REVUE DE STATISTIQUE APPLIQUÉE

ANIS ABI FARAH

# Un estimateur du lexique d'un texte de situation

*Revue de statistique appliquée*, tome 34, n° 2 (1986), p. 63-75

<http://www.numdam.org/item?id=RSA_1986__34_2_63_0>

# UN ESTIMATEUR DU LEXIQUE D'UN TEXTE DE SITUATION

Anis ABI FARAH

*Faculté des Sciences,*
*Université Libanaise, Hadath — Beyrouth*

Un auteur d'un texte de situation, puise ses vocables dans un lexique L de taille $\ell$. Le vocabulaire V du texte T a une taille v inférieure à $\ell$. v peut être déterminée en comptant les différents vocables du texte, mais $\ell$ n'est pas accessible directement. La loi de Waring Herdan fournit un moyen pour estimer $\ell$, [11]. L'estimateur qu'elle donne pour $\ell$ est $\hat{\ell} = v\frac{r+s}{s}$ où r et s sont $\geqslant 0$. Cependant cette estimation est anormalement grande si s est proche de zéro.

Notre objectif dans cet article consiste à présenter un autre estimateur de $\ell$ n'ayant pas l'inconvénient sus-cité et par conséquent ayant de meilleures qualités.

## 1. POSITION DU PROBLÈME

Un vocable u du lexique L appartient au vocabulaire V avec une probabilité $(1 - p_u)$ où $p_u \in ]0,1[$. Nous supposons que $p_u$ ne dépend pas de u; $p_u = p \ \forall u \in L$. Cette hypothèse n'est pas réaliste, mais elle est acceptable en moyenne.

Divisons le texte T en deux sous texte $T_1$ et $T_2$, tels que $T \subseteq T_1 \cup T_2$. $T_1$ et $T_2$ peuvent être disjoints ou non. Aux textes $T_1$ et $T_2$ correspondent les vocabulaires $V_1$ et $V_2$ respectivement. Le lexique L est partagé en quatre parties disjointes et on peut l'écrire sous la forme (avec les notations usuelles de la théorie des ensembles[1])

$$L = (V_1 - V_2) + (V_2 - V_1) + V_1V_2 + C_L(V_1 \cup V_2) \qquad (1)$$

où $C_L$ désigne le complémentaire dans L.

Un vocable $u \in L$, peut être dans $V_1$ avec une probabilité $(1 - p_1)$; $p_1 \in ]0,1[$ et dans $V_2$ avec une probabilité égale à $(1 - p_2)$, $p_2 \in ]0,1[$.

D'où nous déduisons que, si $u \in L$

$$\begin{array}{lll} u \in V_1 - V_2 & \text{avec une probabilité} & (1 - p_1)\, p_2 = r_1 \\ u \in V_2 - V_1 & \text{avec une probabilité} & (1 - p_2)\, p_1 = r_2 \\ u \in V_1 \cap V_2 & \text{avec une probabilité} & (1 - p_1)\,(1 - p_2) = r_3 \\ \text{et} \quad u \in C_L(V_1 \cup V_2) & \text{avec une probabilité} & p_1p_2 = r_4 \end{array} \qquad (2)$$

(1) C'est ainsi que $A - B = A \cap C_L B$, $AB = A \cap B$, $A + B = A \cup B$ si $A \cap B = \varnothing$, etc.



*Mots-clés :* **Corpus, Vocables, Estimateur, Fréquence, expérience aléatoire.**

Chaque vocable de L est considéré comme une expérience aléatoire, qui consiste à déterminer la région de L à laquelle il appatient. Comme card (L) = $\ell$ on a alors $\ell$ expériences aléatoires qu'on suppose indépendantes en probabilités.

Supposons de plus que card $(V_1 - V_2) = N_1$, card $(V_2 - V_1) = N_2$ card $V_1 \cap V_2 = N_3$ et card $C_L (V_1 \cup V_2) = N_4$.

$(N_1, N_2, N_3, N_4)$ est une v·a suivant une loi multinomiale de paramètres $(\ell ; (1 - p_1) p_2, (1 - p_2) p_1, (1 - p_1)(1 - p_2), p_1 p_2]$, avec $N_1 + N_2 + N_3 + N_4 = \ell$ autrement dit $(N_1, N_2, N_3, N_4)$ est un point aléatoire de $R^4$ qui se déplace dans un sous espace de $R^4$, à 3 dimensions.

Mais $\ell$ est inconnu, seule une réalisation de $(N_1, N_2, N_3)$ est connue.

## 2. LOI DE PROBABILITÉ DE $(N_1, N_2, N_3 / N_1 + N_2 + N_3 = n)$

La loi conditionnelle de $(N_1, N_2, N_3 / N_1 + N_2 + N_3 = n)$ est (si $n_4 = \ell - n$) :

$$P[N_1 = n_1, N_2 = n_2, N_3 = n_3 / N_1 + N_2 + N_3 = n]$$

$$= \frac{P[N_1 = n_1, N_2 = n_2, N_3 = n_3, N_1 + N_2 + N_3 = n]}{P[N_1 + N_2 + N_3 = n]}$$

$$= \frac{\dfrac{\ell !}{n_1 ! \, n_2 ! \, n_3 ! \, (\ell - n_1 - n_2 - n_3)!}}{\dfrac{\ell !}{n! \, (\ell - n)!}} \cdot \frac{r_1^{n1} \, r_2^{n2} \, r_3^{n-n1-n2} \, r_4^{n4}}{(r_1 + r_2 + r_3)^n \, r_4^{n4}}$$

$$= \frac{n!}{n_1 ! \, n_2 ! \, n_3 !} \left( \frac{r_1}{1 - r_4} \right)^{n1} \left( \frac{r_2}{1 - r_4} \right)^{n2} \left( \frac{r_3}{1 - r_4} \right)^{n-n1-n2} \qquad (3)$$

avec $n_i \geqslant 0$, $\sum_{i=1}^{3} n_i = n$, $\sum_{i=1}^{3} r_i / (1 - r_4) = 1$.

Donc $(N_1, N_2, N_3 / N_1 + N_2 + N_3 = n)$ est mult.

$$[n ; r_1 / (1 - r_4), r_2 / (1 - r_4), r_3 / (1 - r_4)]$$

## 3. CAS PARTICULIER OÙ $p_1 = p_2 = p$

Dans le cas où les deux sous textes $T_1$ et $T_2$ sont choisis au hasard dans T, on peut supposer que $p_1 = p_2 = p$ c'est-à-dire un vocable de L aura la même probabilité d'être utilisé dans $T_1$ que dans $T_2$. Ce qui fait qu'en remplaçant compte tenu de (2) les $p_i$ par p dans (3) nous obtenons :

$(N_1, N_2, N_3/N_1 + N_2 + N_3 = n)$ est mult.

$$\left[ n; p/(1 + p), p/(1 + p), \frac{1 - p}{1 + p} \right] \qquad (4)$$

De toute façon on peut soumettre au test l'hypothèse nulle $H_0 : p_1 = p_2 = p$ contre l'hypothèse alternative $H_1 : p_1 \neq p_2$.

En effet la loi de probabilité de $N_1$ sachant que $N_1 + N_2 = n$ est binomiale de paramètres

$$\left( n, \frac{p_2(1 - p_1)}{1 - p_1p_2 - (1 - p_1)(1 - p_2)} \right) = (n_1, q).$$

Si l'hypothèse nulle $H_0$ est vraie c'est-à-dire si $p_1 = p_2 \Rightarrow q = 1/2$ et si l'hypothèse alternative est vraie $\Rightarrow q \neq 1/2 : 0 \leqslant q \leqslant 1$.

D'ailleurs nous pouvons vérifier que si $p_2 > p_1 \Rightarrow q > 1/2$ et que si $p_2 < p_1 \Rightarrow q < 1/2$.

La région critique W la meilleure au sens de Neyman pour tester $H_0$ contre $H_1$ est obtenue en écrivant :

$\frac{L_0}{L_1} < c$ dans W, où c est une constante et $L_0$ et $L_1$ sont les fonctions de vraisemblance sous $H_0$ et $H_1$ respectivement :

$$\frac{\left(\frac{n}{k}\right)\left(\frac{1}{2}\right)^n}{\left(\frac{n}{k}\right) q^k (1 - q)^{n-k}} < c \text{ (2) dans W} \qquad (5)$$

Ce qui donne $k < c_1$ si $q < 1/2$ et $k > c_2$ si $q > 1/2$ et la région critique W est définie par :

$[W = A = \{k; c_2 < k < c_1\}$ où A est la région d'acceptation [6]. C'est un test symétrique bilatéral.

## 4. PROPOSITION

$N_3/N_1 + N_2 + N_3 = n$ est un résumé exhaustif au sujet de p.

**Démonstration :**

La fonction de vraisemblance L est :

(2) Avec $\left(\frac{n}{k}\right) = n!/(k!(n - k)!)$.

$$L = \binom{n}{n_1\ n_2\ n_3} \left(\frac{p}{1+p}\right)^{n_1+n_2} \left(\frac{1-p}{1+p}\right)^{n_3} (3)$$

$$= \binom{n}{n_3} \left(\frac{2p}{1+p}\right)^{n_1+n_2} \left(\frac{1-p}{1+p}\right)^{n_3} \binom{n_1+n_2}{n_1} \left(\frac{1}{2}\right)^{n_1+n_2}$$

$$= f(n_3, p) \cdot \varphi(n_1, n_2)$$

$$= P[N_3 = n_3] \times \varphi(n_1 \cdot n_2) \qquad (6)$$

avec $\varphi$ une fonction indépendante de p *c.q.f.d*

## 5. ESTIMATEUR DE MAXIMUM DE VRAISEMBLANCE DE p

Nous pouvons écrire L sous la forme :

$$L = \binom{n}{m_{11}\ m_{12}\ m_{21}} \prod_{i=1}^{2} [(1-p)^i\, p^{2-i}]^{\sum_{j=1}^{\binom{2}{2-i}} m_{ij}} \times \frac{1}{(1-p^2)^n} \qquad (7)$$

en posant $m_{11} = n_1$, $m_{12} = n_2$ et $m_{21} = n_3$ et avec $\binom{2}{2-i} = \frac{2!}{(2-i)!\,i!}$ ($= 2$ si $i = 1$; $= 1$ si $i = 2$). Cette formulation relativement compliquée de L sera utile pour la généralisation donnée au paragraphe suivant :

$$\log L = c - n \log(1-p^2) + \sum_{i=1}^{2} \left( \sum_{j=1}^{\binom{2}{2-i}} m_{ij} \right) [i \log(1-p) + (2-i) \log p]$$

$$\frac{\partial \log L}{\partial p} = \frac{2pn}{1-p^2} + \sum_{i=1}^{2} \left( \sum_{j=1}^{\binom{2}{2-i}} m_{ij} \right) \left[ -\frac{i}{1-p} + \frac{2-i}{p} \right] = 0 \Rightarrow$$

$$2np^2 + \sum_{i=1}^{2} \sum_{j=1}^{\binom{2}{2-i}} m_{ij} [2 - 2p^2 - i(1+p)] = 0$$

$$2np^2 + 2(1-p^2)n - (1+p) \sum_{i=1}^{2} \left( \sum_{j=1}^{\binom{2}{2-i}} m_{ij} \right) i = 0$$

$$\}\leftrightharpoons - (\{ + \coprod) \sum_{i=1}^{2} \left( \sum_{j=1}^{\binom{2}{2-i}} m_{ij} \right) i = 0 \qquad (8)$$

(3) Avec la notation $\binom{n}{n_1\ n_2\ n_3} = \frac{n!}{n_1!\, n_2!\, n_3!}$.

Ce qui donne :

$$2n - (1 + p)[m_{11} + m_{12} + 2m_{21}] = 0$$

et

$$\hat{p} = \frac{n - m_{21}}{n + m_{21}} = \frac{n - n_3}{n + n_3} \tag{9}$$

***Remarque***

Nous constatons que l'estimateur $\hat{p}$ de maximum de vraisemblance de p est une fonction de $n_3$ qui est un résumé exhausif de p.

## 6. GÉNÉRALISATION DU SCHÉMA DU § 1

Supposons qu'on divise le texte T par k sous textes $T_i$; i = 1,...,k; les $T_i$ peuvent être disjoints ou non, $T \subseteq \bigcup_{i=1}^{k} T_i$. A tout sous textes $T_i$ correspond un vocabulaire $V_i$, et le lexique L se trouve être partagé en $2^k$ parties disjointes $\bigcap_{i=1}^{k} V_i, \ldots, [\bigcup_{i=1}^{k} V_i$.

Un vocable u peut être dans $V_i$ avec une probabilité $(1 - p_i)$, où $p_i \in ]0,1[$ i = 1, 2,...,k ce qui nous permet d'écrire, si $u \in L$

$u \in \bigcap_{i=1}^{k} V_i$ avec la probabilité $\prod_{i=1}^{k} (1 - p_i) = r_1$

$u \in \bigcap_{i=1}^{k-1} V_i - V_k$ avec la probabilité $\left[\prod_{i=1}^{k-1} (1 - p_i)\right] p_k = r_2$

.....

$u \in [_L \bigcup_{i=1}^{k} V_i$ avec la probabilité $\prod_{i=1}^{k} p_i = r_{2k}$

Chaque vocable de L est considéré comme une expérience aléatoire, qui consiste à déterminer la région de L à laquelle il appartient. On a alors $\ell$ expériences aléatoires qu'on suppose indépendantes en probabilité.

Supposons de plus que

$$\text{Card}\,(\cap V_i) = N_1,\ \text{Card}\left(\bigcap_{i=1}^{k-1} V_i - V_k\right) = N_2, \ldots,\ \text{Card}\left([_L \bigcup_{i=1}^{k} V_i\right) = N_{2k}$$

$(N_1, N_2, \ldots, N_{2k})$ est une variable aléatoire qui suit une loi multinomiale de paramètres $(\ell; r_1, \ldots, r_{2k})$.

Mais $\ell$ n'est pas connue, seule une réalisation de $N_1, N_2, \ldots, N_{2k-1}$, est connue. D'où, comme on a fait au § 2, on cherche la loi de $(N_1, N_2, \ldots, N_{2k-1}/N_1 + N_2 + \ldots + N_{2k-1} = n)$ et on trouve que cette variable aléatoire suit une loi multinomiale de paramètres :

$$\left( n; \frac{r_1}{1 - r_{2k}}, \ldots, \frac{r_{2k-1}}{1 - r_{2k}} \right),$$

qu'on peut écrire avec les notations de la formule (7) du § 5, et en supposant que $p_i = p \ \forall_i = 1, 2, \ldots, k$ sous la forme L où L ici est la fonction de vraisemblance :

$$L = \binom{n}{m_{11}\ m_{12} \ldots} \prod_{i=1}^{k} [(1-p)^i p^{k-i}]^{\sum_{j=1}^{\binom{k}{k-i}} m_{ij}} \frac{1}{(1-p^k)^n} \qquad (10)$$

## 7. ESTIMATION DU MAXIMUM DE VRAISEMBLANCE DE p

$$\log L = C + \sum_{i=1}^{k} \left( \sum_{j=1}^{\binom{k}{k-i}} m_{ij} \right) [i \log(1-p) + (k-i) \log p] - n \log(1-p^k) \qquad (11)$$

$$\frac{\partial \log L}{\partial p} = \sum_{i=1}^{k} \left( \sum_{j=1}^{\binom{k}{k-i}} m_{ij} \right) \left[ -\frac{i}{1-p} + \frac{(k-i)}{p} \right] + \frac{k n p^{k-1}}{1-p^k} = 0$$

C'est l'équation du maximum de vraisemblance qui se réduit à :

$$n k - (1 + p + \ldots + p^{k-1}) \sum_{i=1}^{k} \left( \sum_{j=1}^{\binom{k}{k-i}} m_{ij} \right) i = 0 \qquad (12)$$

Si $k = 2$ on retrouve (8)
et si $k = 3$, (12) se réduit à :

$$3n - (1 + p + p^2) \sum_{i=1}^{3} \left( \sum_{j=1}^{\binom{3}{3-i}} m_{ij} \right) i = 0$$

qui admet (d'après le résultat plus général donné au § 8) une solution p dans l'intervalle ]0,1[ :

$$p = \frac{-1 + \sqrt{1 + 4\left( \frac{3n}{\sum_i \left( \sum_j m_{ij} \right) i} - 1 \right)}}{2} \qquad (13)$$

## 8. PROPOSITION

L'équation du maximum de vraisemblance (12) a une seule solution p dans l'intervalle ]0,1[

**Démonstration**

Multiplions les 2 membres de (12) par (1 − p) nous obtenons :

$$nk(1-p) - (1-p^k) \sum_{i=1}^{k} \left( \sum_{j=1}^{\binom{k}{k-i}} m_{ij} \right) i = 0 \qquad (14)$$

Mais en tenant compte des inégalités :

$$n < \sum_{i=1}^{k} \left( \sum_{j=1}^{\binom{k}{k-i}} m_{ij} \right) i < nk \qquad (15)$$

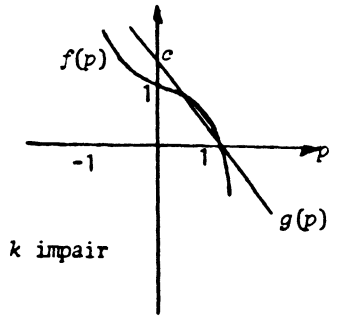
Nous obtenons :

$$c(1-p) - (1-p^k) = 0 \text{ avec } c = \frac{nk}{\sum (\sum m_{ij})\, i} \text{ et } 1 < c < k \qquad (16)$$

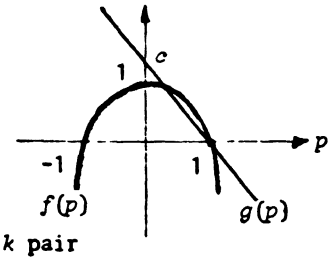
Considérons la fonction $f(p) = 1 - p^k$.

Si k est pair f(p) est symétrique par rapport à l'axe des y, de plus $f'(p) = -k p^{k-1}$, la dérivée est négative quand $p \in [0,1]$, $f''(p) = -k(k-1) p^{k-2}$ la dérivée seconde est aussi négative dans [0,1], et l'on conclut que f(p) est une fonction décroissante dans [0,1] et sa dérivée est une fonction décroissante dans le même intervalle.

Soit g(p) = c(1 − p).




k impair




k pair

C'est une droite dont l'ordonnée à l'origine est $c > 1$ et qui passe par le point (1,0), et sa pente est $-c$.

De même f(p) passe par le point (1,0), et sa pente en ce point est $-k$, plus petite que $-c$, donc pour $p < 1$, p proche de 1, $f(p) > g(p)$ et on a $f(0) = 1 < c$.

On conclut que f(p) et g(p) ont un point commun (1,0), et ne peuvent avoir qu'un seul point commun dans l'intervalle [0,1].

On obtient un résultat similaire pour k impair (voir figure)

## 9. PROPRIÉTÉS DE L'ESTIMATEUR $\hat{p}$ DE p DONNÉ PAR (9)

a) $\hat{p}$ qui est une fonction de $n_3$ qui est un résumé exhaustif pour p, est lui-même un estimateur exhaustif pour p.

b) Moments de $\hat{p}$
$\hat{p}$ admet au voisinage de $E(n_3)$ un développement limité de la forme :

$$\hat{p} = \frac{n - E(n_3)}{n + E(n_3)} + \frac{(n_3 - E(n_3))}{1!}\left(\frac{-2n}{(n + E(n_3))^2}\right) + \frac{(n_3 - E(n_3))^2}{2!}\frac{4n}{(n + E(n_3))^3} + \ldots \quad (17)$$

$$= p - (n_3 - E(n_3))\frac{2n}{(n + E(n_3))^2} + (n_3 - E(n_3))^2\frac{2n}{(n + E(n_3))^3} + \ldots \quad (18)$$

D'où l'on réduit, puisque $n_3$ suit une loi binomiale de paramètres n et $(1 - p)/(1 + p)$

$$E(\hat{p}) \simeq p$$

et

$$V(\hat{p}) \simeq V(n_3)\frac{4n^2}{(n + E(n_3))^4} \quad (19)$$

$$= \frac{1}{n}\frac{(1 - p)\,p(1 + p)^2}{2} \quad (20)$$

$\hat{p}$ est asymptotiquement sans biais; sa variance tend vers zéro lorsque $n \to \infty$, donc il est convergent.

## 10. APPLICATIONS

### a. Premier exemple pratique

Dans [10], Bernard Bénoit décrit l'évolution du lexique du PCF de (1932-1945). Le corpus étudié est constitué d'un échantillon d'éditoriaux de la

revue du comité central. Ce corpus est composé de trois sous corpus correspondant à trois moments de l'histoire du PCF. Chacun des sous corpus est considéré comme un ensemble synchronique de textes groupés autour d'un événement qui le date et lui donne unité et homogénéité. Ces sous-corpus $C_1$, $C_2$ et $C_3$, regroupent respectivement des textes écrits autour des congrès de Paris de 1932, de Villeurbanne de 1936 et de Paris de 1945.

Donc le corpus C étudié est formé par la juxtaposition des sous corpus $C_1$, $C_2$ et $C_3$ de vocabulaires respectifs $V_1$, $V_2$ et $V_3$.

C comporte 60 534 occurences engendrées par 6 846 formes qui sont les éléments du vocabulaire V de C.

Nous pouvons écrire V comme une somme de sept sous ensembles disjoints deux à deux :

$$\underset{6\,846}{V} = \underset{1\,588}{(V_1 - V_2 \cup V_3)} + \underset{1\,203}{(V_2 - V_1 \cup V_3)} + \underset{1\,715}{(V_3 - V_1 \cup V_2)} + \underset{358}{(V_1V_2 - V_3)} + \underset{513}{(V_1V_3 - V_2)} + \underset{492}{(V_2V_3 - V_1)} + \underset{977}{V_1 V_2 V_3} \qquad (21)$$

où le chiffre qui est au-dessous du symbole de l'ensemble représente le cardinal de cet ensemble.

Nous allons donner une estimation du lexique $\ell(C_1, C_2)$ où sont puisés les deux sous corpus $C_1$ et $C_2$. Nous avons les réalisations suivantes des variables $N_i$ : $n_1 = 1\,588 + 513 = 2\,101$, $n_2 = 1\,203 + 492 = 1\,695$, $n_3 = 358 + 977 = 1\,335$ et $n = 5\,131$

$$\hat{p} = \frac{n - n_3}{n + n_3} \qquad \text{d'après (9)}$$

$$= \frac{5\,131 - 1\,335}{5\,131 + 1\,335} \qquad (22)$$

$$= 0.58707$$

Une estimation de $\ell(C_1, C_2)$ est

$$\hat{\ell}(C_1, C_2) = \frac{n}{1 - \hat{p}^2} = \frac{5\,131}{0.65534} = 7\,829 \text{ formes} \qquad (23)$$

Nous faisons les mêmes calculs pour les sous corpus ($C_1$ et $C_3$) d'une part et ($C_2$ et $C_3$) d'autre part, les résultats sont donnés dans les tableaux suivants :

| | 1935 $C_2$ | 1945 $C_3$ |
|---|---|---|
| 1932 $C_1$ | $n_1 = 2\,101$<br>$n_2 = 1\,695$<br>$n_3 = 1\,335$<br>$n = 5\,131$ | $n_1 = 1\,946$<br>$n_2 = 2\,207$<br>$n_3 = 1\,490$<br>$n = 5\,643$ |
| 1935 $C_2$ | | $n_1 = 1\,561$<br>$n_2 = 2\,228$<br>$n_3 = 1\,469$<br>$n = 5\,258$ |

(24)

| | | |
|---|---|---|
| 1932 $C_1$ | $\hat{p} = 0.587$<br>$\hat{\ell} = 7\,829$ | $\hat{p} = 0.582$<br>$\hat{\ell} = 8\,336$ |
| 1935 $C_2$ | | $\hat{p} = 0.562$<br>$\hat{\ell} = 7\,876$ |

Nous avons fait les mêmes calculs en considérant $C_1$ et $C_2$ comme formant un seul texte d'un côté, et le corpus $C_3$ de l'autre côté ce qui à donné :

$$n_1 = 3\,149, \quad n_2 = 1\,715, \quad n_3 = 1\,982 \quad \text{et} \quad n = 6\,846$$

$$\hat{p} = \frac{n - n_3}{n + n_3} = \frac{6\,846 - 1\,982}{6\,846 + 1\,982} = 0.5509$$

et

$$\hat{\ell} = \frac{n}{1 - \hat{p}^2} = \frac{6\,846}{0.6965} = 9\,829 \text{ formes} \qquad (25)$$

C'est une estimation du lexique du discours du PCF pour la période 1932-1945.

De même nous présentons les estimations du même lexique en considérant successivement : b) $C_2$ et $C_3$ réunis contre $C_1$ et c) $C_1$ et $C_3$ réunis contre $C_2$

| | $n_1$ | $n_2$ | $n_3$ | $n$ | $\hat{p}$ | $\hat{\ell}$ |
|---|---|---|---|---|---|---|
| a) | 3 149 | 1 715 | 1 982 | 6 846 | 0.5509 | 9 829 |
| b) | 3 410 | 1 588 | 1 848 | 6 846 | 0.5748 | 10 225 |
| c) | 3 816 | 1 203 | 1 827 | 6 846 | 0.5789 | 10 297 |

Une autre estimation de ce même lexique est fournie par l'application de la formule (13). Nous obtenons en effet les valeurs suivantes des $m_{ij}$ :

| i | $m_{ij}$ | | $i\,m_{ij}$ |
|---|---|---|---|
| 3 | $m_{31}$ | 977 | 2 931 |
| 2 | $m_{21}$ | 358 | 716 |
| | $m_{22}$ | 513 | 1 026 |
| | $m_{23}$ | 492 | 984 |
| 1 | $m_{11}$ | 1 588 | 1 588 |
| | $m_{12}$ | 1 203 | 1 203 |
| | $m_{13}$ | 1 715 | 1 715 |
| | | 6 846 | 10 163 |

$$= \sum_{i=1}^{3} \left( \sum_{j=1}^{\binom{3}{3-i}} m_{ij} \right) i$$

Donc $\hat{p}$ est la solution comprise entre 0 et 1 de l'équation du second degré

$$1{,}02086 - p - p^2 = 0 \qquad (26)$$

Ce qui donne :

$$\hat{p} = 0.6273$$

et

$$\hat{\ell} = \frac{n}{1 - \hat{p}^3} = \frac{6\,846}{0.24687}$$

$$= 9\,090 \text{ formes} \qquad (27)$$

Résultat en concordance avec l'estimation (25).

**b) 2e exemple pratique**

Une autre expérience, faite sur les deux tragédies de Corneille, Sertorius et Sophonisbe, éclaire bien le fonctionnement de la méthode d'estimation de L. Le dépouillement des deux pièces donne les résultats suivant :

| | Sertious (1662) | Sophonisbe (1663) | Ensemble |
|---|---|---|---|
| Nombre d'occurrence N | 17 708 | 16 880 | 34 588 |
| Card (V) | 1 552 | 1 430 | 1 944 |
| Estimation du $\ell$ lexique. (Modèle de Waring Herdan) | 3 200 | 2 800 | 3 300 |

D'autre part les réalisations de $N_1$, $N_2$ et $N_3$ sont :

$n_1 = 514$, $n_2 = 392$, $n_3 = 1\,038$ et $n = 1\,944$
$= \text{Card}(V_1 - V_2)$, $= \text{Card}(V_2 - V_1)$, $= \text{Card}(V_1 V_2)$, $= \text{Card}(V_1 V_2)$

Une estimation de p est :

$$\hat{p} = \frac{n - n_3}{n + n_3} = \frac{1\,944 - 1\,038}{1\,944 + 1\,038} = 0.3038$$

et $\ell$ est estimé par :

$$\hat{\ell} = \frac{n}{1 - \hat{p}^2} = \frac{1\,944}{0.9077} = 2\,142 \text{ formes} \qquad (28)$$

## 11. CONCLUSION

1) Nous constatons sur ce second exemple que l'estimation de $\ell$ fournie par notre modèle est beaucoup plus petite que celle fournie par le modèle de Waring Herdan et calculée par Monsieur Ch. Muller; la différence, de l'ordre de 1 000 vocables, est due à deux raisons principales :

a) Pour estimer les deux paramètres de la loi de Herdan Waring M. Muller utilise deux équations : la 1ère exprime l'égalité entre la moyenne empirique et la moyenne théorique de la fréquence des vocables et la 2e exprime que les fréquences empiriques et théoriques des Hapax (4) sont égales.

Or la moyenne théorique de la fréquence des vocables, suivant le modèle de Waring Herdan n'existe pas toujours, son existence dépend des valeurs de l'un des paramètres du modèle. Voir [11].

b) Comme nous l'avons signalé au début de cet article l'estimation de $\ell$ comporte un dénominateur égal à s, et si s est proche de zéro cette estimation est surévaluée.

2) On peut remarquer aussi que ce modèle est libre des contraintes imposées par le modèle de Waring Herdan, à savoir que $\text{Card}(L_0) > \text{Card}(V_1)$ c'est-à-dire la fréquence des vocables non utilisés dans le texte est supérieure à celle des Hapax (4), hypothèse non vraie dans beaucoup de cas.

## BIBLIOGRAPHIE


[1] G.TH. GUILBAUD. — *Zipf et les fréquences,* Mots n° 1, oct. 1980.

[2] P. LAFON. — *Sur la variabilité de la fréquence des formes dans un corpus.*

[3] M. TOURNIER. — *D'où viennent les fréquences de vocabulaire,* Mots n° 1, oct. 1980.

[4] P. LAFON. — *Statistiques des localisations des formes d'un texte,* Mots n° 2, mars 1981.


(4) Un Hapax est un vocable qui figure une seule fois dans le texte.

[5] CH. MULLER. — *Observation, prévision et modèles statistiques.* Etudes de statistique linguistique III Université de Metz.

[6] CH. MULLER. — *Principes et méthodes de statistique lexicale,* Paris — Hachette — Université 1977.

[7] A. RENYI. — *Calcul des probabilités,* Dunod 1966.

[8] A. ABI FARAH. — Le vocabulaire du livre arabe *DAMA WA IBTIS-SAMA* de Gibran Khalil. *Etudes statistiques,* Fac. des Sciences — Université Libanaise 1983.

[9] A. ABI FARAH. — Un test pour le contrôle de la qualité du travail dans un recensement. *Revue de statistique appliquée,* 1974, vol. XXII, n° 1.

[10] Bénoit ROBERT. — *Le lexique du PCF,* 1932-1946. Mots n° 3, 1981.

[11] A. ABI FARAH. — *Le fondement théorique de la loi de Waring Herdan.*