

F. BENZÉCRI

J.-P. BENZÉCRI

**Analyse du vocabulaire et recherche du thème
dans les articles des volumes XII à XVII de
CAD. (2) Lexiques et grappes**

Les cahiers de l'analyse des données, tome 18, n° 1 (1993),
p. 61-74

http://www.numdam.org/item?id=CAD_1993__18_1_61_0

© Les cahiers de l'analyse des données, Dunod, 1993, tous droits réservés.

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

NUMDAM

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

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

ANALYSE DU VOCABULAIRE ET RECHERCHE DU THÈME DANS LES ARTICLES DES VOLUMES XII À XVII DE CAD (2) LEXIQUES ET GRAPPES

[CAD XII-XVII (2)]

J.-P. & F. BENZÉCRI

3 Lexiques pour décrire le contenu les articles

Avant de présenter les quatre principaux lexiques choisis, nous dirons comment s'est déterminé notre choix.

3.1 Principes de choix

Comme au §2, pour l'analyse des résumés, le choix se fonde sur les notions de mot outil et de mot plein. Mais, de plus, sur des textes comptant plusieurs pages, voire, plusieurs dizaines de pages, apparaît la hiérarchie des rôles dévolus aux diverses formes de mots; ainsi le lexique peut être choisi suivant des critères distributionnels précis, ce qui est impossible avec des résumés de quelques lignes.

3.1.1 Examen préalable de quatre articles

Pour chacun des articles (pris dans les volumes XVII ou XVI, dont la table est donnée au §1), on a rangé, par ordre de fréquence croissante, les formes qui y sont attestées. La liste se termine par des mots outil très fréquents qui sont à peu près toujours les mêmes (cf. §3.2.1); mais certains mots pleins les précèdent de peu: ces mots sont donnés ci-après, avec des repères de fréquence. Selon la longueur, et aussi le genre du texte, il faut placer le seuil plus ou moins bas afin d'avoir une dizaine de mots pleins.

On voit que, dans les quatre cas, ces mots caractérisent bien le contenu de l'article, même si aux mots spécifiques se mêlent des mots génériques: ensemble, groupe, variables...

ñOr : Typologie de textes espagnols de la littérature du Siècle d'Or d'après les occurrences des formes des mots outil:

{15< analyse, auteur, mots, siècle, textes, lexique=20=oeuvres, 30< fragments, formes, 45=chapitres}

@μCh : Compression des images polychromes et sensibilité au contraste chromatique

{ 10< blanc, couleur, période, sin, ensemble, chromatique, 15=stimuli, 20< seuil, cos, fonction, luminance, couleurs, 30=image }

\$for : Compte de salaire et compte de formation, de 1973 à 1988, en France, dans 34 secteurs de l'économie

{ 50=période, analyse, 60=plan, postes, axe, compte, secteur, 80=tableau, formation, 100< entreprises, 132=secteurs }

@Ths : Outils tranchants thessaliens en pierre polie: un réexamen de la typologie de Christos Tsountas

{ 16=haches, plan, tranchant, forme, variables, modalités=20=type, groupes, typologie, 30< tsountas, 50< types, 57=outils }

3.1.2 Réduction d'un lexique \mathcal{L}

Partons du tableau $\mathcal{L} \times L$ croisant avec le lexique \mathcal{L} , de mots non outil, l'ensemble J des articles: $k(i,j)$ = nombre d'occurrences du mot i dans l'article j . Puisqu'il est apparu que le contenu d'un article était bien connu d'après une liste des mots les plus fréquents, il vaut la peine de construire un tableau de rangs, avec $r(i,j)$ = rang de $k(i,j)$ au sein de l'ensemble des nombres $k(i',j)$ contenus dans la j -ème colonne: ainsi, si ' i ' est le mot le plus fréquent de l'article ' j ', on a $r(i,j)=1$; etc. Un mot i de \mathcal{L} sera digne d'être retenu s'il joue le rôle d'indicateur de sens dans un nombre suffisant d'articles.

Reste à fixer ces notions vagues par des nombres. Dans la présente étude, la valeur adoptée pour le seuil de rang, r_s , est 10 (avec quelques essais jusqu'à $r_s=15$); et on convient qu'un mot i de \mathcal{L} est conservé dans le lexique réduit \mathcal{L}_R , s'il existe au moins 5 articles j où $r(i,j) \leq r_s$.

3.2 Quatre lexiques principaux: V, Pl, PIR, XR.

Dans ce §, les lexiques sont décrits en termes généraux: des listes complètes de mots se trouvent au §5, dans les tableaux des classifications.

3.2.0 Inventaire préliminaire du corpus

Sur un ensemble de 191 articles, les 37 formes de fréquence supérieure à 2500 sont données dans la liste suivante, chaque forme étant précédée de sa fréquence:

{2525 pas; 2565 deux; 2582 se; 2600 ne; 2746 avec; 2894 ce; 3137 ou; 3149 plus; 3177 qu; 3190 s; 3219 sont; 3366 n; 3528 nous; 4269 qui; 4361 sur; 4481 il; 4518 au; 5441 pour; 5856 par; 6511 a; 6820 dans; 7012 on; 7417 du; 7633 une; 7830 un; 7972 que; 8727 est; 9480 en; 10985 d; 12657 à; 13641 et; 14168 le; 15453 les; 16206 des; 17177 l; 19716 la; 33781 de }

vient ensuite le premier mot plein, “tableau”, de fréquence 2390; de façon précise, entre les fréquences 1200 et 2500, on rencontre huit mots pleins:

{variables, classes, classe, axe, modalités, ensemble, analyse, tableau}.

3.2.1 Le lexique XR

On a pris pour ensemble X les 936 formes, {amélioration,... classe}, dont la fréquence est comprise entre 70 et 1300. Cet ensemble ne peut, tel quel, offrir la base d'analyses. Il va sans dire que X comporte un grand nombre de mots outil tels que {..., soit, où, ont, non}. On a pu éliminer la plupart de ces mots outil proprement dits, ainsi que les mots pleins qui ne sont pas susceptibles de caractériser le contenu d'un article, en appliquant le critère de rang, avec pour seuil 10; d'où le lexique XR.

La construction de XR a été automatique (cf. §3.1.2), à ceci près qu'on a fixé des seuils arbitraires, notamment pour délimiter X; mais le seuil inférieur, 70, a peu d'importance, dans la mesure où une forme de fréquence très basse offre peu d'intérêt pour l'indexation; quant au seuil supérieur, il est seulement fixé pour éviter que ne figurent parmi les mots les plus fréquents de chaque article des outils universels (articles etc.) qu'on ne songe pas à prendre pour indiquer le contenu.

Quant au seuil de 10, il a été pris d'après la liste des formes de quelques articles et le programme “ranger”, a permis d'essayer d'autres valeurs.

3.2.2 Le lexique Pl et le lexique réduit PIR

Le lexique Pl, de 366 formes, a été choisi dans la même bande de fréquence que X; non par programme, mais au vu de la liste ordonnée des formes créée pour les 191 articles, en éliminant, outre les outils, les mots pleins qui semblaient relever du vocabulaire général plutôt que de thèmes particuliers; par exemple, vers la fréquence 450, les formes génériques: {facteurs, ordre, variable, étude,...}. Plus précisément, on explique, au §4.1, comment, d'après la structure de grappe, on a éliminé de Pl quatre formes qui avaient été initialement admises.

Du lexique Pl dérive le lexique PIR, par la même procédure que XR dérive de X.

3.2.3 Le lexique V de 281 formes de mots vides

Le lexique V comporte exclusivement des formes dont la fréquence est ≥ 100 . Outre les mots outil proprement dits, articles, pronoms, conjonctions, prépositions, formes du verbe être..., on a voulu ne conserver dans V que ce qui paraissait ne pas revêtir de sens technique dans les articles du corpus.

Il n'y a pas de règle stricte qui permette de réaliser un tel choix; mais voici, à titre d'exemple, comment on a traité les formes dont la fréquence varie de 200 à 209:

{200=ans parmi vol méthode chez début mot proche assez etc nettement ni premiers principal ceci enfin rang lettres<210};

7/18 ont été conservées:

{parmi chez assez nettement ni ceci enfin};

“vol” est éliminé comme initiale de Volume (dans les références insérées dans le texte: la bibliographie finale n'est pas prise en compte; cf. §1 *in fine*); “etc” est une simple abréviation; {méthode proche premier principal rang} ont été considérés comme des termes génériques de l'analyse des données; {mot lettres} sont liés spécifiquement à l'analyse des textes; {ans début} se trouvent dans les études diachroniques.

Il a fallu éliminer comme étant mot plein (en notre sens) la préposition “pendant”, qui évoque la diachronie; le nom “genre”, qui se rencontre dans les analyses de textes; l'adjectif “grandes”, qualifiant les grandes entreprises... Mais il est satisfaisant de noter qu'en réduisant le lexique de 294 mots à 281 mots on n'a pas modifié les affectations (obtenues par analyse discriminante) rapportées au §6.1 (*in fine*).

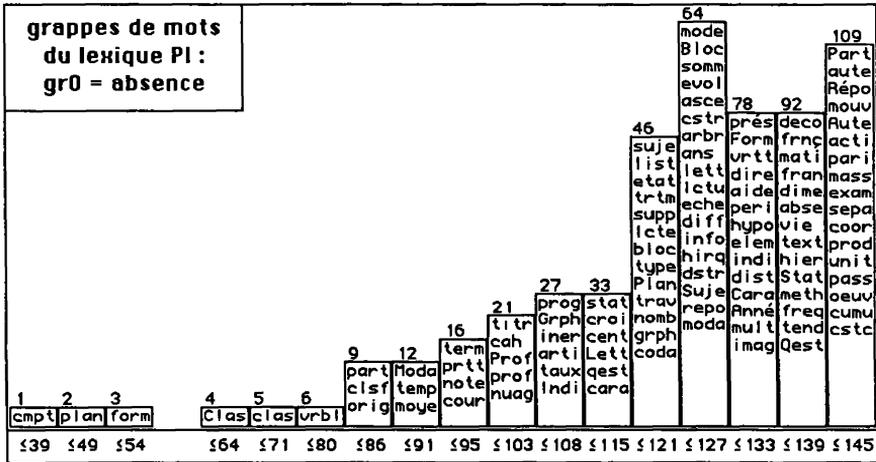
On signalera ici que les mots d'une seule lettre (a, d, l, s, t), d'ailleurs de fréquence très inégale, ont été écartés d'emblée, afin d'éviter toute confusion avec des symboles mathématiques qui subsistent dans le texte, même après qu'on a éliminé les formules isolées (cf. §1 *in fine*). On a même dû écarter “ai”, qui est une forme du verbe avoir; mais que le traitement de texte ne distingue pas de a; !

4 Structure en grappes

Par “grappe” nous entendons ici l'ensemble des occurrences d'une même forme dans un article, sans considérer dans le détail la répartition de ces occurrences. Il s'agira exclusivement du lexique P1 de 370 formes de mots pleins. Partant du tableau P1 × J, on construit (par le programme ‘grap’) un tableau P1 × G défini comme suit.

$G = \{gr0\ gr1\ gr2\ gr3\ gr5\ gr4\ gr6\ gr7\ gr8\ gr9\ G10\ G11\ G12\ G13\ G14\ G15\ G16\ G17\ G18\ \geq 19\}$;

dans la ligne afférente à un mot m , on lit successivement le nombres des textes de J qui contiennent respectivement un nombre d'occurrences de m égal à 0, 1, 2, ..., 18; et, finalement, ≥ 19 ; e.g.: $k(m, gr7)$ = nombre des articles contenant exactement 7 occurrences de m .



4.1 Histogrammes de fréquence des diverses grappes

Avant toute analyse, le tableau $Pl \times G$ peut être observé directement, à l'aide du programme 'zrang'. On considérera les colonnes {gr0, gr1, ≥ 19 }. Afin que les sigles abrégés des mots du lexique trouvent place intégralement dans les créneaux, on a étalé des histogrammes sur deux graphiques.

Si aucune confusion n'est à craindre, le sigle n'est autre que le début du mot, écrit en minuscules; quand il faut distinguer, on met une capitale en tête du sigle d'un pluriel; on abrège le mot en conservant plutôt des consonnes que des voyelles...; et, dans l'exposé, on restitue toujours les formes complètes.

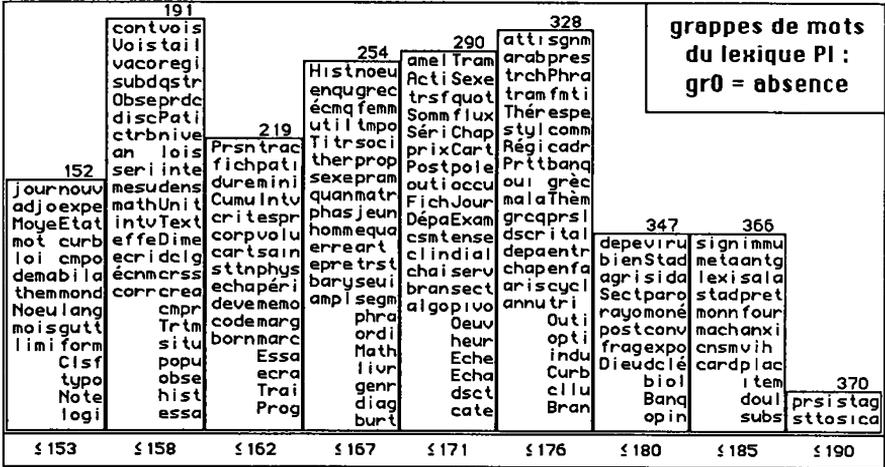
4.1.1 Répartition des mots d'après le nombre des textes d'où ils manquent

Sur la partie gauche de l'histogramme on remarque quelques mots qui ne sont absents que d'un petit nombre de textes (rappelé, ci-après, avec le mot):

{39:compte 49:plan 54:forme 64:classes 71:classe ...};

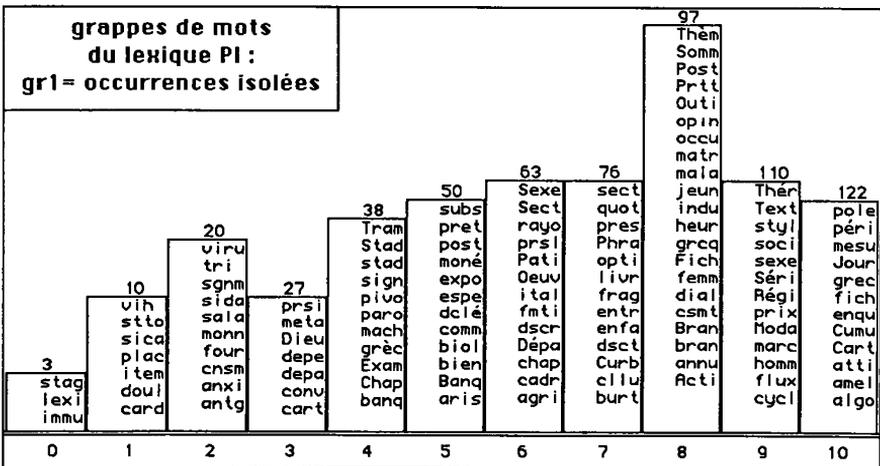
le cas de la forme "compte" est particulier: il peut s'agir d'un terme générique, employé occasionnellement dans tout contexte (rendant compte; tenant compte; et aussi: l'ensemble J compte 17 variables...) ou d'une allusion spécifique à la comptabilité (compte de salaire, compte de formation); "plan" "classe", sont des termes communs en A. des D.; "forme" est soit générique (la forme du nuage; de la courbe), soit spécifique (forme de mot outil), mais il est alors, le plus souvent, au pluriel...

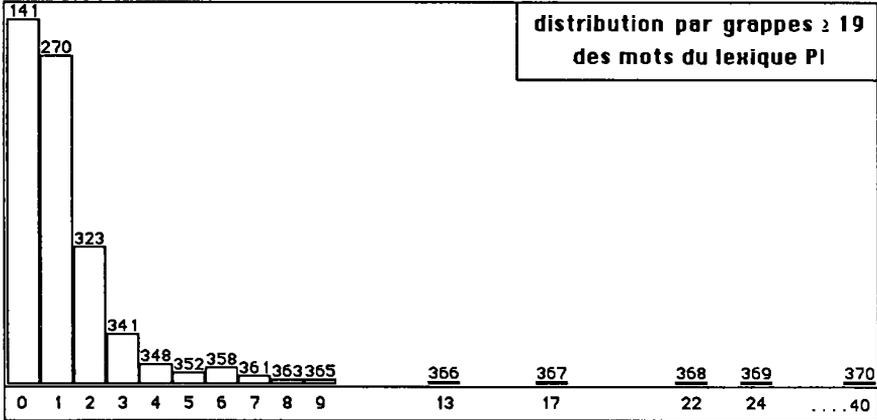
Sur la partie droite, sont les mots qui ne se trouvent que dans un très petit nombre d'articles: "stagiaires" est 71 fois dans @stg (opinion des stagiaires sur



la formation...) et il manque dans les 190 autres articles. De même, les formes “parasites”, “stations”, “SICAV” se rencontrent chacune dans 4 articles au plus; c'est pourquoi on les a éliminées du lexique PI utilisé pour les analyses. Les mots suivants “immunitaire”, “antigénémie”, “salaires”,... présents dans 5 articles au moins, offrent presque tous l'intérêt de signaler, sans ambiguïté un thème déterminé.

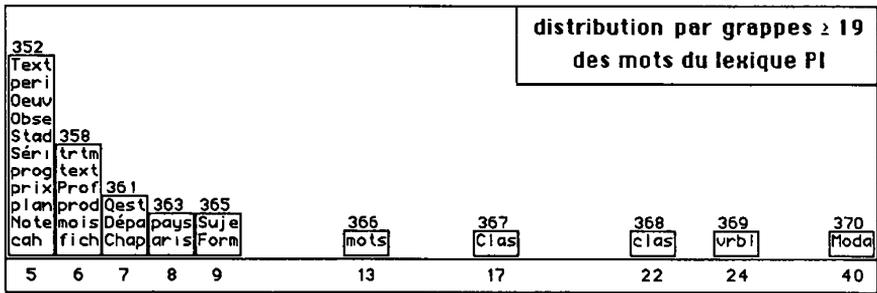
4.1.2 Répartition des mots d'après le nombre des textes où ils figurent une fois





4.1.3 Répartition des mots d'après le nombre des textes où ils figurent plus de 18 fois

Sur l'histogramme global (où ne figure aucun sigle) on voit que, des 370 formes, 141 n'offrent aucune grappe ≥19; 129 en ont une seule; 53 en ont deux; 18 en ont 3; il reste 29 formes offrant au moins quatre grappes ≥19.



La partie droite de l'histogramme est étalée pour qu'on y lise les sigles des mots dont on trouve le plus grand nombre de grappes ≥19:

{ ... sujets:9:formes 13:mots 17:classes 22:classe 24:variables 40:modalités }

le maximum est réalisé par des termes du vocabulaire général de l'A. des D.; le premier terme spécifique étant "mots".

N.B. Dans [IND. DOC.], §3, A. Aït HAMLAT construit un tableau, appelé, MOT × N, analogue au tableau P1 × G considéré ici; elle présente, sous forme d'histogramme, les lignes du tableau afférentes à quelques mots: dans un tel histogramme, le créneau le plus à gauche, qui est également le plus haut, est

proportionnel au nombre des textes d'où le mot est absent; le créneau suivant donne le nombre de textes où le mot figure une seule fois; etc.

Nous ne ferons pas ici de tels histogrammes, nous bornant à l'analyse factorielle et à la CAH.

4.2 Analyse factorielle du tableau PI \times G

```
mots de PI  $\times$  G, effectifs des grappes, de gr0 à  $\geq 19$ ;
trace : 2.713e-1
rang : 1 2 3 4 5 6 7 8 9 10
lambda : 1117 556 193 95 83 71 66 60 56 53 e-4
taux : 4116 2050 713 349 306 260 244 223 206 195 e-4
cumul : 4116 6167 6880 7228 7534 7794 8038 8261 8467 8662 e-4
```

La suite des valeurs propres atteste l'importance du plan (1, 2). On en présente deux graphiques: dans l'un G figure explicitement par les sigles des modalités, tandis que PI est réduit à un nuage de points; dans l'autre, on a, dans la mesure du possible, marqué les sigles des mots; non sans couvrir de hachures les zones trop denses; la forme du nuage PI permet de transporter mentalement G du premier graphique dans le second. À cette fin, les deux graphiques sont présentés l'un en face de l'autre; mais ainsi, ils se trouvent séparés de leur commentaire.

Le nuage a la forme classique, en croissant parabolique, associée au nom de GUTTMAN. Les modalités les plus lourdes {gr0, gr1, ..., gr9} sont régulièrement disposées; au-delà, les fluctuations d'échantillonnage dispersent les points.

Au bord extrême du demi-plan ($F1 > 0$), on trouve les formes déjà signalées au §4.1.1 pour n'être absentes que d'une minorité des articles

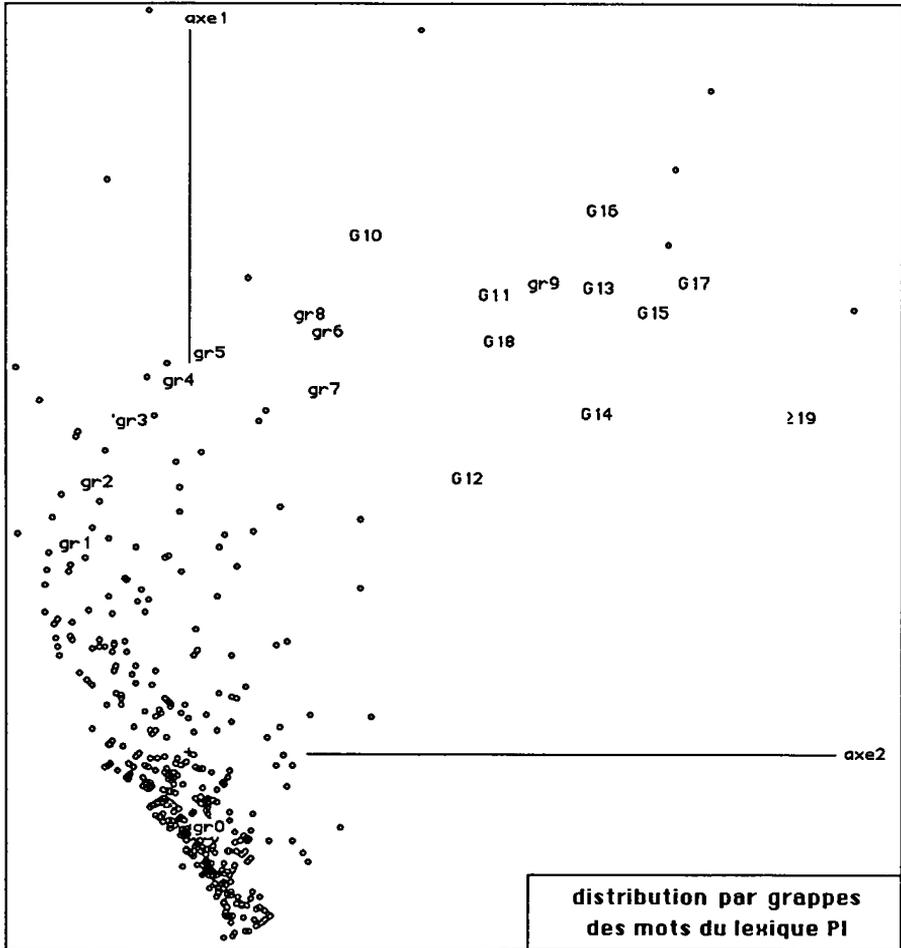
{compte plan forme classes classe} .

Parmi les mots qui ont le plus grand nombre de grappes ≥ 19 , "variables" se détache nettement dans le quadrant ($F1 > 0$; $F2 > 0$). On remarque ensuite, avec de fortes valeurs positives de F1 des formes génériques du vocabulaire de l'A.desD.:

{classification origine partition moyenne parties profil profils nuage individus...} .

De même, dans [IND. DOC.], cf. §3.1.2.2, on trouve "les concepts généraux de l'ensemble du corpus", groupés dans un même quadrant et associés aux plus forts effectifs des grappes.

On a dit que les mots les plus spécifiquement associées à un thème déterminé se rencontrent rarement isolés (gr1) sans offrir toujours des grappes très lourdes. Conformément à ces critères, on trouve de tels mots à l'intérieur de la parabole, mais non avec les plus fortes valeurs positives de F1.

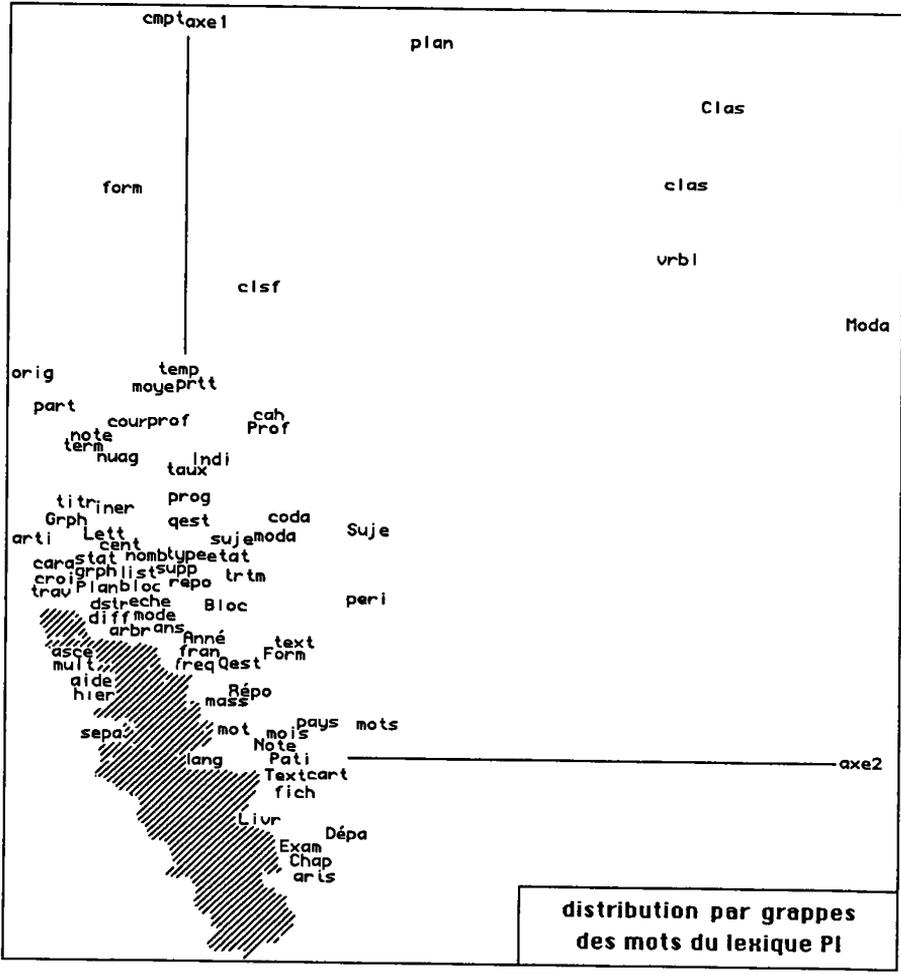


Dans [IND. DOC.], cf. §3.1.2.1, les “mots pertinents représentatifs de la diversité du corpus” occupent une zone “intermédiaire entre ‘0’ et les valeurs élevées” de l’effectif des grappes.

4.3 Classification ascendante hiérarchique

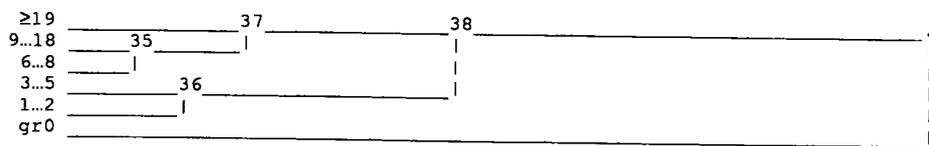
4.3.1 Classification de l’ensemble G des modalités de taille des grappes

On a retenu la partition en 6 classes définie par les 5 nœuds les plus hauts ;

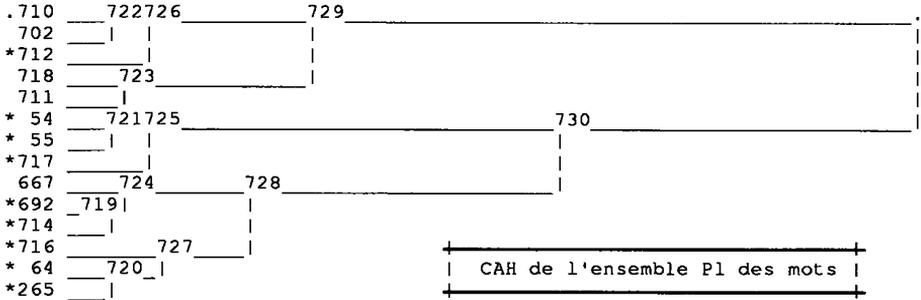


distribution par grappes des mots du lexique P1

sur l'arbre, les classes sont désignées par des sigles dont l'interprétation est claire: e.g. 9...18 = {gr9, G10, G11, ...,G18} ; 1...2 = {gr1, gr2} . La nette



opposition entre gr0 et les autres modalités se voit également sur le plan 1 x 2.



4.3.2 Classification de l'ensemble Pl des mots

On a retenu la partition en 14 classes définie par les 13 nœuds les plus hauts. Afin d'interpréter les classes, on donne, d'après VACOR, le tableau de cumul croisant les partitions retenues sur les ensembles Pl et G, avec, en marge droite, les caractéristiques des classes de Pl.

classes de mots de Pl	de	classes de modalités de G ;	
6	≥19	9...18	6...8
710	26	113	131
702	33	76	120
*712	69	57	45
718	135	182	199
711	102	202	113
*54	22	31	11
*55	17	43	19
*717	64	46	23
667	4	11	39
*692	8	22	32
*714	30	144	144
*716	25	91	107
*64	1	14	20
*265	4	34	27
			3..5
			1...2
			gr0
			5425
			1337
			10729
			248
			1211
			1504
			14712
			848
			16237
			27
			29
			64
			71
			25
			61
			50
			171
			164
			922
			2107
			182
			516
			768
			449
			1122
			2886
			243
			403
			850
			54
			63
			39
			51
			49
			= cdg (1 .2, 3...5>)
			= cdg
			≥19++
			gr0++ 1 2---
			gr0++ 1.2----
			≥19+++ 9 18++++
			≥19++ 9...18+++++
			≥19++++ 9...18++
			1.2+++
			3..5++ 1...2++++
			3..5+ 1.2++
			3..5+++ 1...2++

ci-dessus : Tableau de cumul créé par la procédure VACOR

Sont précédés d'une étoile les sigles des mots conservés dans le lexique réduit PIR: ces mots ne sont pas également distribués dans toutes les classes. Abstraction faite de classes réduites à un ou deux mots, ils prédominent dans 712, 692, 714, 716; ces trois dernières classes sont caractérisées par des grappes d'effectif modéré: de gr1 à gr5.

Au §3.1.2, on a posé comme principe de réduction du lexique qu'un mot est digne d'être retenu s'il joue le rôle d'indicateur de sens dans un nombre suffisant d'articles, où il se manifeste avec une fréquence élevée. Il apparaît ici qu'en dehors de quelques grappes denses qui les ont fait distinguer, les mots retenus offrent, dans un certain nombre d'articles, des grappes d'effectif modéré.

c	Partition en 14 classes : Sigles des formes de la classe numéro c
710	*abse *ans*pari Caté*tend*Part*acti line coor*anné*Trtm*freq*Anné oeuv *exam*mass*fran frnç*Aute deco hier*dime dire cumu hstr*poli mouv Etat meth vie*elem*aute pass vrtt prés arbr indi* hypo Stat
702	them dema adjo date typo cmpo bila gutt trai*carr unit sepa cstc gran mati mang prox ench Sign Dime Type echa Vrtt expe lang limi vaco nive dclg écnm math cmpr hist inte sttn Trai vois epre marg ctrb volu born tail ordi dens Vois code subd situ Moye nouv Clsf prdc an ecri effe *loi *age Homm regi Noeu jour Cumu*lois*mesu corr jeun mond intv Hist
712	*Suje*peri*text*trtm*prod*Qest*Form*pays* Mots
718	*mois*Répo*cart*Note *mot pati*marc*fich*Dépa*Text*sect*csmt*Séri*prix *Oeuv*aris*Chap*crss sexe soci*crit seri deve popu*phys algo cllu enqu Sexe crea ampl erre dure mini écmq pram*form Prsn*Obse sain ecra serv opti Outi ther genr*bary tmpo memo phra segm espr Intv phas quan*Post Jour heur chai prop trsf Cart seui ense oui pole dscr art styl homm bran Thér logi*Unit*curb obse cont*Titr disc Prog diag Math Essa*Pati *qstr*Exam corp*burt Banq Thém
711	dial equa util Bran arab femm quot biol*monn occu tram sign doul stad vih pivo anx entr cadr Curb*paro card grèc sgnm opin frag trch viru post cnsm antg subs pret fmti meta*Stad*plac lexi*comm depe rayo Sect immu sala four matr*Tram conv*ital mach Régi grcq espe*peri Acti grec *depa Eche pres Prtt bien Echa cycl*expo moné*annu*livr cate agri banq item*trac trst*outi amel clin prsl Phra Fich essa*Somm mala enfa atti Dieu flux indu dsct chap tri*noeu*dclé*sida
54	*clas
55	*Clas
717	*vrbl*Moda
667	*imag*dist Cara*somm diff lett aide mult lctu*stat cstr lcte asce hirq croi info trav
692	*titr cara*arti*term*cour*note*orig*part
714	*Bloc*evol*supp*suje*Indi*etat*eche*mode*bloc*type*dstr*list cent*grph Lett Grph Plan*prog*prof*qest*iner*nuag*repo*nomb*taux
716	form*prtt*moda*Prof*coda*clsf *cah*moye*temp
64	*cmpt
265	*plan

4.4 Écrêtement du tableau $PI \times J$

En observant les données, on découvre des grappes d'effectif très élevé: le cas extrême étant celui de "stagiaires" dont les 71 occurrences sont dans un seul article. Certes si une notion tient un grand rôle, le mot qui la désigne ne peut manquer d'être répété; mais les fréquences les plus fortes résultent d'un effet de style: afin d'éviter qu'elles ne perturbent l'analyse, il vaut mieux les écrêter. Ici, on conserve tels quels les nombres de 1 à 10; mais, au-delà, la valeur écrêtée, $k \& (m, j)$, est seulement augmentée de 1 quand le nombre effectif des occurrences, $k(m, j)$ augmente de 5: ainsi {10, 11, ..., 14} sont notés 10; {15, ..., 19} sont notés 12; etc. L'écrêtement est de règle, également, pour les relevés écologiques, où la présence confirmée d'une espèce a la même valeur, que le nombre des individus soit, ou non, très élevé.

Sont exposés au §5 des résultats fondés sur l'analyse de trois tableaux écrêtés, $k&(P1,J)$, $k&(PIR,J)$, $k&(XR,J)$. Le tableau $k(V,J)$ est analysé tel quel, sans écrêtement: et, parce que l'effet de grappe est maximum pour des mots pleins, caractéristiques du thème, lesquels ont été éliminés de V , il n'en résulte pas d'anomalie visible: les valeurs propres issues de $k(V,J)$ sont environ 10 fois plus faibles que celles issues des autres tableaux, pourtant écrêtés; ce qui atteste que le contraste est bien plus faible entre les profils des mots outil qu'entre ceux des mots pleins, même les plus fréquents. D'ailleurs, afin d'écrêter des formes aussi fréquentes que les articles, il faudrait choisir, pour chacune de celles-ci, un seuil permettant d'affaiblir l'effet d'une accumulation anormale; et les données manquent encore sur cette accumulation éventuelle.

N.B. Dans [TEXT. DOC.], §§6-7, il est suggéré de dénombrer les grappes de mots non seulement au sein des textes entiers d'un corpus, mais aussi dans des contextes de longueur déterminée. Ce qui aiderait à distinguer des cas d'accumulation anormale.

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).

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