

J.-P. BENZÉCRI

A. CHABIR

Typologie d'un ensemble de textes arabes d'après les occurrences de formes de mots et de locutions

Les cahiers de l'analyse des données, tome 20, n° 4 (1995),
p. 389-412

http://www.numdam.org/item?id=CAD_1995__20_4_389_0

© Les cahiers de l'analyse des données, Dunod, 1995, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

TYPOLOGIE D'UN ENSEMBLE DE TEXTES ARABES D'APRÈS LES OCCURRENCES DE FORMES DE MOTS ET DE LOCUTIONS

[TEXTES ARABES]

J.-P. BENZÉCRI
A. CHABIR*

1 Composition du corpus

L'article [MOTS ARABES] (in *CAD*, Vol XIX, n°1) rend compte d'une expérience stylométrique portant sur 13 fragments de textes longs, chacun, d'environ 3000 caractères (3k); et tous compris dans le domaine de la philosophie, même s'ils diffèrent quant au niveau de l'exposé.

Le présent travail, porte sur 91 fragments, longs également de quelque 3k, mais d'une plus grande diversité de genre. La cueillette, sans prétendre aucunement embrasser l'infinie diversité de ce qu'on peut lire en arabe, sort de la philosophie pour glaner dans la prose et la poésie classique; et ne fait pas fi de ce qu'ont produit les modernes; non seulement en littérature proprement dite, mais aussi dans la presse quotidienne.

En décrivant le corpus ainsi étendu, nous anticiperons, parfois, sur les résultats des analyses stylométriques.

1.1 Poésie classique et poésie moderne

Du classique, outre les 13 fragments déjà analysés dans [MOTS ARABES], on a pris, d'abord, un recueil de vers et de prose choisi pour exercer à l'épreuve de récitation les élèves de 3-ème année de la Section d'Arabe Littéral de l'Institut National des Langues et Civilisations Orientales: l'INALCO, connu de tous pour être le domaine des Langues'O.

Le Recueil comprend 35 textes, dus à 28 auteurs différents, pour chacun desquels est donnée une brève notice. L'ensemble peut être distribué en trois parties, que nous marquons par les signes { π , ¶, §}.

(*) Maître de langue à l'Institut National des Langues et Civilisations Orientales.

Indice des poètes : فحوس الشعراء

π_{sc1}	الفرزدق	عمر بن أبي ربيعة
π_{sc2}	ابن الرومي	أبو العتاهية
π_{sc3}	ابو فراس الحمداني	أبو الطيب المتنبلي
π_{sc4}	أبو العلاء المعري	أبو الحسن الانباري
π_{sc5}	ابن الفارض	ابن زيدون

Dans { π } sont 12 fragments de poésie ancienne. L'humaniste épris des lettres arabes reconnaîtra ici de très grands noms qui tous méritent l'attention: mais comme la plupart des fragments sont brefs, on s'est résigné à les cumuler en cinq blocs; lesquels, sous les sigles { π_{sc1} , π_{sc2} , π_{sc3} , π_{sc4} , π_{sc5} } répondent, à peu près, aux cinq premiers siècles après l'Hégire (la poésie anté-islamique étant absente).

Dans { η } sont 15 fragments de poésie postérieurs à 1800. La série s'ouvre par η Lac: le poème si bien connu de LAMARTINE, dans une surprenante traduction dont l'analyse stylométrique confirme le strict classicisme.

Suit η shu: quatre fragment de celui qu'on appela le *Prince des poètes*, ahmad šāūqī , أحمد شوقي, (1868-1932); qui, avec l'harmonie de ses vers pesés sur la balance classique, n'a pas le vocabulaire des poètes anciens.

Sort également de ce vocabulaire, malgré son thème, la $qašīdā$, η Arb, où $hāfiẓ ibrahīm$, حافظ ابراهيم, chante la langue du $qāḍ$.

Gibran, $ġibrān kālīl ġibrān$, جبران خليل جبران, libanais de la dispora, est connu de ceux-mêmes qui n'ont rien lu d'arabe: deux fragments de lui sont cumulés dans η gbr.

Aux sept autres fragments, la stylométrie n'a pas trouvé de proche parent: les sigles { η occ, η dpl} rappellent que les notices du Recueil attribuent aux auteurs soit une affinité occidentale, soit l'état de diplomate.

En passant, disons que ces brèves notices, cumulées suivant les trois parties, { π , η , \S }, du Recueil, ont donné: {&Id \S , &Id π , &Id η }.

1.2 Prose classique

{ \S KwD, \S Hdt, \S jhz, \S rbh, \S Mqm, \S sfa, \S twh, \S tah}

Reste les huit proses de { \S }, courant de l'aube du classicisme à notre XX-ème siècle. Nous renonçons à en donner ici plus qu'un aperçu: \S Hdt, Hadith, ou traditions remontant au fondateur de l'Islam; \S rbh: éloge des fortes sciences et des belles lettres; \S twh, marque de la sollicitude d'un calife envers la plèbe de Bagdad; \S sfa, anecdote extraite de l'encyclopédie des "Frères de la pureté et de la sincérité" (أخوان الصفاء)...

La diversité de thème paraît telle que, malgré la brièveté de certains fragments, nous n'en avons pas tenté de cumul.

Et pourtant, de §KwD, *kalīlah wa dimnah*, *كليلة ودمنة*, où *ibn-l-muqaffa'*, *ابن المقفع*, suivant la tradition de l'Inde et de l'Iran, met en arabe des fables dont héritera LA FONTAINE; à la scène familière, *ṣtaḥ*, où, enfant, se découvre aveugle notre futur Docteur de la Sorbonne, *ṭaha ḥusayn*, *طه حسين*; la stylométrie trouve une même prose dont ne s'éloignent ni le hadith ni le récit historique.

En sort seulement, pour suivre la poésie, §Mqm, cette scène précieuse - les arabe disent *maqāmā*, *مقامة*, séance - où un énigmatique jeune homme, d'abord attentif à un débat de lettrés, répond ensuite à ceux-ci en distillant, de chaque poète, quelques gouttes du parfum de son génie (*āl-maqāmāh al-qarīḍiyah*).

بديع الزمان الهذاني : مقامات

Des mêmes *maqāmāt* de *badī' al-zamān al-hamḍānī*, on a saisi, d'autre part, deux pièces: *mqmB*: un obséquieux filou se fait régaler par un paysan, ébloui dans Bagdad (*āl-maqāmāh al-baḡdāḍiyah*); et {*mqm1*, *mqm2*}: un nouveau riche vante sa demeure et tout ce dont il lui a plu de l'orner (*āl-maqāmāh al-maḍiriyah*).

À ce genre des séances, la stylométrie associe encore §jhz: où l'illustre *ḡaḥīz*, *الجاحظ*, met en scène un savant courtois fermant un jour son école pour porter le deuil d'une femme qu'il n'a jamais vue que de loin.

ابن المقفع : الأدب الكبير ؛ رسالة في الصحابة

Sans quitter encore le domaine de {§}, nous avons d'*ibn-l-Muqaffa'*, l'auteur déjà cité, de *kalīlah wa dimnah*, des propos auxquels ne suffiraient pas les langues des bêtes et des oiseaux: d'une part {1Adb, ..., 4Adb}, une grande partie de *al-'adab al-kabīr*, et d'autre part {1Rsl, ..., 4Rsl}, les premiers chapitres de la *risālah fī al-ṣaḥḥbah*: exposés de la norme morale et politique de la société islamique, dans ses débuts.

1.3 Prose moderne

Enfin, la prose moderne a fourni trois blocs homogènes considérables: {Grb, Sgn, Prs}.

Du libanais *Mīka'il Nu'aymah*, *ميخائل نعيمة*, mort quasi centenaire en 1988, on a, subdivisés en {1Grb, ..., 9Grb, aGrb, ..., kGrb}, les trois quarts du livre *al-ḡirbāl*, *الغربال*, - le tamis - : théorie des belles lettres dont le titre reprend cette formule:

فمهنة الناقد إذن هي غربلة الآثار الأدبية .

“la mission du critique est dépasser au tamis les œuvres littéraires”.

De ce même auteur, une poésie, *النهر المتجمد*, “le fleuve pris en glace”, est comptée dans la tranche ¶occ du Recueil.

Dans *siġn al-‘umr*, *سجن العمر* : la prison des jours, (littéralement: de l’âge) dont on n’a saisi que le début {1Sgn, ..., 9Sgn, aSgn,..., eSgn}, *ṭawfīq al-ḥakīm*, *توفيق الحكيم*, un des maîtres du roman égyptien au XX-ème siècle, déroule son autobiographie philosophique, entre le décor et les personnages de la ville et des champs.

Enfin, {1Prs, ..., 9Prs, aPrs, bPrs} est une chrestomathie de la presse égyptienne contemporaine; que nous avons découpée en veillant, dans la mesure du possible, à distinguer les sujets: affaires nationales ou étrangères; et les genres: éditoriaux ou nouvelles brèves.

2 Élaboration des textes et création de tableaux de correspondance

Les données analysées ont exactement le même format que dans [MOTS ARABES] (cité désormais ici:[MA]): il suffit donc de rappeler brièvement les principes adoptés dans ce premier article; en tenant compte de ce que l’on a aujourd’hui, sur Macintosh, (notamment avec Wintext, Teachtext, ou l’éditeur Nasher,) une représentation du texte arabe en octets différente de celle suivie par le premier logiciel ‘Alkaatib’.

2.1 *Scriptio minima* et découpage en formes et locutions

Il est dit, dans [MA:§1.2], que, pour l’analyse des textes arabes, une élaboration totalement automatique, ne requérant lors de la saisie aucune interprétation grammaticale, doit être fondée, non sur la *scriptio plena* (comportant tous les signes diacritiques flottant de part et d’autre de la ligne; et, notamment, les voyelles brèves); mais, au contraire, sur une *scriptio minima*, d’où sont éliminées, avec les signes diacritiques, des marques de distinctions essentielles (telles que celles affectant les variantes de y sans deux points); si ces marques manquent constamment dans maintes éditions classiques; pour ne rien dire de la presse contemporaine.

De plus, notre *scriptio minima* conserve le découpage usuel de la phrase arabe: ainsi, dans le titre du présent article, l’expression “formes de mots et locutions” renvoie aux segments minimaux compris entre deux blancs: les conjonctions de coordination, *ﻭ* et *ﻑ* (wa et fa), se lient au mot suivant à quelque catégorie que celui-ci appartienne; l’article défini se lie au nom; etc... Les locutions ainsi distinguées, ne le cèdent en rien aux formes isolées, prises exclusivement, dans l’étude d’autres langues, pour caractériser le style.

{ ʔ ä a b p t t ġ ċ ħ k d d r z ž s š ş ɗ ʈ z ʕ ğ f
 q ğ k g l m n h w y }

À la *scriptio minima* romanisée sert une police ‘diacr’; laquelle, à des variantes mineures près (issues, pour la plupart, du dictionnaire de Hans WEHR), suit les normes de transcription phonétique internationale. (Les signes {ğ, g, p, ċ, ž} étant réservés pour des mots non arabes: berbères, persans, urdus...).

Voici, par exemple, dans cette transcription, la phrase citée plus haut du “*Tamis*”:

f m h n ä a l n a q d a d n , h y ġ r b l ä a l a ʈ a r a l a d b y ä

2.2 Tri alphabétique des occurrences

Dans [MA], par un programme ‘triarab’, on passe directement du fichier d’octets, créé par ‘Alkaatib’, à la suite ordonnée des formes, romanisée et étiquetée par versets (i.e. phrases ou paragraphes) et chapitres.

Pour les textes saisis par d’autres logiciels, et analysés ci-après, on a bénéficié d’une option ‘Texte seul’, qui, éliminant les variantes de tracé et les ligatures si diverses de la calligraphie arabe, (variantes et ligatures dont une partie subsiste dans le texte imprimé,) donne des fichiers dont la structure est très proche de la *scriptio minima*; et peut y être exactement réduite par un programme approprié, ‘dearab’, avec le code des lettres en octets propre à la police ‘diacr’.

D’après le fichier créé par ‘dearab’, un programme ‘triarap’, crée finalement une liste de formes qui se présente exactement comme celle que crée ‘triarab’ d’après un fichier d’Alkaatib.

Un programme ‘fus’ permet de fondre en une seule les listes afférentes à plusieurs textes; créant ainsi une liste ordonnée qui ne diffère en rien de celle qu’on obtiendrait à partir des textes mis bout à bout (; à supposer que ceux-ci soient saisis dans le même format).

Ainsi, on peut considérer que le corpus décrit au §1, (ou une partie de ce corpus), assimilé à un texte unique, est transformé en la suite, ordonnée alphabétiquement, de toutes ses occurrences de formes; répétées, chacune, autant de fois qu’on la trouve, avec les références, par chapitre et verset, se succédant dans l’ordre du texte.

Ces références ne sont, en fait, utilisées que suivant le découpage du corpus en un ensemble de 91 segments; eux-mêmes déjà décrits, plus haut, avec leurs sigles. On trouve dans [MA:§3.1] le fichier de commande qui définit le découpage du corpus des textes philosophiques en 13 fragments: on a procédé de même ici.

2.3 Dictionnaire et lexiques

Partant du fichier trié des occurrences, le programme 'qamus' crée un dictionnaire du texte, ou suite ordonnée alphabétiquement des formes, chacune accompagnée de sa fréquence. Le programme 'trimu§' range cette dernière suite dans l'ordre croissant des fréquences: c'est d'après le résultat de ce tri que l'on choisit, selon divers critères, un sous-ensemble Δ , ou lexique, de formes destinées à être prises en compte dans l'analyse.

Enfin, le programme 'tridic' crée un tableau de correspondance croisant le lexique Δ retenu; pris comme ensemble des lignes, avec l'ensemble J des segments: $k(i, j)$ étant le nombre des occurrences de la forme i dans le segment j .

2.4 Enchaînement des analyses

Les textes et fragments dont nous disposons différant grandement quant à la taille et au genre, on a cru prudent de n'en approcher que par palliers l'analyse globale.

Dans une première analyse, rentrent les 13 fragments philosophiques de [MA]; et les 22 fragments découpés dans le Recueil, avec {mqmB, mqm1, mqm2} des deux $m\bar{a}q\bar{a}m\bar{a}t$ (cf. *supra*, §1.2); soit un ensemble J1 de 38 fragments. Le lexique Δ_1 , de 120 formes, est choisi, avec un seuil de fréquence à 10, en veillant à éliminer tout ce qui, dans certains fragments, peut prendre valeur de mot plein, lié au contenu. L'axe 1 étant dominé par les poésies modernes de { $\mathbb{f}occ$, $\mathbb{f}dpl$ }, avec $CTR_1(\mathbb{f}Occ)=504\%$ et $CTR_1(\mathbb{f}dpl)=99\%$, on reprend l'analyse du tableau $\Delta_1 \times J_1$ en mettant en supplément les deux colonnes { $\mathbb{f}occ$, $\mathbb{f}dpl$ }: d'où, pour J1 ainsi réduit à J1', une représentation clairement interprétable; tant à l'analyse factorielle qu'à la classification: cf. *infra*, §3.1.

Les 45 fragments distingués dans l'ensemble des trois grands blocs de prose moderne, {Grb, Sgn, Prs} = GSP, ont fait l'objet d'une analyse, $\Delta_2 \times GSP$, fondée sur un lexique Δ_2 de 162 formes, choisies, également, avec 10 pour seuil de fréquence. Analyse factorielle et CAH distinguent nettement les trois blocs: cf. §3.2.

En adjoignant à J1', textes philosophiques et Recueil, les 8 fragments de {Rsl, Adb}, $r\bar{i}s\bar{a}l\bar{a}$ et $a\bar{d}\bar{a}b$, $d'ibn-l-muqaffa'$, le corpus classique est au complet, pour un tableau $\Delta_3 \times Cls$, avec un lexique de 139 mots; le seuil restant à 10. Les trois genres de la philosophie, des belles-lettres (adab) et de la poésie sont bien vus d'après ce tableau: cf. §3.3.

Il semblait difficile de traiter ensemble le classique et le moderne, (i.e. Cls et GSP, ce qu'on notera C&M): car, d'une part, en descendant au seuil de 10, on est conduit à retenir un lexique Δ_5 comprenant 232 formes - ou agrégats - de mots outil; d'autre part, les trois blocs {Grb, Sgn, Prs}, chacun

homogène, relativement à la diversité du reste du corpus, pouvaient dominer l'analyse, imposant les deux premiers axes factoriels, et partageant la totalité des fragments entre trois pôles.

C'est pourquoi, concurremment à $\Delta 5$, on a considéré un lexique plus restreint, $\Delta 4$, comptant 141 formes ou locutions, toutes de fréquence ≥ 20 ; et, dans certaines analyses, on a réduit le poids des colonnes des blocs afférents à {Grb, Sgn, Prs}, en les multipliant par des coefficients; que, sans prétendre suivre une règle certaine, on a pris à (1/4) pour Grb, (1/2) pour Sgn et (3/4) pour Prs.

Au §4, on rend compte, avec plus ou moins de détails des résultats issus des tableaux $\Delta 4 \times C\&M$ et $\Delta 5 \times C\&M$, repondérés ou non. Il nous suffit d'annoncer ici, en bref, que l'analyse et la CAH reconnaissent les principaux genres: philosophie, prose artistique ou poésie, roman et presse; mais avec, entre les variantes, des nuances qui nous font préférer $\Delta 5 \times C\&M$ repondéré.

3 Analyses partielles

3.1 Textes philosophiques et recueil de lecture

mots de $\Delta 1 \times$ fragments de $J 1$: textes philosophiques et recueil, avec { $\mathbb{f}occ$, $\mathbb{f}dpl$ }

trace :	1.917e+0									
rang :	1	2	3	4	5	6	7	8	9	10
lambda :	2498	1757	1459	1128	1011	906	823	782	687	654 e-4
taux :	1303	917	761	589	528	473	430	408	359	341 e-4
cumul :	1303	2220	2981	3570	4097	4570	5000	5407	5766	6107 e-4

Nous nous bornerons à rappeler que, comme on l'a dit au §2.4, l'analyse du tableau $\Delta 1 \times J 1$, 120×38 , distingue les deux fragments ($\mathbb{f}occ$, $\mathbb{f}dpl$), (plus exactement, deux cumuls de fragments) qui créent le 1-er axe et contribuent encore fortement au suivant. Avec { $\mathbb{f}occ$, $\mathbb{f}dpl$ } vont les formes { lst , am , knt , any } qui évoquent le dialogue du poète avec un objet, animé ou non.

Comme il s'agit de poésies modernes, sans parenté étroite avec le reste du corpus, l'analyse est reprise sans celles-ci: soit un tableau $\Delta 1 \times J 1'$, 120×36 .

mots de $\Delta 1 \times$ fragments de $J 1'$: textes philosophiques et recueil, sans { $\mathbb{f}occ$, $\mathbb{f}dpl$ }

trace :	1.726e+0									
rang :	1	2	3	4	5	6	7	8	9	10
lambda :	2094	1504	1094	1001	930	887	781	729	723	589 e-4
taux :	1213	871	634	580	539	514	452	422	419	341 e-4
cumul :	1213	2084	2718	3298	3837	4351	4803	5225	5644	5986 e-4

Sur la suite des valeurs propres et taux, les deux premiers axes se détachent nettement, la décroissance étant, ensuite, lente et régulière. [Relativement à l'analyse de $\Delta 1 \times J 1$, l'inertie totale et les taux afférentes aux trois premiers axes diminuent; mais l'intervalle entre les axes 2 et 3 s'accroît]. C'est pourquoi on publie d'une part, les deux ensembles des mots et des fragments dans le plan (1, 2); et d'autre part la classification de chacun des deux ensembles en correspondance (effectuée dans l'espace des profils rapporté à la totalité des 35 axes extraits par l'analyse).

c	Partition de J1' en 13 classes : Sigles des fragments de la classe c
1	isf†
59	¶jbr rZt3
13	rZt4
49	rZt2 rZt1
31	§sfa
52	¶rbh TfGh ¶shu ¶Arb §Hdt §KwD §tah §twh
48	TfHy Rmq2 Rmq1 Rmq3
56	ghz2 ghz1
4	ghz3
17	πsc1
57	πsc5 πsc2 ¶Lac πsc4 §Mqm πsc3
44	&Id§ &Idπ &Id¶
58	§jhz mqm2 mqm1 mqmB

isf†	F1--	F4--	67	F1< F2< F3>	70
59	F1-	F2-	F3+	64	
rZt4	F1--	F2--	F3++	62	
49	F1-	F2-			
§sfa	F1+	F2+	F4---	65	66
52	F1+			60	
48	F1-	F2+			
56	F1-	F2+++		63	
ghz3	F1-	F2+++	F3---		
πsc1	F1+++	F2+++	F3+++	F4--	61
57	F1+++	F2+	F3++		F2> 69 F1>
44	F1+	F2---	F3--		68 F2<
58	F1++++	F2-	F4+++		

ci dessous: étiquetage sommaire d'après les classes de formes

64: 198+++++ 229++ ; 1: 228+++++ ; 65: 188++ 229++ ; 63: 185++++ 230++ ;
61: 223++++ 226+++ ; 44: 225+++++ ; 58: 217+++++ ;

Au sommet de la hiérarchie, le corpus se sépare en deux branches: j70 et j69. L'étiquetage en terme de facteurs montre que j69 est caractérisée par de fortes valeurs de $F1 > 0$. On a, dans j61, du côté ($F2 > 0$) les 5 siècles de poésie classique, avec la traduction du ¶Lac; et §Mqm, la maqâmâ sur les poètes. Le reste, j69, constitue la classe j68, ($F2 < 0$); se partageant, à un haut niveau, en, d'une part, j58: deux autres maqâmâ et §jhz, الجاحظ, le deuil du savant courtois; et, d'autre part, j44: les notices du recueil, {&Idπ, &Id¶, &Id§}.

La branche j70 comprend j67 et j66. Dans j66, on a, d'une part, j63, Ghazâlî; et, d'autre part, j65: avec j48, philosophie sans appareil technique, et j52, la prose artistique (maqâmâ exceptée) avec des œuvres poétiques modernes: ¶shu, le Prince des poètes; et ¶Arb, à la gloire de la langue arabe.

Enfin j67 comprend les dissertations formelles de philosophie, rZt, ou de théologie, isf† (texte chrétien expliqué dans [MA]). À ces fragments, dont

c	Partition de Δ1 en 16 classes : Sigles des formes de la classe c															
221	akr	aw	kant	alyh	clyh	cliy	ma	fanh	cn							
202	why	hdh	hda	mtl	sayr	anha	wdlk	lan								
222	ay	wlma	hw	aldy	bl	whw	wama	anh	ad	whda	lys	an	nfsh	lanh		
	wkan	kan	bh	bma	cn											

198	awl	ktyra	gmyc	alty	wanma	hy										

211	šy ^o	gyr	fan	kl	wahda											
201	ktyr	m ^c	gha	wahd												

188	mma	flm	šyya													
=====																
185	wada	fhw	dath	ktira	cnh											

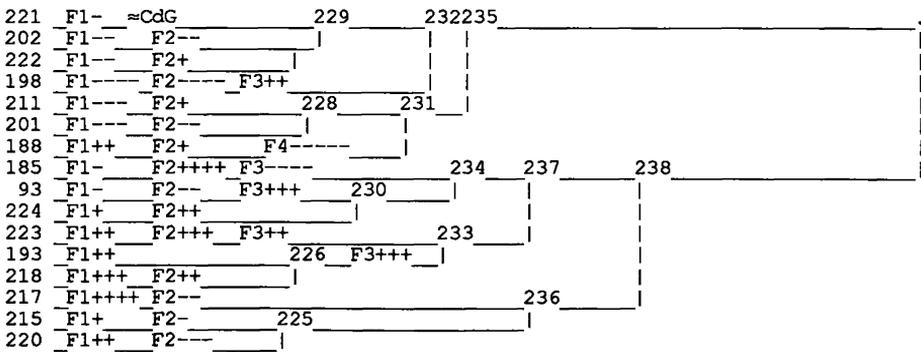
93	hl															
224	wlys	lh	ykw	b ^c d	tlk	wan	lm	qd	fyh	ala	la					

223	wgh	wlkn	lw	ama	fla	wla	km	wlw	wlm	fma	ada					

193	am	knt	fada	any												
218	mnhm	flma	qbl	lk	lma	ly	lha	wma								

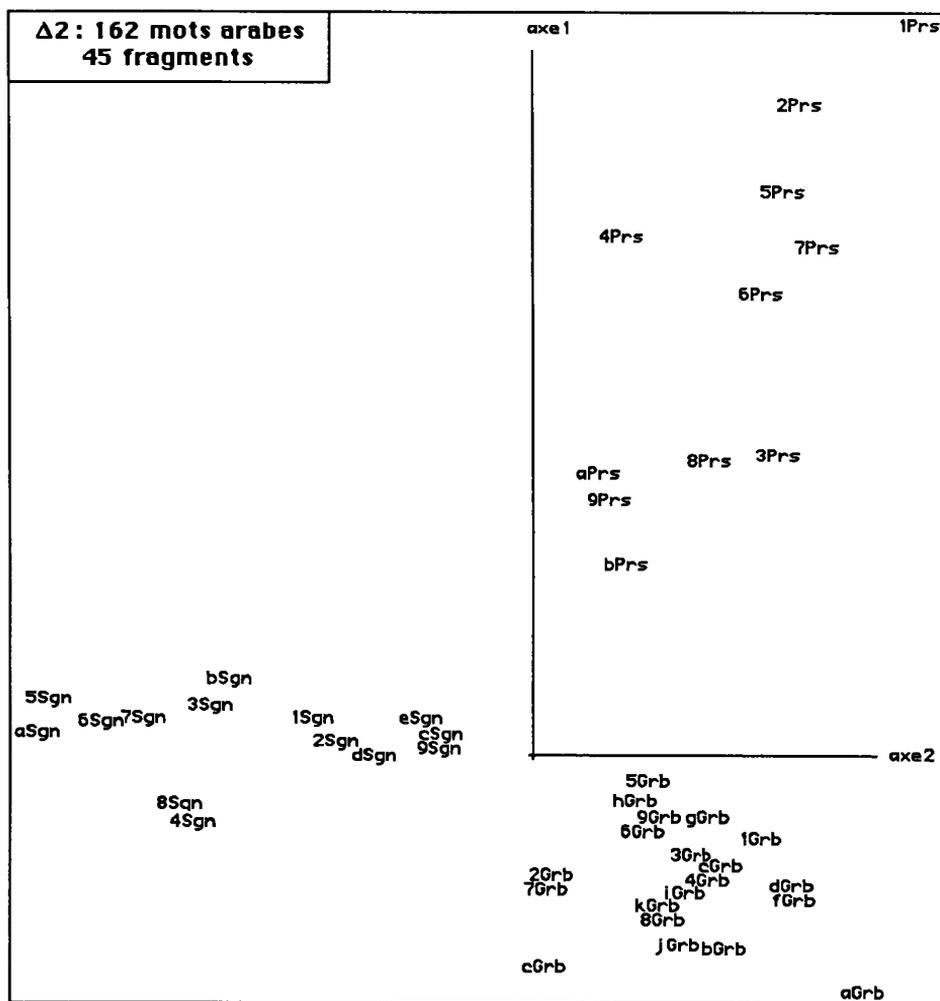
217	clyha	tm	lst	wana	ya	ana	ayn	hty								

215	anma	byn	b ^c d	kma	fy	kyf	wmn	fqd	mn							
220	bha	wqd	dlk	mnh	hyt	wfy	mnha	fyha	aly							



l'association a déjà été signalée dans le demi-plan (F1<0), s'agrange ¶jbr, poésies de Gibran: (F1>0), mais proximité avec rZt3 sur d'autres axes.

Quant à la CAH du lexique Δ1 des formes, nous nous bornerons, en attendant d'autres éléments de comparaison, à la soumettre au lecteur arabisant; en suggérant de l'examiner avec le plan (1, 2) afin de reconnaître les affinités entre textes et locutions; sans négliger l'étiquetage qui signale, entre autres, un lien entre j63, Ghazâlî, et i185.



3.2 Trois blocs de prose moderne

mots de $\Delta 2 \times$ fragments de J2: textes modernes en prose: GSP = {Grb, Sgn, Prs}
 trace : 1.487e+0
 rang : 1 2 3 4 5 6 7 8 9 10
 lambda : 1888 1255 828 715 632 580 530 483 441 413 e-4
 taux : 1269 844 557 481 425 390 356 324 296 278 e-4
 cumul : 1269 2113 2670 3151 3576 3966 4322 4646 4943 5220 e-4

Deux œuvres, Grb et Sgn, cf. *supra* §1.3, et une chrestomathie de la presse, Prs: les trois blocs se voient bien dans le plan (1, 2); et la Classification Ascendante Hiérarchique les distingue sans aucune confusion.

c	Partition de GRB en 13 classes : Sigles des fragments de la classe c	
73	1Grb	gGrb fGrb kGrb 3Grb iGrb 2Grb
72	eGrb	5Grb 6Grb jGrb cGrb
4	4Grb	

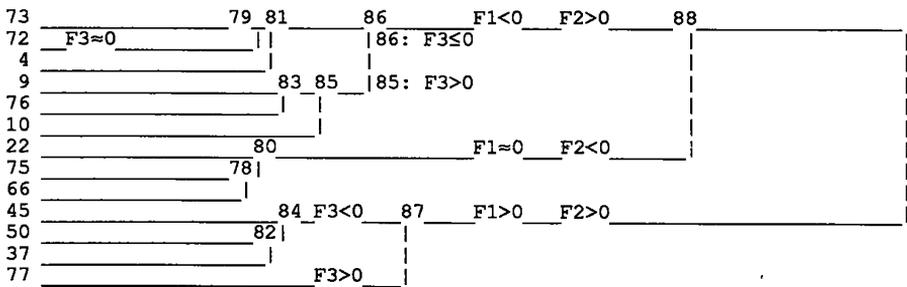
9	9Grb	
76	bGrb	hGrb 7Grb 8Grb dGrb

10	aGrb	

22	2Sgn	
75	cSgn	bSgn 6Sgn 4Sgn eSgn dSgn 9Sgn 3Sgn
66	1Sgn	aSgn 8Sgn 7Sgn 5Sgn

45	bPrs	
50	8Prs	9Prs aPrs
37	3Prs	

77	4Prs	6Prs 7Prs 5Prs 2Prs 1Prs



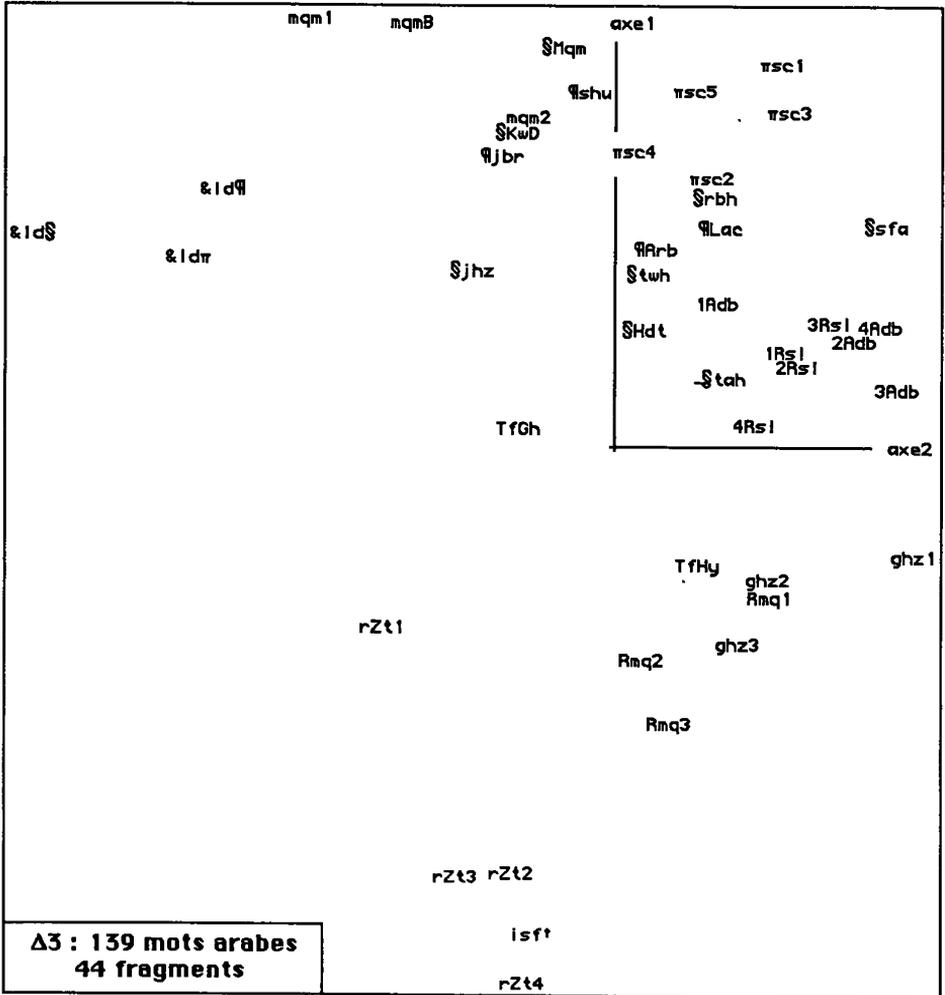
Il apparaît ainsi que la différence de style entre les deux œuvres, Grb et Sgn, se conserve au fil des chapitres; tandis que, malgré la diversité des thèmes abordés, les articles des journeaux n'approchent guères de la littérature proprement dite.

Le bloc de la presse, j87, est partagé en deux; division qu'on retrouvera, (voire un peu modifiée,) dans toutes les analyses globales du §4. La classe j77, la plus écartée vers (F1>0) dans le quadrant (F1>0, F2>0) du plan (1, 2), comprend quasi exclusivement des nouvelles relatives à l'activité du gouvernement égyptien. La classe j84, moins écartée que j77 sur (F1>0) et qui s'oppose, de plus, à celle-ci, sur l'axe 3, comprend, au contraire, des nouvelles sur l'étranger: USA, 3Prs; ou Liban, Palestine, Israel.

À propos du lexique Δ2, composé de 162 formes, nous citerons seulement un détail: la CAH distingue une classe de 4 formes:

{ yqwm cqb m^c klal } { يقوم، عقب، مع، خلال }

{ fait (ou: effectue), après, avec, durant } , fortement associée à j77; et caractérisant cette classe relativement à j84.



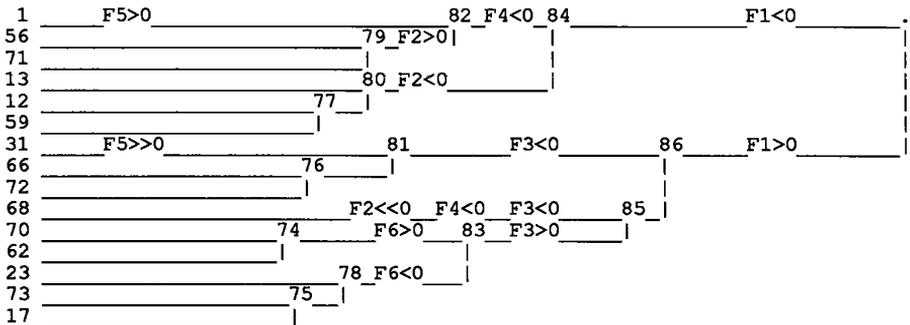
3.3 Philosophie, recueil de lecture et prose classique

mots de Δ3 × 44 fragments de CLS: textes classiques: J1' et {Rsl, Adb}

trace :	1.774e+0									
rang :	1	2	3	4	5	6	7	8	9	10
lambda :	1871	1516	1113	906	861	818	708	678	606	548 e-4
taux :	1054	854	627	510	485	461	399	382	342	309 e-4
cumul :	1054	1909	2536	3047	3532	3993	4392	4774	5116	5425 e-4

Adjoindre à J1', (textes philosophiques et Recueil) les 8 fragments de {Rsl, Adb}, risālā et adab, d'ibn-l-muqaffa^c, modifie peu les agrégats que montrent le plan (1, 2) et la CAH; mais le schéma global de la littérature classique nous paraît meilleur.

c	Partition de Cls en 15 classes : Sigles des fragments de la classe c
1	isft
56	TfHy Rmq2 Rmq1 Rmq3
71	ghz3 ghz2 ghz1
13	rZt4
12	rZt3
59	rZt2 rZt1
31	§sfa
66	\$tah \$twh \$KwD ¶Arb \$Hdt 1Adb ¶shu \$rbh TfGh
72	3Adb 4Adb 1Rsl 2Adb 4Rsl 2Rsl 3Rsl
68	&Id¶ &Idπ &Id\$
70	\$jhz mqm2
62	mqm1 mqmB
23	¶jbr
73	πsc5 ¶Lac πsc4 πsc2 \$Mqm πsc3
17	πsc1



Au sommet, dans j84, la philosophie, avec tous ses genres, se sépare du reste: j86. L'ensemble des belles-lettres en prose, confirmé par 8 fragments, forme une classe j81, agrégée à un bas niveau. Comme au §3.1, ¶shu, le Prince des poètes; et ¶Arb, à la gloire de la langue arabe, vont avec cette prose; où il est juste de trouver TfGh, Ghazâlî vu par Ibn Tofayl.

Du reste, j85, les notices {&Id} se séparent à un niveau élevé. Il n'y a plus, dans j78, que maqâmâ et poésie classique; et ¶jbr, Gibran, et §jhz, الجاحظ, le deuil du savant courtois, sont maintenant mieux placés qu'au §3.1.

Enfin, sans entrer dans les détails, il vaut la peine de relever, dans l'étiquetage de l'arbre, que des facteurs de rang élevé contribuent à distinguer certaines classes. Notamment, maqâmâ et poésie classique ne se séparent que suivant l'axe 6; et des fragments originaux, isft, §sfa, sortent sur l'axe 5.

4 Analyses de l'ensemble des textes classiques et modernes

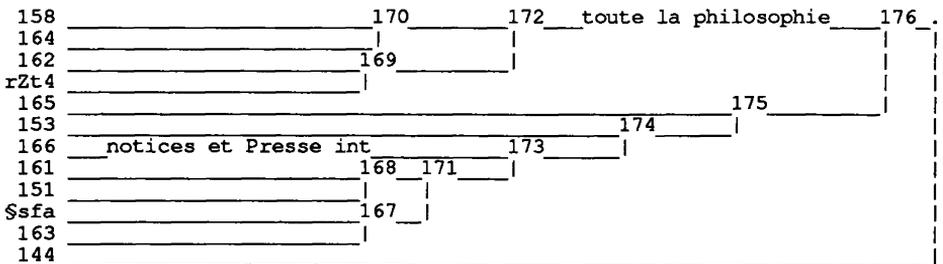
4.1 Analyses fondées sur un lexique restreint de 141 formes

En prenant un lexique restreint, Δ4, ne comptant que 141 formes ou locutions, toutes de fréquence ≥20, on pensait éviter de mettre l'accent sur les particularités peu fréquentes d'un texte. En pondérant les blocs {Grb, Sgn, Prs}, les coefficients étant (1/4) pour Grb, (1/2) pour Sgn et (3/4) pour Prs, on voulait réduire la force attractive de ces blocs pour les autres fragments.

Sans présenter de graphique plan, nous apprécierons, d'après la CAH des fragments, la vue globale du corpus offerte par l'analyse. [Dans les tableaux de partition, des subdivisions interprétables sont marquées entre parenthèses.]

141 mots de Δ4 x 89 fragments de C&M ; lexique au seuil 20 ; corpus non repondéré
 trace : 1.667e+0
 rang : 1 2 3 4 5 6 7 8 9 10
 lambda : 1411 1127 881 808 583 534 479 458 437 408 e-4
 taux : 846 676 528 485 350 320 288 275 262 245 e-4
 cumul : 846 1522 2050 2535 2885 3205 3493 3767 4030 4274 e-4

c	Partition de C&M en 12 classes : Sigles des fragments de la classe c
158	isft ghz2
164	ghz1 ghz3 TfHy Rmq2 Rmq3 Rmq1
162	rZt2 rZt1 rZt3
13	rZt4
commentaire d'Averrhoès sur le livre Z de la Métaphysique	
165	4Grb 1Grb gGrb fGrb eGrb 3Grb kGrb dGrb bGrb ¶jbr cGrb hGrb 7Grb ¶Lac πsc4 9Grb 8Grb aGrb 2Grb 1Grb jGrb 5Grb 6Grb Grb + ...
153	SHdt SKwD 3Adb 1Rsl 4Adb 1Adb 2Adb 4Rsl 2Rsl 3Rsl
Adab	
166	(bPrs 3Prs 8Prs 9Prs aPrs) (&Idπ &Id¶ &Id\$) notices &Id + Presse(international)
161	πsc1 πsc3 πsc5 \$Mqm
151	mqm2 mqmB ¶shu mqml
trois sur cinq des siècles de la poésie classique	
31	\$sfa
163	2Sgn eSgn \$rbh TfGh ¶Arb 1Sgn 9Sgn \$tah 4Sgn dSgn \$twh cSgn πsc2 \$jhz 5Sgn aSgn bSgn 6Sgn 7Sgn 3Sgn 8Sgn Sgn +
144	4Prs 6Prs 7Prs 5Prs 2Prs 1Prs
Presse (nouvelles nationales)	



Après pondération, {isf† rZt} s'isolent dans j174, au sein de la branche j176, complémentaire de celle de la presse; tandis que, dans j173, {Rmq ghz} se retrouvent, associés à {Adb Rsl}; auxquels s'agrègent, opportunément, dans j162, divers fragments de prose classique.

La poésie classique forme une subdivision, comprenant le ¶Lac et §Mqm, le débat de lettrés; les autres maqâmâ en étant proches.

Enfin les blocs de prose moderne, {Grb Sgn}, sont reconnus; mais, avant pondération, s'agrègent à Sgn de multiples fragments qui trouvent ensuite une place sans doute meilleure au sein de l'adab (belles lettres classiques).

232 mots de $\Delta 5 \times 89$ fragments de C&M ; lexique au seuil 10; corpus repondéré

trace :	2.514e+0									
rang :	1	2	3	4	5	6	7	8	9	10
lambda :	1789	1538	1136	934	849	769	724	692	650	604 e-4
taux :	712	612	452	372	338	306	288	275	258	240 e-4
cumul :	712	1323	1775	2147	2484	2790	3078	3353	3612	3852 e-4

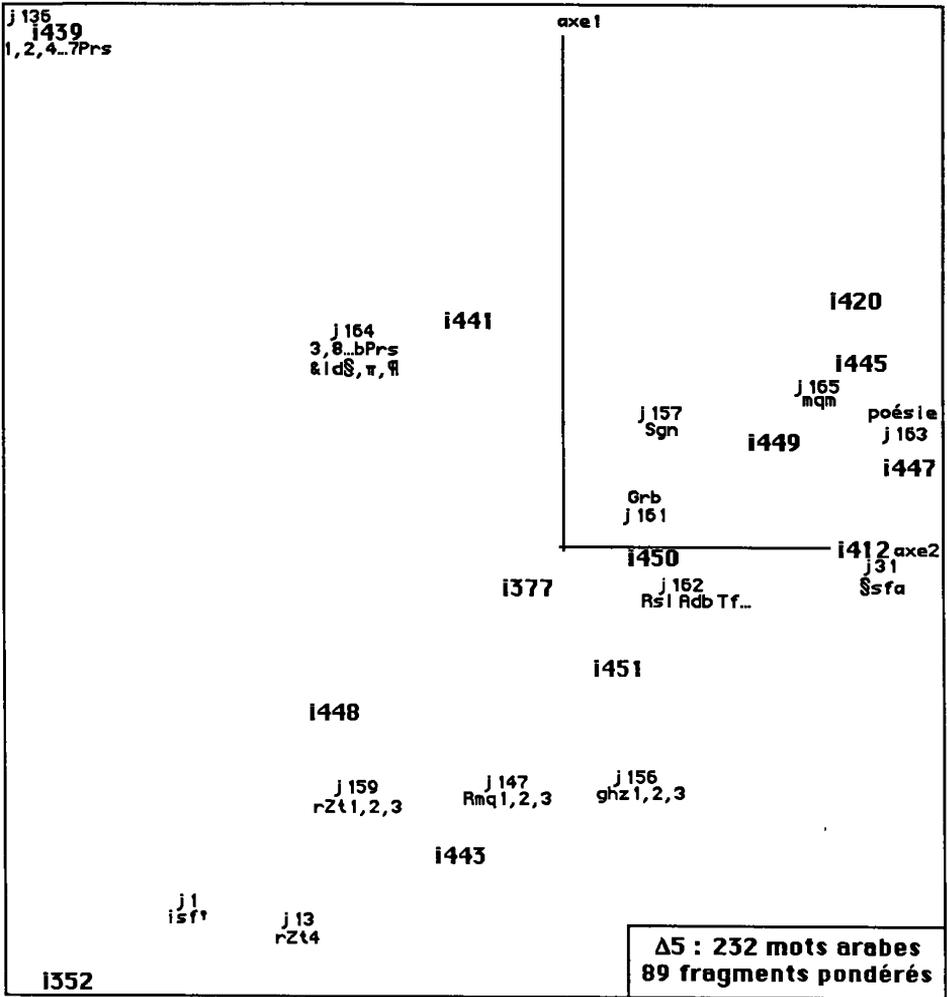
c	Partition de C&M en 13 classes : Sigles des fragments de la classe c		
1	isf†	443+++	352++++
13	rZt4	443++	448++++
159	rZt3 rZt1 rZt2		448++++
147	Rmq1 Rmq2 Rmq3		451++++
156	ghz3 ghz1 ghz2	443++++	451++
31	§sfa		412+++++
162	1Adb §tah TfHy §KwD §twh ¶shu 2Rsl ¶Arb §rbh TfGh 3Adb 3Rsl 1Rsl 4Adb 4Rsl 2Adb		451+ 450++
163	¶sc1 ¶Lac ¶sc3 ¶sc4 ¶sc2 ¶sc5 §Mqm		447+++++
165	mqm2 mqmB §Hdt mqm1		445+++++
161	¶jbr aGrb 7Grb 8Grb dGrb cGrb bGrb fGrb 9Grb 5Grb 6Grb jGrb 1Grb 2Grb 3Grb gGrb iGrb kGrb hGrb eGrb 4Grb		449+++++
157	§jh 4Sgn cSgn 1Sgn 9Sgn eSgn 3Sgn dSgn 2Sgn 6Sgn 8Sgn bSgn aSgn 5Sgn 7Sgn		420+++++
164	(bPrs 3Prs 8Prs 9Prs aPrs) (&Id§ &Id¶ &Id¶)		441++++
136	6Prs 4Prs 5Prs 7Prs 2Prs 1Prs	439+++++	441++++

isf†	_____	174	F1<<_F2<<_F3<<	_____	176.
rZt4	_____	166	livre Z	_____	
159	_____			_____	
147	_____	167	_____	173	F1<_F3>
156	_____		prose	_____	F2>0
§sfa	_____	168	_____		classique (sauf 147)
162	_____		adab etc.	_____	
163	_____	171	_____		F1>_F3<_
165	_____	170	_____		
161	_____	169	_____		
157	_____		Sgn avec jahiz	_____	
164	_____	172	_____		F1>>_F2<<
136	_____		Presse (national)	_____	

c	Partition de $\Delta 5$ en 13 classes : formes de la classe c
449	abda m ^{ch} fhm ant ^{cny} wnhn lna fwa whkga kla hwla lkm em anna
	^{cndna} anfsna swy nhn lkn dak wbyn lyst ^{cndkm} whm hl
420	agn ^{cndma} wknt alyna kdik tht knt mra ly mng faga by nfsy ana
447	bynna lqd wkn lw fla aga b ^{cd} wma lma flma ^{clyk} lk km ^{cndy}
	wlw wfy wlm fma fqd
445	alan kyf kna kga fh ^l wkyf ya lst ldlk ^{clyha} tm ma ^o wana bkl
	twl ana ayn hty
412	šyya bšy ^o flm mma lhm kenwa ^{clyhm}
450	dlk mnh ^{cndh} ga bma wkl ffy lkl fmn ^{cnd} wan fyh ala mnhm fy ^{ma} fan
	dayma alkbyr amam wkant why lha ken kent tlk ykn lm dwn qd ^{cly} mn
	wla ma awla akr aw ^{clyna} mna tkn alak fanh ^{cnya} alyh w ^{cly} ^{clyh} ^{cny}
377	kbyra aldyn bynhm šmy ^c ahd n ^{fs}
451	afdl fkan fhw alyha ^{cny} waga la wlys lh n ^{fs} bd ykwn bhgh bhm
	lmn bgyr hyn lanh hkga wama fama klha bdlk bh swah hm anma qbl
	whgh anhm bha wahda lys gy ^r whw aldy klh an
352	fhy kaša gha wahda wnnhm
443	wqt bl whga bn ^{fs} ad an ^h fana sayr alšy ^o wgh šy ^o alaw wahd ama kl
448	wkan alty wanh bhga hga hgh wdlk fqt lan anha wgyr mtl aktr awl
	wanm hy w ^{lma} ay ktyra hw
441	ktyr mnha ktyra fyha aly nsba h ^{yt} wmn wqd b ^{cd} fy
439	wkaša klal ban hnak byn kma hwl m ^c
449	_____ 453 455 _____ 459 _____ 461 _____ 462 _____
420	_____ _____ _____ _____ _____ _____
447	_____ _____ _____ _____ _____ _____
445	_____ _____ _____ _____ _____ _____
412	_____ _____ 457 _____ _____ _____ _____
450	_____ _____ 456 _____ _____ _____ _____
377	_____ _____ 454 _____ _____ _____ _____
451	_____ _____ _____ _____ _____ _____
352	_____ _____ _____ 458 _____ _____ _____ _____
443	_____ _____ 452 _____ _____ _____ _____
448	_____ _____ _____ _____ _____ _____
441	_____ _____ _____ 460 _____ _____ _____ _____
439	_____ _____ _____ _____ _____ _____

Nous publions ci-dessus le tableau du lexique $\Delta 5$, soumis à la CAH d'après l'analyse du tableau pondéré, $\Delta 5 \times C\&M$.

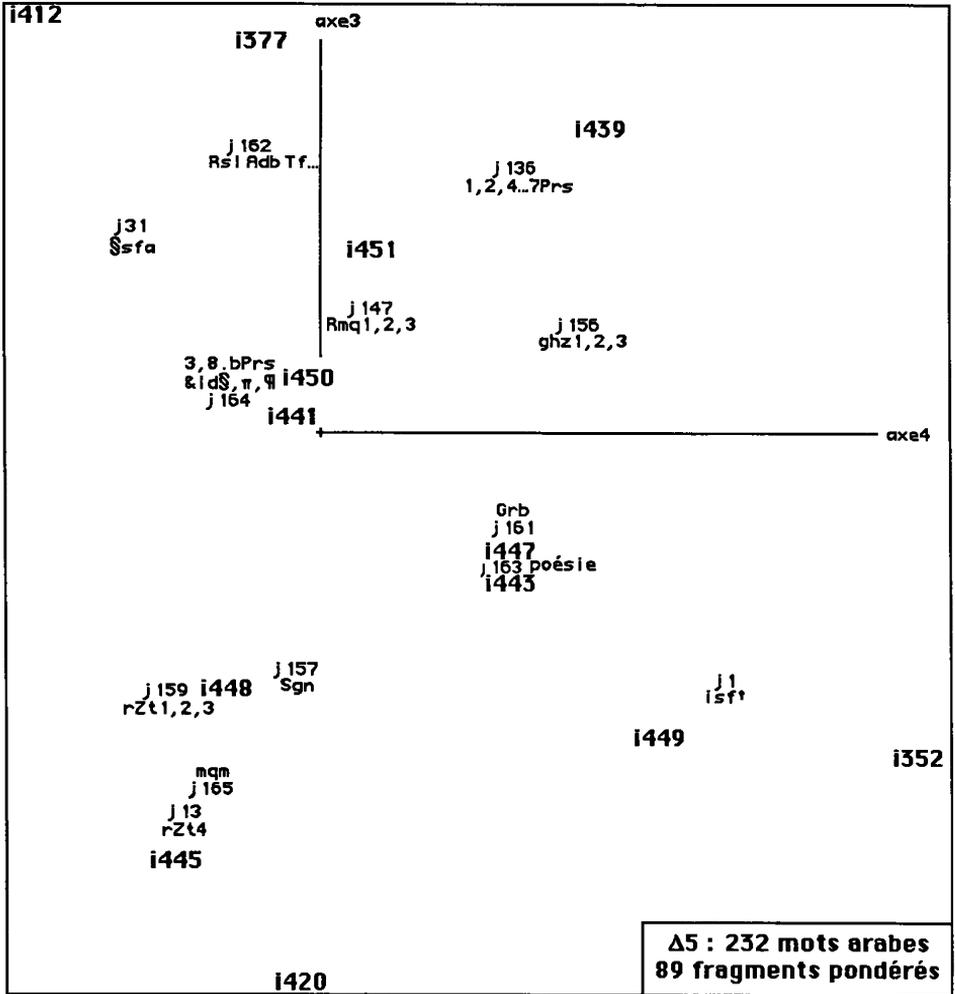
Les 13 classes de la partition retenue ont servi à étiqueter la CAH des 89 fragments; et leurs centres sont marqués, dans les plans (1, 2) et (3, 4), avec ceux des classes de fragments. Le lecteur arabisant appréciera par lui-même les associations ainsi manifestées entre formes ou locutions, d'une part; et genres littéraires, d'autre part.



Dans le plan (1, 2), le nuage des 89 fragments présente un amas, à peu près compris dans le quadrant (F1>0; F2>0) mais voisin du demi-axe (F2>0), et deux pointes; dont l'une, dessinée par la philosophie, tend vers (F1<0; F2<0), isf† étant le fragment le plus écarté; et l'autre, située dans le quadrant (F1>0; F2<0), comprend la Presse et les notices brèves du Recueil; la position extrême étant celle des articles concernant l'activité des ministres égyptiens.

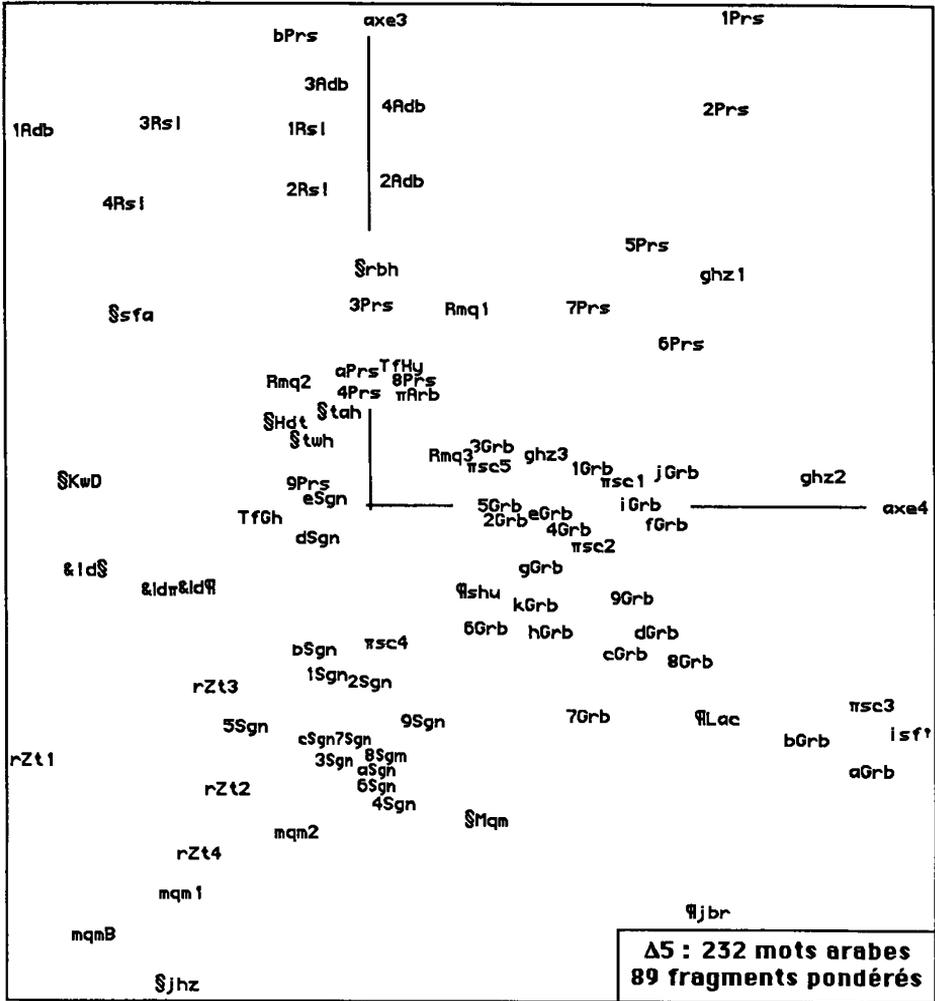
Au sein de l'amas, le facteur F2 est maximum pour la poésie classique (j169); tandis que l'adab (j162) et les œuvres modernes (Grb;j161, Sgn;j157) se projettent sur le demi-axe (F2>0) au voisinage de l'origine.

Soit un triangle {philosophie poésie presse}, dont le côté (F1<0) est vide.



Dans le plan (3, 4), se disperse ce qui formait l'amas du plan (1, 2): j162 (adab) et j167 (Sgn) s'opposent nettement suivant l'axe 3, dans le demi-plan ($F4 < 0$); Grb (j161) et poésie (j163) sont dans le quadrant ($F3 < 0, F4 > 0$). Plus précisément, sur le nuage des fragments individuels, on voit que j163 se distribue dans toutes les directions autour de son centre; tandis que les sections qu'on a distinguées au sein de Grb dessinent plutôt une pointe, vers ($F3 < 0, F4 > 0$).

Reste le cas des poésies modernes { $\{occ, \{dpl\}$ que, depuis le §3.1, on a mises en supplément comme perturbant l'analyse: par le programme 'discr', d'analyse discriminante, on affecté chacun ces points d'une part au centre de



classe le plus proche, d'autre part au fragment individuel le plus proche. Soit:

$$(\text{occ} \rightarrow \text{mqmB})(\text{dpl} \rightarrow 8\text{Sgn}) \quad ; \quad (\text{occ} \rightarrow \text{j157})(\text{dpl} \rightarrow \text{j157}) \quad ;$$

la distance d'affectation étant moins forte pour dpl que pour occ (très excentrique). [On se souviendra que j157 s'identifie à Sgn .]

Nous appellerons enfin l'attention sur les tableaux de valeurs propres qui illustrent le §4. Lorsqu'on passe du lexique $\Delta 4$ à $\Delta 5$, la trace et les valeurs propres augmentent nettement: ce qui atteste un contraste plus prononcé. La pondération augmente également l'inertie, mais moins fortement.

4 Conclusions et perspectives

La présente étude, fondée sur un corpus plus étendu et plus varié que celui considéré dans [MOTS ARABES], confirme qu'une stylométrie de l'arabe peut être fondée sur le dénombrement automatique des formes ou locutions (définies comme suites de caractères délimitées par des blancs ou des signes de ponctuation).

Nous espérons que l'occasion nous sera offerte de traiter un corpus beaucoup plus étendu; comprenant un grand nombre d'œuvres saisies dans leur intégralité; et non seulement de brèves anthologies ou des textes reçus sans avoir été systématiquement choisis.

Références bibliographiques

[MOTS ARABES]: "Sur l'étude des textes arabes d'après les occurrences des formes de mots", in *CAD*, Vol. XIX, n°1, pp. 65-84; (1994).

Textes et Poèmes choisis pour l'épreuve de récitation; Vocalisation, présentation et commentaire de Ayadi CHABIR et Jean TARDY; recueil distribué par la Section d'Arabe Littéral de l'Institut National des Langues et Civilisations Orientales (INALCO) ; Centre d'Asnières; (1995).

Documents pour l'étude et l'histoire de la langue arabe ; 1, prose; par Ayadi CHABIR , INALCO, Section d'Arabe; (1995).