Cyril LABBE
France Télécom RD (28 chemin du vieux Chêne - 38243 MEYLAN Cedex)
cyril.labbe@rd.francetelecom.fr

Dominique LABBE
CERAT-IEP, BP 48 – 38040 GRENOBLE Cedex 9.
Dominique.labbe@iep.upmf-grenoble.fr

QUE MESURE LA SPÉCIFICITÉ DU VOCABULAIRE ?

A la fin des années 1970, P.Lafon a proposé d'appliquer la distribution hypergéométrique à la question de la répartition des "formes" dans un corpus. Sa formule a été appliquée sur de nombreux corpus sous le nom de "spécificités du vocabulaire". Dans cette note, nous présentons, après un rappel de la formule de P.Lafon, une simulation du comportement des résultats de la formule, en fonction de la fréquence des mots et de la taille des parties, ainsi qu'une application à un corpus de textes politiques contemporains. Cette discussion montre la portée et les limites de la notion de "spécificités du vocabulaire".

Mots clefs : Statistique lexicale - Statistique textuelle- Spécificités du vocabulaire - Répartition des formes dans un corpus

Juin 1997

QUE MESURE LA SPÉCIFICITÉ DU VOCABULAIRE ?*

A la fin des années 1970, P.Lafon a proposé d'appliquer la distribution hypergéométrique à la question de la répartition des "formes"[1] dans un corpus (Lafon 1980,1984). Cette formule a été appliquée à de nombreuses reprises mais, à notre connaissance, aucun bilan d'ensemble n'a été dressé depuis celui de B.Habert (1985). Dans cette note, nous présentons, après un rappel de la méthode préconisée par P.Lafon, une simulation du comportement des résultats obtenus en fonction de la fréquence et de la taille des parties, et une application à un corpus de textes politiques contemporains. Cette discussion nous amènera à définir certaines limites à l'utilisation de la méthode.

I. La méthode de P.Lafon.

Cette méthode permet de mesurer les variations de la fréquence dans un corpus découpé en parties et, en fonction d'un seuil choisi par l'analyste, il indique si la fréquence observée dans telle ou telle partie peut-être considérée comme normale ou non. Dans ce dernier cas, P. Lafon propose de baptiser cette forme "spécifique" (de la partie considérée).

On notera par la suite :

T : la longueur du corpus (nombre total de mots de celui-ci)

t_i : la longueur de la partie i

f : la fréquence absolue d'une forme dans le corpus entier.

f_i : fréquence absolue d'une forme dans la partie i

X : la variable aléatoire mesurant le nombre d'apparitions d'une forme dans la partie considérée.

---

* Nous remercions MM. Pierre HUBERT, de l'Ecole des Mines de Paris, Pierre LAFON de l'Ecole Normale Supérieure de Fontenay-Saint-Cloud et Sergio BOLASCO, de l'Université La Sapienza de Rome, qui ont relu ce texte et ont bien voulu nous faire part de plusieurs remarques importantes.

[1] Nous adoptons la terminologie et le symbolisme proposés par le Laboratoire de lexicologie politique de Saint-Cloud et utilisés par P. Lafon dans son ouvrage de 1984 (cf la bibliographie placée à la fin de cette note).

On calcule une probabilité pour qu'une forme de fréquence f apparaisse k fois dans la partie i :

$$(1) \qquad P(X=k)=\frac{\binom{f}{k}\binom{T-f}{t_i-k}}{\binom{T}{t_i}}$$

Cette probabilité atteint son maximum à l'espérance mathématique, c'est-à-dire la "fréquence attendue dans la partie i", qui doit être nécessairement un entier tel que :

$$\frac{(f+1)(t_i+1)}{T+2} - 1 \le f'_i \le \frac{(f+1)(t_i+1)}{T+2}$$

Si la fréquence observée n'est pas égale à cette fréquence attendue, on peut se demander si l'écart entre les deux valeurs est ou non significatif. La probabilité pour qu'on rencontre une fréquence telle que celle observée, sera :

premièrement, avec $fi>f'_i$:

$$S^+ = P(X \ge f_i) = \sum_{k=f_i}^{Min(f;t_i)} P(X=k)$$

Si S+ est plus petit qu'un seuil choisi par l'opérateur — généralement 0,05 ou 0,01 —, on parle alors de *spécificité positive* : le mot est, d'après le calcul hypergéométrique, significativement "sur-employé" dans la partie considérée.

deuxièmement, avec $f_i<f'_i$:

$$S^- = P(X \le f_i) = \sum_{k=0}^{fi} P(X=k)$$

Dans le cas où S- est plus petit que le seuil choisi, on parle de *spécificité négative* : le mot est significativement "sous-employé" dans la partie considérée.

On ne calcule donc pas la probabilité de l'ensemble des possibles ($P(X \le f)$), l'indice ne sera jamais égal à un. Pour que chaque forme ait autant de chances d'être sur-employée que sous-employée, il faudrait même que la valeur maximum de la probabilité soit de 0,5.

Les factorielles de la formule (1) ne peuvent être programmées directement pour les chiffres auxquels on est confronté dans les corpus linguistiques (elles

aboutissent à des nombres extraordinairement grands). P.Lafon propose d'utiliser les logarithmes et aboutit à la formulation suivante[2] :

$$(3)\ \log P(X=f_i) = \log f! + \log (T-f)! + \log t_i! + \log (T-t_i)! - \log T! - \log f_i! - \log (f-f_i)! - \log (t_i-f_i)! - \log (T-f-t_i+f_i)!$$

Pour étudier le comportement de cet indice, on a fait varier ses quatre paramètres : la fréquence totale et la fréquence dans une partie i (f et f_i) puis la taille du corpus et des parties (T et t_i)[3].

II. Comportement de l'indice en fonction de la fréquence.

Pour étudier le comportement de la probabilité en fonction de la fréquence, nous avons réalisé une série de simulations puis une application à un corpus réel.

a. L'asymétrie de S dans les basses fréquences

A priori, la mesure de la répartition des mots dans un texte devrait donner, à chacun de ces mots, une probabilité égale de figurer dans les spécificités positives ou négatives. Dans toutes les applications sur des corpus réels, les effectifs S+ et S- ne sont jamais égaux. Cette asymétrie se manifeste notamment lorsque la fréquence attendue est faible : seule la spécificité positive existe. On ne peut conclure à une spécificité négative que quand :

$$P(X=0)<\text{seuil}$$

Les quatre graphes 1 ci-dessous montrent que la distribution de la probabilité n'est pas symétrique autour de la valeur moyenne et ceci d'autant plus que la valeur attendue est faible. En revanche, mode, moyenne et médiane tendent vers la même valeur quand f grandit :

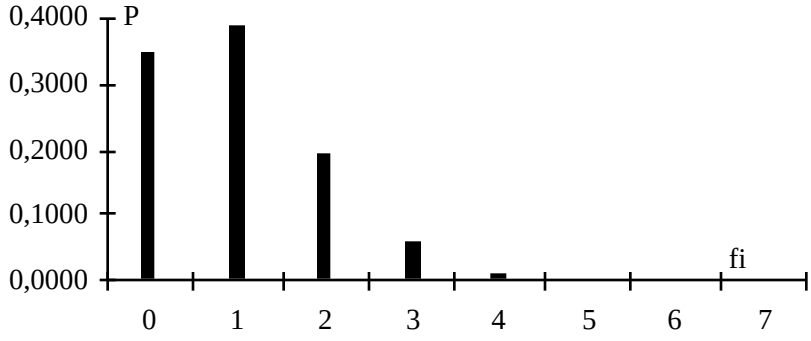
$S_{(f_i=f'_i)} \approx 0{,}5$ (avec f très grand), et
$S_{(f_i=f'_i)} > 0{,}5$ (avec f petit)

---

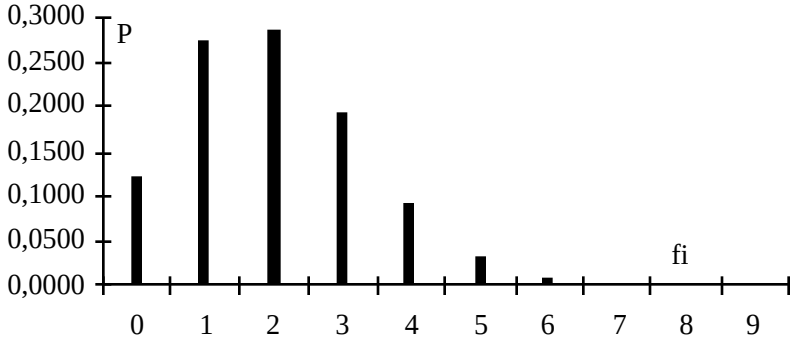
[2] Pour le passage de (1) à (3), nous renvoyons à l'ouvrage de Lafon (p 65-66). Il faut également utiliser la formule de Stirling pour les plus grands nombres. Lafon indique que cette formule donne une approximation à 10^{-8} près. Cela signifie qu'en tout état de cause les comparaisons sur l'indice de spécificité ne doivent pas aller au-delà de 10^{-7}.

[3] P. Lafon a également étudié l'influence de certaines de ces variations (op cit, p 61-64) mais il s'agissait d'établir graphiquement la supériorité de sa formule sur les applications de la loi normale ou de la loi de Poisson. Ses observations ne contredisent pas les résultats de notre simulation

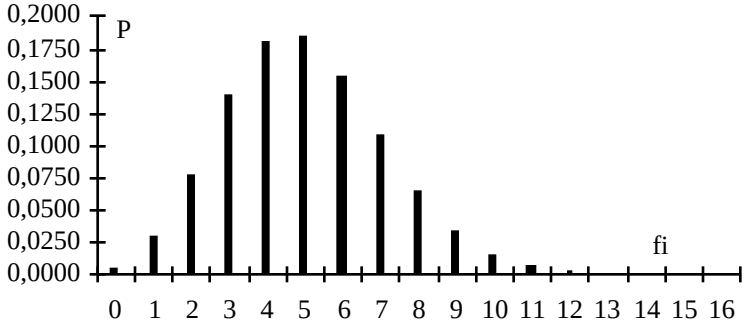
Graphique 1.1 Valeurs de P(X=fi) avec T=100.000 ; t=10.000 et f=10 (f'i=1)



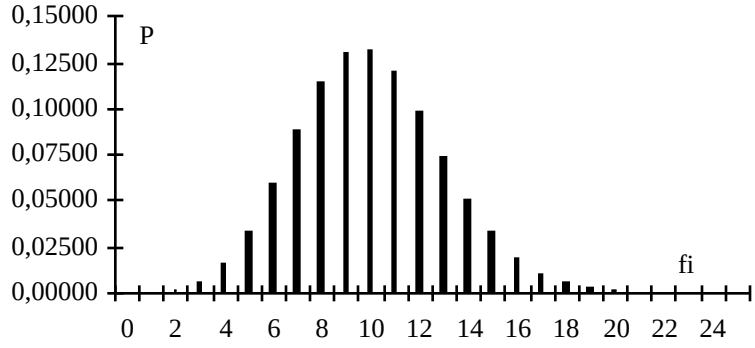
Graphique 1.2 Valeurs de P(X=fi) avec avec T=100.000 ; t=10.000 et f=20 (f'i=2)



Graphique 1.3 Valeurs de P(X=fi) avec avec T=100.000 ; t=10.000 et f=50 (f'i=5)



Graphique 1.4 Valeurs de P(X=fi) avec avec T=100.000 ; t=10.000 et f=100 (f'i=10)



Ces caractéristiques avaient déjà été signalées par A. Salem (Salem 1987, p 214). Le choix des formes soumises au calcul doit donc se porter sur celles dont la fréquence totale est au moins égale à une valeur telle qu'une absence, dans la plus petite des parties du corpus, aboutisse à une spécificité négative (Salem propose de baptiser ce minimum : "seuil d'absence spécifique").

En dessous de ce seuil, le calcul n'a de sens que lorsque la fréquence constatée f_i est supérieure à la fréquence attendue f'_i. La conséquence sera que, plus on descendra dans les basses fréquences, ou plus on découpera le corpus en parties de taille réduite, plus le nombre des spécificités positives excédera celui des spécificités négatives. Il ne faudrait pas attribuer cet artefact à une caractéristique stylistique propre au corpus étudié !

Mais, même au-dessus de ce seuil, la distribution des probabilités n'est pas symétrique autour de l'espérance mathématique (graphique 1.3 et 1.4). Sauf si l'on choisit de ne traiter que les mots à très fortes fréquences, les effectifs S+ et S- seront donc nécessairement différents.

b. La convergence en probabilité

Au-delà de ce problème des basses fréquences, quel est le comportement de la probabilité en fonction de la fréquence ? Ce comportement a été étudié en faisant varier f et f_i. Les autres paramètres du calcul sont maintenus fixes.

Le tableau 2 ci-dessous récapitule les résultats de l'une de ces expériences. On donne à f_i des valeurs entières s'écartant de f'_i, en plus ou en moins et selon un pas régulier. Pour pouvoir superposer diverses expériences, ces valeurs sont exprimées en proportion de f'_i. Le graphique 3 permet de visualiser, pour trois valeurs significatives de f, le comportement de la probabilité en fonction de f_i.

Tableau 2. Valeurs de S avec (T=300.000, t=30.000, f {1.000 -10.000} et fi {0,75 f'i à 1,25 f'i})

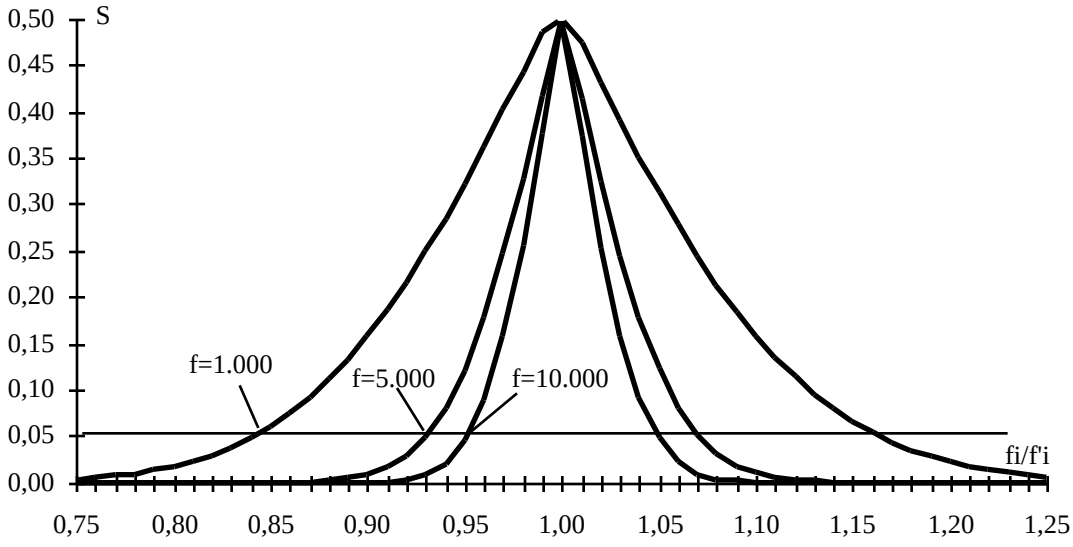
f_i/f'_i f	1000	2000	3000	4000	5000	6000	7000	8000	9000	10000
0,75	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,76	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,77	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,78	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,79	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,80	0,02	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,81	0,02	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,82	0,03	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,83	0,04	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,84	0,05	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,85	0,06	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,86	0,07	0,02	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,87	0,09	0,03	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,88	0,11	0,04	0,01	0,01	0,00	0,00	0,00	0,00	0,00	0,00
0,89	0,13	0,05	0,02	0,01	0,00	0,00	0,00	0,00	0,00	0,00
0,90	0,16	0,07	0,03	0,02	0,01	0,00	0,00	0,00	0,00	0,00
0,91	0,19	0,09	0,05	0,03	0,02	0,01	0,01	0,00	0,00	0,00
0,92	0,22	0,12	0,07	0,05	0,03	0,02	0,01	0,01	0,00	0,00
0,93	0,25	0,16	0,10	0,07	0,05	0,03	0,02	0,02	0,01	0,01
0,94	0,28	0,20	0,14	0,11	0,08	0,06	0,05	0,04	0,03	0,02
0,95	0,32	0,24	0,19	0,15	0,12	0,10	0,08	0,07	0,06	0,05
0,96	0,36	0,29	0,24	0,21	0,18	0,15	0,13	0,12	0,10	0,09
0,97	0,40	0,34	0,30	0,27	0,25	0,22	0,20	0,19	0,17	0,16
0,98	0,44	0,40	0,37	0,35	0,33	0,31	0,29	0,28	0,27	0,26
0,99	0,48	0,46	0,44	0,43	0,42	0,41	0,40	0,39	0,38	0,38
1,00	0,53	0,52	0,52	0,51	0,51	0,51	0,51	0,51	0,51	0,51
1,01	0,47	0,45	0,44	0,42	0,41	0,40	0,39	0,39	0,38	0,37
1,02	0,43	0,39	0,37	0,34	0,32	0,31	0,29	0,28	0,27	0,25
1,03	0,39	0,34	0,30	0,27	0,24	0,22	0,20	0,19	0,17	0,16
1,04	0,35	0,29	0,24	0,20	0,18	0,15	0,13	0,12	0,10	0,09
1,05	0,31	0,24	0,19	0,15	0,12	0,10	0,08	0,07	0,06	0,05
1,06	0,28	0,19	0,14	0,11	0,08	0,06	0,05	0,04	0,03	0,02
1,07	0,24	0,16	0,11	0,07	0,05	0,04	0,03	0,02	0,01	0,01
1,08	0,21	0,12	0,08	0,05	0,03	0,02	0,01	0,01	0,01	0,00
1,09	0,18	0,10	0,05	0,03	0,02	0,01	0,01	0,00	0,00	0,00
1,10	0,16	0,07	0,04	0,02	0,01	0,01	0,00	0,00	0,00	0,00
1,11	0,13	0,06	0,02	0,01	0,01	0,00	0,00	0,00	0,00	0,00
1,12	0,11	0,04	0,02	0,01	0,00	0,00	0,00	0,00	0,00	0,00
1,13	0,10	0,03	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1,14	0,08	0,02	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1,15	0,07	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1,16	0,05	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1,17	0,04	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1,18	0,03	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1,19	0,03	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1,20	0,02	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1,21	0,02	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1,22	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

On observe un resserrement très important de l'intervalle de variation autour de l'espérance mathématique. Par conséquent, plus la fréquence totale s'élève, plus la "zone de spécificité" est importante. Ce qui signifie également qu'une même variation relative de la fréquence sera interprétée différemment selon la valeur absolue de cette fréquence.

Cette observation est logique : elle découle de la "convergence en probabilité" qui est attachée à toute approche probabiliste. On peut donc dire que le calcul proposé par P.Lafon suppose, que plus une forme est employée fréquemment, plus ses variations, en valeur relative, doivent être faibles pour que son comportement soit considéré comme "normal".

Cette caractéristique théorique est-elle vérifiée sur des corpus de textes réels ? C'est ce que nous proposons d'examiner maintenant.

Graphique 3. Valeurs de S avec T=300.000, t=30.000, f {1.000, 5.000, 10.000} et f_i variant de 0,75 f'_i à 1,25 f'_i)



c. Application

L'application porte sur le "corpus Mitterrand" découpé en 4 parties sensiblement égales (Labbé, 1990)[4]. Ce corpus comporte 305.000 mots (T) et 7.000 "vocables" différents dont 1.751 de fréquences (f) supérieures à 11 (pour tenir compte du seuil de spécificité négative défini ci-dessus). Dans un premier temps, nous examinerons les chances pour l'un de ces 1.751 mots d'être

[4] Ces textes sont "lemmatisés" : à chacune des formes sont associées une "entrée de dictionnaire" et une catégorie grammaticale. On parle alors de "mots" et de "vocables" Cette opération permet d'observer l'influence de la langue sur la spécificité du vocabulaire.

spécifique en fonction de sa répartition dans l'ensemble du corpus et de sa fréquence totale.

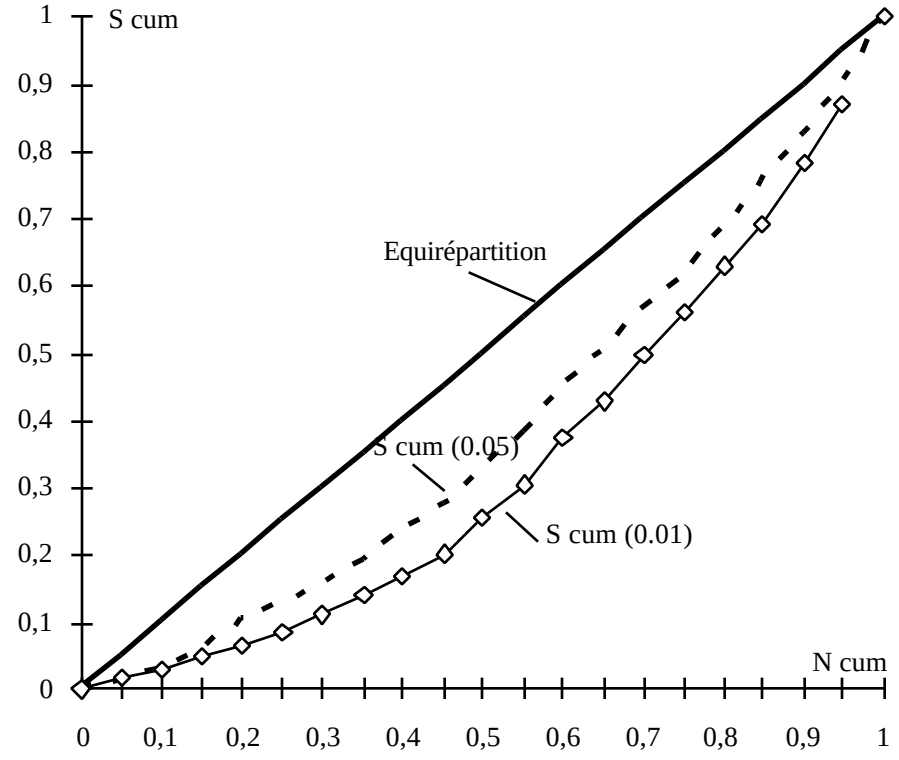
Pour juger de manière synthétique d'une éventuelle influence de la fréquence (f) sur S, nous proposons de construire une courbe de concentration (dite de "Gini-Lorenz") : les mots sont classés par fréquence croissante et, pour chaque classe de fréquence, on compte le nombre de mots qui auront été mesurés spécifiques dans l'une ou l'autre des parties. Puis, l'on compare le poids relatif des différentes classes de fréquences dans l'effectif total et dans les mots spécifiques.

Le tableau 4 et le graphique 5 permettent de visualiser cette expérience.

Tableau 4. Spécificités du vocabulaire en fonction de la fréquence dans le corpus Mitterrand. Classement par fréquence croissante.

Effectifs des classes de fréquence cumulés (N cum)	Spécificités cumulées (seuil de 0.05)	Spécificités cumulées (seuil de 0.01)
0,05	0,01	0,01
0,10	0,03	0,03
0,15	0,06	0,05
0,20	0,10	0,07
0,25	0,13	0,09
0,30	0,16	0,11
0,35	0,19	0,14
0,40	0,23	0,17
0,45	0,28	0,20
0,50	0,33	0,25
0,55	0,38	0,30
0,60	0,45	0,38
0,65	0,50	0,43
0,70	0,56	0,50
0,75	0,62	0,56
0,80	0,69	0,63
0,85	0,75	0,69
0,90	0,83	0,78
0,95	0,90	0,87
1,00	1,00	1,00

Graphique 5 Courbe de concentration de la spécificité en fonction de la fréquence



Dans la première colonne (abscisse du diagramme) : les effectifs relatifs cumulés par classes de fréquence ; dans la seconde et la troisième colonne (ordonnée du diagramme) : la proportion de mots spécifiques en retenant comme seuil 0,05 puis 0,01. Si la fréquence n'influait pas sur la spécificité du vocabulaire, on devrait retrouver sensiblement les mêmes chiffres dans les trois colonnes et, sur le graphique, les points devraient se situer sur la diagonale du carré (équirépartition du caractère sur l'ensemble des classes de fréquence). En revanche, plus ceux-ci s'éloignent de la diagonale, plus le phénomène — ici la spécificité — se concentre sur les mots les plus fréquents, ce qui est parfaitement net ici. En effet, le tableau indique que les 5% de mots dont la fréquence est la plus faible — soit f=12 — ne donnent que 1% des spécificités et qu'à l'opposé, les 100 mots les plus employés en donnent dix fois plus au seuil de 0,05 et 13 fois plus au seuil de 0,01.

Cette expérience montre que plus la fréquence s'élève, plus les mots ont de chances d'être spécifiques[5].

Comment expliquer la corrélation existant entre la fréquence et la spécificité ? On peut penser que la "convergence en probabilité" ne se produit pas, ou moins fortement que ce que la théorie laisserait attendre. Ceci peut être

[5] Dès l'origine, P. Lafon a souligné cette caractéristique, op cit , p 75.

expliqué de la manière suivante. Pour les basses fréquences, le nombre théorique de combinaisons possibles (calculé grâce à la loi hypergéométrique) ne s'éloigne pas trop du nombre de combinaisons effectivement réalisables en respectant les règles du français. Il n'en est pas de même pour les mots de hautes fréquences : les combinaisons statistiquement possibles sont immenses et la plupart ne sont évidemment pas réalisables. Autrement dit, la structuration du discours — sous la double influence de la langue et du projet de l'auteur — se fait d'autant plus sentir que l'on monte dans les fréquences…

Ainsi peut s'expliquer la présence de très nombreux "mots outils" dans la liste des mots spécifiques du corpus Mitterrand que l'on trouvera en annexe de cette note : articles, pronoms, adverbes usuels ou verbes auxiliaires alors que ces mots représentent une proportion très petite du vocabulaire total à la disposition de l'auteur. Ils devraient donc figurer également en petit nombre dans la liste. De plus, — comme il est impossible de ne pas les employer — on s'attendrait à ce que leur utilisation ne fluctue guère chez un même auteur et, dans une certaine mesure, d'un auteur à l'autre. Enfin, cette liste montre que la plupart des probabilités attachées à ces mots très usuels sont proprement infinitésimales et, d'ailleurs, nettement inférieures à la précision du calcul. En tout état de cause, leur répartition devrait être considérée comme extrêmement improbable…

Cependant, au milieu des "mots outils", figurent d'autres verbes, des substantifs, des adjectifs ou des noms propres :

— la plupart ont une fréquence élevée, ce qui explique leur présence dans la liste à côté des mots outils. Naturellement, ce dernier point montre l'intérêt de la méthode de Lafon : elle permet, pour les mots de fortes fréquences, de localiser des écarts relatifs même faibles et de leur affecter un signe ;

— figurent également dans la liste, quelques substantifs moins fréquents. C'est le second intérêt de la méthode : elle permet de détecter des mots peu utilisés mais dont l'emploi est très localisé dans une des parties du corpus (ce qui peut donc être très significatif). Sans le calcul de Lafon, leur répartition particulière risquerait de passer inaperçue puisqu'ils sont perdus dans les profondeurs des listes de fréquences. Ainsi, en tête de la liste, on trouve "chaîne", "équilibre des forces" et "université" qui ne sont employés qu'à un moment particulier du septennat (la création de la cinquième chaîne de télévision, l'affaire des SS20 et de l'équilibre des forces nucléaires en Europe, la contestation étudiante et lycéenne de l'automne 1986). Cette localisation extrême entraîne un S+ dans l'une des parties et trois S- dans les autres.

Cependant, cette brève discussion aura montré qu'il ne paraît guère possible de classer les mots en fonction des valeurs de S puisque le calcul recouvre des phénomènes langagiers et discursifs totalement opposés : fortes contraintes d'un côté ; extrême imprévisibilité de l'autre…

A cela s'ajoute la grande sensibilité de la probabilité à la taille du corpus et des parties.

III. L'influence de la taille sur la spécificité.

Comme précédemment, nous nous proposons d'examiner l'influence de la taille sur la probabilité grâce à une série de simulations théoriques avant de confronter ces résultats avec ceux d'une application.

a. Simulations

L'influence de la taille totale (T) sur la spécificité a été signalée par tous les auteurs : plus le corpus est grand, plus la proportion de mots spécifiques est élevée. Deux éléments sont à prendre en compte :

— Du point de vue probabiliste, l'augmentation de la taille des échantillons doit entraîner un resserrement des observations autour de l'espérance mathématique (voir tableau 2 et graphique 3 ci-dessus).

— Du point de vue linguistique, l'accroissement du vocabulaire est une fonction décroissante de l'allongement du corpus : quand T s'accroît, les fréquences élevées augmentent plus que proportionnellement, ce qui nous ramène au cas de figure précédent.

Autrement dit — pour des raisons qui tiennent au raisonnement probabiliste et à la finitude du vocabulaire —, plus le corpus est vaste, plus les mots fréquents devraient être régulièrement répartis sur l'ensemble du corpus.

L'influence de la taille des parties (t_i) sur la probabilité découle en partie du constat précédent. Elle dépend également de la relation existante entre t_i et T. Nous avons étudié cette relation particulière en faisant varier t_i et en maintenant fixes f et T. Le graphique 6 et le tableau 7 récapitulent les résultats obtenus lors de l'une de ces expériences théoriques.

On observe toujours le même phénomène de gonflement de la zone de spécificité (ou de rétrécissement de la "zone de normalité"). Là encore, le raisonnement probabiliste explique ce résultat. Plus les tailles des parties sont importantes, plus la forme considérée devrait présenter, "normalement", des variations relatives faibles ou encore plus sa fréquence observée (f_i) dans chaque partie devrait s'approcher de la valeur attendue f'_i.

Graphique 6. Evolution de S en fonction de la taille des parties (T=300.000 ; f=10.000 ; t_i {30.000 — 150.000} ; f_i {0,9 f'_i — 1,1 f'_i}

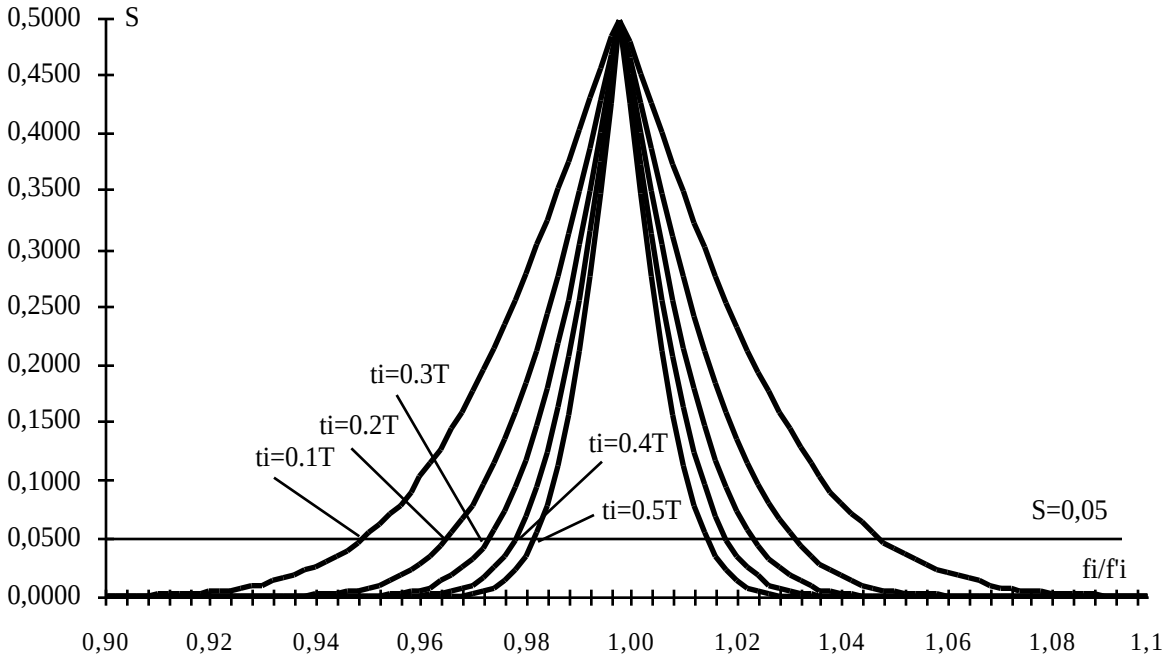


tableau 7. Evolution de S en fonction de la taille des parties (T=300.000 ; f=10.000 ; t_i {30.000 — 150.000} ; f_i {0,9 f'_i — 1,1 f'_i})

t_i= fi/f'i	30.000	60.000	90.000	120.000	150.000
0,90	0,000	0,000	0,000	0,000	0,000
0,91	0,001	0,000	0,000	0,000	0,000
0,92	0,003	0,000	0,000	0,000	0,000
0,93	**0,009**	0,000	0,000	0,000	0,000
0,94	0,021	0,001	0,000	0,000	0,000
0,95	0,046	0,006	0,000	0,000	0,000
0,96	0,090	**0,021**	0,004	0,001	0,000
0,97	0,159	0,065	**0,023**	0,007	0,001
0,98	0,255	0,158	0,093	**0,049**	0,022
0,99	0,375	0,311	0,257	0,206	**0,156**
1,00	0,508	0,506	0,505	0,504	0,504
1,01	0,372	0,309	0,256	0,206	**0,157**
1,02	0,254	0,158	0,094	**0,050**	0,022
1,03	0,159	0,066	**0,024**	0,007	0,001
1,04	0,091	**0,022**	0,004	0,001	0,000
1,05	0,047	0,006	0,001	0,000	0,000
1,06	0,023	0,001	0,000	0,000	0,000
1,07	**0,010**	0,000	0,000	0,000	0,000
1,08	0,004	0,000	0,000	0,000	0,000
1,09	0,001	0,000	0,000	0,000	0,000
1,10	0,000	0,000	0,000	0,000	0,000

b. Application

Pour mesurer l'influence de la taille des parties sur les spécificités du vocabulaire, nous avons découpé le corpus "Mitterrand" en parties égales de plus en plus nombreuses. Les résultats sont donnés dans le tableau 8 et le graphique 9 ci-dessous.

Tableau 8. Proportion du vocabulaire spécifique dans le corpus Mitterrand découpé en parties égales.

Valeur de S inférieures à	Pourcentage du total des mots spécifiques de fréquence supérieure ou égale à 12 pour :								
	2 parties	3 parties	4 parties	5 parties	6 parties	7 parties	8 parties	9 parties	10 parties
-0,05	19,7	15,6	15,3	9,8	8,5	7,5	7,4	5,9	5,4
-0,10	5,0	4,6	5,4	5,8	5,7	4,6	4,4	3,7	3,7
-0,15	3,7	4,7	4,5	4,4	4,6	5,3	4,3	4,0	3,8
-0,20	4,0	4,0	5,2	3,9	4,7	5,5	4,8	4,7	4,1
-0,25	3,0	3,3	4,4	4,2	3,4	3,8	5,2	5,3	4,7
-0,30	2,8	3,7	2,8	4,1	4,2	3,5	4,8	4,7	4,9
-0,35	2,5	3,1	2,7	3,8	3,6	3,8	2,8	4,8	5,4
-0,40	3,5	2,6	3,7	4,4	3,3	3,1	4,3	3,3	3,0
-0,45	3,6	1,6	2,7	2,0	2,0	1,4	1,9	1,3	1,4
≤-0.45	4,7	11,4	8,6	13,5	16,3	18,5	18,6	22,3	23,6
≥0.45	4,7	9,7	8,1	11,2	12,1	12,3	11,0	12,6	13,7
0,45	3,6	1,3	1,3	0,7	0,5	0,7	0,8	0,4	0,4
0,40	3,5	2,8	2,6	1,7	1,5	1,6	1,7	1,1	1,1
0,35	2,5	2,5	2,6	2,8	2,6	2,4	2,1	1,7	1,7
0,30	2,8	2,9	3,1	2,7	2,9	2,8	2,5	2,7	2,3
0,25	3,0	3,7	3,0	3,0	3,3	2,8	2,8	2,5	2,8
0,20	4,0	3,3	3,2	3,3	3,1	3,1	2,9	3,1	2,8
0,15	3,7	4,0	3,9	3,4	3,5	3,2	3,6	3,3	2,9
0,10	5,0	5,0	4,2	4,1	4,4	3,8	3,6	3,8	3,8
0,05	14,7	12,3	12,5	10,7	10,0	10,1	9,5	8,8	8,5
% V spécifiques	34,4	27,9	27,8	20,6	18,5	17,7	16,9	14,7	13,9

Figure 9.1 Diagramme de répartition des spécificités dans le corpus Mitterrand découpé en 2 parties égales

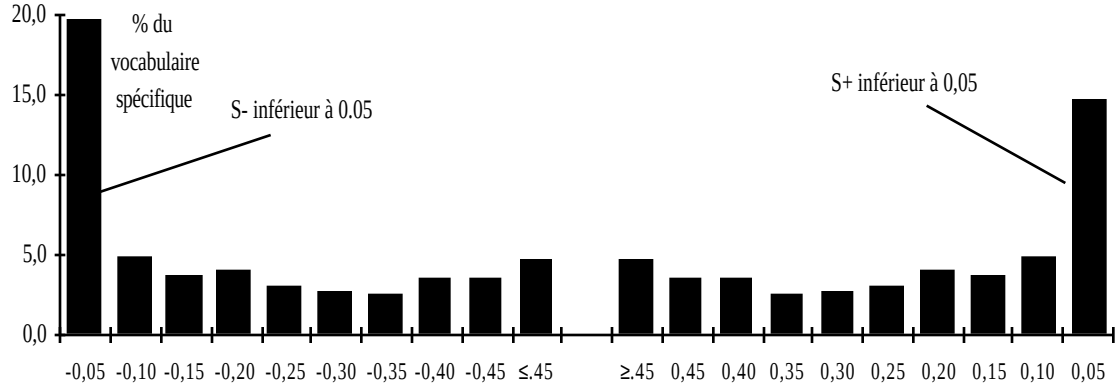
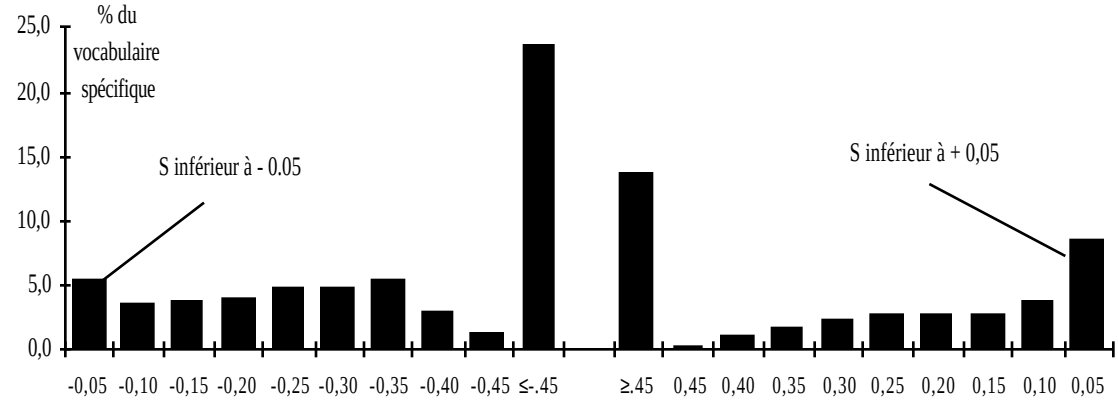


Figure 9.2 Diagramme de répartition des spécificités dans le corpus Mitterrand découpé en 10 parties égales



Avec un découpage en 2 parties égales, plus du tiers des mots sont spécifiques en tel ou tel point du corpus (à un seuil de 0,05)[6]. Avec un découpage en 10 parties égales, cet effectif est légèrement inférieur à 14%, soit une division par 2,5. La taille des parties joue donc un rôle considérable dans les chances qu'auront les mots d'être ou non spécifiques.

---

6 Par rapport au tableau reproduit dans Habert (Habert, p 130), nous n'obtenons jamais des proportions aussi fortes de V spécifiques. Outre l'homogénéité du corpus qui est probablement l'explication principale — même locuteur, période relativement brève, conditions d'énonciation homogènes —, deux éléments peuvent également expliquer cette différence. Lafon effectue ses calculs sur les formes les moins fréquentes seulement quand elles apparaissent concentrées dans une partie (Lafon, p 80-81) alors que nous prenons en compte tous les mots à partir d'un certain seuil de fréquence. En second lieu, nous travaillons ici sur un "corpus lemmatisé" : cela augmente considérablement la population dans les hautes fréquences, donc le nombre de mots sur lequel on peut calculer, mais le procédé peut "lisser" la répartition de certains d'entre eux (les verbes usuels notamment).

Ces résultats appellent quelques remarques complémentaires :

— Il n'y a pas de symétrie entre le haut et le bas du tableau (ou entre les parties gauche et droite des graphiques). Avec deux parties, les spécificités négatives (S^-) sont plus nombreuses que les positives (S^+) ; avec cinq parties et plus, la proportion se renverse et l'écart se creuse progressivement. Les graphes 1 discutés plus haut fournissent l'explication. Pour 2 parties, la plus petite fréquence attendue est de 6 puisque $f \geq 12$: dans une telle configuration, on voit que l'asymétrie des probabilités est assez faible mais qu'elle joue en faveur des S^-. Pour dix parties, les mots, dont la fréquence sur l'ensemble du corpus est inférieure à 30, ne peuvent plus être S^-. L'asymétrie des probabilités joue massivement en faveur des spécificités positives.

— On observe également une asymétrie croissante entre l'effectif à -0,5 et à +0.5. Cette caractéristique tient aussi à l'application du calcul sur des fréquences assez basses…

— Une proportion importante des mots sont spécifiques — et, étant donné leur haute fréquence, ils couvrent la majorité du texte — ce qui pose problème quant au raisonnement probabiliste : la spécificité devient la situation "normale" et la normalité, l'exception !

Enfin, on aura noté que, dans les tableaux 2 et 7, les valeurs de S, lorsque $f_i=f'_i$, sont légèrement supérieures à 0,5 et que l'écart tend à diminuer au fur et à mesure que T,t_i et f augmentent. Ce biais est inhérent au calcul proposé par Lafon. La formule (3) consiste en effet à calculer l'espace sous la courbe entre 0 ou f et la valeur observée *incluse*. La probabilité pour que $X=f'_i$ est donc comptée deux fois. Naturellement, cette caractéristique de la formule ne se voit que lorsque la valeur observée coïncide avec l'espérance mathématique. Elle n'avait pas été aperçue car ces valeurs n'intéressent généralement pas l'observateur qui recherche les spécificités et ne s'intéresse pas à l'ensemble de la distribution.

Conclusions

A l'issue de cette discussion, il nous paraît nécessaire de préciser les limites d'utilisation du calcul hypergéométrique pour la recherche des spécificités du vocabulaire d'un corpus.

En premier lieu, il convient d'y associer un seuil minimal de fréquence en dessous duquel le calcul ne sera pas effectué, seuil qui ne pourra pas être inférieur à la valeur nécessaire pour que la fréquence attendue dans la plus petite partie du corpus soit au moins égale à trois (puisque, en toute hypothèse avec $f'i<3$, la spécificité négative ne peut pas être inférieure à 0,05). Même avec un tel seuil, le nombre des mots S^+ ne sera pas égal à celui des S^-. Pour établir une certaine égalité, la fréquence attendue dans la plus petite des parties doit

probablement être nettement supérieure à 5. Autrement dit, cela limiterait l'application de la formule à quelques mots usuels, dans de grands corpus découpés en parties elles-mêmes vastes. Dans le cas contraire, pour la plupart des mots, le fait d'être S^+ ne pourra pas être considéré comme exactement symétrique de l'événement S^-.

Etant donné l'influence considérable de la fréquence sur la probabilité, il n'est pas souhaitable de comparer, du point de vue de leur spécificité, des mots de fréquences trop différentes. Cela devrait conduire à faire apparaître clairement la fréquence aux côtés de la spécificité et à découper les tableaux en plusieurs compartiments (pour le moins : fréquences fortes, moyennes et faibles).

Plus fondamentalement : puisque la formule mesure indissociablement les effets des choix de l'auteur et les contraintes que la langue fait peser sur ses usagers, il serait souhaitable de comparer des mots de même nature : la spécificité d'un pronom personnel ne devrait s'apprécier que par rapport aux autres pronoms personnels, les substantifs et les adjectifs par rapport aux autres substantifs et adjectifs, de même pour les articles, les numéraux, etc…

Enfin, les parties comparées devraient être sensiblement égales ou, dans le cas contraire, les comparaisons devraient être décomposées pour constituer des sous-groupes de tailles pas trop différentes.

Ces conditions restrictives posées, la présentation des résultats pose un problème aux non-statisticiens : la spécificité est d'autant plus forte que S est faible ! Pour faciliter la compréhension des résultats, deux solutions sont envisageables :

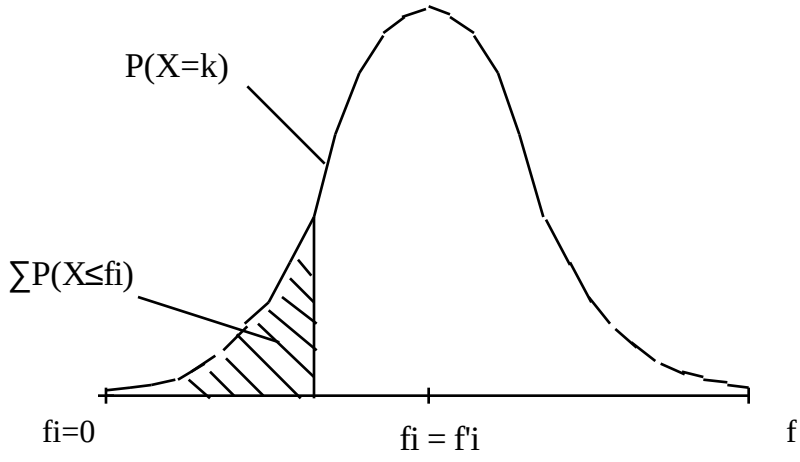
On exprime S sous la forme d'un indice fluctuant entre 0 et ±1 avec une spécificité d'autant plus forte qu'il s'approche de ±1 et une valeur nulle effectivement égale à 0… Ceci peut être obtenu grâce à la formulation suivante :

$$\text{Indice} = \frac{0{,}5 - S}{0{,}5}$$

Mais cette formulation présente un inconvénient : l'indice n'est plus égal à la probabilité (que l'événement f_i ne soit pas dû au hasard). Par exemple avec S = 0,05, l'indice sera égal à 0,90 (et non à 0.95)[7]. C'est pourquoi, à la réflexion, une seconde solution paraît préférable.

Deuxièmement, le calcul de Lafon mesure l'espace compris sous la courbe de probabilité de l'événement f_i :

[7] Dans une première version de ce texte, réalisée en 1994, nous avions préconisé la première formulation. Depuis, Sergio Bolasco a attiré notre attention sur les inconvénients de cet indice.



Pour mesurer cette surface, la formule devient :

$$(1)\ S = \sum_{k=0}^{f_i} P(X=k)$$

Et pour le calcul pratique,

avec $f_i<f'_i$, on utilise la formule (1) ci-dessus,

avec $f_i=f'_i$: $S = 0.5$

avec $f_i > f'_i$, : $S = 1 - \sum_{k=f_i}^{Min(f,t_i)} P(X=k)$

Le mot sera significativement sur-employé, au seuil de 5%, avec un indice supérieur à 0,95 (ou 0,975 selon que l'on considère la moitié de la courbe ou l'ensemble de celle-ci) et à 0,99 (ou 0,995) au seuil de 1%. A l'inverse, il sera significativement sous-employé avec un indice inférieur à 5% ou 1%. La spécificité positive sera d'autant plus significative que S se rapproche de 1 ; la spécificité négative d'autant plus significative que S tend vers 0. Cette formulation est préférable à l'indice ci-dessus ou à la notation adoptée par le laboratoire de lexicométrie de Saint-Cloud, à la suite des travaux de P. Lafon.

En définitive, ce calcul est une application du schéma de l'urne que, il y a plus de 30 ans, C. Muller avait proposé d'appliquer à l'étude du vocabulaire :

"Essayer les méthodes statistiques sur le vocabulaire d'un texte, c'est avouer une croyance, ou tout au moins ne pas refuser une hypothèse : celle d'après laquelle le choix des mots, dans l'exercice du langage, relève des lois du hasard et peut être assimilé à un tirage aléatoire. Cela tout au moins quand on considère

une étendue suffisante de texte, et qu'on l'envisage comme une masse en faisant abstraction de l'ordre d'apparition de ses éléments" (Muller, 1964).

Comme C. Muller en avait eu l'intuition, dès l'origine, ce schéma est un pis-aller : nous sommes condamnés à l'utiliser tant que n'auront pas été développés des modèles théoriques plus adaptés à la réalité du langage.

Grenoble, juin 1997.

Bibliographie

HABERT Benoît, 1983, "L'analyse des formes spécifiques et typologie des énoncés", *Mots*, 7, octobre 1983, p 97-121.

HABERT Benoît, 1985, "L'analyse des formes spécifiques : bilan critique et propositions d'utilisation", *Mots*, 11, octobre 1985, p 127-154.

LABBE Dominique, 1990, *Le vocabulaire de F. Mitterrand*, Paris, Presses de la Fondation nationale des sciences politiques.

LAFON Pierre, 1984, *Dépouillements et statistiques en lexicométrie*, Genève-Paris, Slatkine-Champion.

LAFON Pierre, "Sur la variabilité de la fréquence des formes dans un corpus", *Mots*, 1, octobre 1980, p 127-165.

LEBART Ludovic, SALEM André, 1994, *Statistique textuelle*, Paris, Dunod.

MULLER Charles, 1964, *Essai de statistique lexicale. L'illusion comique de Pierre Corneille*, Paris, Klincksieck.

MULLER Charles, 1977, *Principes et méthodes de statistique lexicale*, Paris, Hachette.

SALEM André, 1987, *Pratique des segments répétés. Essai de statistique textuelle*, Paris, Klincksieck.

Annexe 1. Les 100 indices de spécificités les plus forts du corpus Mitterrand (classement par spécificités décroissantes)

Mots	Fréquence totale	Indices de Spécificité	Mots	Fréquence totale	Indices de Spécificité
le (art)	29559	5,7499E-46	Bretagne	22	1,2172E-10
de	20964	6,8849E-38	industrie	136	1,2414E-10
de	20964	4,5045E-36	missile	37	1,4262E-10
ce (pro)	5260	2,3148E-30	université	24	2,0259E-10
nous	2047	4,5594E-27	cinquième	36	4,3325E-10
impôt	91	3,6982E-25	avoir (verbe)	7713	8,0014E-10
nous	2047	1,5858E-24	musée	25	9,6348E-10
le (art)	29559	4,3662E-22	politique(nom fém.)	405	1,0188E-09
chaîne	76	8,5288E-22	tonne	35	1,3171E-09
Israël	71	2,3725E-19	candidat	42	2,4393E-09
Iran	53	1,0665E-18	consommation	34	3,9324E-09
mille	957	3,2005E-18	palestinien (adj.)	30	4,2587E-09
je	8902	3,3454E-16	soldat	30	4,2587E-09
notre (art)	710	3,6497E-16	je	8902	5,3997E-09
et	5516	9,1155E-16	beaucoup	519	5,6528E-09
réforme	65	1,4064E-15	arabe (adj.)	50	6,6917E-09
à	7228	1,5993E-15	très	627	8,1555E-09
président	303	1,7173E-15	deux	560	1,0907E-08
neuf	821	2,0499E-15	secteur	63	1,3697E-08
politique (nom fém.)	405	3,0575E-15	certain (dét.)	13	1,489E-08
entreprise	218	3,2991E-15	neuf	821	1,6265E-08
étudiant	35	3,5774E-15	département	62	1,7902E-08
être (verbe)	10995	4,8506E-15	mars	61	1,7929E-08
mille	957	5,2554E-15	monsieur	452	2,3313E-08
dix	301	5,7692E-15	en (prép.)	2683	3,3163E-08
nationalisation	57	1,514E-14	Irak	32	3,404E-08
république	330	3,1854E-14	réjouir	30	3,7509E-08
y	1745	3,6265E-14	ce (art.)	2723	4,0497E-08
France	1219	3,6418E-14	trois	314	4,3219E-08
soixante	250	4,6723E-14	pays	742	4,5579E-08
quatre	745	4,8787E-14	croire	500	6,6265E-08
cela	1633	1,2577E-13	Alsace	17	7,4203E-08
on	2587	1,9111E-13	ensemble (adv.)	66	7,4858E-08
Tchad	114	2,1574E-13	vingt	666	8,195E-08
nous	2047	4,3876E-13	force	204	9,9272E-08
Tchad	114	1,0853E-12	croissance	64	9,9586E-08
cent	1200	1,6039E-12	cohabitation	21	1,187E-07
relance	32	2,6447E-12	langue	27	1,2078E-07
majorité	217	3,3374E-12	inflation	83	1,6211E-07
impôt	91	4,2524E-12	payer	50	1,638E-07
le (art)	29559	7,8879E-12	quand	441	1,9586E-07
à	7228	9,4133E-12	juge (nom masc.)	34	2,0153E-07
otage	37	1,3366E-11	emploi	119	2,0361E-07
nous	2047	4,309E-11	réduire	68	2,111E-07
notre (art)	710	4,521E-11	Europe	390	2,1191E-07
équilibre	114	5,061E-11	énergie	48	2,1291E-07

paix	91	5,3937E-11	on	2587	2,1563E-07

Annexe II. Les mots spécifiques du corpus Mitterrand découpé en quatre parties

Mots	Fréquence	Spécificitédans les quatre parties			
chaîne (n.f.)	76	8,5288E-22			8,7005E-05
		5,9929E-06			0,00210703
		5,9929E-06	avoir (v.)	7713	8,0014E-10
		0,00367453			5,02E-07
équilibre (n.m.)	114	5,061E-11			0,0052096
		1,4274E-06	banque (n.f.)	33	5,495E-07
		0,00022548			0,00532218
		0,00024403			0,00532218
force (n.f.)	204	9,9272E-08	Bretagne	22	1,2172E-10
		7,6728E-07			0,00178336
		1,3212E-06			0,00178336
		2,0276E-05	candidat (n.m.)	42	2,4393E-09
non (adv.)	523	2,8239E-07			8,4787E-05
		4,6441E-05			0,00302974
		0,00116901	ce (pro.)	5260	2,3148E-30
		0,00624776			9,7778E-11
nous (pro.)	2047	4,5594E-27			0,00175001
		1,5858E-24	ce (det.)	2723	4,0497E-08
		4,3876E-13			0,00142254
		4,309E-11			0,00420445
réforme (n.f.)	65	1,4064E-15	charge (n.f.)	104	3,1558E-07
		0,0003343			0,00052118
		0,00119105			0,00298387
		0,00359855	chiffre (n.m.)	46	1,9838E-05
son (det.)	868	0,00023928			0,00484486
		0,0011452			0,00484486
		0,00378882	cinquième (det.)	36	4,3325E-10
		0,00759949			3,1762E-05
université	24	2,0259E-10			0,00263694
		0,0090286	consommation (n.f.)	34	3,9324E-09
		0,0090286			5,647E-05
		0,0090286			0,00069655
Dans trois parties :			croissance (n.f.)	64	9,9586E-08
aller (v.)	786	2,2706E-07			1,803E-05
		6,9113E-05			1,803E-05
		0,00599067	de (prep.)	20964	6,8849E-38
alliance (n.f.)	42	9,2296E-06			4,5045E-36
		8,4787E-05			0,00858815
		0,00062563	disposer (v.)	162	3,2708E-07
Alsace	17	7,4203E-08			0,00115378
		0,00751583			0,00727124
		0,00751583	dix (det.)	301	5,7692E-15
arabe (adj.)	50	6,6917E-09			8,9458E-07
					0,00030253

domaine (n.m.)	152	6,5136E-06
		0,00112968
		0,00421396
économie (n.f.)	87	0,00059008
		0,00806731
		0,00806731
énergie (n.f.)	48	2,1291E-07
		0,00320499
		0,00320499
entreprise (n.f.)	218	3,2991E-15
		4,6584E-05
		0,00025781
et (conj.)	5516	9,1155E-16
		2,4826E-05
		0,0001277
être (v.)	10995	4,8506E-15
		5,8657E-05
		0,00121785
étudiant (n.m.)	35	3,5774E-15
		4,2351E-05
		0,00053652
Europe	390	2,1191E-07
		7,5572E-07
		0,00418295
Habré	25	0,00021438
		0,00702209
		0,00702209
Hissein	25	0,00021438
		0,00702209
		0,00702209
il (pro.)	4976	3,5691E-06
		0,00015046
		0,00806081
impôt (n.m.)	91	3,6982E-25
		4,2524E-12
		2,0649E-05
investissement	54	4,1045E-07
		0,00019632
		0,00019632
Irak	32	3,404E-08
		0,00117145
		0,0067059
Iran	53	1,0665E-18
		4,454E-06
		4,0996E-05
Israël	71	2,3725E-19
		3,2503E-06
		3,2503E-06
le (art.)	29559	5,7499E-46
		4,3662E-22
		7,8879E-12
Liban	65	6,2324E-07
		1,4142E-05
		0,00359855
Libye	41	0,00011054
		0,00637569
		0,00637569
libyen (adj)	25	7,3565E-06
		0,0007523
		0,00702209
Louvre	26	0,00038961
		0,00545454
		0,00545454
majorité (n.f.)	217	3,3374E-12
		3,5351E-06
		0,00032183
mars (n.m.)	61	1,7929E-08
		0,00019115
		0,00736562
missile (n.m.)	37	1,4262E-10
		0,00031765
		0,00893878
monsieur (n.m.)	452	2,3313E-08
		8,7741E-05
		0,00447171
musée (n.m.)	25	9,6348E-10
		0,00702209
		0,00702209
nationalisation (n.f.)	57	1,514E-14
		9,6733E-05
		0,00176672
nécessaire (adj.)	125	7,6194E-05
		0,00024838
		0,00563478
neuf (det.)	821	2,0499E-15
		1,6265E-08
		7,5788E-05
notre (det.)	710	3,6497E-16
		4,521E-11
		0,0082605
nouveau (adj.)	269	3,6595E-05
		0,00405375
		0,00866057
otage (n.m.)	37	1,3366E-11

		2,382E-05			0,00017627
		0,00031765			0,00079586
Palestinien (n.m.)	28	2,2071E-06	social (adj.)	227	1,1981E-05
		0,00327961			0,00032152
		0,00327961			0,00807394
pas (adv.)	4042	4,6071E-07	soixante (det.)	250	4,6723E-14
		2,5712E-06			0,00011633
		0,00928358			0,00202481
président (n.m.)	303	1,7173E-15	Tchad	114	2,1574E-13
		2,2595E-05			1,0853E-12
		0,00041266			0,00061691
quatre (det.)	745	4,8787E-14	tonne (n.f.)	35	1,3171E-09
		2,3998E-07			4,2351E-05
		0,00907428			0,00053652
que (pro.)	2326	0,00852945	Union Soviétique	84	8,8629E-07
		8,9538E-05			2,1522E-06
		0,00040214			0,00030124
référendum (n.m.)	27	6,0441E-06	vingt (det.)	666	8,195E-08
		0,00423187			9,6621E-06
		0,00423187			0,00241435
relance (n.f.)	32	2,6447E-12	y (pro)	1745	3,6265E-14
		0,0001004			7,9962E-07
		0,0067059			0,00246845
république (n.f.)	330	3,1854E-14			