LES CAHIERS DE L'ANALYSE DES DONNÉES

J.-P. BENZÉCRI
A. CHABIR

Essais pour une stylométrie appliquée aux textes arabes

Les cahiers de l'analyse des données, tome 13, n° 1 (1988), p. 69-80

<http://www.numdam.org/item?id=CAD_1988__13_1_69_0>

Les Cahiers de l'Analyse des Données
Vol. XIII - n°1 - 1988 - pp. 69-80

ESSAIS POUR UNE STYLOMÉTRIE APPLIQUÉE AUX TEXTES ARABES

[STYL. ARAB.]

J.-P. BENZÉCRI
*A. CHABIR**

1 Des données au tableau analysé

1.1 Origine des données

Dans la première livraison des Cahiers d'études arabes, le professeur Gérard Lecomte publie des tableaux statistiques d'un intéret exceptionnel. Les dépouillements ont porté sur 12 Textes, ou groupes de textes, pour chacun desquels l'auteur a considéré un échantillon ramené à l'effectif normalisé de 10000 mots. Au lieu de se contenter, comme il est d'usage, de dénombrer les parties du discours en suivant une nomenclature plus ou moins classique, G. Lecomte a mis à profit le système morphologique de la langue arabe, et, non content de recenser les formes, il a tenu compte des fonctions, notamment pour les emplois des cas. Même si l'interprétation de l'analyse de telles statistiques ne peut être faite validement que par celui qui les a colligées, nous pensons que les premiers résultats de nos calculs méritent du moins d'être soumis à G. Lecomte.

1.2 L'ensemble des textes

Sans entrer dans le détail, nous énumérerons les 12 textes, ou échantillons, sur lesquels ont porté les relevés. Anticipant sur les résultats des analyses, nous rangerons les textes en quatre groupes, en conservant les numéros d'origine.

Q Poésie :

1, Préislamique; 3, Ummayade; 9, Andalouse.

H Coran et Logia:

2, Coran; 4, al-Bukhari; 5, Ibn Hanbal; 7, Ps. Ibn Qutayba.

(*) Chargé d'enseignement à l'Institut national des langues et civilisations orientales.

A Adab:

6, Adab; 8, Prose andalouse; 10, Historiens tardifs; 11, Prose moderne, (Les jours, de Taha Husayn).

J Presse:

12, deux journaux.

Dans les analyses, ces textes ont d'abord été désignés par les sigles T01 à T12. Puis, compte tenu des résultats, on a utilisé diverses lettres, (autres que T), pour rappeler le genre de chaque texte: Q, initiale de qaSIda pour la poésie; A pour l'adab; C pour le Coran; H pour les Hadiths et Ps. Ibn Qutayba; J pour les journaux.

1.3 L'ensemble des catégories de formes (*Partes Orationis Arabicæ*)

Le relevés sont rangés dans 18 tableaux, non compris un tableau récapitulatif. Sans entreprendre d'expliquer ces tableaux, nous en énumérerons les titres, en précisant les formes retenues dans une première analyse, avec, entre accolades, les sigles choisis pour les désigner.Dans ces sigles, on a suivi approximativement les conventions alphabétiques des orientalistes, adaptées par A. Chabir:]=hamza, Majuscule= emphatique ou longue, etc.

1: Verbes avec les dérivées nominaux: Les emplois des verbes sont dénombrés par formes; nous avons seulement retenu les formes I à VIII et X des Trilitères, (les autres formes étant peu fréquentes): {I, II, III,..., X}.

2: Aspects des verbes: Nous avons éliminé l'inaccompli énergique, qui est rare et très caractéristique de la langue coranique: {mADi, mrf[, mnSb, mjzm,]amr}.

3: Voix: Nous avons retenu les deux voix, active et passive; mais, afin d'éviter les doubles comptes, nous avons retranché des valeurs du tableau 3, celles données au tableau 4 pour les participes: {Actf, Psif}.

4: Participes: L'auteur distingue 4 subdivisions de base, selon qu'il s'agit de l'actif ou du passif; et de la forme I ou d'une forme dérivée (x): {pra1, prax, prp1, prpx}.

5: Masdars de formes dérivées: Sont dénombrés sans distinction de forme; les masdars de la forme I allant avec les noms: {mSdr}.

6: Substantifs, "adjectifs" et masdars de la forme I: On distingue, outre les formes quadrilitères, (Nom4), 27 balances sur base trilitère, dont nous avons retenu les plus fréquentes désignées par les sigles {Noma, Nomb,..., Nomy}.

Noma fa[Il(a)	Nomd maf[il(a)	Nomg fa[l	Nomh fa[la
Nomi fu[l	Nomj fu[la	Nomk fi[l	Noml fi[la
Nomm fa[al	Nomn fa[ala	Nomo fa[Al	Nomp fa[Ala
Nomq fi[Al	Nomr fi[Ala	Noms fu[Al	Nomw fa[Ul(a)

7: Pluriels: Diverses balances, dont nous avons retenu neuf, notées {Plua, Plub...}.

Plua	mafA[il	Plub	mafA[Il	Plud	af[Al
Pluh	fu[Ul	Pluj	fu[ul	Plul	fi[Al
Pluo	fu[al	Plux	-Un	Pluy	-At

8: Thèmes nominaux à vocation qualificative ("adjectifs"):

On distingue entre épithète et attribut: {na[t, khbr}.

9: Démonstratifs: Proche et Lointain: {hic, ille}.

10: Pronoms personnels: Sujet, CD après verbe, autre CD, CI après préposition, CI après nom: {hwa, huv, hux, hi, hu}.

11: Prépositions: {bi, min, li, fI, [alA, ilA, [an, ma[a, [ind}.

12: Particules de coordination: {wa, fa, tumm, aw}.

13: Particules de subordination: {]an,]ann,]ida,]in, law, lamm, llad(relatif)}.

14: Particules négatives: {mA, lA, lam, illA}.

15: Particules à valeur affective: {yA!, la!, qad}.

16: Emplois du nominatif: Sujet et Attribut: [Sujt, Attr}.

17: Emplois du cas direct: Complément direct, attribut de kân, après cas direct,.., Vocatif: {Cdir, CkAn, CCdr, CHAl, Cprp, Cloc, C!!}.

18: Emplois du cas indirect: Après préposition, (Harf) et en annexion: {jrrH, jrrD}.

Au total, on a retenu un tableau comprenant 95 lignes, (*partes orationis arabicæ*), et 12 colonnes, (textes ou échantillons de textes).

2 Les méthodes statistiques

2.1 L'analyse des correspondances

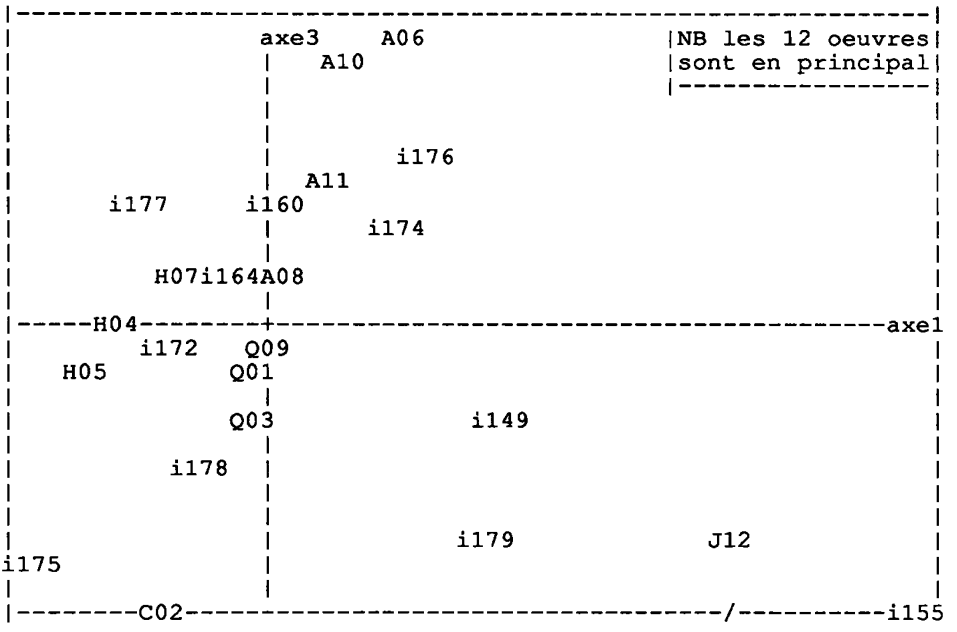
Comme dans de nombreuses études antérieures, le tableau de base a été soumis à l'analyse des correspondances. Outre des tableaux numériques, certes essentiels, mais que le linguiste non entrainé ng consulte pas volontiers, la

méthode fournit des graphiques plans sur lesquels figurent à la fois les lignes et les colonnes, représentées par leurs sigles.

Sur ces graphiques, on a, entre les éléments d'un même ensemble, des proximités d'autant plus grandes que, par leur composition en pourcentage, ces éléments sont plus voisins. Par exemple, on attend que dans H04 et H05, qui sont deux recueils de Hadith, les parties du discours soient employées dans des proportions similaires; ce qui est en effet le cas. Il est moins clair, a priori, pour le profane que fa et thumma ont proportionnellement des répartitions voisines, (même si fa est 11 fois plus fréquent que tumm).

De même, les points représentatifs d'un texte et d'une partie du discours sont proches l'un de l'autre si, en bref, dans le texte la fréquence d'emploi de cette partie est relativement élevée.

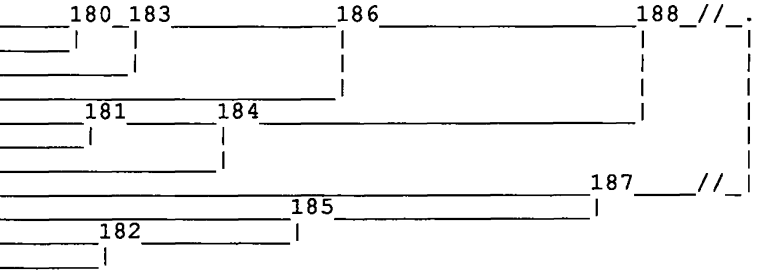
Il faut cependant noter que la consultation des graphiques, pour directe qu'elle soit, offre quelque difficulté. En effet, un seul plan ne peut suffire à rendre compte du système complexe des proximités entre textes et parties du discours: en réalité, il s'agit d'une structure spatiale, multidimensionnelle, qui n'est complétement décrite que par ses projections sur plusieurs plans, comme c'est le cas pour un dessin technique. D'autre part, il n'est pas facile de consulter un graphique où figurent les 95 parties du discours. Dans ces conditions, la classification complète utilement l'analyse factorielle.



2.2 La Classification Ascendante Hiérarchique, (CAH)

La CAH fournit une représentation arborescente de chacun des deux ensembles, textes et parties du discours, fondée sur la même réalité spatiale que pour l'analyse factorielle. En bref, les points les plus proches sont agrégés entre eux pour former de petites classes qui elles mêmes s'agrègent en classes plus grandes. On a retenu pour les parties du discours une partition en 11 classes; sur les graphiques plans, au lieu de représenter les 95 éléments, on a seulement figuré les centres des classes. Ainsi, la lecture des résultats se fait en deux étapes: par exemple, la CAH nous apprend que mSdr, Na[t, Pluo (pluriels en fu[al) constituent la classe i155; et sur le graphique plan on voit que cette classe est très caractéristique de la presse (J12): ce qui ne surprend pas, au moins pour les épithètes, multipliées à l'exemple des langues occidentales.

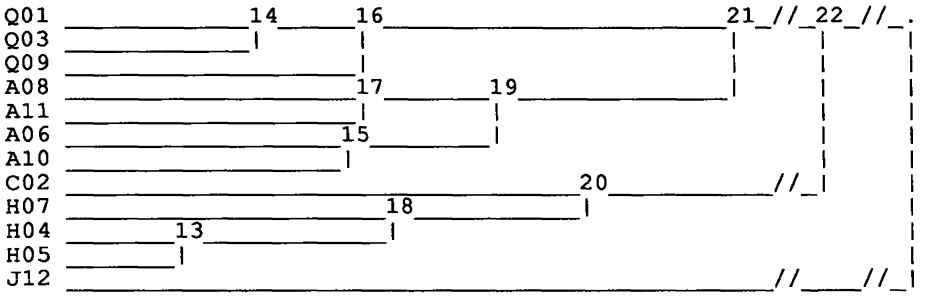
```
=============================================================
  c | Partition en  11 classes : Sigles des individus de c
-------------------------------------------------------------
 172| I Actf Cloc CCdr hux mADi Ille huv hi
 178| Cdir Cprp qad ]ida mrf[ CHAl IV prax ]in mjzm
 175| yA! mA  C!! la! law ]amr Attr Khbr
 177| aw lA Nomw  hwa [alA illA ]an mnSb ma[a lamm tumm fa
 164| hu ]ann Noms Psif Sujt Nomg Plul
 160| [an Nomk Nomp Nomi Nomo
 149| Nom4 prpx Pluh Plua
 155| Pluo Na[t mSdr
 179| Plux llad Pluy VIII X VI pral Pluj II Nomd [ind bi III
    | li
 174| Plub CkAn lam Nomn VII ilA  min fI Nomh Hic
 176| Plud Nomj Nomm wa Nomq Noma JrrH V JrrD Nomy Nomr prp1
    | Noml
=============================================================
```



ci dessus l'arbre de la partition des parties du discours en 11 classes, (étude avec la Presse moderne)

Pour interpréter avec toute la précision souhaitable une classification ascendante hiérarchique, il ne suffit pas de consulter les graphiques plans donnant la représentation simultanée des textes et des classes de parties du discours: il faut recourir à des tableaux de nombres, notamment aux divers

listages Vacor d'aide à l'interprétation. Ces listages donnent, pour chaque classe, les éléments, (ou les classes), de l'autre ensemble par lesquels elle est caractérisée, soit du fait d'un excès, soit du fait d'un défaut. Les principales informations contenues dans ces listages peuvent servir à étiqueter l'arbre de la CAH, comme on le fera au §3.2.



ci dessus la CAH des œuvres, avec la Presse moderne

3.0 Enchainement des analyses

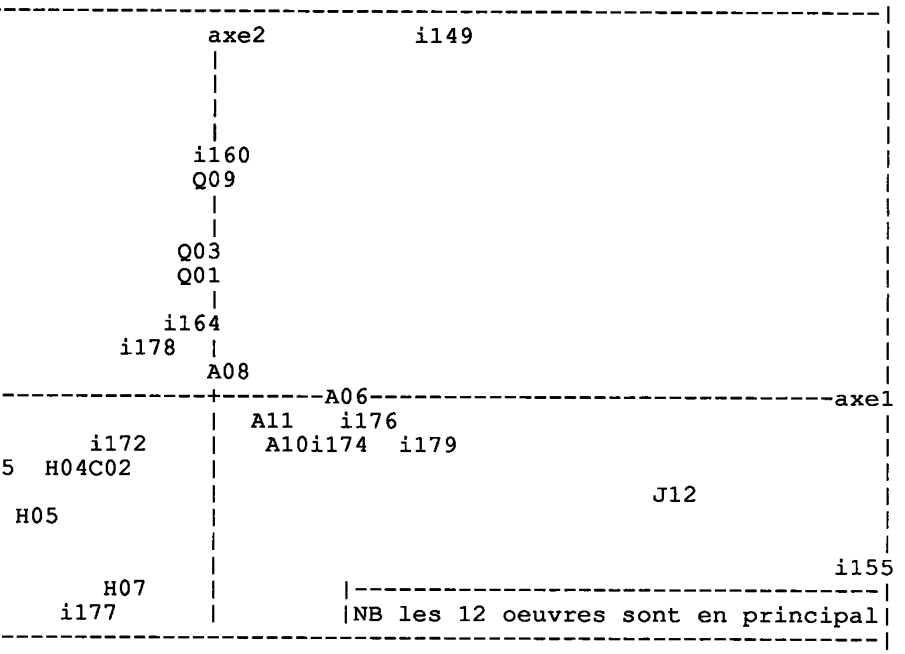
Une première analyse, suivie de CAH, a porté sur l'ensenble des 12 textes: l'opposition entre la presse d'une part et l'ensemble des textes littréraires d'autre part, domine cette analyse, au point de dissimuler, dans une certaine mesure, la diversité des textes littéraires. On a donc repris l'analyse en mettant en supplémentaire la presse, (J12). C'est à dire que l'analyse a été faite sans J12, qui a toutefois été projeté, a posteriori, sur les graphiques.

3.1 Analyse sur l'ensemble des 12 textes

Partons du plan (1,2): sur l'axe 1, (axe horizontal), J12 est très écarté à droite, suivi à bonne distance de la prose littéraire (adab etc). Quant aux parties du discours, on a du même côté, outre la classe i155, très excentrique, déjà signalée, les classes i179, i174, i176 qui contiennent notamment toutes les formes verbales à l'exception de I et IV. La prose andalouse, c'est à dire A08, le *Collier de la colombe*, se place très proche de l'origine des axes, comme s'il représentait une forme moyenne de la langue arabe: ce qu'on peut expliquer partiellement par le caractère composite de ce texte qui compte des fragments de poésie.

La classification des textes est très satisfaisante: la presse s'oppose d'abord à tout le reste; même si, comme on le voit à l'analyse factorielle, elle est plus proche de l'adad que des autres textes. Ensuite, le Coran, associé aux Hadiths et à H07, s'oppose aux textes littéraires proprement dits; et ceux-ci sont partagés en prose et poésie.

A quelques remarques près qui semblent s'imposer, la classification des *partes orationis* doit être laissée au professeur G. LECOMTE. On constatera seulement que les nuances qu'il a introduites dans le dénombrement des formes nominales sont, *a posteriori*, pleinement justifiées par leur répartition éclatée dans toutes les classes de la CAH. (Voir aussi, *in fine*, le *Post Scriptum*).



3.2 Analyse sans la presse moderne

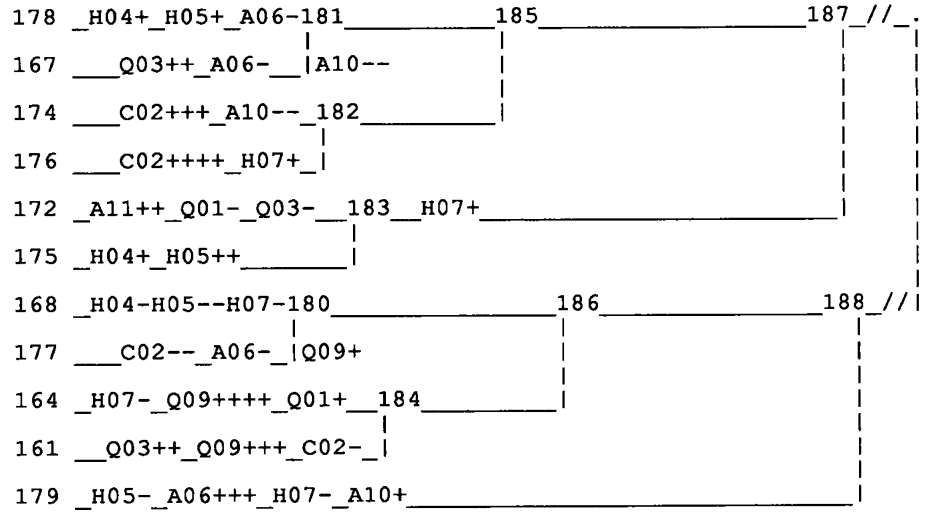
Dans l'ensemble cette analyse s'accorde avec la précédente quant aux proximités qu'elle montre entre textes. Mais pour les parties du discours, nous présumons qu'elle est bien plus intéressante, car les formes affectionnées par la presse sont maintenant réparties selon leur place dans la tradition littéraire.

Il vaut la peine d'expliquer comment s'interprète l'étiquetage d'un arbre Considérons d'abord la classification des parties du discours. Commençons par la classe i178: celle-ci contient les deux voix Actf et Psif, les formes verbales I et III, l'aspect accompli, (mADi); les pronoms personnels au cas direct, après verbe ou ailleurs, (huv et hux); le pronom personnel au cas indirect après préposition, (hi); le nominatif sujet; le cas direct employé comme circonstanciel ou après cas direct, (Cloc, CCdir); la préposition li:

i178 = {I, Actf,huv,Cloc,CCDr,hux,mADi,Psif,Sujt.III,hi,li}.

Cette classe, i178, est employée avec une fréquence élevée dans les deux recueils de Hadiths; d'où les mentions H04+ et H05+; et elle est relativement peu employée par Ibn Qutayba: A06-.

```
=============================================================
  c | Partition en  11 classes : Sigles des individus de c
-------------------------------------------------------------
 178| I Actf  huv Cloc CCdr  hux mADi Psif Sujt III hi li
 167| Cdir la!  law  qad Cprp ]ida
 174| ]in IV Pluj CHAl mjzm prax mrf[ Pluy llad
 176| Plux Khbr ]amr Attr
 172|  Hic  ]an mnSb ma[a lamm  ilA
 175|  yA!   mA  C!! [alA  hwa illA aw lA Nomw tumm fa
 168| Noma Nomq  X II hu [ind bi  min  V VIII Plul Na[t pral
 177| Nomh   fI Plub CkAn  lam Nomn  VII ]ann Noms Nomd Nomg
 164| Nomp  [an   VI Nomk Nomi Nomo
 161| prpx Nom4 Pluo Plua Pluh
 179| Nomr prp1 Noml Nomy Plud mSdr Nomj Nomm   wa Ille JrrH
    |JrrD
=============================================================
```



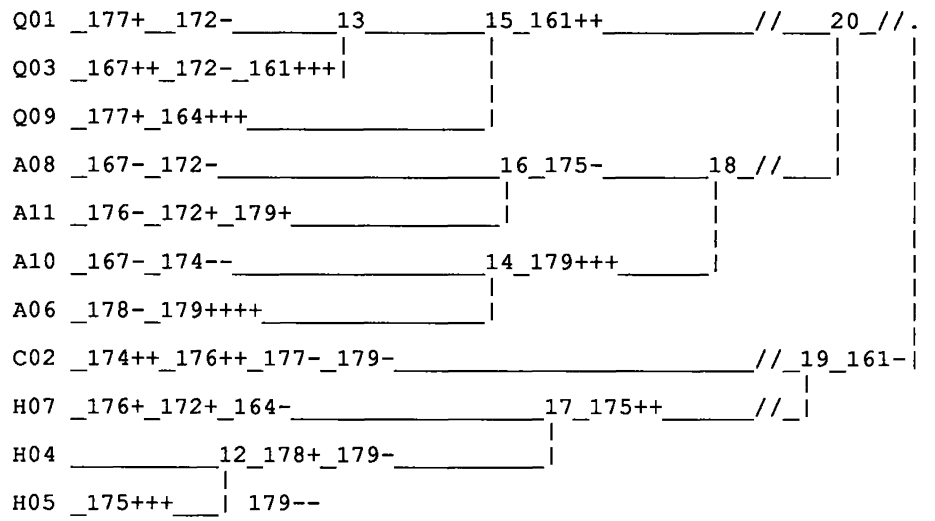
ci dessus l'arbre de la partition des parties du discours en 11 classes, (étude sans la Presse moderne)

NB On prendra garde à ce que, dans la présente classification, les mêmes numéros sont attribués à des classes autres que celles considérées au §3.1.

Les classes i174 et, plus encore, i176 sont d'un grand emploi dans le Coran: C02+++, C02++++; on ne s'étonnera pas de trouver dans ces classes,

l'impératif,]amr, et l'inaccompli apocopé, mjzm. D'autre part, la classe i174 est nettement évitée par les Historiens tardifs, A10--; et i176 est plutôt affectionné par le Pseudo Ibn Qutayba, H07+.

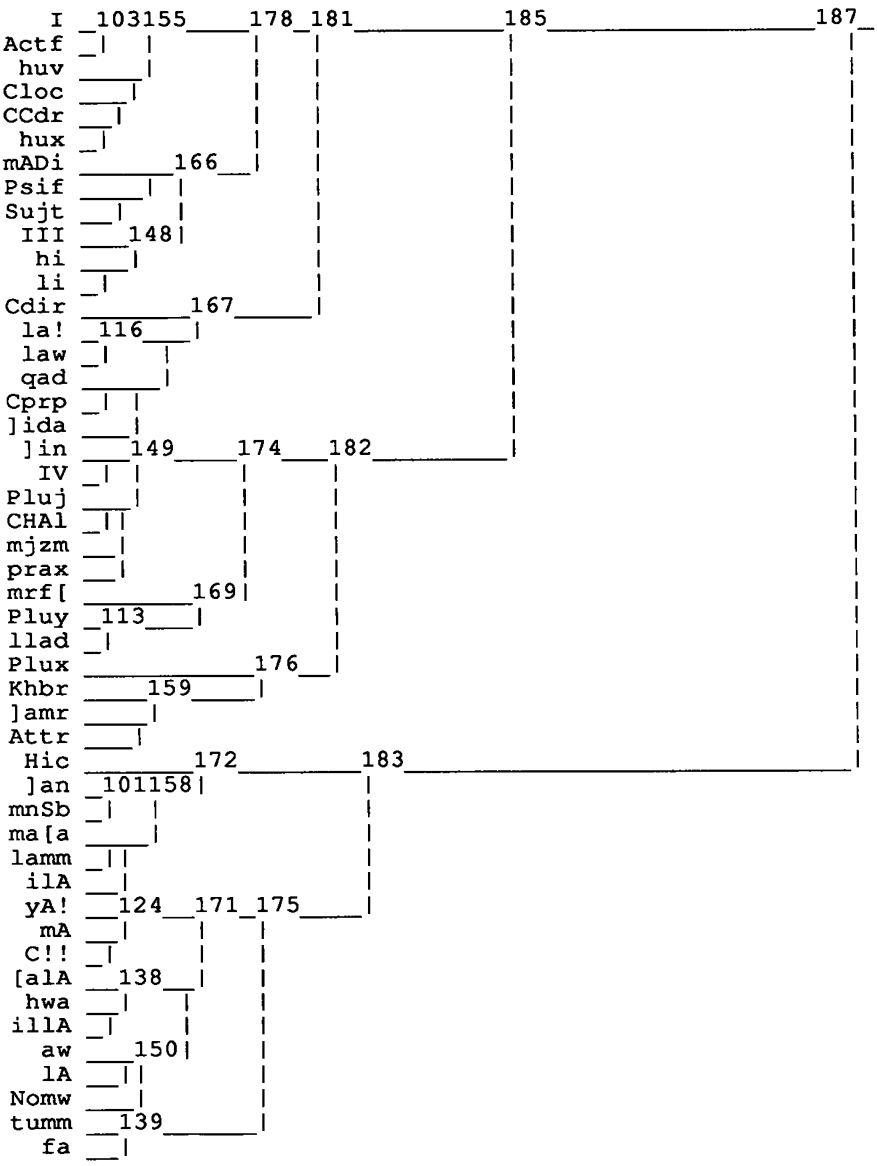
NB Le détail de la classification des parties du discours est donné sur deux pages, sur chacune desquelles est représentée l'une des deux classes, 187 et 188, en lesquelles est scindé l'ensemble I.



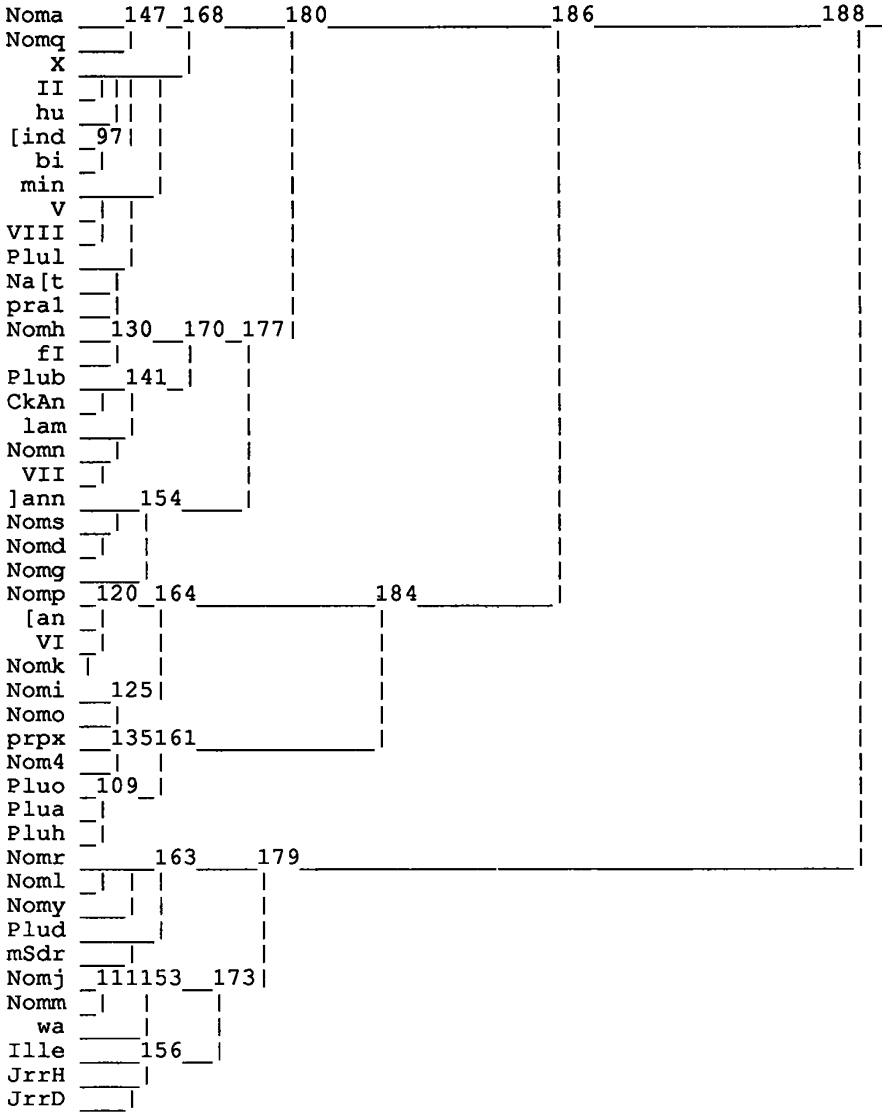
ci dessus l'arbre de la CAH des œuvres, (étude sans la Presse moderne)

On interprète de même l'étiquetage de la CAH des œuvres. On lit, par exemple, sur la ligne C02, les mentions 174++, 176++: ce qui nous rappelle la prédilection du Coran pour les parties du discours rangées dans les classes i174 et i176. Les mentions portées sur les deux arbres ne sont toutefois pas exactement les mêmes: en effet, sur la CAH des *partes orationis*, on porte les affinités qui sont le plus caractéristiques des classes i1178, i167, etc; tandis que pour étiqueter l'arbre des œuvres, on choisit les affinités le plus caractéristiques de celles-ci. De plus, sur la ligne C02 on lit les mentions négatives 177-, 179-. Enfin, la mention 161-, inscrite à droite du nœud 19, vaut pour les quatre œuvres comprises sous celui-ci; c'est à dire d'une part le Coran C02; et d'autre part les Hadiths et le Ps. Ibn Qutayba: H04, H05, H07.

ci dessous l'arbre de la CAH générale des parties du discours, (étude sans la Presse moderne) : représentation de la classe 187



ci dessous l'arbre de la CAH générale des parties du discours, (étude sans la Presse moderne) : représentation de la classe 188



Référence bibliographique

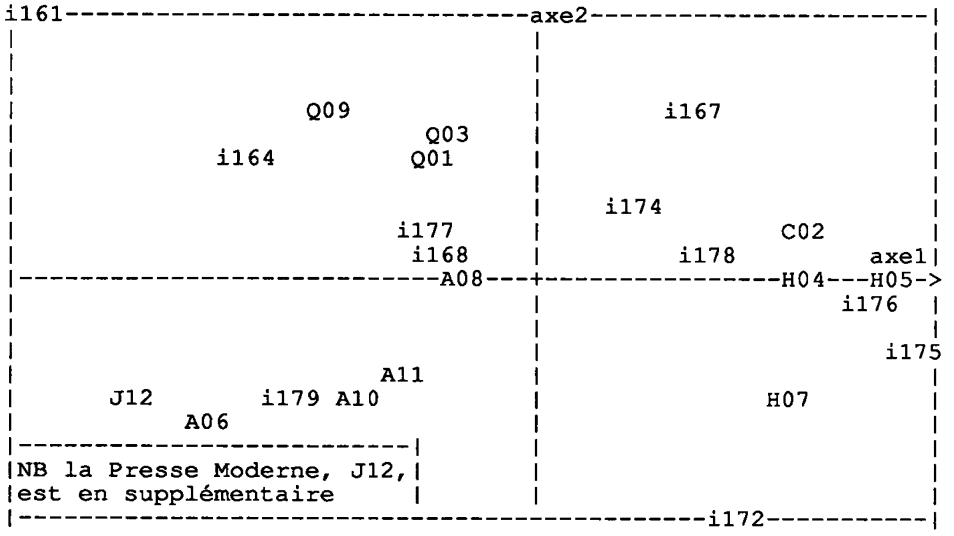
Gérard LECOMTE, "Éléments pour une stylométrie appliquée aux textes arabes", *Cahiers d'études arabes,* publiés sous les auspices de l'Institut national des langues et civilisations orientales, N°1, Paris 1987.

Post Scriptum

Une épreuve du présent travail a été soumise, avant publication, à Monsieur Auguste FRANCOTTE, Administrateur du Comptoir d'Escompte de Belgique et orientaliste distingué, qui nous a communiqué la remarque suivante:

Feu Philippe Marçais, qui fut mon maître, avait le sentiment que, dans l'arabe contemporain, la construction kAna avec complément au cas direct progressait par rapport à la phrase nominale...

Effectivement, dans l'étude avec la presse moderne, CkAn se place dans la classe i174, laquelle est du côté positif de l'axe 1 où l'on trouve, très à l'écart, la presse moderne J12, suivie de l'adab, y compris Taha Husayn; tandis que Attr, (attribut au nominatif de la phrase nominale) est dans i175, à l'extrémité négative de l'axe 1, avec le Coran et les logia.



De même, dans l'analyse où J12 est en supplémentaire, Attr est dans la classe i176, et CkAn dans i177; et l'opposition entre i176 et J12 s'accorde avec l'intuition de Philippe Marçais.