Note sur l'approximation de la loi hypergéométrique par la formule de Muller

Pierre Hubert, Dominique Labbé

Pierre HUBERT
(Ecole des Mines de Paris)

Dominique LABBE
(Université de Grenoble II)

Note sur l'approximation de la loi hypergéométrique par la formule de Muller

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. 77-91.

Résumé
Le raisonnement part de l'estimation de la probabilité d'absence d'un vocable dans un échantillon exhaustif prélevé dans un corpus, connaissant la distribution des fréquences des vocables qui constituent ce corpus. C'est la formule qui a été proposée il y a plus de vingt ans par Charles Muller et qui est ici comparée avec la loi hypergéométrique. Deux applications sont examinées : le calcul de l'accroissement du vocabulaire dans des corpus et le prélèvement aléatoire d'un grand nombre d'échantillons exhaustifs sur ces corpus. On démontre ainsi, théoriquement et empiriquement, que la *formule de Muller* représente une bonne approximation de la loi *hypergéométrique*. On montre également la nécessité d'associer aux valeurs calculées un écart type qui permettra d'estimer l'intervalle de confiance attaché aux valeurs obtenues grâce à cette formule de Muller.

Abstract
The argument which is developed here starts from the computation of the probability that a word will be absent from an exhaustive random sample drawn from a corpus whose complete frequency distribution is known. This probability is the basis of the formula put forward, more than 20 years ago, by C. Muller. Muller's formula is compared here to its equivalent in the hypergeometric model. Two studies were carried out: first the computation of vocabulary increase in corpuses and, secondly, the comparison between Muller's values and averages obtained by drawing a large number of random samples from several corpuses. It is thus demonstrated that this formula is a good approximation of the hypergeometric law. The need for associating standard deviations to the computed values is also emphasised since confidence levels have to be taken into account.

Les travaux de Guiraud, Muller, Evrard, Brunet... ont introduit et acclimaté le raisonnement probabiliste en statistique lexicale ; l'utilisation du schéma d'urne est devenu pratique courante dans la recherche contemporaine. Il n'est pas sûr toutefois que les conséquences de ce raisonnement aient été toutes clairement tirées. Certes la discussion n'est pas neuve : par exemple, une controverse s'est déroulée, il y a quelques années, à propos de la "répartition" des occurrences d'un vocable dans un corpus donné. Peut-on assimiler ce phénomène à une "densité de probabilité" ? A Paul Bratley qui répondait par la négative, Etienne Brunet a opposé de solides arguments (Brunet : 1982, 1983a et b). Mais, une fois admis la validité de ces arguments en faveur du schéma probabiliste, quelle loi utiliser ? ici le débat oppose les tenants du strict respect de la statistique - grâce à la loi hypergéométrique (Lafon : 1984) - et les défenseurs de la loi normale (Brunet : 1982) et de l'écart réduit (Muller : 1981a et 1981b).

Nous proposons d'examiner le problème d'un point de vue plus général : si l'on admet le schéma d'urne, tout segment d'un corpus peut être décrit comme un échantillon exhaustif prélevé au hasard dans cette urne. Nous montrerons dans cet article que, en toute rigueur, le modèle hypergéométrique devrait être utilisé mais que l'on commet une erreur négligeable en lui substituant la formule proposée par Charles Muller il y a maintenant plus de vingt ans (Muller : 1964 ; pour l'application à l'oeuvre de Corneille, Muller : 1967).

COMPARAISON DES MODELES HYPERGEOMETRIQUE ET DE MULLER

Un texte ou un groupe de textes sont donc considérés comme une urne contenant une population de N mots. Au sein de cette population on désigne un vocable particulier de fréquence absolue *f*. Si l'on procède à un tirage aléatoire exhaustif de *N'* mots dans l'urne, la fréquence absolue du vocable désigné dans l'échantillon apparaît comme une variable aléatoire *F'* susceptible de prendre une valeur *f'* comprise entre 0 et f. En toute rigueur, la distribution de probabilité de cette variable aléatoire peut être décrite par la loi hypergéométrique.

On peut alors écrire :

$$\text{Prob}[F' = f'] = \frac{C_f^{f'} \, C_{N-f}^{N'-f'}}{C_N^{N'}} \text{ pour } 0 \leq f' \leq f$$

Et l'on peut vérifier que :

$$\sum_{f'=0}^{f'=f} \text{Prob}[F' = f'] = 1$$

Un cas particulièrement intéressant, dans l'étude de l'accroissement du vocabulaire, est celui où f' est nul, c'est-à-dire que le vocable étudié n'apparaît pas dans l'échantillon d'effectif N'. Dans ce cas, on peut écrire :

$$\sum_{f'=0}^{f'=f} \text{Prob}[F'=0] = 1$$

Cette expression n'a de sens que si : $N' < N - f$

Au delà de cette valeur, il est certain que l'échantillon de taille N' comprendra le vocable désigné et donc que Prob [F' = 0] sera nul.

L'expression ci-dessus peut-être développée :

$$\text{Prob}[\text{F'}=0] = \frac{(\text{N} - f)!}{\text{N'}!(\text{N} - \text{N'} - f')!} \frac{N'!(N-N')!}{N!} = \frac{(N-f)!}{N!} \frac{(N-N')!}{(N-N'-f)!}$$

Après simplification, la première fraction comporte au dénominateur le produit des f nombres entiers de $(N - f + 1)$ à N. Son numérateur est égal à l'unité. De façon analogue, le second membre de l'égalité comporte au numérateur le produit des f nombres entiers compris entre (N-N'-f+1) et (N-N1), le dénominateur étant égal à un.

On peut alors écrire :

$$\text{Prob}[F'=0] = \prod_{i=1}^{i=f} \left[\frac{N-N'-i+1}{N-i+1} \right]$$

Connaissant N, N' et f, cette expression est aisément programmable. Si l'on pose : $u = N'/N$, ce qui correspond au "taux de sondage", l'expression ci-dessus peut être réécrite sous la forme :

$$\text{Prob}[F'=0] = \prod_{i=1}^{i=f} \left[\frac{1 - u - \frac{i-1}{N}}{1 - \frac{i-1}{N}} \right]$$

Nous appellerons $Q_f^*(u,N)$ cette probabilité d'absence d'un vocable de fréquence f dans un échantillon exhaustif extrait d'une population d'effectif N ; la taille de l'échantillon N' étant égale à $u.N$. En toute rigueur, cette probabilité Q_f^* dépend des deux paramètres u et N. Il est important de comparer cette expression avec celle généralement utilisée en statistique lexicale et qui a été formulée par Charles Muller :

$$Q_f(u) = (1 - u)^f$$

Nous désignerons cette expression dans la suite de cette étude sous le nom de "formule de Muller".

Pour mener à bien la comparaison proposée ci-dessus nous reprendrons l'expression :

$$Q_f^* = \prod_{i=1}^{i=f} \left[\frac{N-N'-i+1}{N-i+1} \right] \text{ avec } f \le N' < N - f$$

Nous remarquerons que, quel que soit i,

$$\frac{N-N'-i}{N-i} < \frac{N-N'}{N}$$

Dans ces conditions, on peut écrire :

$$\left(\frac{N-N'-f+1}{N-f+1}\right)^f < Q_f^* < \left(\frac{N-N'}{N}\right)^f = Q_f$$

L'erreur δ_f qui serait commise en substituant Q_f à Q_f^* est donc telle que :

$$\delta_f < \left(\frac{N-N'}{N}\right)^f - \left(\frac{N-N'-f+1}{N-f+1}\right)^f$$

soit encore,

$$\delta_f < \left[\left(\frac{N-N'}{N}\right) - \left(\frac{N-N'-f+1}{N-f+1}\right)\right]\left[\left(\frac{N-N'}{N}\right)^{f-1} + \left(\frac{N-N'}{N}\right)^{f-2}\left(\frac{N-N'-f+1}{N-f+1}\right) + \ldots + \left(\frac{N-N'-f+1}{N-f+1}\right)^{f-1}\right]$$

Ou, en utilisant de nouveau l'inégalité :

$$\frac{N-N'-i}{N-i} < \frac{N-N'}{N}$$

$$\delta_f < \frac{(N-N')(N-f+1)-(N-N'-f+1)N}{N(N-f+1)} f\left(\frac{N-N'}{N}\right)^{f-1} = \left(\frac{N-N'}{N}\right)^{f-1} \frac{N' f(f-1)}{N(N-f+1)}$$

Et, en utilisant l'égalité $u = N'/N$, il vient :

$$\delta_f(u,N) < u(1-u)^{f-1} \frac{f(f-1)}{N-f+1}$$

L'utilisation de Q_f simplifie grandement la démarche puisque cette probabilité ne dépend que de la variable u contrairement à Q_f^* qui, elle, dépend de u et de N'.

Il faut se souvenir que le vocabulaire de tout texte (d'une longueur minimale) est toujours composé d'un grand nombre de vocables utilisés chacun un petit nombre de fois, de telle sorte que le nombre d'apparition de n'importe quel vocable (même le plus fréquent) reste très petit par rapport à la longueur du texte ($f << N$). Il est donc possible de comprendre intuitivement que, à condition que N' soit supérieur à la valeur la plus grande possible de f (c'est-à-dire le

nombre d'apparitions du vocable le plus utilisé) et inférieur à ($N - f$), on aura toujours $f' << N'$ et, par conséquent, un $\delta_f(u, N)$ très petit.

Pour le vérifier empiriquement, nous avons appliqué les deux formules sur le vocabulaire de J. Racine tel qu'il évolue au cours de son œuvre (Bernet, 1983). Le tableau I donne, à l'issue du dépouillement de chaque pièce, le nombre théorique de vocables apparus depuis le début de l'œuvre avec ces deux formules. Nous vérifions ainsi que les conséquences pratiques de cette erreur sont négligeables, au niveau de la modélisation de l'accroissement du vocabulaire, et que les différences observées sont toujours effectivement inférieures à Δ calculé à partir de l'estimation des δ_f .

Tableau 1 Comparaison de la formule de Muller et de la loi hypergéométrique : calcul de l'accroissement du vocabulaire dans l'œuvre de J. Racine (ordre chronologique, d'après le dépouillement de C. Bernet)

	Valeurs théoriques obtenues par les modèles : Hypergéométrique ()	C. Muller	Δ
La Thébaïde	1656,57	1656,56	0,0255
Alexandre	2111,57	2111,56	0,0129
Andromaque	2382,61	2382,61	0,0075
Britannicus	2576,04	2576,04	0,0048
Bérénice	2726,70	2726,70	0,0034
Bazajet	2850,18	2850,17	0,0023
Miîhridate	2954,78	2954,78	0,0016
Iphigénie	3045,50	3045,50	0,0010
Phèdre	3125,57	3125,56	0,0006
Esther	3197,20	3197,20	0,0003
Athalie	3262,00	3262,00	0,0000

L'expression $(1-u)^f$ fournit donc une excellente estimation de la probabilité d'absence d'un vocable de fréquence f dans un échantillon exhaustif de taille N', c'est-à-dire de la loi hypergéométrique. Comme indiqué ci-dessus, cela tient au profil de la distribution des fréquences observée dans les textes. En particulier, dans les classes de fréquences les plus hautes, aucun vocable ne dépasse 7% de N. Ainsi se trouve justifiée la limite inférieure - de l'ordre de $0{,}1N$ pour N' - énoncée empiriquement par C. Muller pour l'emploi de sa formule dans la mesure de la croissance du vocabulaire ou pour la comparaison de textes inégaux par "raccourcissement" du plus long de ces textes à la dimension du plus court (Muller, 1964).

Comme l'avait déjà montré expérimentalement E. Brunet, pour le cas de la loi normale (Brunet : 1982), on commet donc une erreur négligeable en utilisant la formule de Muller au lieu du modèle hypergéométrique. Cependant la nature des résultats obtenus n'a pas toujours été clairement discutée.

DISTRIBUTION D'ECHANTILLONNAGE DANS UN TEXTE

Le problème réside dans la nature de l'opération réellement effectuée lorsque l'on simule le tirage d'un échantillon de *N'* mots dans un corpus de taille *N*. Deux optiques sont possibles.

D'une part, nous pouvons estimer qu'il s'agit d'un raisonnement par analogie servant à réaliser une stricte réduction proportionnelle du corpus. Autrement dit, l'opérateur fabrique une sorte de "maquette" : il postule que, si l'auteur avait employé *N'* mots au lieu de *N*, il aurait réalisé la même œuvre dans un format plus réduit. L'emploi du modèle probabiliste représente une commodité transitoire ; les résultats obtenus ne sont entachés d'aucune incertitude et toute différence constatée, aussi mince soit-elle, devient significative. C'est ainsi que l'on raisonne habituellement et, en particulier, lorsque l'on compare des textes de longueur différente ou des auteurs entre eux.

D'autre part, on peut considérer que le raisonnement précédent pèche doublement :

- il repose sur un postulat très discutable : un texte ou une œuvre ne sont pas homogènes et prélever au hasard des individus dans cette population c'est s'exposer à des fluctuations plus ou moins importantes ;
- de manière plus sérieuse encore, nous ferons remarquer que le modèle probabiliste est un tout : on ne peut lui emprunter ses outils quand ils semblent utiles pour ignorer ensuite les conséquences de cet emploi quand elles vont à l'encontre de la commodité...

Poursuivons donc le raisonnement probabiliste. Le texte se constitue de *N* mots prélevés au hasard dans une urne. Si l'on veut respecter le schéma d'urne, il faut bien admettre que les *N* tirages successifs, dont est issu le corpus étudié, ont été soumis à des fluctuations d'échantillonnage normales en pareille circonstance. Une telle idée serait susceptible d'expliquer, au moins en partie, un fait d'expérience : lorsqu'on étudie l'apparition des vocables nouveaux, au long d'un texte ou d'une œuvre, on constate que ce phénomène n'est pas régulier : en certains passages, il se produit des afflux et, dans d'autres au contraire, un ralentissement de l'apport en vocables neufs. Le raisonnement probabiliste permet de poser que le tirage provoque, au moins en partie, ces fluctuations. Par conséquent, il faut associer un écart type et un intervalle de confiance aux mesures effectuées sur ce corpus.

Une bonne estimation de l'écart type peut-être obtenue aisément en développant le modèle ci-dessus. On sait que le texte considéré comprend *N* mots représentant *V* vocables dont V_i de fréquence absolue *i* (*i* variant de 1 à *n*). On a vu qu'il est possible d'attacher à un vocable particulier de fréquence absolue *i*, une probabilité de non apparition dans un échantillon exhaustif de taille *N'* :

$$Q_i(u) = (1 - u)^i \text{ avec } u = N'/N$$

Pour le tirage d'un échantillon exhaustif de taille N', nous associerons, à chaque vocable du texte, une variable aléatoire X_v ($v = 1, 2, \ldots, V$) suivant une loi de Bernouilli, c'est-à-dire telle que :

$$\text{Prob}[X_v = 0] = Q_{i(v)}(u)$$
$$\text{Prob}[X_v = 1] = 1 - Q_{i(v)}(u)$$

L'espérance de cette variable est :

$$E(X_v) = \left[1 - Q_{i(v)}(u)\right]$$

Sa variance est :

$$Var(X_v) = Q_{i(v)}(u)\left[1 - Q_{i(v)}(u)\right]$$

Si nous définissons *V* comme le nombre de vocables contenus dans un échantillon exhaustif de *N'* mots, *V* apparaît comme une variable aléatoire que nous exprimerons en fonction des X_v :

$$V'(u) = \sum_{v=1}^{v=V} X_v$$

On suppose que les X_v sont indépendants. Il ne s'agît que d'une approximation. Pour s'en convaincre il suffit de considérer l'étendue de la distribution de *V* : sous la forme qui vient d'être écrite, *V* pourrait prendre toutes les valeurs de 0 à *V*. Or *V* est limité intérieurement puisqu'un échantillon de taille *N'* comprend nécessairement un certain nombre de vocables, dépendant de *N'* et de la structure lexicale du texte (si *f* était la fréquence absolue maximale et si *N'* était inférieur à *f*, cette limite inférieure serait égale à 1). De même, *V* est limité supérieurement au minimum de *V* et de *N'* qui peut être évidemment inférieur à *V*. Nous examinerons plus bas, grâce à des simulations d'échantillonnage, les limites de cette approximation.

Admettant l'indépendance des X_v , on peut alors calculer l'espérance de *V* comme :

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

Si nous remarquons qu'il existe V_i vocables de fréquence *i* cette expression devient :

$$E\left[V'(u)\right] = \sum_{i=1}^{i=n} V_i \left[1 - Q_{i(v)}(u)\right] = V - \sum_{i=1}^{i=n} V_i \, Q_i(u)$$

où l'on retrouve la formule classique de C. Muller.

Nous pouvons également calculer la variance de V :

$$Var\left[V'(u)\right] = Var\left[\sum_{v=1}^{v=V} X_v\right] = \sum_{v=1}^{v=V} Var\left[X_v\right]$$

Si nous remarquons, là encore, qu'il existe *Vi* vocables de fréquence *i* cette expression devient :

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

On notera que nous travaillons avec le nombre de classes de fréquences observé sur le texte entier. Etant donné que l'étendue réelle de la distribution de *V'* est très probablement plus réduite, on doit obtenir une variance légèrement surévaluée[1]. Remarquons également que la variable aléatoire *V* a été définie comme la somme de nombreuses variables aléatoires supposées indépendantes et de même ordre de grandeur ; on doit donc s'attendre à ce que la distribution des réalisations de *V* soit approximativement normale.

On comparera cette formule à celle que propose Charles Muller dans son manuel (1977 : p 104) :

$$Var[V'(u)] = V\left(\frac{E(V'(u)}{V}\right)\left(1-\frac{E(V'(u)}{V}\right)$$

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

En pratique, cela équivaut à admettre que tous les vocables ont la même probabilité d'apparaître dans l'échantillon de taille $N' = u.N$. Il n'en est rien puisque cette probabilité dépend de la classe de fréquence à laquelle appartient le vocable considéré, cette propriété étant d'ailleurs le fondement du modèle d'accroissement lexical proposé par C. Muller. La substitution du produit des sommes à la somme des produits simplifie les calculs mais conduit la plupart du temps à une surestimation qui peut dépasser 25% de la valeur de σ pour les plus faibles valeurs de *u* comme on peut le constater à la lecture du tableau II qui donne les résultats des calculs menés sur l'ensemble des tragédies de J. Racine (rappelons au passage que notre formule conduit déjà à une surestimation).

[1] Voir Serant 1988 qui confirme cette légère surestimation et indique son ampleur probable. D. Serant propose également un mode de calcul précis de la variance.

Tableau II. Espérance mathématique du nombre de vocable cumulés (*V'*) dans la suite des tragédies de J. Racine selon la formule de Muller. Ecart type associé (σ) et comparaison avec l'écart type de Muller (σCM) (d'après !e dépouillement de C. Bernet, 1983).

Tragédies	*V'*	σ	σ CM	(σCM) /σ
La Thébaïde	1656,57	20,59	28,55	1,39
Alexandre	2111,57	20,01	27,29	1,36
Andromaque	2382,61	19,19	25,34	1,32
Britannicus	2576,04	18,25	23,27	1,28
Bérénice	2726,70	17,18	21,14	1,23
Bazajet	2850.18	15,95	18,97	1,19
Mithridate	2954,78	14,52	16,68	1,15
Iphigénie	3045,50	12,80	14,22	1,11
Phèdre	3125,57	10,66	11,43	1,07
Esther	3197,20	7,69	7,97	1,04
Athalie	3262,00	0,00	0,00	-

Il paraît donc souhaitable de ne pas retenir l'expression de la variance proposée par Muller. Il vaut mieux combiner le calcul de celle-ci avec celui des $V_iQ_i(u)$ opéré sur chaque classe de fréquence.

Pour illustrer les conséquences pratiques de cette discussion nous donnons, dans les tableaux III et IV, les résultats d'un calcul effectué sur les deux débats télévisés qui opposèrent, d'une part, V. Giscard d'Estaing et F. Mitterrand en mai 1981 et, d'autre part, J. Chirac et L. Fabius en octobre 1985.

Tableau III. Espérance mathématique du nombre de vocables cumulés dans un texte selon la formule de Muller et écart type associé (débat Giscard-Mitterrand, mai 1981, textes découpés en tranches de 500 mots)

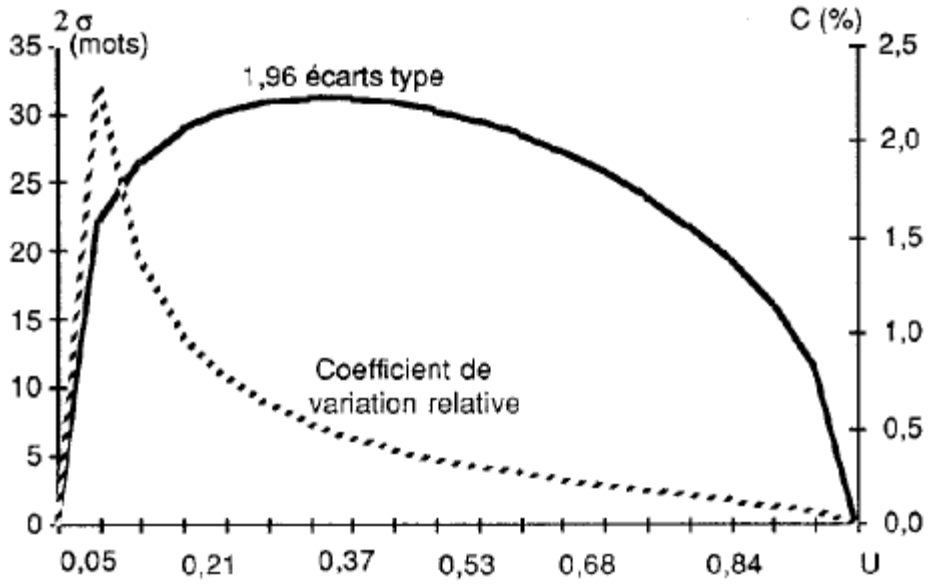
	F. Mitterrand		V. Giscard d'Estaing	
N'	V’	σ	V’	σ
500	219,11	11,37	214,92	11,26
1000	354,74	13,57	436,69	13,32
1500	463,95	14,72	450,72	14,36
2000	557,59	15,38	538,30	14,97
2500	640,47	15,74	614,67	15,31
3000	715,31	15,88	682,81	15,49
3500	783,84	15,86	744,55	15,53
4000	847,24	15,70	801,16	15,47
4500	906,36	15,42	853,54	15,33
5000	961,83	15,02	902,34	15,11
5500	1014,14	14,52	948,09	14,83
6000	1063,70	13,89	991,19	14,47
6500	1110,80	13,14	1031,97	14,05
7000	1155,71	12,24	1070,70	13,56
7500	1198,65	11,16	1107,59	13,00
8000	1239,81	9,85	1142,83	12,35
8500	1279,33	8,19	1176,58	11,61
9000	1317,35	5,90	1208,98	10,75
9500	1354,00	0,00	1240,12	9,76

Tableau IV. Espérance mathématique du nombre de vocables cumulés dans un texte selon la formule de Muller et écart type associé (débat Chirac-Fabius d'octobre 1985, textes découpés en tranches de 500 mots)

	J. Chirac		L. Fabius	
N'	*V'*	σ	*V'*	σ
500	216,95	11,21	216,79	11,00
1000	351,12	13,24	346,04	12,90
1500	459,20	14,17	448,85	13,75
2000	551 ,63	14,57	536,07	14,10
2500	633,12	14,62	612,59	14,15
3000	706,34	14,42	681,17	13,96
3500	773,02	14,02	743,58	13,57
4000	834,35	13,43	801,02	13,00
4500	891,21	12,63	854,32	12,23
5000	944,29	11,62	904,09	11,25
5500	994,11	10,34	950,82	10,00
6000	1041,12	8,66	994,85	8,37
6500	1085,65	6,28	1036,50	6,06
7000	1128,00	0,00	1076,00	0,00

Ces textes ont été dépouillés selon la norme dite de "C. Muller" et découpés en tranches de longueur égale suivant un pas de 500 mots. On a calculé *V'* pour des fractions croissantes des textes, *N*' variant de 500 à *N*. De chacune des valeurs théoriques, on déduit un écart type dont les valeurs sont données dans les tableaux III et IV. Pour rendre la comparaison lisible nous avons rapporté les valeurs des σ aux *V'* correspondants pour le texte de Jacques Chirac. Les résultats de ce calcul sont présentés dans le graphique 1 : en ordonnées ont été portées les valeurs du coefficient de variation relative (C) en fonction de u.

Graphique I. Evolution de l'écart type et du coefficient de variation relative (C) en fonction de u (interventions de J. Chirac face à L. Fabius)



On voit que la taille de *N'* a une influence directe sur l'importance relative de σ : plus *N'* est proche de *N*, moins grande sera l'incertitude pesant sur les calculs. Ce constat entraîne plusieurs conséquences. Par exemple, comparer des textes entre eux, en raccourcissant !e plus long à la dimension du plus petit, revient à tester l'hypothèse nulle selon laquelle ces deux textes obéissent à la même loi de probabilité : on les suppose extraits de deux populations normales ayant même moyenne et même écart type.

Pour réaliser ce test, il faut associer à la valeur théorique *V'* - obtenue par cette opération de raccourcissement du texte le plus long - un intervalle de confiance égal, par exemple, à plus ou moins 1,96σ (en acceptant 5% de chances d'erreur). Les valeurs empiriques (ici le nombre de vocables différents contenus dans le texte le plus court) ne pourront être considérées comme significativement différentes des valeurs calculées que si elles se situent en dehors de cet intervalle.

Les observations empiriques de la statistique lexicale viennent-elles conforter ce raisonnement ? Pour le vérifier, nous nous sommes livrés à une expérience simple sur les quatre textes des deux débats télévisés mentionnés ci-dessus. Grâce au générateur de nombres au hasard de l'ordinateur, nous avons prélevé aléatoirement un certain nombre d'échantillons exhaustifs dans ces textes et nous avons mesuré le nombre de vocables différents contenus dans chacun des échantillons (tableau V).

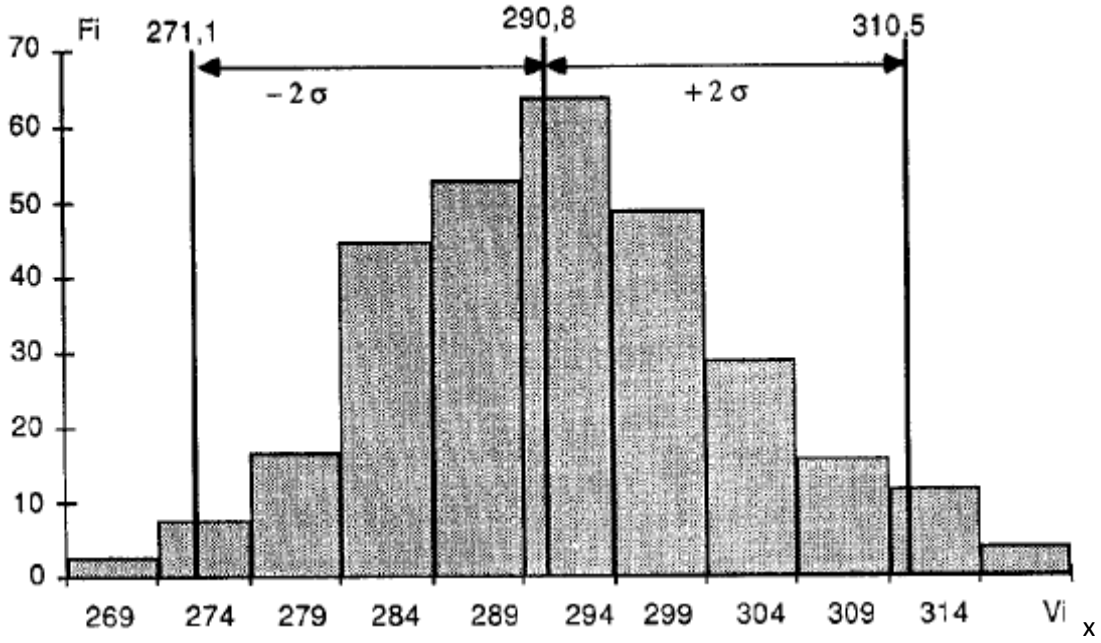
Tableau V. Espérances mathématiques du nombre de vocables différents dans une tranche de *N'* mots d'un texte et moyennes pour des échantillons exhaustifs de même taille prélevés aléatoirement dans ce texte.

	V.G.E.	F.M.	J.C.	L.F.
N' = 500				
Valeurs théoriques				
V'	214.92	219,11	216,95	216,79
Ecart type	11,26	11,37	11,21	11,00
Echantillons = 300				
Moyenne	214.10	218,78	215,57	215.46
Ecart type	9.74	9,33	9,20	8,73
N' = 1000				
Valeurs théoriques				
V'	346,69	354,74	351,12	346,04
Ecart type	13,32	13,57	13,24	12.90
Echantillons = 200				
Moyenne	345,17	354,45	351,17	345,53
Ecart type	11.14	12,00	11,50	11,29
N' = 2000				
Valeurs théoriques				
V'	538,30	557,59	551,63	536,07
Ecart type	14,97	15,38	14,57	14,10
Echantillons = 100				
Moyenne	537,70	555,94	549,31	533,18
Ecart type	13,71	14,05	11,87	12,36

Autrement dit, nous avons réalisé empiriquement l'opération que simule théoriquement la formule de Muller. L'expérience ayant été renouvelée un grand nombre de fois, pour différentes tailles d'échantillon, nous avons obtenu des moyennes d'échantillons et des écarts types autour de ces moyennes.

Pour comprendre la nature de l'opération nous présentons, pour l'une de ces expériences, la manière dont se répartissent les échantillons en fonction du nombre de vocables différents qu'ils contiennent (graphique 11). La normalité de la distribution d'échantillonnage est attestée à la fois par la forme en cloche du polygone de fréquence et par le groupement des valeurs centrales (moyenne, mode et médiane).

Graphique II. Distribution d'échantillonnage : nombre de vocables différents observés dans 300 échantillons exhaustifs de 750 mots prélevés aléatoirement dans le texte des interventions de F. Mitterrand (débat avec V. Giscard d'Estaing)



Il est intéressant de comparer les valeurs empiriques observées dans nos échantillons avec les valeurs théoriques obtenues par la formule Muller (tableau V). Aux moyennes d'échantillons sont donc associés des intervalles de confiance. Compte tenu de ces intervalles on voit que, pour la totalité de nos textes, il n'y a pas de différence significative entre la moyenne empirique et la valeur calculée. La même remarque peut être faite à propos de l'écart type (notons que le calcul de l'écart type théorique a été réalisé suivant la méthode présentée ci-dessus et non avec la formule de Muller). On remarquera cependant que la valeur observée pour l'écart type est toujours inférieure à la valeur calculée, ce qui vérifie la légère surestimation que nous avions prévue plus haut.

En conclusion, nous pouvons donc affirmer que la formule de Muller estime bien l'espérance mathématique du nombre de vocables contenus dans un échantillon prélevé au hasard et sans remise dans un texte (échantillon exhaustif). Autrement dit, cette formule donne une bonne estimation de la loi hypergéométrique et c'est à tort qu'on l'a baptisé "binomiale". De plus, cette expérience confirme le raisonnement probabiliste : il est nécessaire d'associer aux valeurs calculées un écart type qui permettra d'inscrire dans un intervalle de confiance l'estimation obtenue grâce à la formule de Muller.

Nous soulignerons qu'il convient d'accepter l'idée selon laquelle, en statistique lexicale, toute valeur théorique est inscrite dans une certaine incertitude qui interdit de tirer des conclusions à partir d'écarts relatifs trop faibles. Dans un tel cas, il faut compléter les mesures par des intervalles de confiance ou par le calcul de l'écart réduit.

Cette discussion laisse une question pendante : le modèle hypergéométrique s'adapte-t-il bien aux textes réels que doit traiter la statistique lexicale ? Dans ce cas, les valeurs observées

dans les corpus devraient se situer presque toutes dans les intervalles de confiance obtenus grâce à la formule de Muller. Nous verrons qu'il n'en est pas toujours ainsi, ce qui conduit à modifier le modèle probabiliste (cf. Hubert & Labbé 1988).

REFERENCES

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

Brunet Etienne (1982). "Loi hypergéométrique et loi normale. Comparaison dans les grands corpus". *Actes du second colloque de lexicologie politique*. Paris : Klincksieck, tome III, p. 253-264.

Brunet Etienne (1983a). Le viol de l'urne. In Colette Charpentier et Jean David. *La recherche française par ordinateur en langue et littérature*. Genève-Paris : Slatkine-Champion, 1985, p. 253-264.

Brunet Etienne (1983b). L'hydre de l'urne. *Cahiers de lexicologie*, 43. 1983-2, p 3-31.

Hubert Pierre & Labbé Dominique (1988). Un modèle de partition du vocabulaire. In Labbé Dominique, Thoiron Philippe et Serant Daniel (Ed.). *Etudes sur la richesse et la structure lexicale*. Paris-Genève: Slatkine-Champion, p 93-114.

Lafon Pierre (1984). *Dépouillements et statistiques en lexicométrie*. Genève-Paris: Slatkine-Champion.

Muller Charles (1964). "Calcul des probabilités et calcul d'un vocabulaire". Reproduit dans : *Langue française et linguistique quantitative*. Genève-Paris: Slatkine-Champion,1979, p 167-176.

Muller Charles (1967). *Etude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille*. Paris : Larousse. (réédition : Genève-Paris, Slatkine-Champion, 1979).

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

Muller Charles (1981a). Sur les répartitions lexicales. Reproduit dans : *Langue française, linguistique quantitative, informatique.* Genève-Paris : Slatkine-Champion, 1985, p. 87-101.

Muller Charles (1981b). La répartition lexicale : problèmes et solutions. Reproduit dans : *Langue française, linguistique quantitative, informatique.* Genève-Paris : Slatkine-Champion, 1985, p. 103-113.

Serant Daniel (1988). A propos des modèles de raccourcissement de textes. In Labbé Dominique, Thoiron Philippe et Serant Daniel. *Etudes sur la richesse et la structure lexicale*. Paris-Genève: Slatkine-Champion, p 115-124.