LES CAHIERS DE L'ANALYSE DES DONNÉES

J.-P. BENZÉCRI
F. BENZÉCRI

Analyse du vocabulaire et recherche du thème dans les articles des volumes XII à XVII de CAD. (3) Typologie et discrimination

Les cahiers de l'analyse des données, tome 18, n° 1 (1993), p. 75-96

<http://www.numdam.org/item?id=CAD_1993__18_1_75_0>

Les Cahiers de l'Analyse des Données
Vol XVIII - 1993 - n°1 - pp. 75-96.

ANALYSE DU VOCABULAIRE ET RECHERCHE DU THÈME DANS LES ARTICLES DES VOLUMES XII À XVII DE *CAD* (3) TYPOLOGIE ET DISCRIMINATION

[*CAD* XII-XVII (3)]

J.-P. & F. BENZÉCRI

5 Analyse factorielle et classifications sur le corpus de 191 articles croisés avec quatre lexiques principaux

Nous considérons les quatre lexiques dans l'ordre {V, Pl, PlR, XR}, qui est celui dans lequel ont été effectuées les analyses; et est l'inverse de l'ordre d'importance suivi, au §3.2, pour présenter les lexiques.

5.1 Analyses fondées sur un lexique V de 283 mots vides

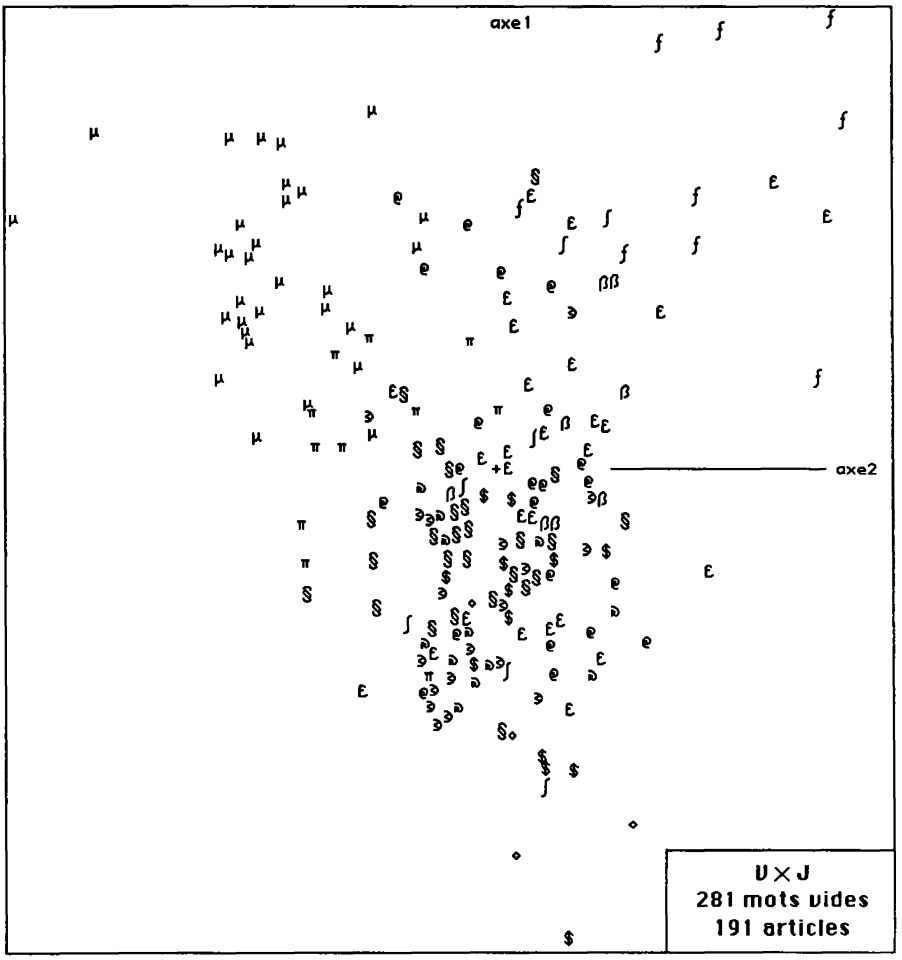
D'une telle analyse, d'après l'expérience acquise dans l'étude des textes littéraires, on attend, a priori, une typologie stylistique. Reste à voir ce que peut être le style dans nos articles. Il apparaît qu'ici, le style interfère avec le genre et le thème, non sans déceler la marque de certains auteurs.

291 mots de V × 191 articles de CAD

trace :	3.504e-1									
rang :	1	2	3	4	5	6	7	8	9	10
lambda :	288	218	136	103	92	79	72	67	63	61 e-4
taux :	821	621	387	295	262	225	207	190	180	175 e-4
cumul :	821	1442	1829	2124	2386	2612	2818	3008	3189	3363 e-4

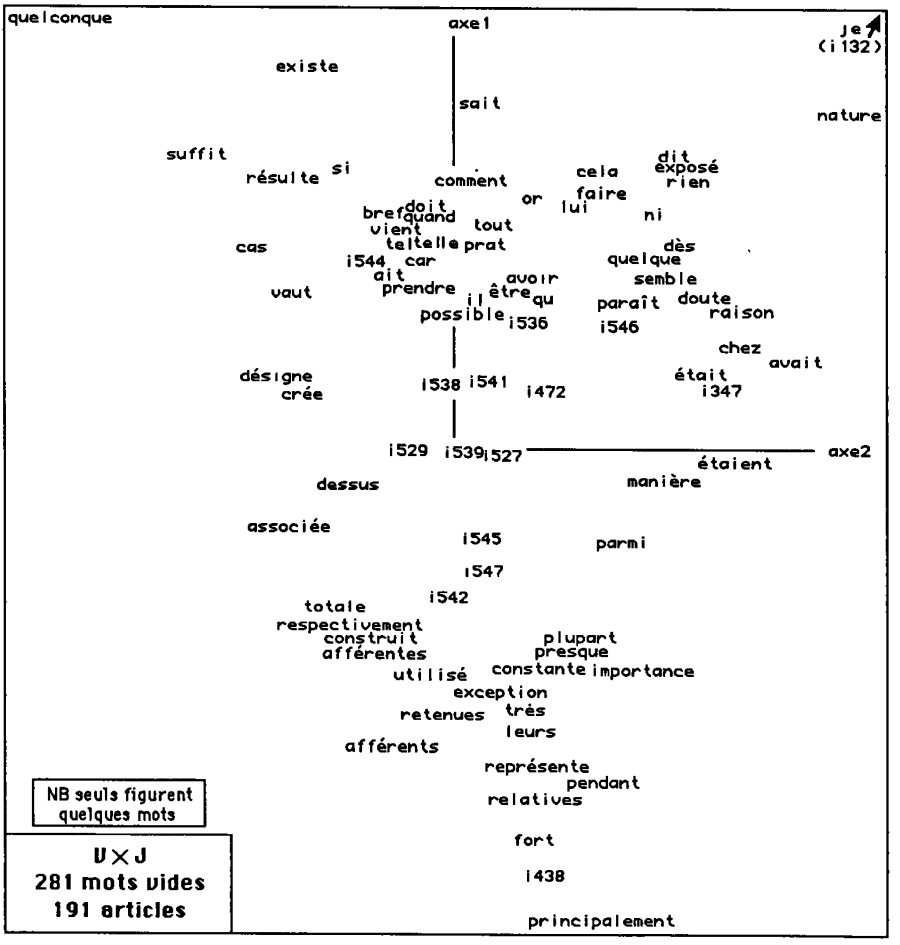
Ainsi, dans la partition en 15 classes définie par les 14 nœuds les plus hauts, un tiers environ des articles se trouvent compris dans des classes aisément interprétables. Il s'agit principalement des subdivisions de la classe j380. Considérons, d'abord, le plan (1,2), nettement distingué par une dénivellation dans les taux d'inertie entre les facteurs 2 et 3. Le nuage J des articles y dessine un triangle; dont la pointe est en (F1<0); tandis qu'adossé au côté opposé, j380 occupe, à peu près, le demi-plan (F1>0); et c'est suivant l'axe 2 que j380 se subdivise en j370 et j378. Les articles étant marqués, chacun, par la première lettre de son sigle, on en voit aussi bien le contenu sur le plan que dans le tableau de la CAH.

Le quadrant (F1>0;F2<0) contient un amas de 'μ' figurant les articles de mathématiques qui composent la classe j370. Il n'y a dans j370 que deux articles



non mathématiques: {®Vin F1<0; @ngr F1≈0}, mais ceux-ci ne sont pas à l'intérieur du quadrant; et seuls six des articles dont les sigles comportent la lettre 'μ' manquent dans j370: encore ne s'agit-il que d'articles dont le thème majeur est l'écologie, (@μCt); la démographie, (©μar); la psychophysique, {@μch, @μCh}; ou la phyqique, {μEPR, μEnt}.

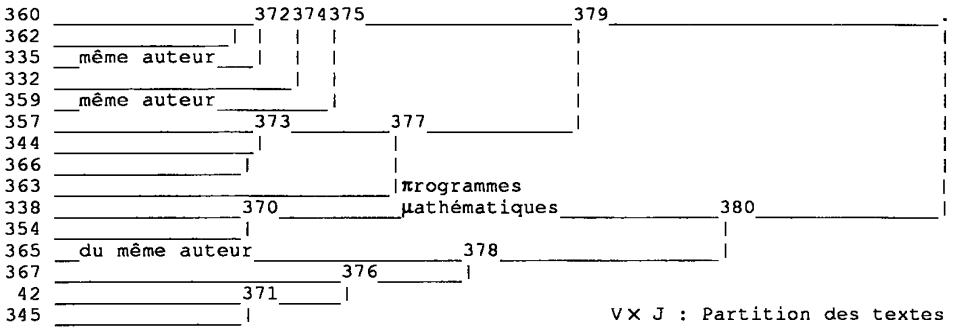
L'étiquetage de la CAH par VACOR, signale les classes de mots dont use à profusion la dialectique du mathématicien: i544, i541, voire i538; et les plus écartés des ces mots se lisent dans un plan, à la périphérie du nuage des centres de classes.



Dans le quadrant (F1>0;F2>0), avec la classe j378, c'est une autre dialectique, fondée sur i536, que nous offrent des articles dont le thème principal est souvent la philosophie, '*f*' (aucun sigle marqué de l'initiale '*f*' n'est en dehors de j378), ou la linguistique, '£'; avec des considérations générales sur la méthode.

Quant au détail, on notera la classe j365: trois articles où le même auteur principal, Chr. RUTTEN, suit, avec minutie, l'expression de la pensée d'Aristote dans les textes. Dans j371, {*f*Prê, j345}, une fréquence élevée de "je" est la marque, en général, de la rédaction d'un exposé oral. Dans *f*Prêt, le temps

```
 c | V× J : Partition en 15 classes : Sigles des textes de la classe c
---+-------------------------------------------------------------------
 360| ◊fbr ◊Oxy ∫tkl                                          547+++++
 362| £Aur £prl £frL £Dia @chê ◊Loi @Ths ®rch @str           547++
 335| @stg @Src              347+                             547++
 332| ©Grc ©Pét              542+++
 359| $Lé2 $Lé1              545+++  438+++                   547++
-----------------------------------------------------------------------
 357| ®Imm $for $cmp ©Mr∫ ©Mrt $$Nr ◊vir ©Ent ©E1z $Arb ®$Vi ®Caf ®Cao ®étu
    | £Rep $&DA $@Br $@ct πPCc £©F© £Chn ®$Qb $$âg $ADN ©Sal ©$Tr ®Exm ©pub
    | $Esp ®com                                        547+
 344| $MzT ®Arb £Tr2 ©vgn $Cmc @rch £SMF $$pr @Jrd ©Pnt $Sé2 $Sé1 $cdr ©$SF
    | £Par ©1Pr @prf ßJap £HxG @Mor @Alg £NT2 £NT1 £Lat £ñOr
 366| $ECG $sST ∫tki $rét @Prd πpop @man @µCt ∫tk2 $Se4 $Tnd $hm2 $hm3 $d&d
    | $@Bq $usi £Arb $eff ©Cal $hm1 $&nx $gel $2sé $HLT ©Par £Gns ©Chô $SeH
    | $Pra π©dg ©stt ßSIC ®Thr ©µar £cry $gên @pGr @&Ps £mot $&hX ∫ana
-----------------------------------------------------------------------
 363| π©rt πCor πInd πlin πprs πCum £Tr1 πCré     538+++
=======================================================================
 338| µbl2 µBl3                                          544+++++ 541+
 354| µbl1 µnq0 µMlt µdéc µnqT µInB µrec µInM µPrt µGut µblð µRel ®Vin µQs1
    | µcah µvoi µbrd µΔel µméd @ngr µVAC πµdé µinh µprs µSim µbar µniv µfus
    | µCv2 µCv1                                      538+ 544+++   541+
-----------------------------------------------------------------------
 365| fAÉc £Mt2 £Mt1                546+++                    541++++  536++
 367| fOmn fusu fqal fLgS ßFrn ßcib ßvoi @Goo ©Alg ∫Imm ßpôl ßHKg $not ∫exp
    | @fum @µch £sty @µCh £etr £ibr µEPR @&ps fOrd µEnt @Hst $Imp £prl £vox
    |                                              538+          541+     536++
  42| fPrê                    347+++              je++                    536++
 345| fðia fAri £hxL ßMrt ∫ðia                    je+++           541++    536++
---+-------------------------------------------------------------------
```



imparfait, i347={avait étaient était}, sert à mettre en perspective des étapes successives de la pensée d'Aristote.

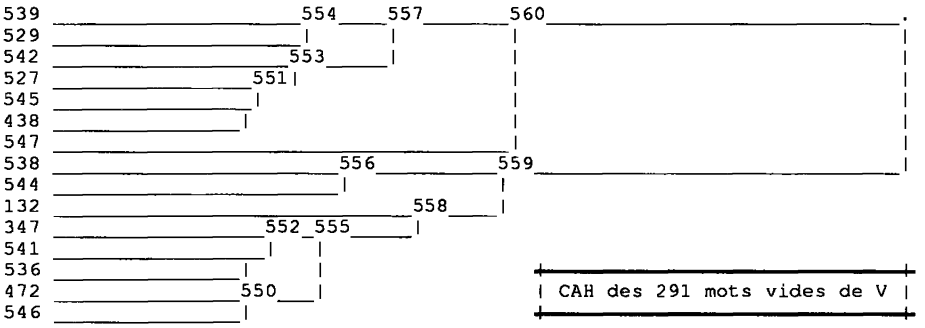
Après la branche j380, étalée dans le demi-espace (F1>0), il reste à considérer la branche complémentaire j379. Se signale la subdivision j363, où, partout (y compris dans £Tr1), il s'agit d'expliquer les fonctions de programmes ('π'). Approximativement renfermée dans le quadrant (F1>0;F2<0) (avec les mathématiques), j363 se détache véritablement dans (F4>0;F6<0). Les autres subdivisions comprennent des études de cas rentrant dans les domaines de

```
 c | VX J : Partition en 15 classes : formes de la classe c
---+-------------------------------------------------------------------------
 539| à mettre mêmes celles ceux autant celui ci selon autre part en soumis
    | suivi ayant le compris présent outre fois dans tous ainsi déjà dont mais
    | au avec de
 529| afin ici considérer parce même notera peine vaut exactement noter étant
    | on pour suivant considère tenu considéré respecivement chacun toujours
    | conclusion effet précise façon montre sein entre sur
-------------------------------------------------------
 542| pendant totale contre
 527| exception comprend représente plupart presque role constante attention
    | introduction aussi qui la
 545| terme vers du et construction importance ses également toutes leurs
    | autres relative relativement grande grand très
 438| principalement fort
-------------------------------------------------------------------------------
 547| contraire telles quel diversité associé mal près détail peu va quant
    | nettement particulièrement afférentes afférents enfin pu diverses divers
    | présente tandis aux donne construit chaque chacune des nombreux
    | certain certaines notamment retenues remarque utilisé les différentes
    | ces sont relatives été ont elles seront leur
===============================================================================
 538| ou partir peuvent permet par crée jusque eventuellement comporte choisi
    | bref prendre faut vient ensuite puis une un tels directement laquelle
    | paraît serait pour doit être peut
 544| désigne associée suffit résulte quelconque cas alors puisque précisément
    | ceci sera donc quand sait tout toute tel car soient donné telle soit
    | existe différents si
-------------------------------------------------------------------------------
 132| je
-------------------------------------------------------
 347| avait étaient était
 541| quasi non particulier dessus maintenant où que est comme se après
    | cependant fait cet cette dire avoir donner faire ce généralemnt apparaît
    | plutôt quelques assez voir delà qelle beau pourquoi moins fin celle bien
    | or nos lesquelles met mieux là semble
 536| ne pas ait eux encore seulement qu aucune aucun seule pris suit seul
    | sans possible pratique comment il doute manière souvent ils raison elle
    | ni sa lui
-----------------------------------------
 472| avons nous
 546| nature chez constitue dès lors cela parmi agit vu toutefois font
    | lesquels objet ailleurs rien quelque notre lequel prend dit exposé
---+-------------------------------------------------------------------------
```

l'économie, de la géographie, de la médecine et parfois de la linguistique. Ici encore, on peut signaler l'agrégation d'articles dus à un même auteur: j359={$Lé2 $Lé1}, j335={@stg @Src}; et se rapportant au même sujet ou à des sujets similaires.

Quant à la CAH des mots, nous invitons le lecteur à remarquer, dans le tableau du contenu des classes, des couples de formes consécutives (donc vraisemblablement agrégées à un bas niveau) qui relèvent d'un même paradigme, ou composent une locution usuelle. Dès la première ligne, on a:

{... celles ceux ... celui ci ... autre part ...};



dans i529, {notera peine vaut ... noter}, évoque la locution "il vaut la peine de noter..."; dans i547 on a les paradigmes {afférentes afférents ... diverses divers ... chaque chacune ... certain certaines ...}; et dans i544 {tout toute}. On a déja noté l'imparfait de l'indicatif dans i347: {avait était étaient}.

5.2 Analyses fondées sur un lexique Pl de 367 mots pleins

```
corr(axp1,axP1) = .92 ; corr(axp1,aXe1) = .91 ; corr(aXe1,axP1) = .93
corr(axp2,axP2) = .89 ; corr(axp2,aXe2) = .91 ; corr(aXe2,axP2) = .88
corr(axp3,axP3) =-.84 ; corr(axp3,aXe3) = .81 ; corr(aXe3,axP3) =-.77
```

N.B. *On a noté respectivement* axP, axp, aXe, *les facteurs issus des analyses des* §§5.2, 5.3 et 5.4; *les corrélations sont calculées sur l'ensemble des* 191 *chapitres.*

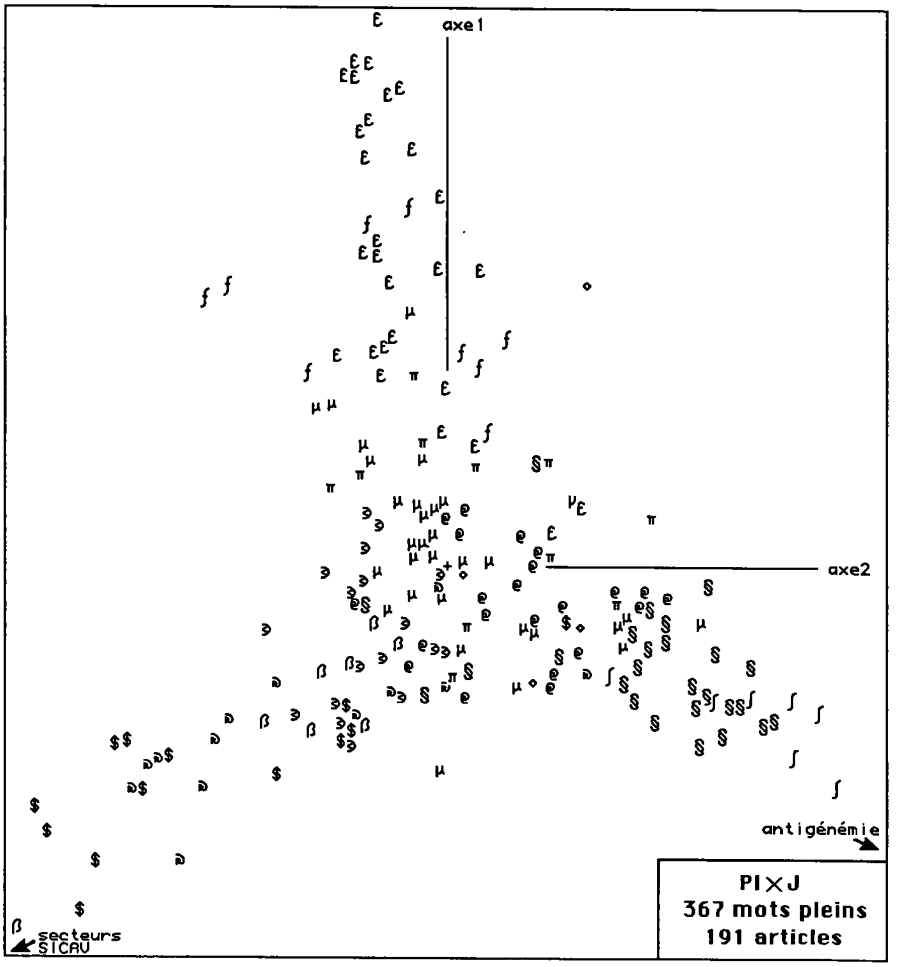
Dans les §§5.2, 5.3 et 5.4, on rend compte d'analyses fondées sur un lexique composé exclusivement, ou principalement, de mots pleins: le tableau ci-dessus atteste qu'il y a, entre ces analyses une grande similitude

```
367 mots de Pl x 191 chapitres de CADXII-XVII (tableau écrêté)
trace   :  7.536e+0
rang    :     1     2     3     4     5 ..  10 ..  15 ..  25 ..  50 .  75 ...
lambda  :  3388  3200  2754  2450  2145  1522  1198   780   456   286  e-4
taux    :   450   425   365   325   285   202   159   103    60    38  e-4
cumul   :   450   874  1240  1565  1849  3018  3869  5113  7071  8270  e-4
```

Sans prétendre être exhaustif nous donnerons d'abord quelques résultats issus du tableau croisant le lexique Pl avec l'ensemble des 191 chapitres (ou articles).

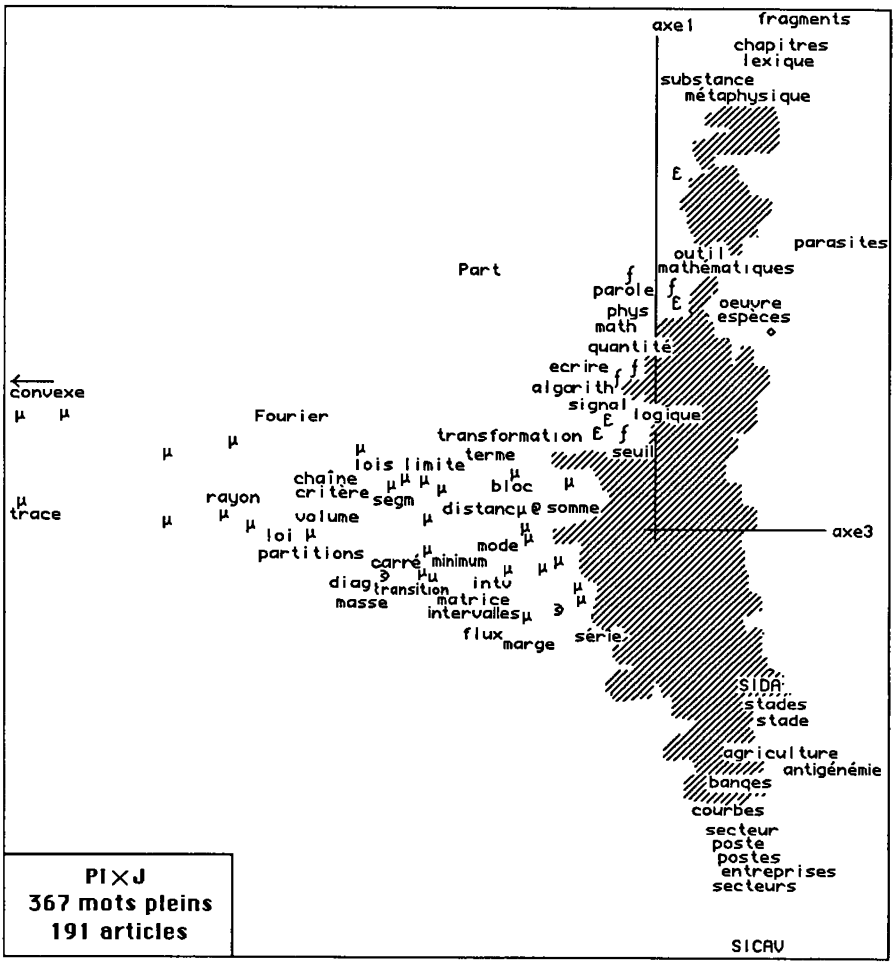
Les valeurs propres sont dix fois plus fortes que dans l'analyse du §5.1, fondée sur un vocabulaire V de mots vides (une semblable remarque a été faite au §2, à propos des analyes des résumés fondées sur différents lexiques). La décroissance des taux est à peu près la même; à ceci près que le plan (1, 2) n'est pas nettement signalé par une forte dénivellation entre F2 et F3.

Dans l'espace engendré par les axes 1 à 3, le nuage des chapitres (ainsi que celui des mots) affecte la forme d'un oursin à quatre pointes.



On a, dans le plan (1,2), trois pointes orientées respectivement suivant le demi axe (F1>0) et les demi bissectrices (F1<0;F2>0), (F1<0;F2<0). Chaque texte étant marqué, sur le graphique, par la première lettre de son sigle, il apparaît que la première pointe contient des textes littéraires ou philosophiques (£,*f*); tandis que dans la deuxième prédominent les études cliniques, y compris celles relatives au SIDA (§,ſ); la troisième pointe étant occupée par l'économie et, notamment, par la géographie et la cartographie économique ($,®,©).

Dans le plan (1,3), on a, d'une part, le long de l'axe 1, une distribution dense de points qui est la projection des trois pointes du plan (1,2); et, d'autre



part, sur le demi axe (F3<0), une nouvelle pointe exclusivement occupée par des sigles en 'μ', mathématiques.

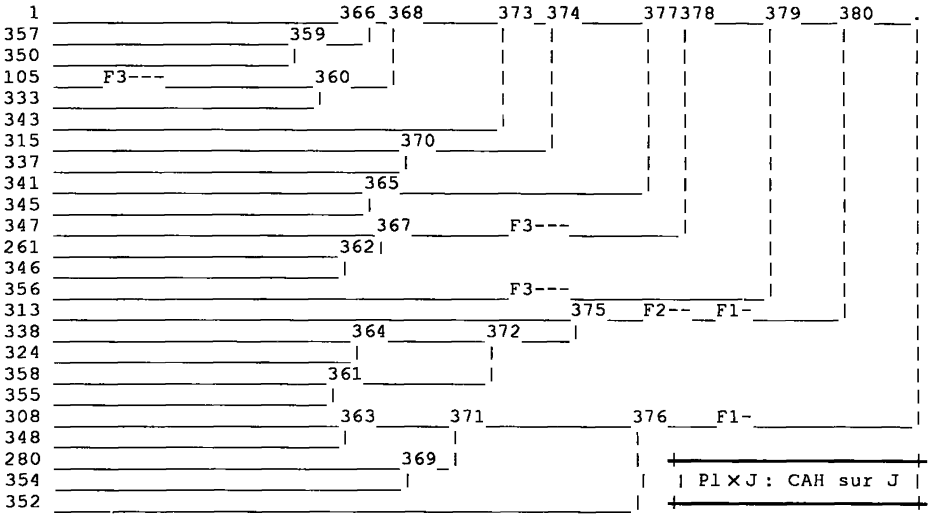
Du nuage des mots, le plan (1,2) ne montre que les individus les plus excentriques, sur chacune des deux pointes du demi-plan (F1<0). Dans le plan (1,3), on a tenté de marquer tous les termes mathématiques afférents à la pointe (F3<0).

En analogie avec la représentation spatiale, la classification arborescente des textes n'offre pas un système harmonieux de dichotomies successives

```
 c | P1×J : Partition en  24 classes : Sigles des articles de la classe c
---+-------------------------------------------------------------------------
  1| ◊fbr
-----------------
357|353:πpop @ngr @µch @µCh µEnt µEPR µinh £Par µméd µSim µvoi µcah ƒ∂ia ƒOrd
   | £Gns ƒOmn                                          µath prédomine
   |351:◊Loi @µCt ◊Oxy ◊vir @fum @chê µVAC $usi £Chn ®Exm ®µVi ®Arb ©Pét @str
350| @man πPCc @Ths @Hst £Aur £ibr £Rep ©vgn £©F© £SMF £Dia £hxL @prf £sty
   | £HxG £Arb £frL                                    £angage principalement
----------------------
105| µRel                                              Relativité (s. de Fourier)
333| £prl £vox £Vox                                    la Parole (périodique)
-----------------------------------------
343| £NT1 £NT2 £ñOr £Lat µfus £Tr2 £Tr1 £etr £cry      £inguistique
-----------------------------------------
315| ƒLgS ƒAri ƒqal £Mt2 £Mt1                           Philosophie
337| ƒAÉc ƒusu ƒPrê                                     économie et philosophie
-------------------------------------------------------------------------------
341| ®Vin ©Alg ©Elz ©lPr ©Cal ©Sal ©Ent $$pr ©pub ©Pnt ©Mrt ©Chô ©$Tr ©!
345| πlin πCum πInd πCré πCor π©dg ©stt π©rt           πrogrammes (dont ©artog.)
-------------------------------------------------------------------------------
347| µrec µPrt @Goo @Jrd µniv µQs1 µbar µMlt µprs µGut
261| µbl∂ µbll µbl2 µBl3
346| ©Mrƒ ©µar µnq0 µInB µInM µnqT                      µathématique (peu d'excep)
-------------------------------------------------------------------------------
356| µbrd µΔel µCv2 µCv1                                Géométrie
-------------------------------------------------------------------------------
313| $cdr $cmp $Lé? $Lé1                               économie par secteurs
-----------------------------------------------------
338| ®Imm ©Par ßcib ßFrn ßJap ßMrt ßvoi ßpôl           ßourse et immobilier
324| ßHKg µdéc πµdé $Sé2 $Sé1                          séries chronologiques
-----------------------------------------
358| $Esp $@Bq $for
355| ©Grc ®$Qb $Arb ßSIC @Prd ®rch ®étu ®Cao ®Caf ®com $Cmc  éco par rég.
-------------------------------------------------------------------------------
308| @Src @stg                                           enquêtes et échelles
348| $&nx $&hX @Mor $&DA @&ps @rch πprs @Alg @pGr £mot $not @&Ps $@Br
-----------------------------------------
280| ©$SF $$Nr $$âg
354| ®Thr $rét $@ct $MzT $sST $Tnd $d&d $eff $SeH $gên $gel $Se4 $hm3 $ADN
   | $2sé $Pra $Imp $ECG $HLT $hm1 $hm2                  médecine ($@ct excepté)
-----------------------------------------------------
352| ∫tk2 ∫tk1 ∫exp ∫tki ∫∂ia ∫ana ∫Imm                    SIDA (tous les '∫')
---+-------------------------------------------------------------------------
```

équilibrées: le schéma global est celui de branches se détachant successivement d'un noyau central (ou, plutôt, s'y agrégeant: la classification ayant été construite par voie ascendante). Mais la plupart des classes de la partition retenue ont une interprétation nette, marquée sur le tableau des sigles.

À l'exception de {§$pr, ®§Qb}, qui concernent l'économie de santé, tous les sigles comportant notre caducée '§' sont dans la classe j376. Des 46 articles de j376, 34 ont dans leur sigle '§' ou '∫' (SIDA); quant aux 12 autres, on a dans j354, ®Thr, "sources thermales"; et dans j348, {@Mor, @&ps, @&Ps}, "questionnaire sur l'expérience de pré-mort" et "échelles en psychiatrie..."; à quoi s'agrègent 7 articles traitant d'échelles ou d'enquêtes en dehors du domaine médical.

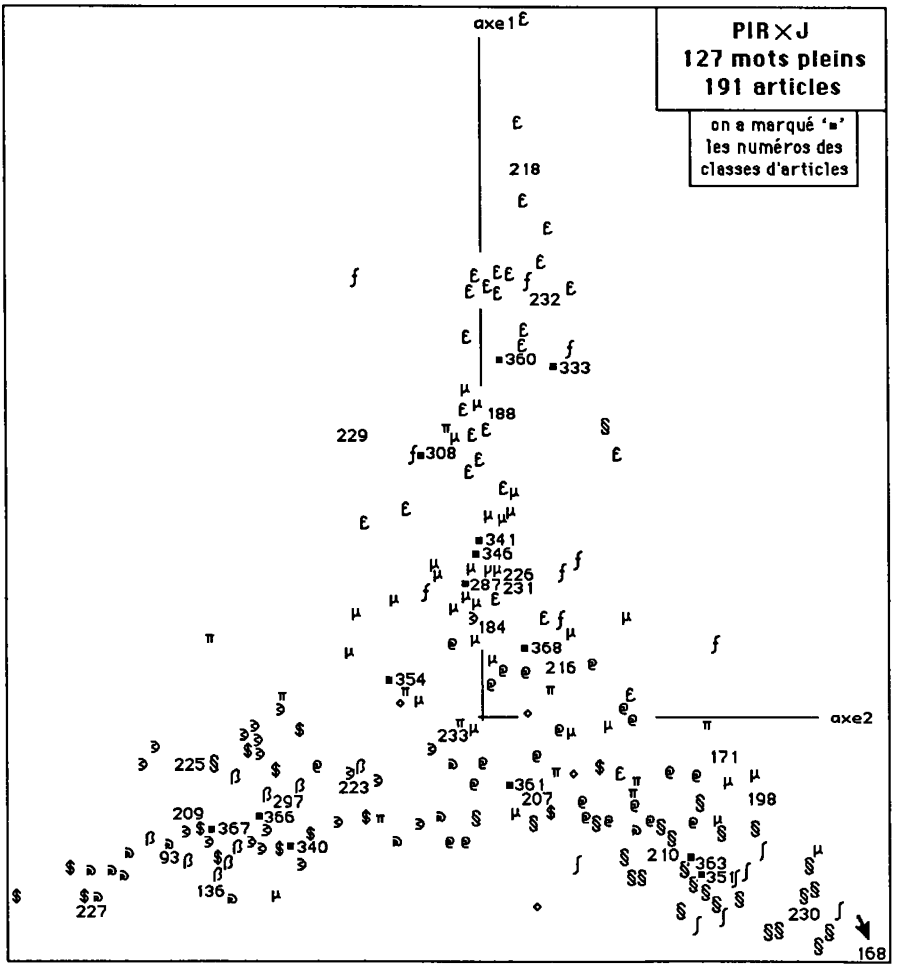


Dans la quasi-totalité des articles de j375, il s'agit d'économie. Les exceptions s'expliquent aisément: {μdéc, πμdé} exposent, sur des cas modèle, la méthode des séries décalées; ®étu analyse la distribution des étudiants, à Paris VI, par nationalité. En dehors de j375, on ne fait à l'économie, que des allusions sporadiques; excepté dans j341, ©artographie, et dans 337 (histoire de la pensée).

Le domaine des mathématiques est partagé entre j356, j367, et la première subdivision, j353, de j357. La place de ce que l'on en trouve ailleurs est explicable: μfus (associativité du tri par fusion), rejoint la £inguistique, dans j343; μRel (relativisation du continuum spatio-temporel) a en commun avec j333 l'étude de la décomposition en série ou intégrale de Fourier (ce nom figure dans le lexique Pl). Sur le listage des facteurs, les sigles en 'μ' se signalent par un F3 négatif (cf. supra).

La subdivision j345 contient les articles décrivant le logiciel MacSAIF et son utilisation, notamment en cartographie. Seul manque πprs, "Codage linéaire par morceaux et Équation personnelle", qui va avec les échelles, j348; et πμdé, "séries chronologiques décalées", qu'on trouve dans j324. À j345 s'agrège j341: les études illustrées de cartes.

Reste j374: dans cet agrégat, on distingue encore des subdivisions interprétables. Dans les subdivisions {j350, j333, j343, j315, j337}, F1 est nettement positif; y prédomine un type de discours, philosophie, linguistique...; ce qui correspond à la pointe observée, dans les plans (1,2) ou (1,3), sur le demi-axe (F1>0).



5.3 Analyses fondées sur un lexique réduit PlR de 127 mots pleins

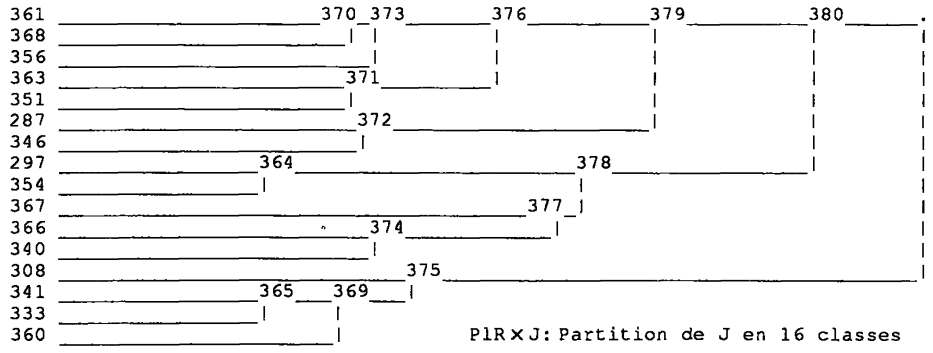
```
127 mots de PlR (Pleins; réduit) × 191 chapitres; tableau écrêté
trace  : 4.917e+0
rang   :    1     2     3     4     5    10 …  15 …  25 …  50 …  75 …
lambda : 3423  3182  2694  2362  1888  1210   906   590   253   119  e-4
taux   :  696   647   548   480   384   246   184   120    52    24  e-4
cumul  :  696  1343  1891  2371  2755  4185  5238  6658  8603  9485  e-4
```

De l'analyse factorielle, on présente seulement le plan (1,2), avec les lettres initiales des sigles des chapitres, et les numéros des classes des deux partitions.

```
  c | PlR×J: Partition de J en 16 classes : les articles de la classe c
----+-------------------------------------------------------------------
 361| ◊fbr πpop πlin @Prd ◊Oxy @Ths $usi ◊vir $cdr µVAC $@Bq @prf @µch @µCt
    | ◊Loi ®Exm @fum $ECG ©Mrƒ ©Mrt @chê £Chn @str ®Thr $@ct $MzT $rét $sST
    | $$âg ©$SF $$Nr                                              ≈CdG
 368|!@man µRel @µCh µinh µEnt µEPR (ƒOmn @Goo ƒ∂ia ƒOrd) µvoi µcah µSim µbar
    | µMlt µPrt µGut | µ∆el µbrd µCv1 µCv2    !!!!          216+++    188++
-------------------------
 356| @Jrd @Src @stg µprs µniv µQs1 | @&ps πPCc £mot @Mor @pGr @Alg πprs $@Br
    | $&nx @rch $not @&Ps           (enquêtes, échelles)   198+++    171++++
-----------------------------------------
 363| $hm1 $hm2 $hm3 $&DA $&hX $gel $2sé $ADN $eff $gên $SeH $HLT $Se4 $Tnd
    | $d&d $Imp                                                198++     230+++++
 351| ∫tki ∫exp ∫tk2 ∫tk1 ∫Imm ∫∂ia ∫ana        (stades sida) 168+++    230++
-------------------------------------------------
 287| µbl∂ µbl2 µbl1 µBl3            structure de blocs diag        226+++++
 346| µnq0 µrec ©µar µnqT µInM µInB  reconstit de d. manq  231+++++  226+
-------------------------------------------------------------------------
 297| ©Chô ©pub ©Sal ©Ent ©Pnt $$pr ©1Pr ©Cal ©vgn                  225+++++
 354| µfus πCré πInd πCor π©dg ©stt π©rt                  184+++  225++++
-----------------------------------------------
 367| $Arb ®Caf ®Cao $Cmc ®com $Lé2 $Lél ®Arb ©Pét ®$Vi ®étu ßSIC ©Grc ©Alg
    | ©El z ®Vin ®rch $for $cmp $Esp ®$Qb                  209++   227++++
-------------------------------------------------
 366| ©$Tr ®Imm ©Par | ßFrn ßcib ßMrt ßpôl ßJap ßvoi   (prix) 93+++ 223++++
 340| @ngr ßHKg µdéc $Sél $Sé2 πµdé              (décalées séries)  136+++++
-----------------------------------------------------------------------------
 308| ƒAÉc fusu ƒPrê          (Aristote, politique, non Italie!)  229+++++
 341| £prl £vox £Vox                                (parole…)  232+++++
 333| £Mt2 £Mtl ƒLgS ƒAri ƒqal                          218+++ 229+++
 360| £Par!µméd £©F© £SMF!@Hst £Aur £Rep £ibr πCum £Tr1 £Tr2 £cry!$Pra £Gns
    | £Dia £NT1 £NT2 £Lat £ñOr £etr £Arb £sty £frL £HxG £hxL  218+++ 232++++
----+-------------------------------------------------------------------
```

Après avoir considéré la partition de l'ensemble J en 14 classes définie par les 13 nœuds les plus hauts, on a demandé de subdiviser les classes 364 et 365, afin de distinguer A(364), 297=©artographie, de B(364), 354=notices de πrogrammes; et A(365), 341=parole, de B(365), 333=philosophie.

Mettons à part la classe 361, très hétérogène (et dépourvue de caractère franc, étant proche du centre de gravité du nuage): on peut donner de chacune



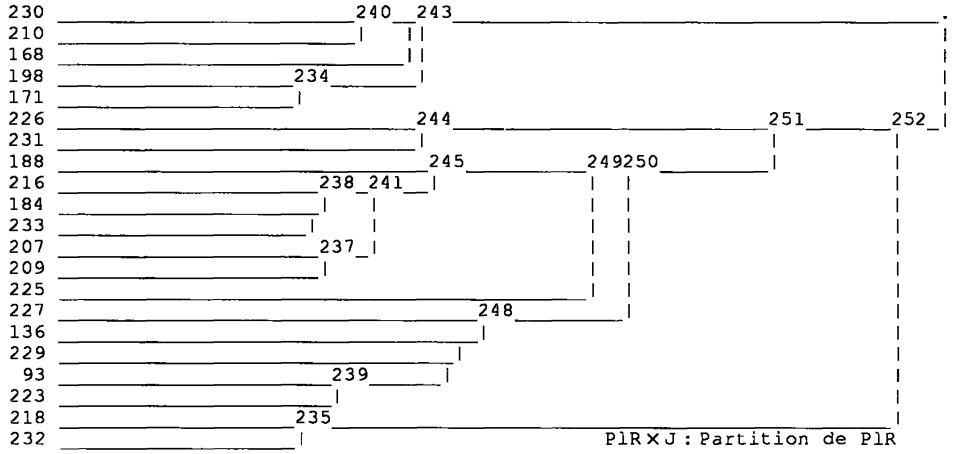
c	PlR×J: Partition de PlR en 21 classes : mots de la classe c
230	absence patients observations état examens examen traitement traitements placebo
210	burt modalité modalités variables codage individus barycentrique
168	stades sida
198	sujet sujets statistique note échelle notes
171	réponse question questionnaire réponses questions
226	distance critère élément bloc blocs
231	masse loi lois trace
188	dimension parties
216	distribution mode hypothèse mesure termes carré image
184	fichier listage programme
233	cah partition classes classification classe | mois physique type nombres forme sommet partie taux activité supplémentaires graphique plan compte origine article moyenne inertie nuage profil profils
207	age ans consommation
209	commerce secteur noeud produits
225	sommets unités France carte trames département départements
227	postes croissance année années annuels pays exportations
136	décalées séries
229	monnaie Aristote Italie politique
93	prix
223	Paris évolution cours titre temps période tendance courbe périodes marché titres
218	livre chapitres oeuvres auteurs auteur textes
232	parole outil fréquence mot texte formes mots

des autres classes une interprétation qui souffre peu d'exceptions. L'étiquetage par le listage VACOR permet de rendre compte des classes d'articles par les classes de mots du lexique PlR. Notre commentaire se bornera à signaler quelques particularités.

La classe 368 comprend principalement des articles μathématiques; l'article @man, "Implantation dans un atelier...", bien qu'exposé de façon abstraite est une exception manifeste; la subdivision {*f*Omn, @Goo, *f*∂ia, *f*Ord} comprend, outre l'article méthodologique @Goo, trois exposés épistémologiques où les arguments de logique formelle sont présentés dans un style mathématique; les quatre derniers articles, signalés par '!!!!', {μΔel, μbrd, μCv1, μCv2}, tous consacrés à la géométrie constituent la classe B(368), benjamin de 368.

La classe 356 comprend des analyses d'enquêtes, des exposés sur la méthode des échelles et des cas modèles marqués 'μ', du fait de leur style mathématique; l'étiquetage par 198+++, 171++++ va de soi.

La classe 371 se subdivise en 363=A(371) et 351=B(371). Cette dernière classe contient tous les titres se rapportant au SIDA; tandis que dans 363 il s'agit, plus généralement, d'études cliniques avec essais thérapeutiques.



Les articles des classes 287 et 346, respectivement Aîné et Benjamin de 372, traitent de deux problèmes théoriques: la reconnaissance par la CAH de la structure en blocs diagonaux d'un tableau de correspondance, et la reconstitution des données manquantes; l'article ©µar comprend une application à la représentation ©artographique de données issues de recensements faits à la Martinique. Ici encore, l'étiquetage s'impose.

La classe 378 a pour Aîné et Benjamin 364 et 377: il s'agit, presque partout, de géographie ou d'économie.

Dans les deux subdivisions {297, 354} de la classe 364, les termes de cartographie (225) sont fréquemment employés; de plus, on trouve dans 354 184={fichier, listage, programme}. Les articles de la classe 297 (y compris §$pr: "praticiens à faibles recettes en France"), sont tous illustrés de ©artes. La classe 354 comprend un exposé de théorie des algorithmes, µfus, et des notices de programmes dont plusieurs se rapportent directement à la cartographie, ou se fondent sur un exemple de données ventilées par département.

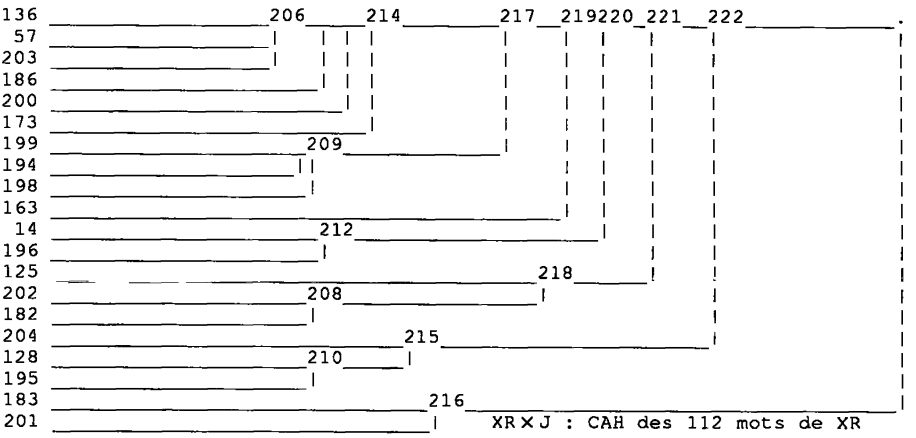
Les trois subdivisions retenues de la classe 377 sont 367, 366 et 340. Dans 367, la plupart des articles font référence au développement temporel, notamment de flux internationaux. Dans 366 et 340 il s'agit de finance: le mot 'prix' caractérise 366; avec d'une part les terres et l'immobilier, et d'autre part la ßourse. La méthode des séries décalées est dans 5/6 des articles de 340: quant à @ngr, il concerne un processus industriel: la vérification d'engrenages en temps réel.

Le tableau de la partition proposée se termine par les articles de £inguistique compris dans les subdivisions de la classe 375. On ne signalera qu'une anomalie: 308 est bien étiquetée par 229, {monnaie, Aristote, Italie, politique}; à ceci près que l'agrégation d'"Italie' à 229 s'explique par des articles de politique ou d'économie dont aucun n'est dans 375.

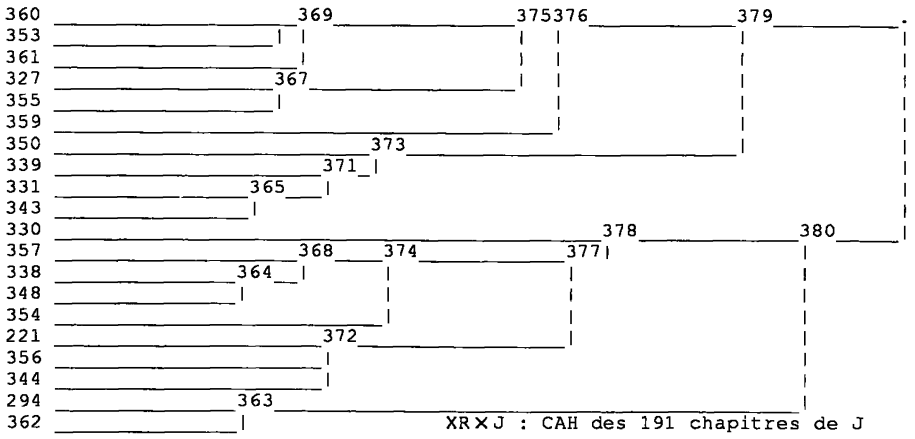
5.4 Analyses fondées sur un lexique réduit XR de 112 mots, pleins pour la plupart, mais avec quelques outils

```
XR x J  : 112 mots de XR(réduit) x 191 chapitres de J (tableau écrêté)
trace   : 2.031e+0
rang    :    1     2     3     4     5    10 ..  15 ..  25 ..  50 ..  75 ..
lambda  : 1610  1361  1282  1036   892   506   365   233    95    39  e-4
taux    :  793   670   631   510   439   249   180   115    47    19  e-4
cumul   :  793  1463  2094  2605  3044  4609  5689  7085  8905  9667  e-4
```

Compte tenu de leurs fortes corrélations avec ceux issus des précédentes analyses (cf.§5.2, in principio), on ne présente pas les résultats de l'analyse factorielle de XR x J.



Afin de publier face à face les tableaux des partitions issues de la CAH des deux ensembles XR et J, on a dû les séparer des arbres hiérarchiques.



```
 c | XR×J : Partition en 20 classes : Sigles des articles de la classe c
---+-------------------------------------------------------------------------
360| ◊fbr @chê @Ths @µCt ◊Loi $ECG $@Bq ®Exm @fum ©Par @prf ®Thr £prl $@ct
   | ©Mrt ©Mrƒ $cdr ®Imm $cmp ◊vir µrec µVAC @Jrd @stg @Src (@Alg $Pra $not
   | ∫exp £Rep $MzT ©$SF $$Nr $$âg)      ????????    ≈CdG   203+  136+
---------------------------------------------------------------------------
353| $rét $ssST @Prd ◊Oxy πpop @str(olabe)  études de cas ?  203+  186+++++
361| @man ®Arb ©Grc ©Pét $Lél $usi $Lé2 $for  géogr et économ    202++++
---------------------------------------------------------------------------
327| πlin πInd πCré πCor πCum                          πrogrammes    200++++
355| ©Elz ©vgn ©Alg ©stt π©rt π©dg ©Sal ©Chô ©$Tr ©Pnt ©pub ©Ent $$pr
   | ©Cal ©1Pr                                        ©artographie  163+++++
---------------------------------------------------------------------------
359| ®$Vi $Esp ®rch ®Vin $Arb ®$Qb ®étu ®Caf ®Cao ®com $Cmc
   |                                        ®égions (pays)  202++++  182+++++
---------------------------------------------------------------------------
350| ∫tk2 ∫tkl ∫tki ∫Imm ∫ana ∫∂ia                              SIDA    204+++++
339| πprs !πPCc @&ps £mot @Mor !£Chn      (!! 2 exceptions)
   | @pGr $@Br @rch @&Ps $&nx          échelles psy, sujets 128++ 195++++
331| $gên $SeH $Se4 $Imp $d&d $eff $HLT $Tnd   médecine et essais 195+++++
343| $hm3 $hml $hm2 $&hX $&DA $gel $2sé $ADN     thérapeutiques   204++++
===========================================================================
330| $Sél $Sé2 ßHKg πµdé µdéc                       séries décalées = 125+++++
---------------------------------------------------------------------------
357| µvoi µcah µEPR µCv2 µRel µbrd µΔel µCvl µSim µméd  géométrie 198+++++
338| @Hst @Goo µinh @µch @µCh µfus £vox £Vox µEnt (∫∂ia ∫usu ∫Omn ∫Ord !!!)
   |                                                      198+  194+++
348| @ngr ßSIC ßcib ßMrt ßFrn ßpôl ßvoi ßJap        ßourse        199+++++
-------------------------------------------
354| £Mtl £Mt2 ∫Prê ∫qal ∫LgS ∫AÊc ∫Ari    Aristote : 201+ 194+++  183+++++
---------------------------------------------------------------------------
221| µbl2 µbll µbl∂ µBl3          structure de blocs   blocs+++++ 194+++
356| µprs µniv µQsl µbar µMlt µGut µPrt     cas modèle 198+ 128++  196++++
344| µInM µInB ©µar µnqT µnq0             trace et d. manquantes   196+++++
---------------------------------------------------------------------------
294| £hxL £frL £sty £HxG                                          201+++++
362| £Lat £ñOr £NT2 £NTl £Gns £cry £Tr2 £Trl £Par £Aur £©F© £Arb £SMF £etr
   | £Dia £ibr                                        £inguistique   201++++++
---+-------------------------------------------------------------------------
```

L'originalité de la présente analyse réside dans son lexique : celui-ci (cf. §3.2.1) a été choisi, de façon quasi automatique, suivant un critère de rang, au sein d'un ensemble de 936 formes. Par le jeu de ce critère, la plupart des mots outils universels ont été éliminés, mais quelques uns de ceux-ci subsistent: 20 dans la classe 194; 4 ou 5 dans la classe 203; et le pronom 'je', associé à Aristote, dans183. (On ne s'étonnera pas de nous voir compter "4 ou 5", si l'on pense que la notion d'outil n'est pas définie sans laisser place au doute.)

Il vaut la peine de considérer d'abord le rôle de ces outils dans l'étiquetage de la CAH des 191 articles. Comme au §5.1, 'je' est associé à j354, i.e. à des discussions portant sur la doctrine d'Aristote. Les outils de la dialectique, compris dans 194, vont avec ce même j354, les textes mathématiques de j221, et la classe j338; classe où, à des textes matématiques (y compris ceux de {£vox, £Vox}, consacrés à l'analyse de FOURIER de la voix), s'agrègent des exposés d'épistémologie, (dont le style quasi mathématique a été vu au §5.3, dans j368).

c	XR×J : Partition de XR en 20 classes : formes de la classe c
136	age ans
57	mois
203	coté niveau analysé plan avons très été ont base facteurs suivant résultats chaque nombre données individus codage variable variables
186	type groupe groupes
200	tableaux colonnes lignes ligne programme fichier
173	cah classes prtition classification classe
199	titres cours temps valeur valeurs
194	recherche ordre éléments produit sous où non donc tout soit cas bien elle son forme selon autres fait ci ainsi ici part autre sans après
198	fonction modèle vers centre droite espace point points
163	carte départements
14	blocs
196	trace masse loi
125	séries décalées
202	années période profil profils produits secteur
182	annuels pays
204	patients état examen examens observations stades
128	réponses questions
195	traitement traitements sujets sujet notes
183	!!je Aristote
201	chapitres oeuvres textes texte mot mots formes

La classe 203, quant à elle, contribue peu à l'étiquetage.

Au reste, la classification des chapitres est analogue à celle du §5.3. Mis à part un résidu, j360, proche du centre de gravité, ainsi que j353, les classes sont cohérentes et l'étiquetage en explique clairement l'agrégation.

Quant à l'organisation d'ensemble des classes, c'est-à-dire quant à l'interprétation des divisions supérieures de la hiérarchie, l'analyse du §5.3 nous paraît quelque peu préférable à celle du §5.4. Mais, d'une part, l'introduction dans XR des outils de la dialectique nous intéresse en ce qu'elle permet de retrouver les liens entre style et thème déjà considérés au §5.1; et , d'autre part, on a souligné l'intérêt du critère de choix de XR, où la notion imprécise d'outil n'intervient pas explicitement.

Même si un corpus de 191 articles n'offre pas une base suffisante pour définir la typologie sémantique d'un domaine de la science, les expériences de discrimination, objet du §6, permettront de comparer, d'un point de vue pratique, les mérites des lexiques.

6 Analyse discriminante

6.0 La méthode et ses variantes

Le but pratique de l'analyse des textes scientifiques et techniques est l'indexation automatique. Il importe donc d'apprécier dans quelle mesure analyse factorielle et CAH permettent d'insérer un document nouveau dans un corpus déjà élaboré. À cette fin, nous avons appliqué l'analyse discriminante, comprise comme l'affectation d'un ensemble d'individus à un ensemble de centres, dans l'espace engendré par les axes factoriels.

Le procédé offre de multiples variantes. L'ensemble des centres peut être un ensemble de textes individuels (articles) ou un ensemble de centres de gravité de classes, constituées par la CAH; on peut, afin de déterminer le centre le plus proche d'un individu nouveau, prendre en compte un nombre plus ou moins élevé de facteurs; les résultats dépendent, évidemment, du lexique d'après lequel les textes sont décrits; ainsi que de l'écrêtement éventuel des grappes (cet écrêtement '&' sera toujours fait, sauf avec le vocabulaire V; cf. *supra* §4.4).

Il est apparu que l'affectation obtenue était d'autant meilleure qu'on prenait pour centres un ensemble d'articles et utilisait le plus grand nombre de facteurs: ainsi, c'est en considérant dans leurs détails les similitudes entre textes qu'on devine le mieux ce que recèle un texte nouveau.

De façon précise, quant aux centres et individus à affecter, nous rendons compte de trois expériences; en considérant, pour chacune, les quatre lexiques {V, Pl, PlR, XR}.

6.1 Affectation d'un quart des articles du corpus de base aux articles restants

L'ensemble J des 191 articles est partagé en J1 et J2: J2 comprend les articles dont le rang, divisé par 4 a pour reste 3, i.e. les 48 articles de rang 3, 7, ...,191; J1 comprend les 143 articles restants. Le tableau ci-joint donne, pour chacun des articles de J2, l'article de J1 qui en a été trouvé le plus proche, le lexique étant PlR.

affectation des j2 aux j1; lexique PlR; espace des axes 1...93;

(£prl->£vox) (πpop->πlin) (µvoi->µVAC) (µinh->@Src) (@stg->@Src) (@prf->$@ct)
($Sé2->$Sél) (£Mtl->£Mt2) (£etr->£Arb) (@pGr->@Alg) (ſusu->ſPrê) ($$Nr->©$SF)
($@Bq->@Src) ($hml->$gel) (πCré->πInd) (®Vin->®rch) (©Cal->©vgn) (@str->ßcib)
(©pub->©Pnt) ($hm2->$&hX) (@chê->@Src) ($ECG->$HLT) ($hm3->$Se4) (£Trl->£Tr2)
(µfus->πCor) ($2sé->$&hX) (©Ent->©vgn) (∫ana->∫∂ia) (ſ∂ia->ſOrd) (µCv2->µCvl)
(®$Vi->®étu) ($$âg->©$SF) (£NT2->£NTl) ($gên->$&DA) ($ADN->$&DA) (µBl3->µbll)
(@Jrd->@Src) (©Mrt->©Mrſ) ($Lé2->$Lél) (£Lat->£ñOr) (@&Ps->$@Br) (£cry->£Tr2)
($Cmc->®com) (@µCh->£vox) ($&nx->@rch) (®Thr->$@ct) (µniv->µQsl) (£©F©->£SMF)

La plupart des affectations sont satisfaisantes. Considérons en détail la première ligne. On a successivement: deux articles relatifs à l'analyse de la parole; deux notices de programme; deux études mathématiques; un modèle de réseau nerveux affecté à une enquête sociologique (cette affectation est douteuse;

mais µinh n'a pas d'équivalent dans le corpus); deux enquêtes, analysées par le même auteur; deux recueils d'analyses de plusieurs tableaux provenant d'enquêtes... Et ainsi de suite, jusqu'aux deux dernières lignes où la seule paire qui ne satisfait pas pleinement est (®Thr->$@ct): mais sans doute n'est-il pas facile d'apparier une brève note relative à un ensemble de stations thermales, alliant la clinique à la chimie, non sans rappeler, au passage, la méthode de découpage des variables. La fin est particulièrement heureuse: car il s'agit de deux articles considérant d'après leurs titres un corpus de publications.

L'affectation est un peu moins bonne, mais reste appréciable, avec les lexiques Pl (non réduit) et XR (réduit, mais automatiquement, sans révision pour éliminer les outils). Il semblerait que le lexique V de mots vides ne permît pas d'apprécier le contenu des articles; il réussit pourtant dans de nombreux cas. On en jugera d'après la dernière ligne, seule publiée ici, du tableau d'affectation.

```
($Cmc->£SMF) (@µCh->£vox) (§&nx->§&hX) (®Thr->§&hX) (µniv->µSim) (£©F©->£SMF)
```

{§&nx, µniv, £©F©} sont bien à leur place; et @µCh, article relatif à des stimuli visuels périodiques, est heureusement affecté à £vox qui, à propos de la chaîne parlée, évoque également la décomposition de FOURIER.

6.2 Affectation au corpus de base d'articles d'un cahier, non pris en compte

On peut craindre que le succès rapporté au §6.1 ne résulte, en partie au moins, de ce que l'ensemble J2 des articles à affecter a contribué à créer les axes de référence. Or, à l'exception de @pGr (thermomètre de sympathie vis-à-vis de personnalités politiques grecques) les articles du cahier (XIII,n°3), dont le sommaire figure au §1, ne sont pas pris dans le corpus de base des 191 articles: on les a affectés à J.

```
affectation aux j ; lexique V ; succès:  (4/8)
(@Chr->£Rep)° (@Erg->@&Ps)+ (µSta->@&Ps)   (πVPr->©stt)°
(@πrc->@&Ps)  (@Pré->£NT2)° ($EsL->@&Ps)* (@∫sd->$@ct)*
affectation aux j ; lexique PlR ; succès:  (5/8)
(@Chr->@Ths)  (@Erg->@&Ps)+ (µSta->µrec)+ (πVPr->µRel)?
(@πrc->πCré)* (@Pré->$&nx)? ($EsL->@&Ps)* (@∫sd->®§Vi)*
affectation aux j ; lexique Pl ; succès: (5/8)
(@Chr->ßcib)° (@Erg->@&Ps)+ (µSta->®§Vi)   (πVPr->πCré)+
(@πrc->£mot)° (@Pré->@Alg)+ ($EsL->$Imp)+ (@∫sd->@pGr)*
affectation aux j ; lexique XR ; succès: (6/8)
(@Chr->£vox)  (@Erg->@&Ps)+ (µSta->µméd)+ (πVPr->πCré)+
(@πrc->πCré)* (@Pré->◊fbr)  ($EsL->$Imp)+ (@∫sd->@Alg)*
```

On marque d'une '*' les succès; d'un '+', les meilleurs succès; d'un '°', des rapprochements non dénués d'intérêt: e.g. entre @Chr, "chronologie de l'insertion professionnelle", et £Rep, "le repas idéal: analyse de réponses libres en trois langues" (autre questionnaire). Les taux de succès sont bons; de quelque manière qu'on les compte. Le lexique PlR, déjà distingué au §6.1, offre ici l'avantage de ne se tromper jamais de très loin.

6.3 Affectation au corpus de base de textes divers

L'expérience rapportée ici est plus hasardeuse que la précédente: les textes à affecter sont cinq articles, parus dans *CAD*, mais en dehors de la période du corpus (4 de ces articles figurent dans le volume "Pratique de l'A. des D. en Médecine...": MA n°1, n°2; MC1n°1; MC2n°1); et un exposé, non publié.

§Cor [INT. CORR. MÉD.]:V3: Introduction à l'a des correspondances d'après données médicales
§Chl [SYST. CHOL.]:X4: Le système du cholestérol chez le rat: analyse simultanée de 4 expér.
§Esc [ESCULAPE I]:XI2: Efficacité thérapeutique et effic. méthodologique: la révolte d'Esculape
ƒBrg [BERGSON]:VII4: Qualité et quantité: grandeur et espace selon Bergson et en a. des d.
ƒMus [DE MUSICA]:non publié: Analyse du "De Musica" de Boèce.

On soumet au lecteur, avec le tableau des affectations, la qualité de représentation dans l'espace engendré par les axes 1 à 10: on a dit, au §6.0, que les meilleures affectations sont fondées sur des similitudes de détail entre articles et requièrent qu'on considère le maximum de facteurs; mais la qualité dans l'espace (1...10) pourrait mesurer l'intégration d'un texte supplémentaire dans l'ensemble du corpus.

Affectations avec 93 facteurs: QLT dans l'espace des axes 1-10 : (cumul) ;

V :	(ƒBrg->ƒqal)	(ƒMus->£Lat)	($Ana->@&Ps)	($Cor->®Arb)	($Chl->®Arb)	($Esc->@&Ps)
QLTx:	440	155	143	57	174	261
Pl& :	(ƒBrg->ƒqal)	(ƒMus->£mot)	($Ana->@&Ps)	($Cor->µVAC)	($Chl->@&Ps)	($Esc->@&Ps)
QLTx:	125	56	203	95	14	205
PlR& :	(ƒBrg->ƒqal)	(ƒMus->πPCc)	($Ana->@&Ps)	($Cor->µVAC)	($Chl->µVAC)	($Esc->@&Ps)
QLTx:	358	78	282	171	86	262
XR& :	(ƒBrg->ƒqal)	(ƒMus->£Vox)	($Ana->∫∂ia)	($Cor->◊fbr)	($Chl->£vox)	($Esc->@&Ps)
QLTx:	467	14	387	217	71	343

L'article ƒBrg est toujours associé à ƒqal; ce qui est très satisfaisant; les affectations de ƒMus (πPCc excepté) sont interessantes à divers titres; §Ana, exposé général, illustré d'exemples divers, ne va pas mal avec @&Ps, qui passe en revue les emplois des échelles, et touche à la médecine; §Cor n'est bien affecté qu'à µVAC: il s'agit, ici comme là, de calculer sur un tableau, des coordonnées euclidiennes, des projections orthogonales; §Chl n'est jamais reconnu; §Esc, illustré de multiples exemples cliniques, peut aller avec @&Ps.

7 Perspectives et conclusions

Dans [IND. DOC.] est analysé un corpus de 268 comptes rendus de visites écrits par des chercheurs du groupe Elf Aquitaine. Chaque compte-rendu tient sur une seule page et l'ensemble du corpus comprend quelque 51000 occurrences de 8100 mots. Bien que travaillant dans un grand centre de calcul, l'auteur, en 1984, considère comme difficilement réalisable une analyse croisant avec les 268 documents l'ensemble des 4500 mots présentant au moins deux occurrences. Assurément ce projet semble de peu d'intérêt pour l'étude du vocabulaire, mais il est aujourd'hui à la portée d'un ordinateur de bureau.

C'est sur un tel ordinateur (un Macintosh SE/30) qu'on a réalisé les calculs du présent article, portant sur un nombre équivalent de documents, mais vingt

fois plus longs que ceux du corpus d'[IND. DOC.]. Les principales conclusions de cet important travail se trouvent ici confirmées.

L'analyse des répétitions (nous disons, au §4: "de la structure en grappe",) guide efficacement le choix du lexique des formes dénombrées: le lexique XR, défini quasi automiquement offre une base satisfaisante à la classification et à la discrimination; le lexique PlR, choisi en tenant compte, également, d'une appréciation subjective de la notion de mot plein, fournit des résultats quelque peu préférables. D'autre part, des essais ont montré l'utilité de l'écrêtement des fréquences élevées (introduit au §4.4). Il reste cependant à préciser la distribution des grappes, selon les suggestions de [TEXT. DOC.].

L'expérience préalable, objet du §3.1.1, a montré que le thème d'un article est bien indiqué par les quelque dix mots pleins qui s'y rencontrent avec la fréquence maxima: cette remarque peut aider les auteurs à mettre des mots en vedette pour l'indexation de leur travail.

Nous employons ici comme synonymes "mots" et "formes". En l'état actuel des recherches, il n'y a aucun avantage à lemmatiser, i.e. à ramener toute forme à un mot de base (singulier pour nom, masculin singulier pour un adjectif, infinitif pour un verbe). En effet, d'une part, on ne sait pas lemmatiser automatiquement sans commettre d'erreur; et d'autre part, en recensant les formes brutes, on prend en compte, implicitement, les effets de sens et les locutions qu'on ne sait pas non plus reconnaître par des algorithmes infaillibles. En dénombrant les formes brutes, on construit sans difficulté des tableaux de correspondance dont l'analyse n'a point déçu nos espoirs.

Bien que de longueur modérée, les textes analysés ici ne sont pas tous homogènes: en commentant les CAH, on a signalé que certains articles comprennent à la fois un exposé mathématique et un exemple d'application. Il conviendrait d'expérimenter sur une affectation de chacun des §, analogue à celle appliquée, au §6, à des articles. D'ailleurs, on ne devrait pas se contenter d'affecter tout texte à une classe: il faut, plus généralement, reconnaître, en quelque sorte, ses coordonnées sur des axes sémantiques que la classification des formes du lexique permettrait de définir (cf. [IND. DOC.], §3.3.3).

Dans [S.M.F. – LIOUVILLE], sont analysés les titres des articles de deux périodiques mathématiques: mais le but est de suivre, au cours d'un siècle, la répartition des thèmes traités; non de classer validement les titres pris un par un pour eux-mêmes.

L'expérience du §2 sur ces textes très courts que sont les résumés en a donné une typologie assez satisfaisante; et le traitement parallèle de l'original français avec sa version anglaise a, d'autre part, confirmé que (comme l'avait vu Ch. ARBACHE) l'analyse des données peut réussir à aligner les vocables de deux langues. Nous nous proposons donc de reprendre l'expérience sur une

plus grande échelle en tentant de décrypter, d'après un livre bilingue, une langue dont nous ne connaîtrions presque rien!

Après plusieurs analyses stylistiques de corpus en diverses langues (dont la plus récente, [SIÈCLE D'OR.], contient une bibliographie relative à nos méthodes et à leurs applications), nous avons traité ici un corpus documentaire de taille équivalente. Deux cents articles n'offrent certes pas une base suffisante pour définir la typologie sémantique d'un domaine de la science: mais nous ne voyons pas d'obstacle qui nous empêche de poursuivre dans la même voie pour élaborer un corpus d'un volume dix ou cent fois supérieur.

Références bibliographiques

A. Aït HAMLAT : "Analyse des répétitions et indexation automatique des documents", [IND. DOC.], in *CAD*, Vol IX, n°2, pp. 173-204, (1984).

Ch. ARBACHE : "Distribution des vocables dans le texte hébraïque de la Bible et dans la traduction grecque des Septante", [TEXTE BIBLE] §9, in *CAD*, Vol XI, n°1, pp. 19-23; (1986).

J.-P. BENZÉCRI : "Description des textes et analyse documentaire", [TEXT. DOC.], in *CAD*, Vol. IX, n°2, pp. 205-211; (1984).

J.-P. & F. BENZÉCRI : "Typologie de textes espagnols de la littérature du siècle d'or, d'après les occurrences des formes des mots outils", [SIÈCLE D'OR.], in *CAD*, Vol. XVII, n°4, pp. 425-464; (1992).

J.-P. & F. BENZÉCRI, G. D. MAÏTI et collaborateurs : *Pratique de l'Analyse des Données en Médecine, Pharmacologie et Physiologie clinique*, éditeur: Statmatic, 6/8, Avenue S. Allende, 93804, Épinay sur Seine; (1992).

SAGOMBAYE NODJIRAM : "Analyse du vocabulaire mathématique des titres des articles de deux périodiques: un siècle du Bulletin de la S.M.F. et du Journal de Liouville", [S.M.F. – LIOUVILLE], in *CAD*, Vol. XVII, n°2, pp.133-158; (1992).