LES CAHIERS DE L'ANALYSE DES DONNÉES

A. M. ALKAYAR

Classification d'un ensemble varié de textes français d'après les occurrences de mots pleins

Les cahiers de l'analyse des données, tome 18, n° 2 (1993), p. 239-244

<http://www.numdam.org/item?id=CAD_1993__18_2_239_0>

Les Cahiers de l'Analyse des Données
Vol XVIII 1993 - n°2 pp 239 244

CLASSIFICATION D'UN ENSEMBLE VARIÉ DE TEXTES FRANÇAIS D'APRÈS LES OCCURRENCES DE MOTS PLEINS

[TEXT. VAR. FR.]

*A. M ALKAYAR**

On a déjà analysé des tableaux de correspondance croisant un ensemble de textes avec un lexique de mots outil ou de mots pleins. Ainsi la totalité du texte grec du Nouveau Testament a été considérée avec quelques livres de l'Ancien Testament, des textes philosophiques en langue classique et des fragments très divers d'époque hellénistique. On a constitué une anthologie latine allant de la prose et de la poésie classique jusqu'à la littérature chrétienne la plus récente; et, pour la seule période du Siècle d'Or, un florilège espagnol où figurent la plupart des genres. Le corpus de six années des *Cahiers de l'Analyse des Données*, a servi de matière à une expérience en analyse documentaire. À l'horizon, on imagine, comme dans un mirage, que l'ensemble des genres et des thèmes, présentés dans quelques milliards de caractères, puisse être compris dans un ordre commun...

La présente note rend compte d'une expérience portant seulement sur un corpus de quelque 22000 occurrences, en 120000 caractères; mais avec des textes fort variés. Il apparaîtra que cette diversité, à l'échelle de la présente expérience, facilite grandement la discrimination des genres.

1 Le corpus

Il comprend trois parties principales et des compléments divers. Nous en détaillons la composition, en énumérant les sigles des fragments.

Les réponses de huit personnalités à un questionnaire ont fait l'objet de l'étude [PHARM. QUEST.]: avec, en plus des huit auteurs, les questions elles-mêmes et la synthèse des réponses, on a un texte de quelque 40000 caractères, partagé en dix items ou fragments (cf. [PHARM. QUEST.], §1.4):

{Qest, Résu, Rcrd, dRPR, Cntr, dUDC, CNI, FN, dPS, PRép }.

(*) Étudiant en Doctorat à l'Université Pierre et Marie Curie, Paris VI; et جامعة بغداد.

Du "Traité de la Vraie Dévotion à la Sainte Vierge", composé au début du XVIII-ème siècle par Saint Louis-Marie Grignon de Monfort, on a saisi l'introduction et le premier chapitre, soit 44000 caractères, le tout découpé en six fragments, en respectant le plan de l'auteur: {Gr†1, Gr†2, Gr†3, Gr†4, Gr†5, Gr†6}.

Le Traité de Maastricht, tel qu'il a été soumis à ratification par référendum, mais sans ses annexes, a fourni 16000 caractères de prose diplomatique, où sont distingués un préambule et deux "Titres": {Maa1, Maa2, Maa3}.

En 1807, "Sa Majesté l'Empereur de toutes les Russies et sa Majesté l'Empereur des Français, Roi d'Italie, Protecteur de la Confédération du Rhin, étant animés d'un égal désir de mettre fin aux calamités de la guerre" avaient, à Tilsit, ratifié un traité, dont le texte nous a également fourni 16000 caractères en trois fragments: {Til1, Til2, Til3}, dont le dernier contient des clauses secrètes. En effet, à Tilsit, encore que sur d'autres bases qu'à Maastricht, Alexandre et Napoléon, entendaient pourvoir à l'ordre européen. Mais l'Angleterre renaclant au projet, il fut signé, en même temps et même lieu, un "Traité d'alliance offensive et défensive entre la Russie et la France": {Til4}.

Sept ans plus tard, "Leurs Majestés l'Empereur d'Autriche, le Roi de Prusse et l'Empereur de Russie par suite des très grands événements qui ont signalé en Europe les cours des trois dernières années..., ayant acquis la conviction intime que la marche précédemment adoptée par les puissances dans leurs rapports mutuels doit être absolument changée...", Alexandre prend l'initiative d'un "Projet d'Acte de Sainte Alliance": {SteA}. Est saisi le texte même du projet de 1814, et non le document finalement signé à Paris, par Leurs Majestés, le 26 Septembre 1815.

À cela on a adjoint, de la prose baroque de notre cher VOITURE, deux pages galantes: "Métamorphose de Julie en diamant" et "Métamorphose de Léonide en perle": {Voit}.

Soit, en tout, 25 fragments.

2 Choix du lexique des mots pleins

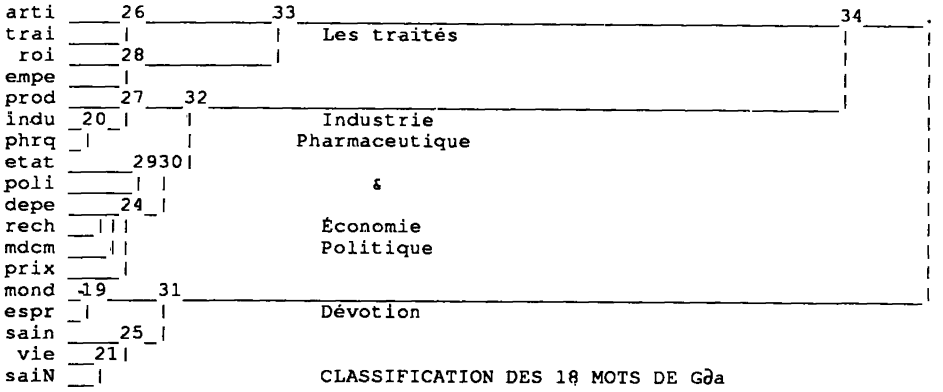
Par le programme 'trigalac', le corpus, considéré comme un texte unique, a été rangé en une liste alphabétique donnant, après chaque occurrence, son adresse par numéro de fragment et de paragraphe. Le programme 'qamus' a produit, d'après la liste des occurrences, une liste alphabétique des formes; où chacune de celles-ci figure une fois, sans indication d'adresse, mais avec le nombre de ses occurrences. Cette liste ayant été triée par fréquences, il a été facile de constituer une liste des formes de mots pleins dont la fréquence est ≥ 10. Cette liste de 92 formes, rangées alphabétiquement, constitue un premier lexique, 'Gə'; que par 'tridic', on a croisé avec l'ensemble J des 25 fragments; k(i, j) étant le nombre des occurrences de la forme *i* dans le fragment *j*.

article	mdcm medicament	recherche
depenses	Mdcm medicaments	remboursement
dvpm developpement	membres	roi
empereur	mere	saint
esprit	monde	saiN sainte
etat	ordre	sante/
force	pays	traite/
Homm hommes	phrq pharmaceutique	vie
industrie	politique	vierge
majeste/	prix	volumes
marche/	produits	

Ce tableau n'a toutefois pas été analysé directement; mais, par 'ranger' on y a cherché les mots qui, dans quatre fragments au moins, sont, selon l'ordre des fréquences, parmi les 10 (resp. les 15) premiers de Gə: d'où les sous-lexiques Gəa et Gəf. Ci-dessus, est publié Gəf; et ci-dessous, le tableau de correspondance croisant Gəa avec l'ensemble J des 25 fragments du corpus.

mots de Gəa x chapitres de GrPhD

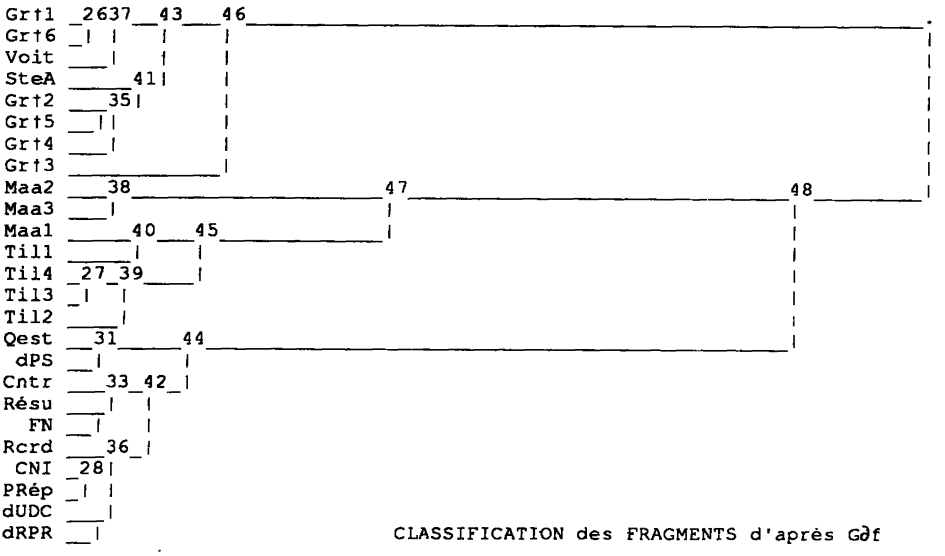
25	Qest	Résu	Rcrd	dRPR	Cntr	dUDC	CNI	FN	dPS	PRép	
Grt1 Grt2 Grt3 Grt4	Grt5	Grt6	Til1	Til2	Til3	Til4	Maa1	Maa2	Maa3	Voit	SteA
article	0	0	0	0	0	0	0	0	0	0	
0 1 0 1	0	0	3	21	16	11	0	8	21	0	3
depenses	1	1	5	3	0	1	0	3	4	4	
0 0 0 0	0	0	0	0	0	0	0	0	0	0	0
empereur	0	0	0	0	0	0	0	0	0	0	
0 0 0 0	0	0	12	13	10	5	0	0	0	0	2
esprit	0	0	0	0	0	0	1	0	0	0	
5 5 3 7	2	4	0	0	0	0	0	0	0	2	0
etat	0	0	5	2	2	0	0	2	2	1	
0 0 0 0	0	0	2	2	0	0	1	3	0	0	0
industrie	2	1	4	2	2	0	1	2	8	2	
0 0 0 0	1	0	0	0	0	0	0	0	1	0	0
mdcm medicament	1	1	1	3	4	3	1	2	6	4	
0 0 0 0	0	0	0	0	0	0	0	0	0	0	0
monde	0	0	1	0	0	0	0	0	0	1	
6 4 3 4	4	3	0	0	0	1	1	0	0	2	0
phrq pharmaceutique	2	2	2	1	0	0	1	2	8	1	
0 0 0 0	0	0	0	0	0	0	0	0	0	0	0
politique	2	3	3	10	3	2	1	2	1	3	
0 0 0 0	0	0	0	1	0	0	2	2	9	0	0
prix	1	2	9	5	2	4	4	2	4	10	
0 1 0 0	0	0	1	0	0	0	0	0	2	2	0
produits	0	0	6	4	1	0	1	0	10	8	
0 0 0 0	0	0	0	0	0	0	0	0	0	0	0
recherche	1	3	4	4	3	1	2	6	2	4	
0 0 0 0	0	0	0	0	0	0	0	0	1	0	0
roi	0	0	0	0	0	0	0	0	0	0	
2 0 0 2	0	0	5	17	10	2	4	0	0	0	1
saint	0	0	0	0	0	0	0	0	0	0	
9 7 5 22	3	3	0	0	0	0	0	0	0	0	0
saiN sainte	0	0	0	0	0	0	0	0	0	0	
4 6 8 20	9	1	0	0	0	0	0	0	0	0	4
traite/	0	0	1	0	0	0	0	0	0	0	
0 0 0 0	0	0	3	4	7	2	2	8	12	0	0
vie	0	0	0	0	0	0	0	0	0	0	
4 3 3 5	3	2	0	0	1	0	0	0	1	0	2



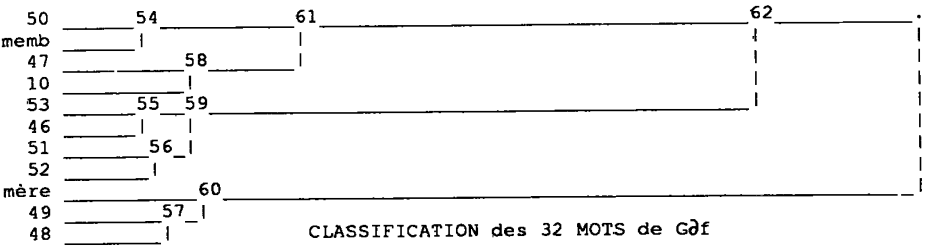
Afin de faciliter la lecture de ce tableau, où les informations afférentes à un mot ne peuvent tenir dans la largeur de la page, on a placé sur une première ligne les 10 fréquences dans les fragments issus du questionnaire; puis décalées, sur une deuxième ligne, les fréquences chez GRIGNON de MONFORT, dans les traités de Tilsit et de Maastricht; et, enfin chez VOITURE et pour la Sainte Alliance.

3 Analyse de correspondance et classification

On a analysé les deux tableaux G∂a × J, 18 × 25, et G∂f × J, 32 × 25; et, après ces analyses, on a effectué une classification ascendante hiérarchique pour chacun des lexiques; et deux classifications pour l'ensemble J des fragments.



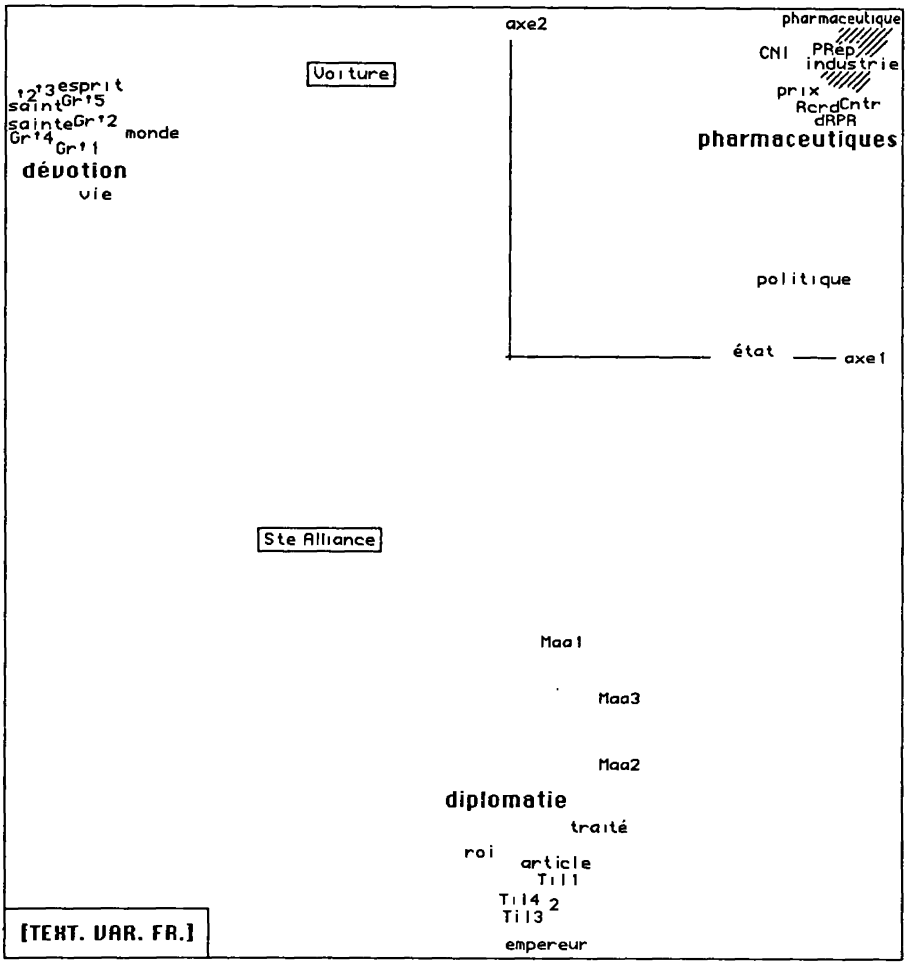
c	Partition de Gəf en 11 classes: mots de la classe numéro c
50	article traité
14	membre
47	empereur roi
10	majesté
53	ordre pays état
46	développement marché politique
51	dépenses santé recherche médicament remboursement prix
52	produits médicaments industrie volume pharmaceutique
15	mère
49	hommes vie force monde esprit
48	saint sainte vierge



Tous les résultats obtenus concordent. Pour faciliter la lecture, nous présentons d'abord, face à face, ceux des classifications.

Le corpus rassemble trois genres: *Dévotion*, *Diplomatie*, *Pharmaceutiques*. Cette structure est bien reconnue. Reste à considérer les détails. Le projet de Sainte Alliance, où (cf. tableau) se rencontrent les mots {article, empereur, roi} mais aussi {sainte, vie}, s'agrège, du fait de ces derniers, plutôt avec la *dévotion* qu'avec la *diplomatie*. Réduite à {esprit, monde, prix}, toute la finesse de VOITURE se place entre *dévotion* et *pharmaceutiques*, mais plus près de la première. Des clauses ou motifs de Maastricht tendent vers les préoccupations *pharmaceutiques*; mais le préambule, Maa1, avec son cortège de chefs d'état, non tous dépourvus de couronne, suivis de leurs ministres, rejoint l'ouverture de Tilsit, Til1. Les dix items *pharmaceutiques* sont étroitement groupés; mais c'est par la correspondance avec le vocabulaire, plus étendu, de Gəf, qu'on en retrouve le mieux la structure, analysée en détail dans [PHARM. QUEST.]; et c'est pourquoi on a choisi de publier l'arbre issu de Gəf × J.

Quant aux mots, le vocabulaire propre à la diplomatie ou celui de la dévotion à la Sainte Vierge sont bien à part; mais il y a une classe où pharmacie et diplomatie se réjoignent en matière de politique et d'économie. La classification de Gəf ne fait que compléter celle de Gəa.



Cumulant environ les 2/3 de l'inertie totale, le plan (1,2) issu de l'une ou l'autre analyse présente l'ensemble de la structure; pour plus de clarté, on publie le graphique, moins dense, issu de l'analyse G∂a × J. On notera que, dans ce plan, SteA est plus proche des traités que de Gr†; mais, dans l'espace, SteA s'agrège à Gr†: une image plane ne peut être, en tout, fidèle à l'espace.

Références bibliographiques

Le présent cahier donne dans [PHARM. QUEST.] une bibliographie relative aux méthodes. Quant aux textes, nous rendrons hommage aux éditeurs de Moscou qui ont publié l'admirable prose française d'Alexandre I-er et de ses ministres ou correspondants.