit plus haut, λ est un coefficient dont la valeur exprime le rapport entre les vocables généraux et les vocables locaux. Nous présenterons plus loin la manière dont il peut être estimé. Nous avons appliqué ce modèle aux textes des deux débats dont il a déjà été question plus haut. Les résultats de ces expériences, pour les 9.500 premiers mots de la rencontre télévisée entre F. Mitterrand et V. Giscard d'Estaing, sont retracés dans le tableau I ainsi que dans le graphique III.

Tableau I : Vocables cumulés dans le débat V. Giscard – F. Mitterrand découpé en tranches de 500 mots (V = valeurs observées et V = valeurs calculées avec le modèle de partition du vocabulaire)

N’	F. Mitterrand		V. Giscard d'Estaing	
	V’	*V*	*V’*	*V*
500	164,7	204	160,3	178
1000	276,7	303	266,6	270
1500	371,9	385	354,8	360
2000	457,3	448	432,3	425
2500	535,9	515	502,6	483
3000	609,5	586	567,5	541
3500	679,0	667	628,2	610
4000	745,3	732	685,6	694
4500	808,9	792	740,3	746
5000	870,2	853	792,6	798
5500	929,4	927	842,9	847
6000	987,0	979	891,5	902
6500	1042,0	1040	938,7	940
7000	1097,6	1088	984,4	991
7500	1150,9	1143	1029,0	1040
8000	1203,2	1190	1072,5	1086
8500	1254,4	1258	1115,1	1123
9000	1304,6	1310	1156,7	1147
9500	1354,0	1354	1197,6	1210

Graphique III : croissance du vocabulaire dans les interventions de V. Giscard d'Estaing (débat avec F. Mitterrand, mai 1981). Valeurs observées et calculées à l'aide de la formule de Muller et du modèle de partition.



On voit, dans ce tableau et ce graphique, que le modèle réalise un ajustement assez fin des valeurs empiriques. De plus, les écarts se distribuent également des deux côtés des courbes théoriques et non pas toujours en dessous comme c'était le cas avec la formule de Muller. On peut donc en conclure que la partition du vocabulaire est un phénomène avéré et que notre modèle en réalise une estimation assez proche de la réalité.

Naturellement, ce modèle serait d'une faible utilité, si nous en restions là, puisque nous avons travaillé sur un cas d'école : des textes découpés en segments égaux. Dans la réalité, il n'en est jamais ainsi et la statistique lexicale étudie des textes ou des fragments inégaux.

GENERALISATION DU MODELE DE PARTITION DU VOCABULAIRE

Examinons le problème devant lequel se trouve habituellement placé le linguiste : la comparaison de textes ou de fragments de textes de longueurs inégales. Le modèle, que nous venons de proposer ci-dessus, peut répondre à cette préoccupation à condition d'y apporter quelques reformulations. Tout d'abord, nous définirons un nouveau paramètre :

$$p = \frac{Km}{M + Km}$$

II s'agit de la fraction de mots (et de vocables) du texte appartenant aux ensembles associés aux différentes tranches du texte (vocabulaires locaux). Ce paramètre peut s'exprimer en fonction de λ si nous reformulons *m* et *M*, eux aussi, en fonction de *N* et de λ.

Il vient :

$$p = \frac{K}{M + \lambda} \quad \text{et q = 1 – p et donc} \quad \lambda = \left(\frac{q}{p}\right) K$$

On peut alors donner une nouvelle expression de *V'*, nombre de vocables attendus dans la tranche d'ordre *K* et de longueur *k*,

$$V'_p(k,K) = \left(p\,\frac{k}{K} + q\right) V - q \sum_{i=1}^{i=n} V_i Q_i \left(\frac{k}{K}\right)$$

On remarque que cette expression ne fait plus intervenir que le rapport *k*/*K*, c'est-à-dire la relation entre la position du fragment considéré dans le texte et le nombre de fragments contenus dans le texte ou dans l'œuvre, sans qu'il soit nécessaire que les k soient égaux ou découpés régulièrement. Posant $u = k/K$, il vient :

$$V'_p(u) = (pu + q)V - q \sum_{i=1}^{i=n} V_i Q_i(u)$$

Sous cette forme, on remarque que l'expression devient indépendante du découpage considéré plus haut. Elle ne dépend que de *u*, de *V* et des V_i qui sont des paramètres mesurés

sur l'ensemble du texte et que le dépouillement standard des corpus permet de connaître. Le paramètre p est, a priori, inconnu mais nous l'estimerons en ajustant la formule aux données empiriques dont on dispose.

Remarquons d'abord que pour p = 0 :

$$V_0(u) = V - \sum_{i=1}^{i=n} V_i Q_i(u)$$

On retrouve ici la formule classique, inspirée du schéma probabiliste, qui fournit une excellente approximation de la loi hypergéométrique. Comme le remarquait C. Muller, cette formule est donc valable pour un texte où ne se ferait sentir aucune spécialisation lexicale, un texte dont chaque vocable aurait une espérance mathématique constante tout au long de son déroulement.

A l'autre extrémité, pour p = 1, nous avons : $V'_1(u) = uV$.

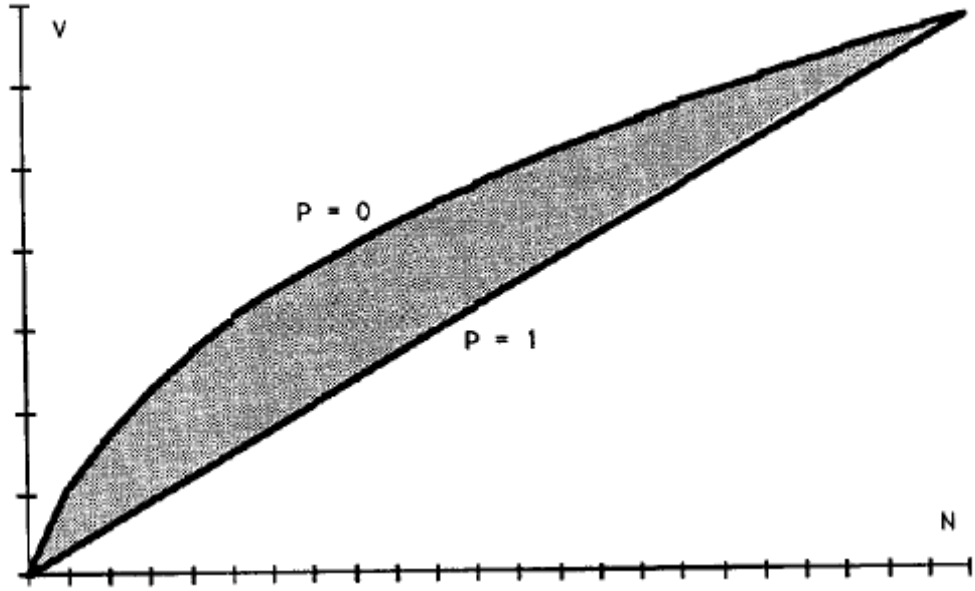
Ce qui donne une croissance linéaire du vocabulaire tout au long du texte : cette situation correspond à un texte (ou une œuvre) dont chaque fragment serait totalement spécialisé c'est-à-dire qu'il n'aurait aucun vocable en commun avec les autres fragments ; donc un texte sans vocabulaire général.

Ces deux cas extrêmes, et aussi théoriques l'un que l'autre, peuvent être mis en évidence en réécrivant V (espérance du nombre de vocables en fonction de u) :

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

Le graphique IV illustre les allures possibles de V : on voit que les courbes représentatives se situent dans un intervalle compris entre une courbe et une droite. D'une part, la limite supérieure de l'intervalle correspond à un coefficient p nul : c'est le cas d'un texte ou d'une œuvre sans aucune spécialisation. Les points de la courbe sont obtenus par application de la formule de Muller (c'est-à-dire de la loi hypergéométrique) à la distribution des fréquences observée sur l'ensemble du texte. D'autre part, l'accroissement linéaire correspond à un texte dont les fragments sont totalement spécialisés et sans aucun lien entre eux. La droite et la courbe sont calculées sur les mêmes valeurs empiriques.

Graphique IV : valeurs possibles de V en fonction de p



On comprend alors la signification de *p* : ce coefficient estime le rapport entre la partie spécifique du vocabulaire et le vocabulaire total. Ou encore, il mesure la spécialisation plus ou moins forte du vocabulaire d'un auteur en fonction des sujets traités, des thèmes qu'il aura abordés... Le coefficient *p* estime donc une qualité intrinsèque au texte : le partage entre le vocabulaire général et les vocabulaires spécialisés qui ont servi à le produire. Plus précisément, il permet de mesurer l'importance de cette partition. Telle est la raison pour laquelle nous proposons de nommer *p* : "coefficient de spécialisation du vocabulaire" ou, mieux, "coefficient de partition du vocabulaire".

Pour identifier et estimer ce coefficient, nous disposons généralement d'une série de valeurs empiriques observées sur le texte. Il s'agit d'abord de *V* : nombre total de vocables différents relevés dans le texte entier. On dispose également de la distribution de fréquences des V_i. Enfin, on connaît, au niveau de certaines césures, le nombre de vocables différents apparus depuis le début du texte, soit :

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

Nous considérerons que la meilleure valeur de *p* sera celle qui minimisera la somme des écarts quadratiques entre les valeurs observées - les $V'_*(u_k)$ - et les valeurs calculées : les $V'_p(u_k)$.

Soit $\Psi(p)$ cette somme des écarts quadratiques :

$$\Psi(p) = \sum_{k=1}^{k=K} \left[V'_p(u_k) - V'_*(u_k)\right]^2$$

On minimisera $\Psi(p)$, fonction du seul *p*, en annulant sa dérivée première.

Il vient alors :

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

En théorie, une seule valeur de *V* est nécessaire pour ce calcul. Mais il est évident que *p*, obtenu dans ces conditions, sera peu fiable. En pratique, comme tout paramètre, sa valeur sera influencée par le nombre et la qualité des mesures utilisées pour le calculer. Cette réserve admise, on peut se demander si *p* est bien indépendant de la taille et du nombre des segmentations opérées dans le corpus étudié. Par construction le modèle doit assurer cette stabilité puisqu'il ne prend pas en compte la dimension absolue de chacune des *k* parties mais simplement le rapport *u*. Des vérifications empiriques ont été conduites sur plusieurs corpus : elles permettent de conclure que *p* n'est pratiquement pas influencé par la modification du découpage même quand celui-ci devient très fin. Pour le montrer nous donnons, dans le tableau ci-dessous, les résultats d'une expérience consistant à découper, en tranches égales de 500 puis de 100 mots, le texte des deux débats télévisés de mai 1981 et octobre 1985. On voit que les fluctuations du paramètre sont inférieures à un pour cent.

Tableau II : Valeurs de *p* obtenues sur les débats Giscard-Mitterrand et Chirac- Fabius.

	Mitterrand	Giscard d'E.	Chirac	Fabius
Tranches de 100 mots	0,39215	0,34937	0,19908	0,21847
Tranches de 500 mots	0,39457	0,34890	0,19721	0,21016

II faut également noter que, du point de vue de l'analyse lexicale, le phénomène mesuré sera influencé par la plus ou moins grande hétérogénéité des ensembles étudiés. En effet, plus le corpus s'allonge plus la probabilité sera forte que l'auteur ait fait appel à des vocabulaires spécialisés nombreux. Autrement dit : dans nos formules, *N* n'a aucune influence sur la valeur de *p* mais, dans la pratique, ce coefficient ne sera pas indépendant de la longueur. En toute rigueur, la comparaison des valeurs de *p*, pour des auteurs différents, ne pourra probablement se faire que sur des textes de tailles pas trop différentes. D'autres applications du modèle devront être entreprises pour éclaircir ce point.

APPLICATIONS

Pour illustrer l'intérêt du modèle de partition du vocabulaire nous voudrions évoquer succinctement certaines de ses applications possibles et, plus particulièrement, trois d'entre elles. Notre modèle peut apporter une information synthétique sur le style d'un auteur ou d'une œuvre. Il permet également de localiser les variations thématiques au sein d'un corpus. Enfin, il complète et précise les mesures de richesse du vocabulaire.

1. Une mesure de la diversité du vocabulaire

Par exemple, le tableau II permet de formuler un jugement sur l'étendue relative des vocabulaires généraux et spécialisés chez les quatre hommes politiques étudiés. F. Mitterrand et V. Giscard d'Estaing ont puisé plus du tiers de leurs vocables dans des registres spécialisés ce qui signifie que de nombreux thèmes ont été abordés au cours du débat et que les deux candidats ont mobilisé un vocabulaire étendu. En effet, ces valeurs sont importantes et excèdent nettement les performances de J. Chirac et L. Fabius. On remarquera également que les deux valeurs obtenues pour chaque débat sont relativement proches. Autrement dit, dans la mobilisation du vocabulaire, les conditions d'énonciation du discours jouent un rôle probablement plus décisif que la personnalité des individus ou les stratégies qu'ils suivent.

Autre exemple : sur l'ensemble de ses tragédies, J. Racine atteint un p égal à 0,3304 ce qui est l'indice d'une diversité relativement grande dans l'œuvre et au sein de certaines pièces elles-mêmes (cela explique le profil du graphique I et l'écart important qui sépare les données empiriques de celles obtenues grâce au modèle hypergéométrique...) On se souviendra en effet que le genre était enfermé dans des règles assez strictes et que l'auteur avait à sa disposition un lexique moins étendu que celui offert par le français contemporain.

L'application du modèle à des corpus étendus devrait permettre d'obtenir un nombre suffisant de valeurs de p pour approfondir ces comparaisons que nous souhaitons riches d'enseignements.

2. Localisation des variations thématiques dans un corpus.

Le modèle donne, pour chaque segment d'un corpus, des valeurs théoriques V auxquelles sont associés des intervalles de confiance (l'écart type est calculé sur la partie probabiliste du vocabulaire, c'est à dire sur les seuls vocables "généraux"). Nous donnons ci-dessous les valeurs calculées sur les tragédies de J. Racine (tableau III) que l'on pourra comparer avec les résultats obtenus grâce à la formule de Muller (cf. Hubert & Labbé 1988).

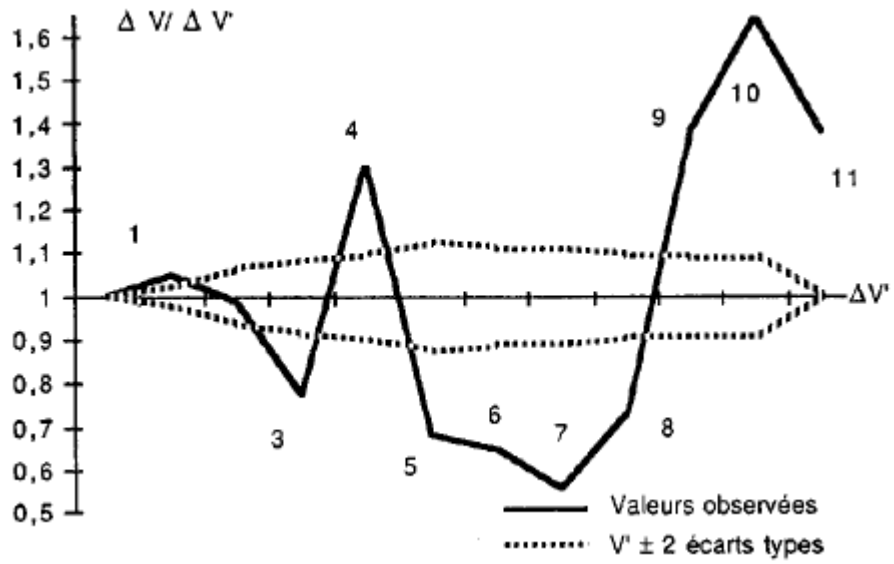
Tableau III : Nombre de vocables cumulés dans l'œuvre de J. Racine (*V* : valeurs observées, *V'* valeurs calculées avec le modèle de partition) et écart type associé

Tragédies	*V*	*V'*	σ
La Thébaïde	1244	1184,3	13,8
Alexandre	1638	1583,9	13,4
Andromaque	1868	1880,5	12,9
Britannicus	2185	2123,7	12,2
Bérénice	2310	2306,4	11,5
Bazajet	2435	2498,5	10,7
Mithridate	2533	2674,2	9,7
Iphigénie	2659	2847,1	8,4
Phèdre	2867	2997,4	6,9
Esther	3052	3110,1	5,3
Athalie	3262	3262,0	0,0

Comme pour le débat Giscard-Mitterrand, on notera la qualité de l'ajustement obtenu sur les données cumulées. La courbe passe au centre du nuage de points : les valeurs observées se répartissent également des deux côtés des valeurs calculées et se trouvent à l'intérieur ou à proximité de l'intervalle de confiance. Les valeurs calculées peuvent donc être considérées comme une bonne estimation de ce qu'aurait été l'œuvre de J. Racine si le style et les thèmes traités étaient demeurés inchangés de la première à la dernière tragédie.

Pour vérifier cette hypothèse, on peut étudier l'accroissement du vocabulaire d'une pièce à l'autre (*ΔV*/ *ΔV'*) : graphique V. Les valeurs de *ΔV* sont portées en abscisses. On leur a associé un intervalle de confiance de ± 2 σ : si la loi de progression du vocabulaire observée sur l'ensemble de l'œuvre était également à l'œuvre dans chacune des pièces toutes les valeurs de *ΔV*/ *ΔV'* se trouveraient inscrites dans le fuseau entourant l'axe horizontal. Comme il n'en est pas ainsi nous pouvons conclure à des variations stylistiques (ou thématiques) significatives.

Graphique V : Apport en vocables nouveaux dans chaque tragédie de J. Racine, observé (*ΔV*) et calculé à l'aide du modèle de partition du vocabulaire (*ΔV'*).



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

Le graphique met en lumière des faits déjà notés par les commentateurs de J. Racine :

- tout d'abord, on constate une tendance moyenne à l'enrichissement du vocabulaire au fur et à mesure du vieillissement et, tout particulièrement, dans les quatre dernières tragédies ;

- on observe également que *Britannicus* occupe une place à part. Elle "est la première pièce romaine et apporte donc les vocables nouveaux nécessaires pour créer l'atmosphère et le cadre propres à cette tragédie" (Bernet 1983 : p. 121) ;

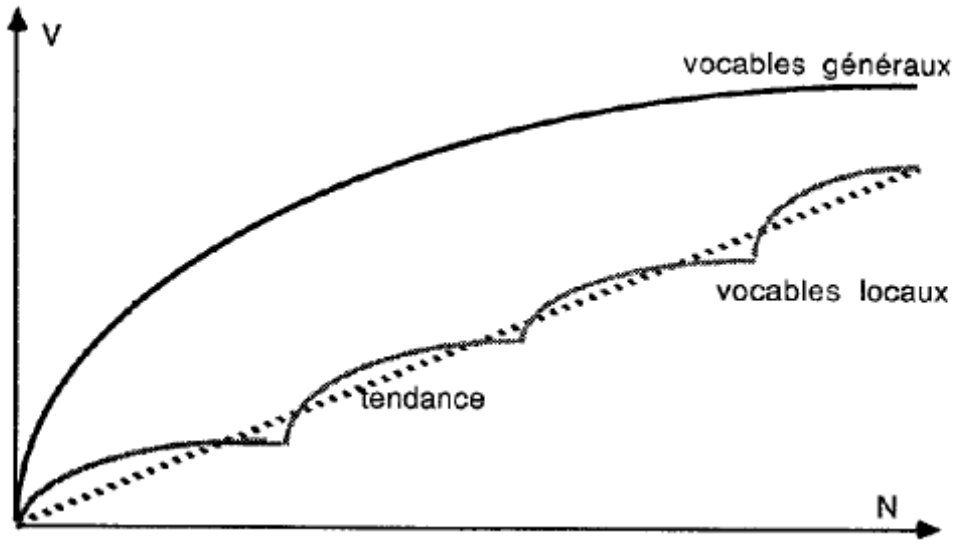
- enfin, *Iphigénie* marque le début du redressement de la courbe : elle annonce donc le recours à un registre nouveau. Comme le notent les commentateurs, le "vocabulaire d'*Iphigénie* a relativement plus de points communs avec *Phèdre*, *Esther* et *Athalie* qu'avec les pièces qui la précédent...", "En réalité, il n'y a pas deux mais quatre tragédies sacrées de Racine : *Iphigénie*, *Phèdre*, *Esther* et *Athalie*. Son génie a puisé aux deux sources, du merveilleux païen et du merveilleux chrétien" (Bernet 1983 : p 118). Cela se traduit, au niveau de cette pièce, par un premier afflux de vocables nouveaux, afflux qui se prolonge et s'amplifie dans les deux pièces suivantes avant de commencer à décliner sans que J. Racine semble épuiser ce nouveau registre comme cela avait été le cas pour l'antiquité gréco-romaine dans ses huit pièces antérieures.

Imaginons maintenant notre modèle placé dans la situation du lecteur ou du spectateur. Le texte ou la pièce se déroule devant lui et il note au fur et à mesure l'apparition des vocables suivant qu'ils proviennent du réservoir général ou des vocabulaires locaux. Le comportement des vocables généraux nous est connu : nous avons vu qu'ils suivent une courbe rapidement asymptotique. Leur entrée en scène est très rapide au début du texte et se fait de plus en plus rare au fur et à mesure qu'il s'allonge. Pour l'instant, nous avons considéré que les vocables locaux viennent de manière régulière et continue.

Nos observations y conduisaient : le texte est découpé en fragments auxquels sont associés autant de réservoirs, ce qui amène une tendance linéaire.

Si l'on pouvait traiter l'apparition des vocables comme un phénomène continu nous obtiendrions le graphique VI.

Graphique VI : schéma théorique d'apparition dans un texte des vocables selon qu'ils appartiennent au vocabulaire général ou au vocabulaire local.



La droite symbolisant l'accroissement des vocables locaux devient une série de petites courbes d'un profil identique à celui du vocabulaire général. Quand l'auteur ouvre un nouveau registre, on observe un afflux de vocables spécialisés, puis, progressivement, le réservoir s'épuise et l'apport décline : la courbe passe en dessous de la tendance jusqu'à ce qu'un nouveau vocabulaire soit mobilisé ce qui se manifestera par un nouvel afflux... Dans la réalité, il y aura naturellement des chevauchements : en général, un auteur n'attend pas d'avoir épuisé un registre pour aller chercher d'autres sources d'inspiration

3. Les mesures de richesse du vocabulaire.

Nous voudrions enfin donner un dernier exemple d'application possible de notre modèle : il s'agit du problème traditionnel de la "richesse du vocabulaire". Sur cette question, le schéma mis au point par C. Muller atteint une grande précision (Muller : 1970). Il présente toutefois un inconvénient : comme nous l'avons dit ci-dessous, le modèle considère que le texte ne connaît pas de spécialisation lexicale, de telle sorte que les valeurs calculées sont, de manière quasi-systématique, supérieures aux valeurs observées. Dès lors, la technique du "raccourcissement" - du texte le plus long à la taille du plus court -"avantage" le premier et cet avantage sera d'autant plus grand que leur différence de longueur est importante ou que la spécialisation lexicale y est intense. Parfois, ce biais n'a pas de conséquences pratiques, mais il n'en est pas toujours ainsi comme nous allons le montrer dans un petit exemple. Nous donnons, dans le tableau ci-dessous (tableau IV), les valeurs de *V* et de *V'* pour les

textes des deux débats télévisés. Les textes de F. Mitterrand, V. Giscard d'Estaing et L. Fabius ont été raccourcis à la longueur de celui de J. Chirac (7.432 mots).

Tableau IV Comparaison de la richesse du vocabulaire chez MM. V. Giscard d'Estaing, F. Mitterrand, J. Chirac et L. Fabius (opération de raccourcissement)

	V observés (7432 mots)	Raccourcissements (7 432 mots) : C. Muller V'	Modèle de partition V'	Modèle de partition Intervalle de confiance
J. Chirac	1181			
F. Mitterrand	1132	1193	1141	1123-1158
L Fabius	1119	1120	1118	1117-1119
V. Giscard d'Estaing	1037	1103	1024	1003-1044

Nous donnons successivement les valeurs observées en arrêtant le dépouillement au 7.432ème mot, puis les valeurs calculées à l'aide du modèle de Muller et du modèle de partition qui vient d'être exposé ci-dessus. Aux valeurs obtenues à l'aide de ce dernier sont associés des intervalles de confiance de ± 2 écarts types,

On voit que les valeurs calculées grâce au modèle de Muller conduisent à classer F. Mitterrand devant J. Chirac, du point de vue de la richesse du vocabulaire, en contradiction avec les données observées (qui peuvent naturellement être accidentelles...) Notre modèle aboutit cependant à confirmer l'observation directe et place le discours de J. Chirac devant celui de F. Mitterrand. Deux éléments expliquent ce chassé-croisé :

- la longueur du texte de F. Mitterrand (il excède d'un tiers la dimension des propos de J. Chirac) : la formule de Muller avantage donc F. Mitterrand ;

- le fait que le futur président fasse beaucoup appel à des registres spécialisés. En effet, le tableau II indique que *p*, calculé sur ses interventions, atteint une valeur élevée : le candidat socialiste a donc puisé environ 40% de ses vocables dans des vocabulaires "spécialisés" alors que, chez J. Chirac, cette proportion est inférieure de moitié. Par conséquent, une forte composante linéaire se trouve à l'œuvre dans la croissance du vocabulaire de F. Mitterrand : le phénomène s'éloigne beaucoup de la courbe théorique dont l'équation est donnée par la formule de Muller...

Pour la comparaison d'œuvres ou d'auteurs entre eux, la méthode de C. Muller vise donc un cas particulier : celui de textes sans spécialisation lexicale (ou de longueurs très proches). En pratique, un tel cas se rencontre rarement et la méthode de C. Muller conduit à surestimer le nombre de vocables du texte qui subit l'opération de "raccourcissement". Cette surestimation sera d'autant plus grande que :

- *u* (N'/N) est faible, c'est-à-dire que la différence de longueur entre les deux textes ou fragments de textes comparés est importante ;

- la spécialisation lexicale est forte dans le texte le plus long.

Dans ces conditions, la méthode Muller peut être conservée lorsqu'elle amène à conclure en faveur du texte ou fragment le plus court : dans ce cas, on peut même penser que l'écart est sous-estime et que cette sous-estimation est fonction des deux conditions énoncées ci-dessus. En revanche, la situation inverse doit être considérée avec prudence et il sera préférable de recourir au modèle que nous venons d'exposer.

REFERENCES

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

Brunet Etienne (1978). *Le vocabulaire de Jean Giraudoux. Structure et évolution*. Genève-Paris : Slatkine-Champion.

Hubert Pierre & Labbé Dominique (1988). Note sur l'approximation de la loi hypergéométrique par la formule de Muller. In Labbé Dominique, Thoiron Philippe et Serant Daniel. *Etudes sur la richesse et la structure lexicale*. Paris-Genève : Slatkine-Champion, p 77-91.

Muller Charles (1970). "Sur la mesure de la richesse lexicale. Théorie et expériences". Reproduit dans *: Langue française et linguistique quantitative*. Paris-Genève : Slatkine-Champion, 1979, p 281-307.

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