

A. AÏT-HAMLAT

Analyses comparatives d'un ensemble de données lexicales

Les cahiers de l'analyse des données, tome 8, n° 2 (1983),
p. 199-204

http://www.numdam.org/item?id=CAD_1983__8_2_199_0

© Les cahiers de l'analyse des données, Dunod, 1983, 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/>

ANALYSES COMPARATIVES D'UN ENSEMBLE DE DONNÉES LEXICALES

[COMP. LEX.]

par A. Aït-Hamlat (1)

1 Données et analyse : Pour illustrer la méthode de classification descendante hiérarchique (C.D.H.) qu'il a conçue, M. Reinert traite un exemple de données lexicales (cf. [C.D.H. LEXICALE] § 3 ; C.A.D. Vol VIII n° 2 pp 187-198). L'auteur lui-même a soumis ces données à la C.A.H. ; et compare les résultats obtenus à ceux de la C.D.H. : mais il reste bien des analyses complémentaires à faire ; et nous en présentons quelques unes ici.

Rappelons que le tableau de base est un tableau en (0,1), croisant un ensemble I de 160 vocables avec un ensemble J de 46 militants membres du parti communiste ; avec $k(i,j) = 1$ si le militant j a usé du vocable i en répondant au cours d'un entretien à la question "Être communiste, ça veut dire quoi?". L'ensemble J des militants est d'autre part subdivisé en 3 sections : Entreprise (n° 3) ; Quartier (n° 2) ; Université (n° 1).

Ces données, aimablement communiquées par M. Reinert, sont les seules dont nous disposons : n'ayant pas participé aux entretiens, nous ne tenterons aucune interprétation des résultats quant à leur sens, nous bornant à en commenter la structure interne révélée par l'analyse.

Nous présentons successivement : une analyse factorielle du tableau de correspondance $I \times J$ (§ 2) ; des C.A.H. sur I et J effectuées d'après les 6 premiers facteurs issus de cette analyse ; à la différence de la C.A.H. de M. Reinert, qui a traité directement le tableau $I \times J$ (§ 3) ; et une analyse factorielle complémentaire croisant des partitions provenant des deux C.A.H. précédentes (§ 4).

2 Analyse du tableau de correspondance entre vocables et sujets :

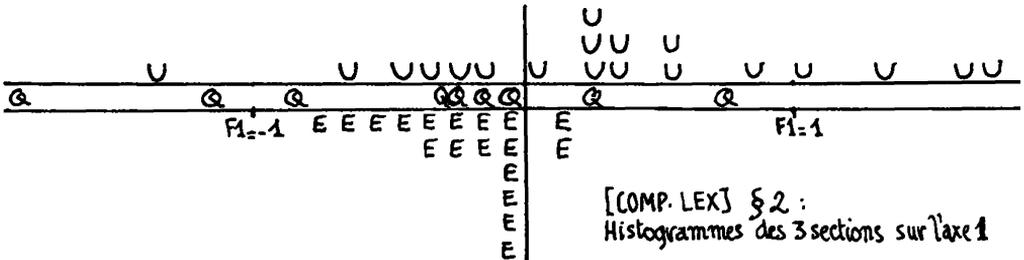
Comme le montre le tableau ci-dessous, les valeurs propres sont assez élevées ; mais elles décroissent lentement et les taux d'inertie sont faibles :

rang	1	2	3	4	5	6
λ	0,26	0,22	0,21	0,19	0,19	0,17
τ	5,5%	4,7%	4,4%	4,1%	4,0%	3,6%

L'appartenance des sujets à 3 sections distinctes, offre une interprétation à l'axe 1. En effet les 18 sujets de la section Entreprise ont tous un facteur 1 compris entre -0,70 et +0,19 (avec

(1) Etudiante en thèse de 3° cycle au laboratoire de statistique.

deux sujets seulement du côté positif) tandis que des 19 sujets de la section Université, 12 ont un facteur F1 supérieur à + 0,20 ; les 7 autres s'étalent jusqu'à -1,32 ; enfin des 9 sujets de la section Quartier, 2 seulement ont un F1 positif ; les autres s'étalant du côté négatif jusqu'à -1,83.



La concentration sur l'axe 1 des sujets de la section Entreprise, est corroborée par deux autres particularités. Ces sujets sont en moyenne proches du centre de gravité du nuage : le carré de la distance spatiale (distance du χ^2) variant de 1 à 8 pour la section E ; de 5 à 20 pour la section Q de 3 à 20 pour la section U. Les sujets de la section E ont de plus les poids les plus élevés : or (cf. M. Reinert) le poids d'un sujet n'est autre que le nombre des mots qu'il a employés parmi les 160 mots retenus (mots pleins employés par au moins 4 sujets) : ce nombre peut aller de 6 à plus de 70 ; il est ordinairement de 20 à 30.

Il est donc remarquable que les sujets de la section E aient en moyenne un poids double de ceux de la section U et triple de ceux de la question Q ; ceci étant dû vraisemblablement non à un discours plus abondant dans E que dans U ; mais au caractère plus central du vocabulaire de E.

Quant aux vocables on trouve à l'extrémité positive de l'axe 1 (prédominance de la section U) les mots :

"efficace ; adhère ; théories ; contradictions ; idéologie ; passées ; combat" ; (qui constitue la classe 291 de la C.A.H. sur I ; cf. *infra* § 3) et de plus "engager et participation".

Les vocables situées du côté $F1 < 0$ sont moins écartées et apportent de moindres contributions que ceux trouvés de l'autre côté on notera :

"logement ; amélioration ; situation ; augmentation ; brime".

Sur les axes 2, 3, 4 etc. ; il n'apparaît pas de séparation des sujets par section ; quant aux vocables on note, à l'extrémité positive de l'axe 2 :

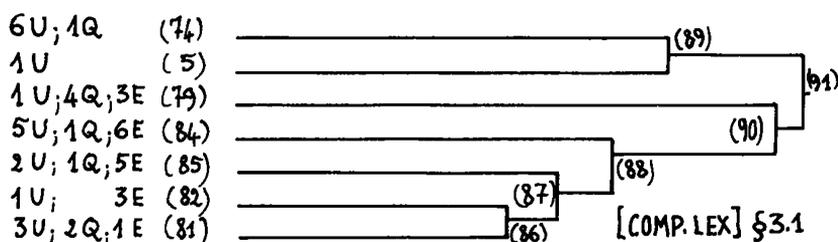
"argent ; travailleur ; meilleur ; classe ouvrière ; capitalisme" ; (vocables dont les 4 premiers rentrent dans la classe 288 de la C.A.H. sur J : cf. § 3). Et à l'extrémité de l'axe 3 :

"Inégalité ; supprimer ; bonheur ; système ; idéal ; injustices". Enfin sur l'axe 4 plus de 40% de l'inertie provient d'un seul individu (de la section U) associé à des vocables :

"Socialiste ; idéal ; engager ; faire avancer..." (qui constituent principalement la classe 303 de la C.A.H. sur J ; cf. § 3).

3 Classifications ascendantes hiérarchiques sur les vocables et les sujets : Ces classifications sont faites dans l'espace rapporté aux 6 premiers axes issus de l'analyse du § 2.

3.1 C.A.H. sur les individus : On figure ci-dessous une partition en 7 classes ; ce niveau étant choisi d'après la relative homogénéité quant aux sections : e.g. la classe 74 comprend 6 sujets de la section U et 1 de la Q (Quartier).



Il est remarquable qu'après le filtre de l'analyse factorielle la C.A.H. donne des résultats s'accordant mieux avec la C.D.H., que ceux de la C.A.H. sur le tableau de correspondance $I \times J$ (rapportés par M. Reinert). Identifions, comme M. Reinert les 4 classes les plus hautes des 3 hiérarchies à comparer ; on a les tableaux de concordance ci-dessous :

C.A.H de M. REINERT sur le tableau $I \times J$		C.A.H sur les 6 facteurs (87) (84) (79) (89)							
C.D.H. de M. REINERT	1	16	4		1	14	6	1	
	2	6	2		1	3	6		
	3			6	1			7	
	4	1			8	1			8

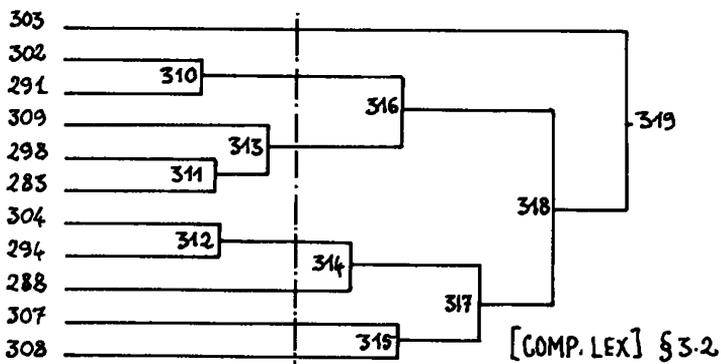
[COMP.LEX] §3.1

En particulier, la classe 2 de la C.D.H. de M. Reinert est bien mieux retrouvée dans la présente C.A.H. : car sur 9 individus de la classe 2, 6 sont dans la (84) dont ils forment la moitié (les 6 autres individus de la classe (84) proviennent de la classe 1 de la C.D.H.).

Compte tenu des résultats de l'analyse factorielle, on ne sera pas surpris que la classe (89), dont 7 des 8 militants viennent de la section Universitaire, se projette à l'extrémité + de l'axe 1; l'individu 5 signalé par l'analyse factorielle comme créant l'axe 4, s'unit très haut à la classe (74) pour former (89). La classe (79) se détache du côté $F2 > 0$.

Quant à l'axe 3, il est remarquable que les 5 individus les plus écartés du côté $F3 < 0$ (il s'agit de 4U et 1Q) soient distribués par la C.A.H. dans plusieurs classes : deux dans (82) ; deux dans (84) ; et le dernier isolé (5). Au reste l'interprétation des classes sera complétée au § 4 en mettant en correspondance une partition sur I et une partition sur J.

3.2 C.A.H. sur les vocables : On représente la partie supérieure de l'arbre avec une partition en 11 classes ; dont plusieurs correspondent à des groupes de mots déjà signalés par l'analyse factorielle (cf. § 2). Au § 3 on retiendra seulement une partition en



7 classes ; dont les mots apparaîtront disposés avec les classes d'individus qui les emploient le plus.

4 Correspondance entre classes de mots et classes de sujets :

Soit respectivement C et D des partitions de deux ensembles I et J entre lesquels est donnée une correspondance $k(i,j)$; on peut noter :

$$k(c,d) = \sum \{k(i,j) \mid i \in c ; j \in d\}$$

dans le cas présent, $k(c,d)$ est le total, sur l'ensemble des mots i de la classe c , des nombres $k(i,d)$ de sujets j de la classe d qui les emploient ; ou encore le total sur d des nombres $k(c,j)$ de mots de la classe c employés respectivement par chaque sujet j .

On a pris pour C la partition de I en 6 classes définies par les 6 noeuds supérieurs de la hiérarchie présentée au § 3.2 ; i.e. :

$$C = \{303, 310, 313, 312, 288, 307, 308\} ;$$

quant aux sujets, la partition D retenue est celle même présentée au § 3.1 :

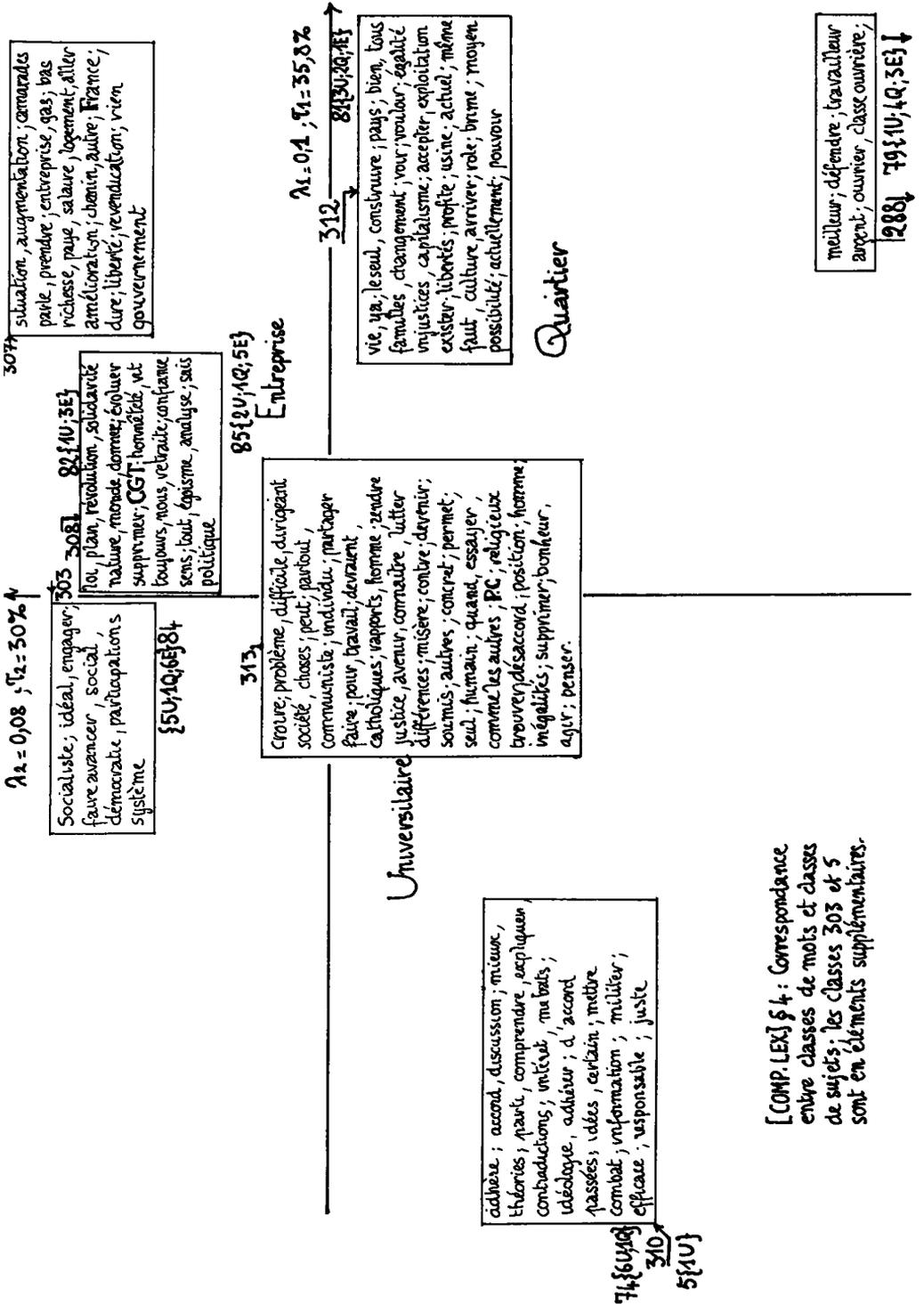
$$D = \{74, 5, 79, 84, 85, 82, 81\}.$$

A D, on a adjoint trois éléments supplémentaires E, Q, U associés aux trois sections Entreprise, Quartier, Université, et définis par cumul des sujets ; de façon précise, on a pour E :

$$k(c, E) = \sum \{k(i,j) \mid i \in c ; j \in E\},$$

(où on a noté $j \in E$, l'appartenance du sujet j à la section E).

Une première analyse de la correspondance $C \times D$ montre un premier axe créé par l'association de la classe 5 de D (classe formée d'un seul sujet de la section U) avec la classe de vocables 303. On a donc fait deux autres analyses : dans la première de celles-ci seul 5 est mis en élément supplémentaire ; dans la deuxième, sont écartés à la fois 5 et 303. C'est cette dernière analyse qu'on présentera ici : elle ne diffère de la précédente que quant à la place de



l'élément supplémentaire (sujet) 5 : lequel est attiré par la classe de vocables 303 lorsque celle-ci est en principal ; tandis qu'il va avec la classe 74 (6U ; 1Q) quand 303 est en supplémentaire.

On s'est borné à dessiner le plan 1×2 qui cumule 66% de l'inertie ; en attachant au point figurant chaque classe c , la liste des vocables dont elle se compose.

Considérons d'abord le triangle des éléments supplémentaires U, E, Q : sur l'axe 1, Université s'oppose à Entreprise et Quartier ; puis sur l'axe 2 E se sépare de Q. Des trois points, E est le plus proche de l'origine, ce qui s'accorde avec les faits remarqués au § 2 quant à cette section.

A l'extrémité négative de l'axe 1 (et avec $F_2 < 0$) on trouve la classe 74 des sujets (6U, 1E) associée à la classe 310 des vocables : le listage FACOR d'aide à l'interprétation du § 3 signale que ces classes étaient dans l'analyse du § 2 à l'extrémité positive de l'axe 1 : ce qui confirme une forte association avec U.

Très à l'écart dans le quadrant ($F_1 > 0, F_2 < 0$) on trouve la classe 79 (1U, 4Q, 3E) associée à la classe 288 : cette association était déjà visible à l'extrémité positive de l'axe 2 issue de l'analyse du § 2.

Le reste des classes de sujets ou de vocables rentre dans le quadrant ($F_1 > 0, F_2 > 0$) ; ou s'en écarte peu.

En bref, le plan 1×2 issu de l'analyse du § 4, offre de l'ensemble des résultats de l'analyse factorielle initiale du § 2, et des C.A.H. du § 3 une vue schématique d'ailleurs conforme, tant pour les sujets que pour les vocables, aux résultats de la C.D.H. publiés par M. Reinert.

Ainsi qu'on l'a dit au § 3.1 la classe 3 de M. Reinert (C.D.H.) s'identifie à la classe 79) de la C.A.H. : on retrouve en effet les mots caractéristiques données par M. R. pour 3, dans la classe 288 qui sur le plan se projette avec (79). La classe 1 de la C.D.H. correspond à (87) partagée ici en (81), (82), (85) : avec des classes d'individus, vont les classes de mots 312, 307, 30 où on trouve les caractéristiques de 1. La classe 4 de la C.D.H. se reconnaît dans (74) : et les caractéristiques sont dans 310.

La classe 2 de la C.D.H. est la plus difficile à identifier dans la C.A.H. : toutefois on peut rapprocher (2) de (84) : et la plupart des caractéristiques données par 2 sont dans la classe centrale 313 qui est proche de (84).

Il serait facile de projeter sur ce plan, en éléments supplémentaires, les vocables et les sujets individuels : mais le schéma y perdrait en netteté ce qu'il gagnerait en nuances. Les résultats détaillés étant déjà obtenus par l'analyse $1 \times J$ et les C.A.H. (présentées sommairement ici aux §§ 2 et 3), nous nous arrêterons à ce schéma.