Les Cahiers de l'Analyse des Données
Vol. XX - 1995 - n°4 - pp. 389-412.

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.

Les cahiers de l'analyse des données 0339-3097/95/04/$ 4.00/© Gauthier-Villars

Indice des poètes : فهرس الشعراء

πsc1	الفرزدق		عمر بن أبي ربيعة
πsc2	ابن الرومي	أبو العتاهية	أبو نواس
πsc3	ابو فراس الحمداني		أبوالطيب المتنبي
πsc4	أبو العلاء المعري		أبو الحسن الانباري
πsc5	ابن الفارض		ابن زيدون

Dans {π} sont 12 fragments de πoésie 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 {πsc1, πsc2, πsc3, πsc4, πsc5} répondent, à peu près, aux cinq premiers siècles après l'Hégire (la poésie anté-islamique étant absente).

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

Suit ¶shu: quatre fragment de celui qu'on appela le *Prince des poètes*, aḥmad šawqī, 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 qaṣīdä, ¶Arb, où ḥāfiẓ ibrāhīm, حافظ ابراهيم, chante la langue du ḍād.

Gibran, ǧibrān ḵalīl ǧibrān, جبران خليل جبران, libanais de la disapora, est connu de ceux-mêmes qui n'ont rien lu d'arabe: deux fragments de lui sont cumulés dans ¶gbr.

Aux sept autres fragments, la stylométrie n'a pas trouvé de proche parent: les sigles {¶occ, ¶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, ¶, §}, du Recueil, ont donné: {&Id§, &Idπ, &Id¶}.

1.2 Prose classique

{§KwD, §Hdt, §jhz, §rbh, §Mqm, §sfa, §twh, §tah}

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

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

Et pourtant, de §KwD, kalīlah wa dimnah, كليلة ودمنة , où ibn-l-muqaffaʿ, 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, §tah, où, enfant, se découvre aveugle notre futur Docteur de la Sorbonne, taha 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 maqāmä, 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-maqāmah al-qariḍiyyah).

بديع الزمان الهمذاني : مقامات

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

À ce genre des séances, la stylométrie associe encore §jhz: où l'illustre ǧaḥiẓ, 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-Muqaffaʿ, 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-ʾadab al-kabīr, et d'autre part {1Rsl, ..., 4Rsl}, les premiers chapitres de la risālah fī al-ṣaḥābah: 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 Mīḵaʾil Nuʿaymah, ميخائل نعيمة , mort quasi centenaire en 1988, on a, subdivisés en {1Grb, ..., 9Grb, aGrb, ..., kGrb}, les trois quarts du livre al-ġirbāl, 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 depasser 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 siğn al-ʿumr, سجن العمر : la prison des jours, (littéralement: de l'âge) dont on n'a saisi que le début {1Sgn, ..., 9Sgn, aSgn,..., eSgn}, tawfīq 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, w et f (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.

{ᵓ ä a b p t ṯ ğ č ḥ ḵ d ḏ r z ž s š ṣ ḍ ṭ ẓ ᶜ ḡ f
ḳ ġ 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 {ġ, g, p, č, ž} étant réservés pour des mots non arabes: berbères, persans, urdus...).

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

fmhnä alnaḳd aḏn, hy ḡrblä alaṯar aladbyä

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 maqāmāt (cf. *supra*, §1.2); soit un ensemble J1 de 38 fragments. Le lexique Δ1, 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 {¶occ, ¶dpl}, avec CTR1(¶Occ)=504‰ et CTR1(¶dpl)=99‰, on reprend l'analyse du tableau Δ1 × J1 en mettant en supplément les deux colonnes {¶occ, ¶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, Δ2 × GSP, fondée sur un lexique Δ2 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}, risālä et adab, d'ibn-l-muqaffaᶜ, le corpus classique est au complet, pour un tableau Δ3 × 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 Δ5 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 Δ5, on a considéré un lexique plus restreint, Δ4, 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 Δ4 × C&M et Δ5 × 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 Δ5 × C&M repondéré.

3 Analyses partielles

3.1 Textes philosophiques et recueil de lecture

mots de Δ1 × fragments de J1 : textes philosophiques et recueil, avec {¶occ, ¶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

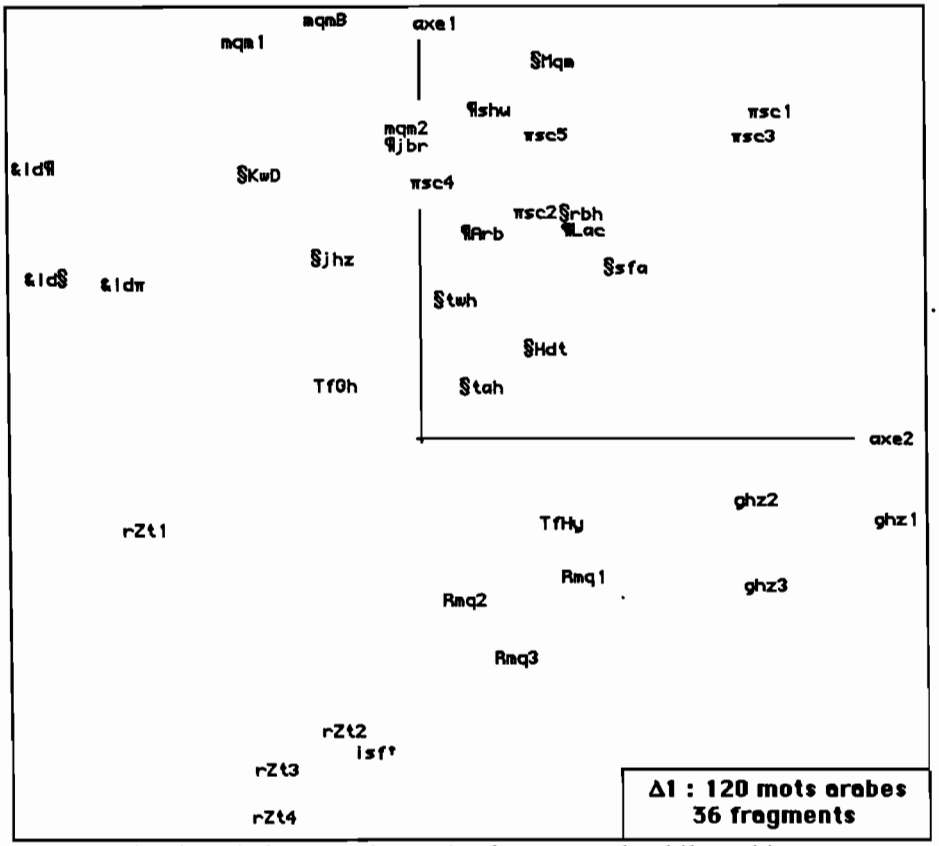
Nous nous bornerons à rappeler que, comme on l'a dit au §2.4, l'analyse du tableau Δ1 × J1, 120 × 38, distingue les deux fragments {¶occ, ¶dpl}, (plus exactement, deux cumuls de fragments) qui créent le 1-er axe et contribuent encore fortement au suivant. Avec {¶occ, ¶dpl} vont les formes {lst, am, knt, 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 Δ1 × J1', 120 × 36.

mots de Δ1 × fragments de J1': textes philosophiques et recueil, sans {¶occ, ¶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

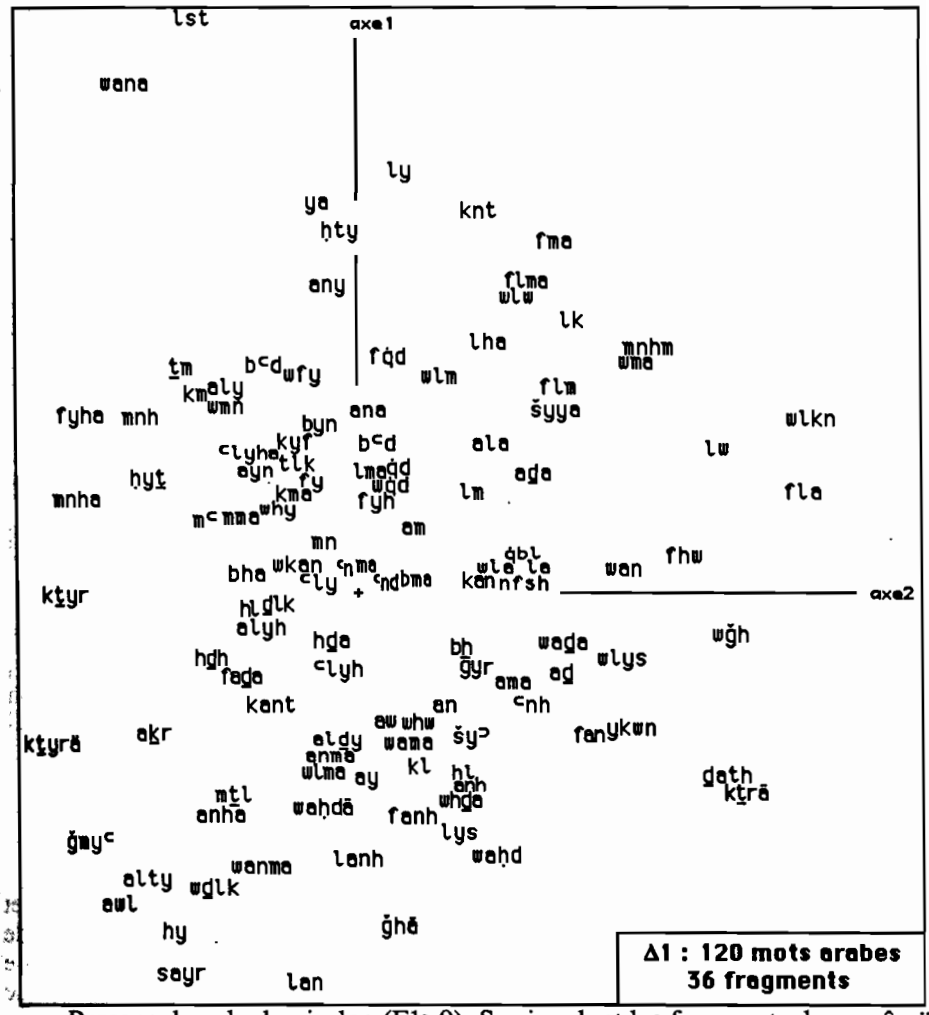
Sur la suite des valeurs propres et taux, les deux premiers axes se détachent nettement, la décroissance étant, ensuite, lente et régulière. [Relativement à l'analyse de Δ1 × J1, l'inertie totale et les taux afférentes aux trois premiers axes diminuent; mais l'intervalle entre les axes 2 et 3 s'acccroît]. 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).



Dans le plan (1, 2), on voit que les fragments de philosophie ont tous un facteur F1 négatif; à l'exception de TfGh, où Ibn Tofayl, spéculant sur les profondes pensées de l'Imam Ghazâlî, peint aussi celui-ci, se dérobant au vulgaire, et ne se livrant même pas sans réticence à ses proches.

Dans TfHy, l'autre passage du même auteur, ḥayy, le "philosophe autodidacte", seul habitant de l'île où il a grandi, médite sur la Divinité. Aucun trait pitoresque ne se mêle à la pensée: mais comme le penseur n'a pas fréquenté les écoles, les propos qu'on lui prête sont plus proche de la vulgarisation, représentée par le "traité décisif" d'Averrhoès, Rmq, que du commentaire rigoureux que ce même auteur a écrit de la métaphysique d'Aristote, rZt. La distinction avait déjà été vue dans [MA].

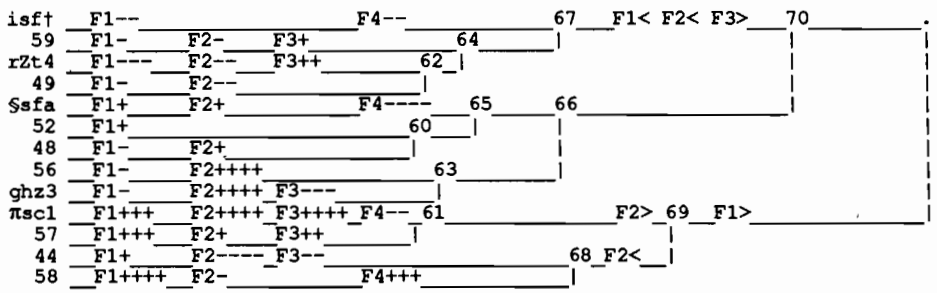
Le tahāfut, "effondrement des philosophes", œuvre polémique de Ghazâlî, est dans le quadrant (F1<0, F2>0); mais vers l'extrémité de l'axe2.



Passons dans le demi-plan (F1>0). Se signalent les fragments de maqâmä (séances...) vers l'extrémité de l'axe 1. Les trois cumuls, par parties, des notices du Recueil, {&Idπ, &Id¶, &Id§} sont dans le quadrant (F1>0, F2<0). Sans être étroitement groupés, les cumuls de poésies par siècle, {πsc1,... , πsc5}, sont dans le quadrant (F1>0, F2>0), à l'exception de πsc4 (F2≈0). La prose du recueil, {§}, n'occupe pas de position excentrique (§Mqm excepté).

Mais parce que le plan (1, 2) ne rend compte que de 21% de l'inertie, il convient de chercher dans la classification une vue systématique du corpus fondée sur l'ensemble des associations entre fragments et formes.

```
  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
```



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

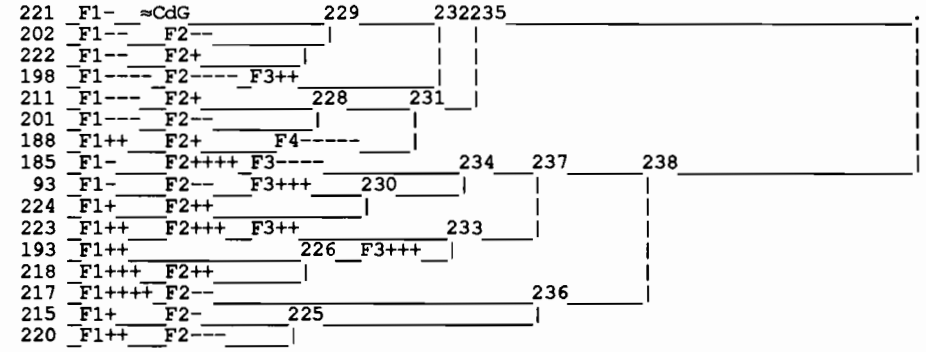
```
64: 198+++++ 229++  ;  1: 228++++++  ; 65: 188++ 229++ ; 63: 185++++ 230++ ;
61: 223++++  226+++ ; 44: 225+++++   ; 58: 217++++++   ;
```

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 maqâmä 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 maqâmä 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 (maqâmä exceptée) avec des œuvres poétiques modernes: ¶shu, le Prince des poètes; et ¶Arb, à la gloire de la langue arabe.

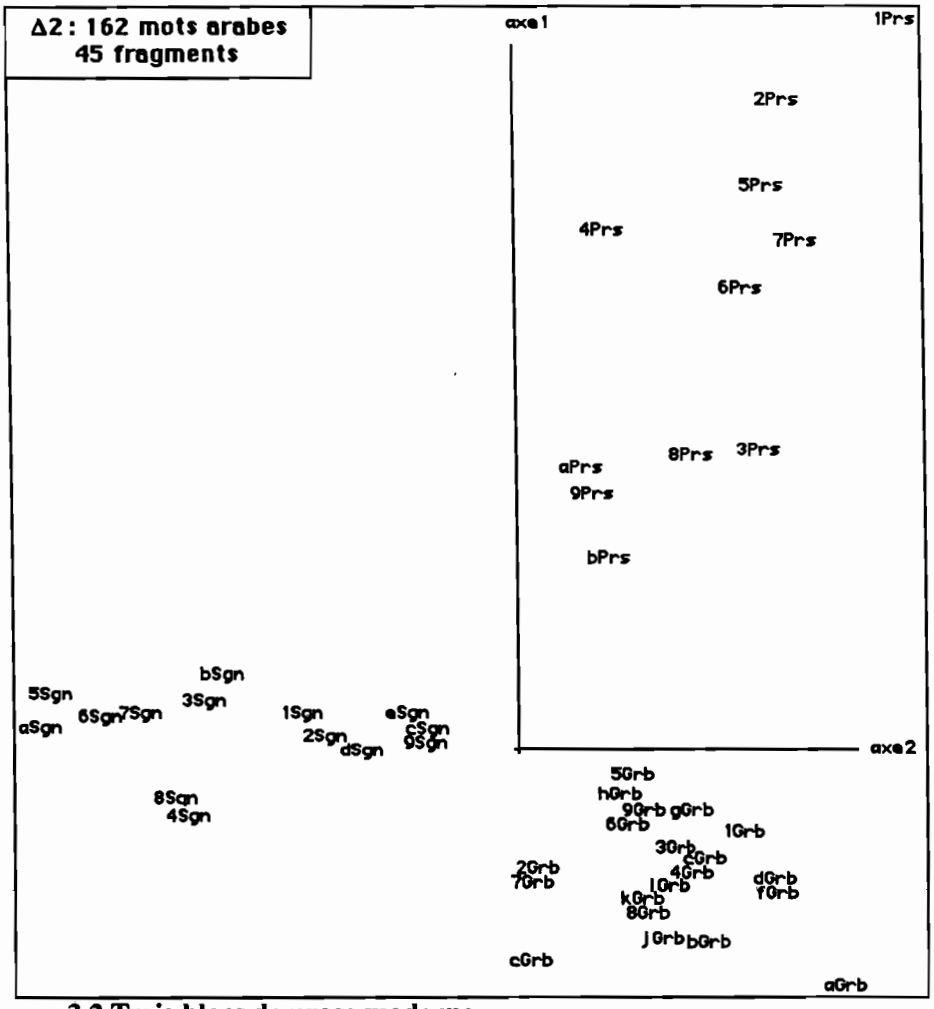
Enfin j67 comprend les dissertations formelles de philosophie, rZt, ou de théologie, isf† (texte chrétien expliqué dans [MA]). À ces fragments, dont

```
 c | Partition de Δ1 en 16 classes : Sigles des formes de la classe c
-----------------------------------------------------------------------
221|  akr  aw kant alyh ᶜlyh  ᶜly  ma fanh  ᶜn
202|  why  hgh  hda  mtl sayr anha wglk  lan
222|   ay wlma  hw algy  bl  whw wama  anh  ag whda  lys  an nfsh lanh
   | wkan  kan  bh  bma ᶜnd
-----------------------------------------------------------------------
198|  awl ktyrä ğmyᶜ alty wanma  hy
-----------------------------------------------------------------------
211|  šyᵖ  ğyr  fan   kl waḥdä
201| ktyr   mᶜ  ğhä waḥd
-----------------------------------------------------------------------
188|  mma  flm šyya
=======================================================================
185| wada  fhw dath kträ  ᶜnh
-----------------------------------------------------------------------
 93|   hl
224| wlys   lh ykwn  bᶜd  tlk  wan   lm   ḍd  fyh  ala   la
-----------------------------------------------------------------------
223|  wğh wlkn   lw  ama  fla  wla   km  wlw  wlm  fma  ada
-----------------------------------------------------------------------
193|   am  knt fada  any
218| mnhm flma  ḍbl   lk  lma   ly  lha  wma
-----------------------------------------------------------------------
217| ᶜlyha  tm  lst wana   ya  ana  ayn  hty
-----------------------------------------------------------------------
215| anma  byn  bᶜḍ  kma   fy  kyf  wmn  fḍd   mn
220|  bha  wḍd  dlk  mnh  hyt  wfy mnha fyha  aly
-----------------------------------------------------------------------
```



l'association a déjà été signalée dans le demi-plan (F1<0), s'agrège ¶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'étquetage qui signale, entre autres, un lien entre j63, Ghazâlî, et i185.



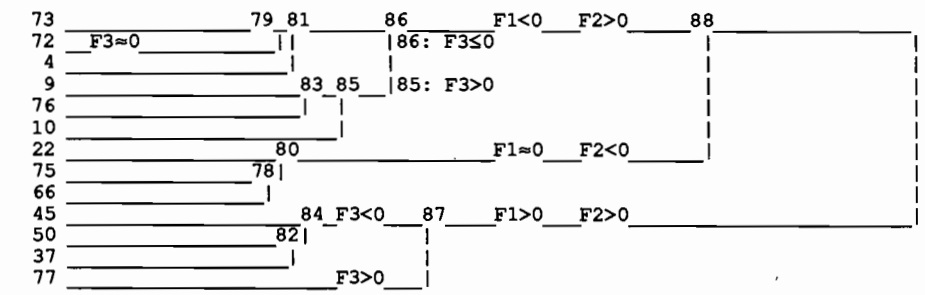
3.2 Trois blocs de prose moderne

mots de Δ2 × 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 Ascendente 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



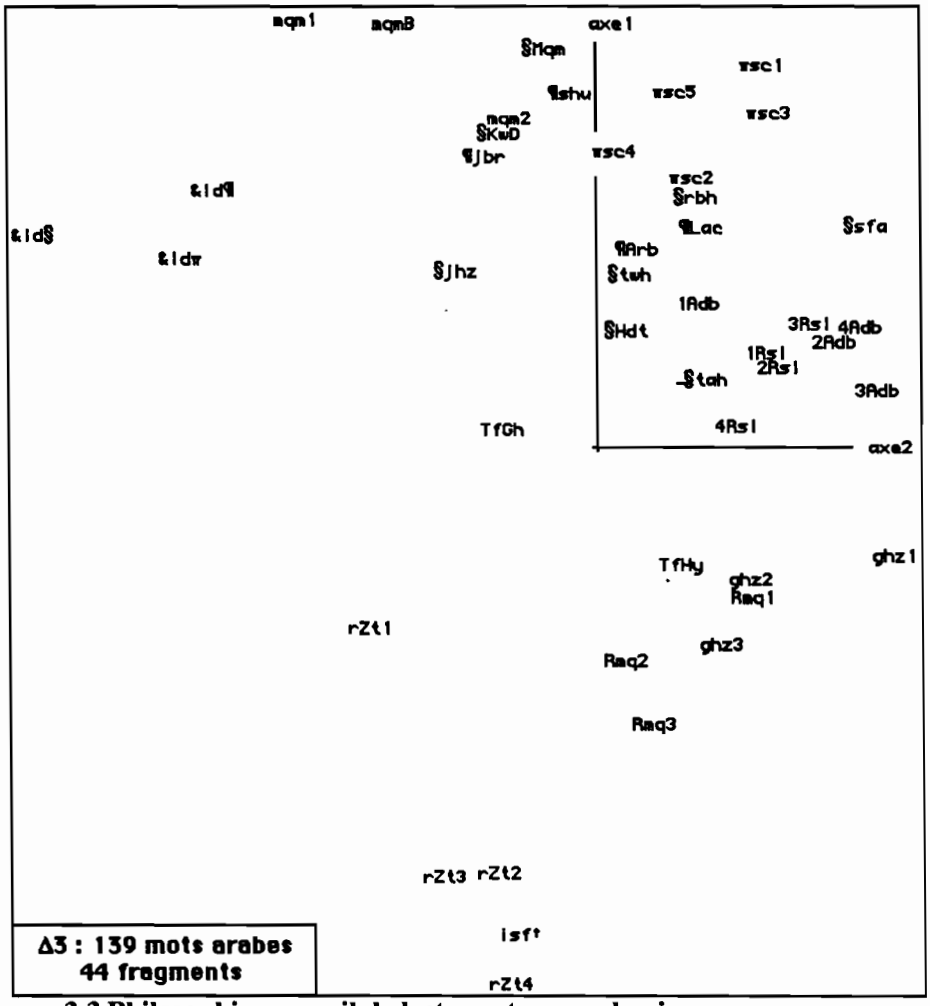
Il apparaît ainsi que la différence de style entre les deux œuvres, Grb et Sgn, se conserve au fil des chapitres; tandis que, malgré la diversité des thèmes abordés, les articles des journeaux n'approchent guères de la littérature proprement dite.

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 mᶜ klal} { يقوم، عقب، مع، خلال}

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

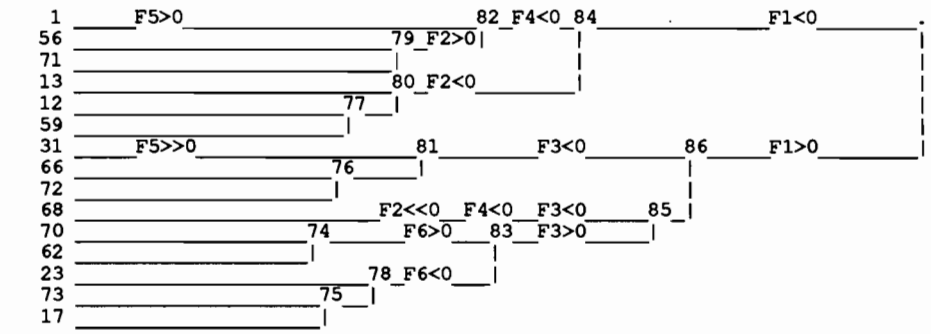


3.3 Philosophie, recueil de lecture et prose classique

```
mots de Δ3 x 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}, risālä et adab, d'ibn-l-muqaffaᶜ, 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	isf†
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 maqâmä 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 élévé contribuent à distinguer certaines classes. Notamment, maqâmä et poésie classique ne se séparent que suivant l'axe 6; et des fragments originaux, isf†, §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

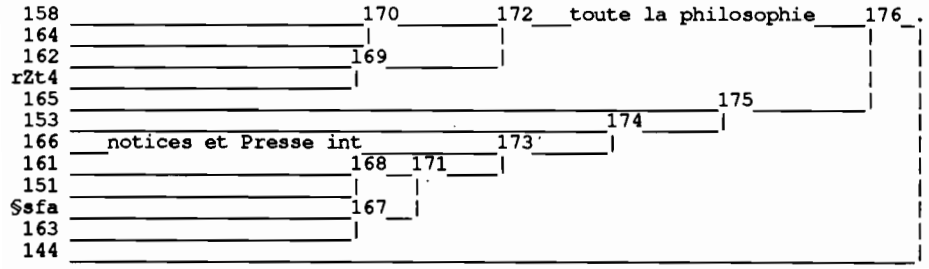
En prenant un lexique restreint, Δ4, ne comptant que 141 formes ou locutions, toutes de fréquence ≥20, on pensait éviter de mettre l'accent sur les particularités peu fréquentes d'un texte. En pondérant les blocs {Grb, Sgn, Prs}, les coefficients étant (1/4) pour Grb, (1/2) pour Sgn et (3/4) pour Prs, 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 × 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	isf† ghz2
164	ghz1 ghz3 TfHy Rmq2 Rmq3 Rmq1
162	rZt2 rZt1 rZt3 — commentaire d'Averrhoès sur le livre Z de la Métaphysique
13	rZt4
165	4Grb 1Grb gGrb fGrb eGrb 3Grb kGrb dGrb bGrb ¶jbr cGrb hGrb 7Grb ¶Lac
	πsc4 9Grb 8Grb aGrb 2Grb iGrb jGrb 5Grb 6Grb — Grb + …
153	$Hdt $KwD 3Adb 1Rsl 4Adb 1Adb 2Adb 4Rsl 2Rsl 3Rsl — Adab
166	(bPrs 3Prs 8Prs 9Prs aPrs) (&Idπ &Id¶ &Id$) — notices &Id + Presse(international)
161	πsc1 πsc3 πsc5 $Mqm — trois sur cinq des siècles de la poésie classique
151	mqm2 mqmB ¶shu mqm1
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)



L'ensemble des textes philosophiques de tout genre est reconnu: après pondération, il s'oppose même à tout le reste, en formant une branche.

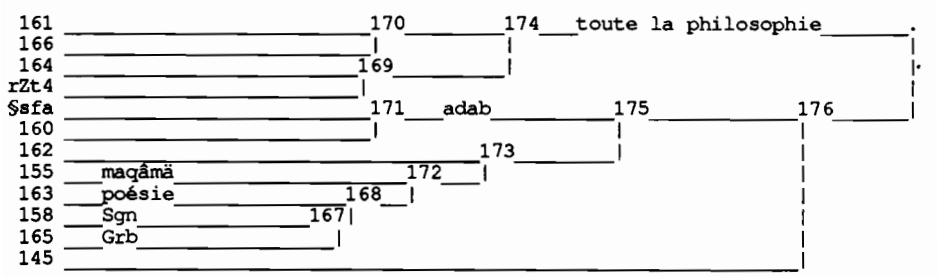
On retrouve la Presse séparée en deux blocs: l'un, {1 2 4 5 6 7}, surtout relatif à l'activité du gouvernement ; l'autre à l'étranger. Mais ces blocs ne sont pas directement agrégés. Avec le second, vont les brèves notices du Recueil.

Particulièrement dans l'analyse sans pondération, l'ensemble des chapitres de Sgn attire des fragments divers; dont le genre ne laisse pas présumer une telle similitude. En partie de ce fait, les cumuls par siècle des poésies classiques sont dispersés; et le rapport entre poésie et prose n'est pas net.

Cependant, la plupart des segments du corpus sont bien vus.

```
141 mots de Δ4 × 89 fragments de C&M ; lexique au seuil 20 ; corpus repondéré
trace  :   1.812e+0
rang   :      1     2     3     4     5     6     7     8     9    10
lambda :   1636  1322   992   818   673   644   573   545   522   466 e-4
taux   :    903   729   547   451   371   355   316   301   288   257 e-4
cumul  :    903  1632  2180  2631  3002  3358  3674  3974  4262  4520 e-4
```

```
  c | Partition en  12 classes : Sigles des fragments de la classe c
----+-------------------------------------------------------------------------
161| isf† ghz2
166| ghz1 ghz3 TfHy Rmq3 Rmq2 Rmq1
-----------------------------------
164| rZt2 rZt1 rZt3
 13| rZt4
=========================================================================
 31| §sfa
160| §Hdt §KwD §tah 3Adb 1Adb 4Rsl 2Adb 1Rsl 4Adb 2Rsl 3Rsl
-------------------------------------------
162|(bPrs 3Prs 8Prs 9Prs aPrs)((§rbh ¶Arb TfGh)(&Idπ &Id¶ &Id§))
-----------------------------------
155| mqm2 mqmB ¶shu mqm1
----------------------------
163| πsc1 πsc3 πsc5 §Mqm
----------------------
158| §jhz §twh 9Sgn cSgn eSgn dSgn πsc4 ¶Lac πsc2 2Sgn 1Sgn 4Sgn 7Sgn 3Sgn
   | aSgn 5Sgn bSgn 6Sgn 8Sgn
--------------------
165| ¶jbr cGrb bGrb 9Grb hGrb 7Grb 8Grb dGrb aGrb jGrb 6Grb 5Grb gGrb eGrb
   | 4Grb 1Grb 2Grb iGrb fGrb 3Grb kGrb
-------------------------------------------------------------------------------
145| 4Prs 7Prs 6Prs 5Prs 2Prs 1Prs
=========================================================================
```



4.2 Analyses fondées sur un lexique étendu de 232 formes

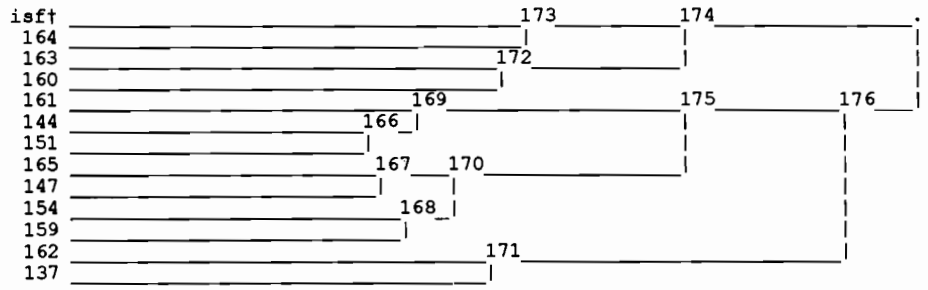
Afin de faciliter la comparaison avec l'autre lexique, nous considérons d'abord les classifications de l'ensemble C&M des fragments, classiques et modernes, fondées sur les tableaux brut ou repondéré.

La structure d'ensemble de la presse est bien vue dans les deux cas (à ceci près qu'avant pondération, manque le fragment 3Prs); les notices du Recueil allant avec les nouvelles internationales, plus concises. Après pondération, la presse s'oppose même à tout le reste, en formant une branche.

Nous savons qu'il y a, en philosophie deux niveaux: le plus rigoureux étant dans rZt; l'autre, dans {Rmq ghz}. Avant pondération, le tout forme une branche 174, où se trouvent également deux traités en belle forme, {Adb Rsl}.

```
232 mots de Δ5 × 89 fragments de C&M ; lexique au seuil 10 ; corpus non repondéré
trace  :   2.329e+0
rang   :      1     2     3     4     5     6     7     8     9    10
lambda :   1529  1314  1045   944   707   640   622   580   567   541 e-4
taux   :    657   564   449   405   304   275   267   249   243   232 e-4
cumul  :    657  1221  1669  2075  2378  2653  2920  3169  3412  3645 e-4
```

```
  c | Partition en  13 classes : Sigles des fragments de la classe c
----+-------------------------------------------------------------------
  1| isf†
164| rZt4 rZt3 rZt1 rZt2          commentaire d'Averrhoès sur le livre Z de la Métaphysique
---------------------------------------------
163| Rmq1 Rmq3 Rmq2 ghz3 ghz1 ghz2       Traité décisisif et Effondrement des Philosophes
160| 1Adb 3Adb 3Rsl 1Rsl 2Rsl 4Rsl 4Adb 2Adb                    adab el kabîr et risâlä
=============================================================================
161| fGrb 9Grb 5Grb 6Grb 3Prs kGrb 3Grb iGrb gGrb hGrb jGrb 1Grb eGrb 4Grb
144| ¶jbr aGrb
151| ¶Arb 7Grb 8Grb dGrb cGrb bGrb
-----------------------------------------------
165| §sfa 2Sgn TfHy 2Grb §Hdt §tah 9Sgn ¶shu §twh §KwD dSgn §rbh TfGh cSgn
   | 4Sgn 1Sgn eSgn 3Sgn
147| §jhz 6Sgn 8Sgn bSgn aSgn 5Sgn 7Sgn
-------------------------------
154| πsc1 ¶Lac πsc3 πsc4 πsc2 πsc5 §Mqm
159| mqmB mqm1 mqm2
-----------------------------------------------------------------------------
162| &Idπ &Id¶ &Id§ bPrs 8Prs 9Prs aPrs   manque 3Prs!
137| 6Prs 4Prs 5Prs 7Prs 2Prs 1Prs
----+-------------------------------------------------------------------
```



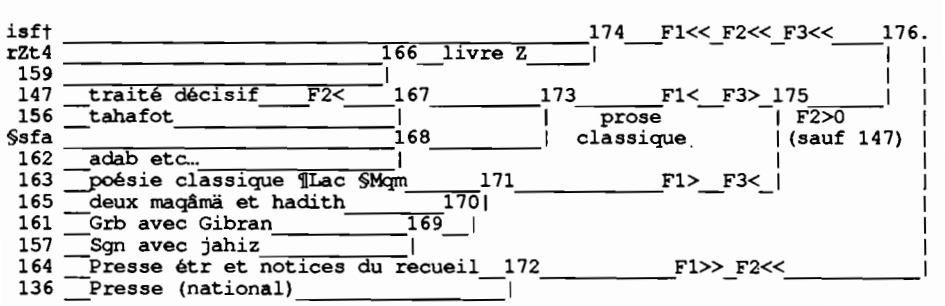
Après pondération, {isf† rZt} s'isolent dans j174, au sein de la branche j176, complémentaire de celle de la presse; tandis que, dans j173, {Rmq ghz} se retrouvent, associés à {Adb Rsl}; auxquels s'agrègent, opportunément, dans j162, divers fragments de prose classique.

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

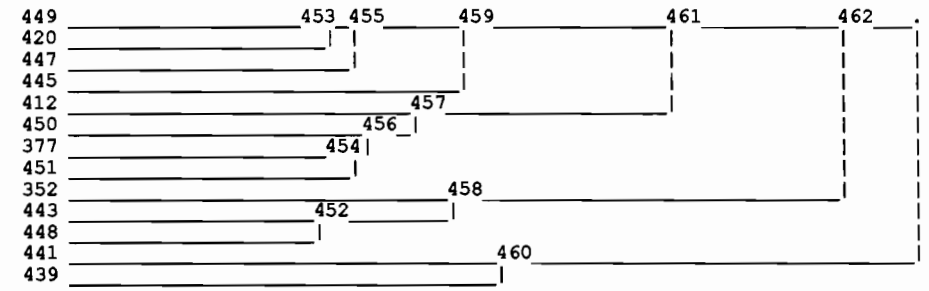
Enfin les blocs de prose moderne, {Grb Sgn}, sont reconnus; mais, avant pondération, s'agrègent à Sgn de multiples fragments qui trouvent ensuite une place sans doute meilleure au sein de l'adab (belles lettres classiques).

```
232 mots de Δ5 × 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	πsc1 ¶Lac πsc3 πsc4 πsc2 πsc5 §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	§jhz 4Sgn cSgn 1Sgn 9Sgn eSgn 3Sgn dSgn 2Sgn 6Sgn 8Sgn bSgn aSgn 5Sgn 7Sgn		420+++++
164	(bPrs 3Prs 8Prs 9Prs aPrs) (&Id§ &Idπ &Id¶)		441++++
136	6Prs 4Prs 5Prs 7Prs 2Prs 1Prs	439+++++	441++++

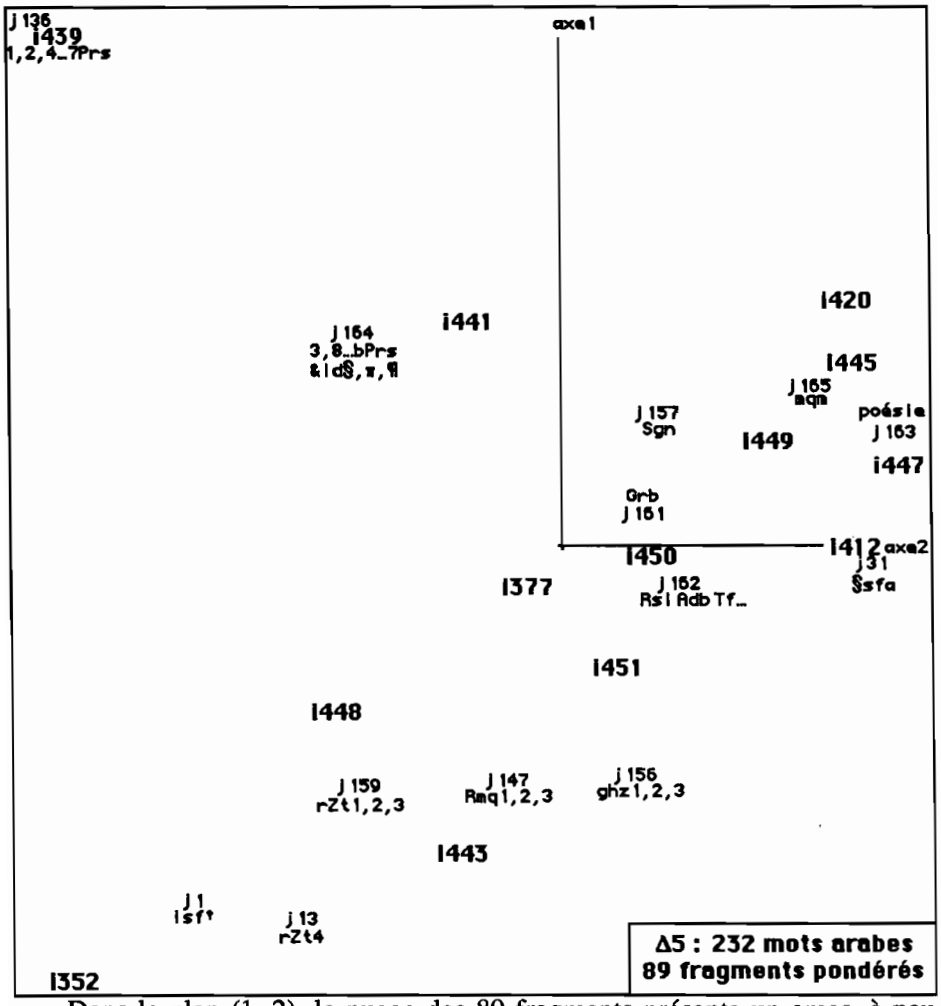


c	Partition de Δ5 en 13 classes : formes de la classe c
449	abda mᶜh fhm ant ᶜny wnḥn lna fwą whkḡa kla hwla lkm am anna
	ᶜndna anfsna swy nḥn lkn ḡak wbyn lyst ᶜndkm whm hl
420	aḡn ᶜndma wknt alyna kḡlk tḥt knt mrä ly mnḡ faḡa by nfsy ana
447	bynna lą̈d wlkn lw fla aḡa bᶜḍ wma lma flma ᶜlyk lk km ᶜndy
	wlw wfy wlm fma fą̈d
445	alan kyf kna kḡa fhl wkyf ya lst lḡlk ᶜlyha ṯm maᵓ wana bkl
	ṭwl ana ayn ḥty
412	šyya bšyᵓ flm mma lhm kanwa ᶜlyhm
450	ḡlk mnh ᶜndh ḡa bma wkl ffy lkl fmn ᶜnd wan fyh ala mnhm fyma fan
	dayma alkbyr amam wkant why lha kan kant tlk ykn lm dwn ą̈d ᶜly mn
	wla ma awla aḵr aw ᶜlyna mna tkn alaḵ fanh ᶜnha alyh wᶜly ᶜlyh ᶜn
377	kbyrä alḡyn bynhm ğmyᶜ aḥd nfsh
451	afą̈l fkan fhw alyha ᶜnh waḡa la wlys lh nfs bd ykwn bhḡh bhm
	lmn bḡyr ḥyn lanh hkḡa wama fama klha bglk bh swah hm anma ą̈bl
	whḡh anhm bha waḥda lys ḡyr whw alḡy klh an
352	fhy ḵaṣä ğhä waḥdä wmnhm
443	wą̈t bl whḡa bnfsh aḡ anh fana sayr alšyᵓ wğh šyᵓ alaw waḥd ama kl
448	wkan alty wanh bhḡa hḡa hḡh wḡlk fą̈ṭ lan anha wḡyr mṯl akṯr awl
	wanm hy wlma ay kṯyrä hw
441	kṯyr mnha kṯyra fyha aly nsbä ḥyṯ wmn wą̈d bᶜd fy
439	wḵaṣä ḵlal ban hnak byn kma ḥwl mᶜ



Nous publions ci-dessus le tableau du lexique Δ5, soumis à la CAH d'après l'analyse du tableau pondéré, Δ5 × C&M.

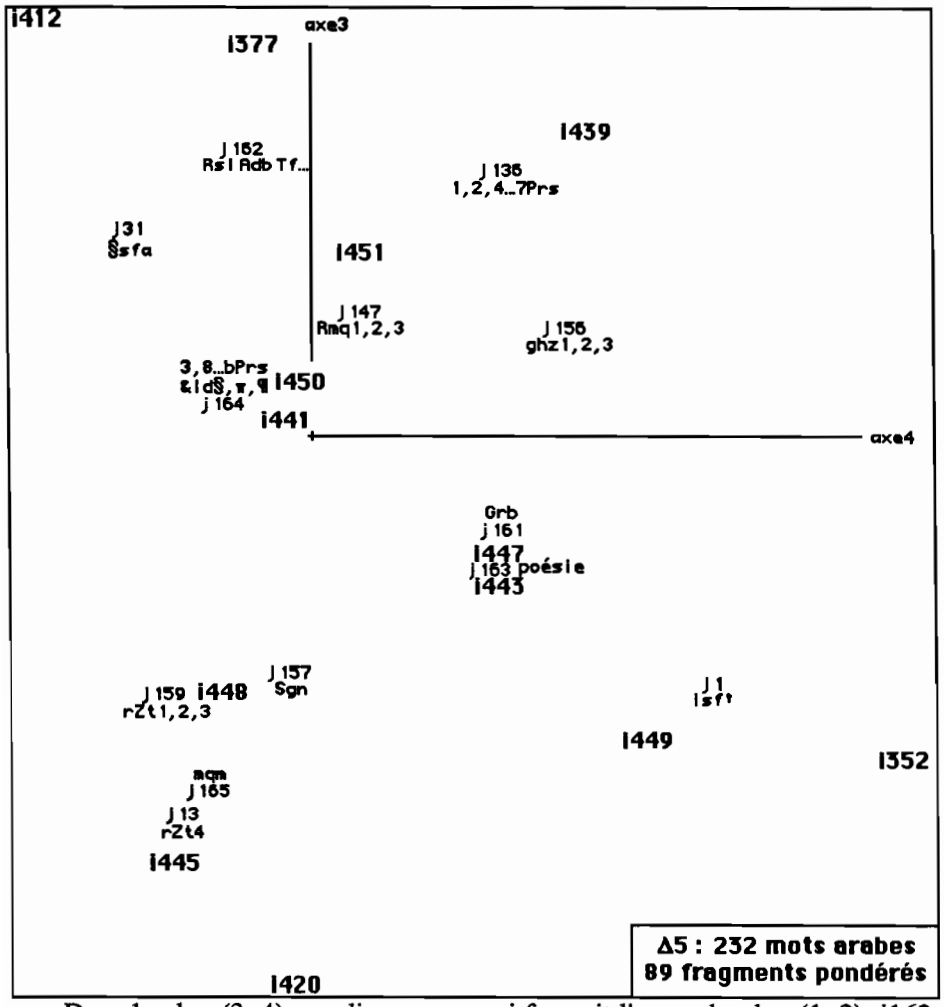
Les 13 classes de la partition retenue ont servi à étiqueter la CAH des 89 fragments; et leurs centres sont marqués, dans les plans (1, 2) et (3, 4), avec ceux des classes de fragments. Le lecteur arabisant appréciera par lui-même les associations ainsi manifestées entre formes ou locutions, d'une part;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

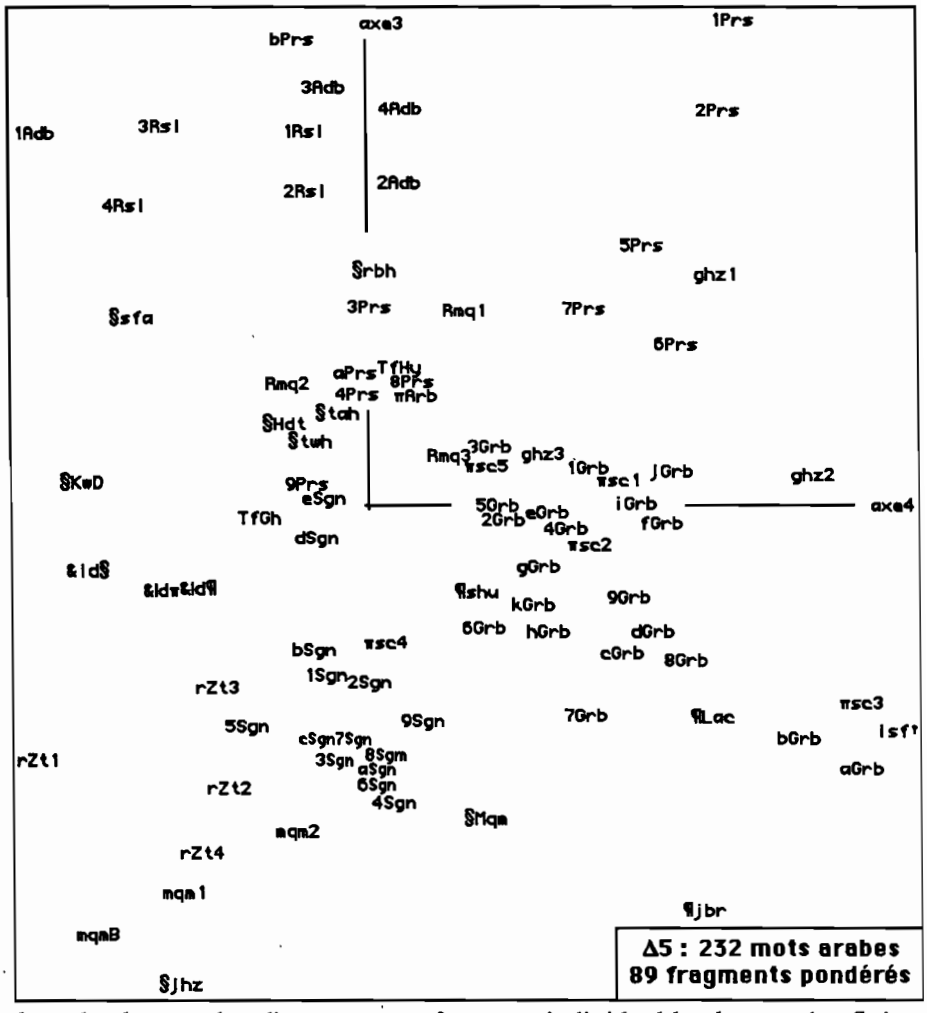
Au sein de l'amas, le facteur F2 est maximum pour la poésie classique (j169); tandis que l'adab (j162) et les œuvres modernes (Grb:j161, Sgn:j157) se projettent sur le demi-axe (F2>0) au voisinage de l'origine.

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 disitribue 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 'discri', 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:

(¶occ->mqmB)(¶dpl->8Sgn) ; (¶occ->j157) (¶dpl->j157) ;

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 Δ4 à Δ5, 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).