UNE OBSERVATION A PROPOS DE LANGAGE ET CLASSIFICATIONS

Ugo Berni Canani

Dans quelques-uns des articles contenus dans le volume *Langue française et linguistique quantitative*, M. Charles Muller parle de l'application de la série de Waring

$$\frac{1}{x-a} = \frac{1}{x} + \frac{a}{x(x+1)} + \frac{a(a+1)}{x(x+1)(x+2)} + \ldots$$

à la distribution des fréquences des mots dans un texte, expliquant le modèle et quelques variantes et en donnant plusieurs exemples. On considère les termes de la série

$$1 = \frac{x-a}{x} + \frac{(x-a)a}{x(x+1)} + \frac{(x-a)a(a+1)}{x(x+1)(x+2)} + \ldots$$

comme représentant la probabilité pour un mot d'avoir dans un texte la fréquence 1,2,3 etc... Sur la base des données

N = nombre des occurrences des mots du texte

V = nombre des mots utilisés dans le texte

V_1 = nombre des mots qui ont la fréquence 1 dans le texte, on calcule les valeurs de x et de a et on construit la série de valeurs théoriques pour le nombre de mots qui ont la fréquence 2,3,4 etc. J'ai été frappé par la précision des approximations obte-

nues par M. Muller et, convaincu comme je le suis de l'existence de fortes analogies entre les problèmes touchant à la structure du lexique et les problèmes de classification, j'ai pensé à la possibilité d'appliquer la même série aux catalogues des bibliothèques. Il s'agit de considérer à la place de la distribution des mots dans un texte celle des fiches bibliographiques dans les différentes rubriques d'un catalogue. J'ai fait un essai sur des documents juridiques (résumés de sentences civiles) classifiés par le Bureau du Fichier de la Cour de Cassation Italienne, et j'ai reporté les résultats sur trois tableaux correspondant respectivement aux documents de 1977 (7839), de 1977 et 1978 (16.568) et de 1977, 1978 et 1979 (26.121). Dans les tableaux la première colonne indique la fréquence, jusqu'à un certain seuil. Les fréquences cumulées dépassant ce seuil sont représentées par le signe +. La deuxième colonne indique le nombre des rubriques du catalogue (schéma de classification civile de la Cour de Cassation) utilisées dans la classification des documents respectivement 1, 2, 3 etc. fois. La troisième colonne fournit l'estimation des données de la deuxième colonne obtenue avec la série de Waring. La quatrième une estimation obtenue avec la variante utilisée initialement par G. Herdan et qui donne apparemment une meilleure approximation.

Comme vous pouvez le constater, les approximations sont plutôt satisfaisantes.

1977: n. doc. = 7839

fréq.			
1	1463	1463	1463
2	542	576	531
3	281	282	261
4	154	158	150
5	115	97	96
6	58	63	65

7	47	44	47
8	43	31	35
9	33	23	27
10	21	17	21
+	86	89	147

V = 2843 N = 7743
a = 2,08358 x = 4,29248

1977 + 1978: n. doc. = 16568

fréq.	obs.	est. 1	est. 2
1	1696	1696	1696
2	757	815	783
3	433	456	435
4	295	281	270
5	197	186	181
6	117	130	128
7	97	94	95
8	84	71	72
9	63	54	56
10	39	43	45
11	44	34	37
12	29	28	30
13	22	23	26
14	26	19	22
15	31	16	18
16	13	14	16
17	19	12	14
18	14	10	12
19	15	9	11
20	11	8	10
+	86	89	131

V = 4088 N = 16253
a = 2,681079 x = 4,58245

1977 + 1978 + 1979: n. doc. = 26121

fréq.	obs.	est. 1	est. 2
1	1843	1843	1843
2	851	947	925
3	520	558	542
4	327	359	349
5	251	246	241
6	173	177	174
7	140	132	131
8	121	101	101
9	90	80	80
10	81	64	65
11	68	52	53
12	43	43	46
13	47	36	38
14	31	31	32
15	35	26	28
16	33	23	24
17	32	20	21
18	22	17	19
19	18	15	16
20	12	13	15
+	218	173	213

V = 4956 N = 25716
a = 2,830414 x = 4,506114