

## OCCURRENCES DES MOTS OUTIL ET STYLE DES ARTICLES SCIENTIFIQUES

### [STYLE ARTICLES]

M.-M. THOMASSIN\*

HASSAN HAMOUD Anwar\*\*

#### 1 Des données aux analyses

##### 1.1 Le corpus des articles

“Analyse du vocabulaire et recherche du thème dans les articles des volumes XII à XVII de *CAD*”, font l’objet de trois articles, [*CAD* XII-XVII (1,2,3)], du n°1 du Vol. XVIII de *CAD* (1993). Certaines analyses sont fondées sur des lexiques où prédominent les mots pleins; d’autres, sur un lexique V de 283 mots dits vides. Mais, en acceptant comme vide tout ce qui n’évoque pas directement l’origine précise des données d’un article, on reste dans le domaine propre de *CAD*; sans aborder la stylistique générale des publications scientifiques.

L’occasion s’offre ici de considérer un ensemble de textes relevant de méthodes scientifiques assez variées pour présenter des différences de style dans l’usage même des mots outil qui leur sont communs. Nous adjoignons au corpus des 191 articles de [*CAD* XII-XVII], 63 articles publiés par l’Orstom dans les *Cahiers des Sciences Humaines*, *CSH*; avec, en complément, une partie de la thèse, h@b, de Mme M. BLANCHET: “*Le rapport des habitants à l’espace physique et à l’espace social de Marne-la-Vallée. Perception et images*”. Ce travail rentre dans le domaine des sciences humaines, mais comporte des développements théoriques et des descriptions minutieuses qu’on ne trouve pas, d’ordinaire, dans les articles d’une revue.

Le corpus de *CAD* est complètement décrit dans [*CAD* XII-XVII (1)], Quant au genre, les articles peuvent relever des mathématiques, de la programmation, des études de cas, voire de la philosophie et de la méthodologie générale. Les thèmes les plus fréquemment rencontrés sont la médecine, l’économie, la linguistique, la représentation cartographique de données ventilées par départements... On se bornera à rappeler la valeur des sigles par lesquels ces articles sont désignés; principalement des lettres

(\*) Géographe à l’O.R.S.T.O.M.

(\*\*) Étudiant en Doctorat: Université Pierre et Marie CURIE et جامعة بغداد .

20 AVR. 1995

ORSTOM Fonds Documentaire

N° : 41307 ex 1

Cote : 5

initiales qui seules figurent sur les graphiques de la présente étude. Ces rappels sont donnés plus loin, suivant la commodité de la mise en page.

Des 63 articles de *CSH* (Orstom), nous énumérons les titres. Le lecteur notera que certains thèmes - langues vernaculaires, santé, migrations... - font l'objet d'une série d'études: en effet, *CSH* publie des numéros spéciaux. Pour les distinguer de *CAD*, les articles de *CSH* sont désignés par un sigle numérique de \*01 à \*65 (manquent les numéros \*46 et \*47; attribués initialement aux articles \*24, \*25). Les articles de *CSH* sont, en général, plus longs que ceux de *CAD*; mais du même ordre de grandeur: ≈40k octets.

- \*01 Travail industriel et identité de classe à Bahia, Brésil
- \*02 Emploi dans le bâtiment et insertion des migrants à Bogota, Colombie
- \*03 L'emprise urbaine: Famille familialisme modernité à Bahia, Brésil
- \*04 La ville et le volcan: Quito, entre Pichincha et Cotopaxi, Équateur
- \*05 droit au logement à Brasilia: enjeu démocratique ou clientélisme
- \*06 Comment être pauvre et citoyen: l'Église au Brésil
- \*07 Femmes et politique dans les milieux populaires en Équateur
- \*08 Du salarié au petit producteur libre: expériences de réforme agraire
- \*09 l'élection présidentielle de 1989: géographie électorale du Brésil
- \*10 Expansion de la frontière économique, accès au marché et transformation de l'espace rural en Amazonie brésilienne
- \*11 La vallée de Mexicali : un satellite agricole de la Californie ?
- \*12 Réforme agraire et fiction municipale au Mexique (Veracruz)
- \*13 Retour en arrière : les haciendas dans leur gloire, Veracruz
- \*14 Une figure peu connue du monde rural: le rancho du Mexique
- \*15 Extractivisme en Amazonie brésilienne: système en crise d'identité
- \*16 Mutations au Brésil. Vue d'Amazonie: 1980-1990
- \*17 La lecture des traces dans des sociétés sans écriture
- \*18 Gandhisme: traditions de résistance et résistance au colonialisme
- \*19 Trajectoires sociales, mobilité spatiale et accumulation paysanne en Amazonie brésilienne
- \*20 Migration et intensification: dynamique agraire au N-E ivoirien
- \*21 Tunisie: petite entreprise et grande banque: un prêt d'aide au développement
- \*22 Économies et États en Afrique francophone: l'interventionnisme
- \*23 Groupements villageois et processus de transition
- \*24 Du risque au développement. Anthropologie sociale et nutrition
- \*25 Médecins et anthropologues: une rencontre à venir
- \*26 Réduction de la mortalité maternelle en Equateur: Sciences sociales et santé publique
- \*27 Le conflit mauritano-sénégalais: les Peul de la haute vallée du Sénégal
- \*28 Les Peul "refusés". Les Peul mauritaniens réfugiés au Sénégal
- \*29 La santé publique entre soins de santé primaires et management
- \*30 La sécurité alimentaire à l'heure du néo-libéralisme
- \*31 Irrigation, émigration et sécurité alimentaire sur le fleuve Sénégal
- \*32 La réhabilitation des petits périmètres irrigués à Madagascar : une alternative aux importations de riz
- \*33 Crises et régulation en pays Soninké (Sénégal) depuis ≈1850
- \*34 L'agriculture urbaine: activité créatrice d'emplois et survie à Lomé
- \*35 Restauration populaire et sécurité alimentaire à Abidjan
- \*36 Les prix agricoles: coûts de production en milieu non monétarisé
- \*37 Vers une évaluation biologique du développement économique

- \*38 Pouvoir d'achat du salaire minimum indicateur socio-économique et alimentaire
- \*39 Population sous-alimentée à Abidjan: Crise et solidarité familiale
- \*40 Développement agricole et déforestation en Amazonie brésilienne
- \*41 Riz de l'oncle Sam et marché mondial: "compétitivité" à coup de \$
- \*42 Nécessité alimentaire quotidienne et menace socio-politique en Haïti
- \*43 Accès aux aliments, état et communautés paysannes au Pérou
- \*44 Inégalités alimentaires dans une région calcaire (Mexique)
- \*45 Irrigation et problématique foncière dans la vallée du Sénégal
- \*48 La contribution des langues au développement: domaine de la santé
- \*49 Compte rendu de livre: Vers une révolution francophone
- \*50 Langues vernaculaires et exogamie linguistique: le cas du Vaupés colombien
- \*51 La jeunesse face au multilinguisme en Casamance (Sénégal)
- \*52 Linguistique et développement rizicole dans le nord du Cameroun
- \*53 Les Populations Océaniques aux XIXème et XXème siècles
- \*54 L'émigration dans la littérature égyptienne contemporaine
- \*55 Campagnes et villes de l'Inde : deux cultures antagonistes ?
- \*56 L'attraction urbaine. Soixante-cinq ans d'observations sur Abidjan
- \*57 Légende à revoir: l'ouvrier du bâtiment brésilien sans feu ni lieu
- \*58 Émancipation des jeunes et migrations interafricaines
- \*59 Citadins en mouvement: pratiques résidentielles à Quito
- \*60 Migrations saisonnières chez les Mafas, montagnards du Nord Cameroun: solution au surpeuplement et frein à l'émigration définitive
- \*61 Migrations de travail et migrations de mariage dans une communauté marchande d'une petite ville indienne
- \*62 Navette et migration en ville: la mobilité dans l'ouest de l'Inde
- \*63 Migrations internationales, emplois et urbanisation: villes moyennes de l'État de Jalisco (Mexique)
- \*64 Aspects politiques, sociaux et économiques de la migration dans une région industrielle récente de l'Inde
- \*65 Mobilité marchande et urbanisation: cas en Côte-d'Ivoire et Niger

De la thèse h@b on a retenu 95 fragments (de longueur 3k à 6k), dont les sigles comprennent une capitale, de A à K, désignant le chapitre, et un numéro de §. Nous ne décrivons pas ici ce texte, dont le contenu doit faire l'objet d'une étude séparée.

## 1.2 Choix d'un lexique de mots outil

Voici comment on a choisi un lexique  $\Delta\&$  de quelque 90 mots outil d'usage commun.

Pour chacun des deux sous-corpus, *CAD* et *CSH* (Orstom), on a considéré un lexique des  $\approx 200$  formes les plus fréquentes: soit  $\Delta c$ ,  $\Delta o$ ; dans la réunion  $\Delta c+o$  de ces lexiques, on distingue trois segments:  $\{\Delta c-o \Delta o-c \Delta o\&c\}$ ; chacun de ces trois segments compte environ 100 formes; il est remarquable que  $\Delta c-o$  (mots de  $\Delta c$  qui ne sont pas dans  $\Delta o$ ) comprend essentiellement des mots pleins;  $\Delta o-c$  (mots de  $\Delta o$  qui ne sont pas dans  $\Delta c$ ) de même; au contraire, l'intersection  $\Delta c\&o$  (mots qui sont à la fois dans  $\Delta c$  et dans  $\Delta o$ ) comprend quasi exclusivement des outils.

Pour fixer un lexique  $\Delta\&$  qu'on puisse considérer comme un lexique d'outils d'usage commun, il a suffi d'éliminer de  $\Delta c\&o$ : d'une part les formes

réduites à une lettre unique {a j l n...} qui dans des articles de *CAD* peuvent être des symboles et non des mots; et, d'autre part, la quinzaine de formes:

{compte cours dire ensemble espace etat exemple groupe groupes niveau nombre pays periode produits temps}.

De ces formes, certaines: {compte cours ensemble groupe ...}, ont tantôt une valeur générale d'outil, tantôt une acception technique; on peut, de même, s'interroger sur la valeur de {part permet place} qu'on a inclus dans  $\Delta\&$ , comme des outils; et, des quelques outils qui sont dans  $\{\Delta c-o \Delta o-c\}$ , certains sont d'un usage général, même s'ils se trouvent à un rang supérieur à 200 dans l'un ou l'autre des deux sous-corpus. Mais les résultats des analyses fondées sur le lexique  $\Delta\&$  ne nous paraissent pas dépendre de l'acceptation ou du rejet de ces quelques formes.

On crée donc un tableau croisant, avec un lexique  $\Delta\&$  de 93 formes, l'ensemble des 254 articles du corpus (*CAD*+*CSH*). Une première analyse montre, dans le plan (1, 2), une nette différence entre les deux sous-corpus; mais on s'interroge sur les contributions de quelques mots.

Les formes {nous on} apparaissent nettement associées à *CAD*; or elles apportent, ensemble, à l'axe 1  $\approx 20\%$  de son inertie: ayant vérifié que la distinction entre *CAD* et *CSH* (Orstom) subsiste quand {nous on} sont en supplément, on a décidé de conserver ces formes dont l'usage n'est pas lié à une méthode ou à un thème. On retrouvera, au §3, des analyses où {nous on} sont en supplément.

La forme {sous} apporte à l'axe 6 la moitié de son inertie: sont en cause deux articles de *CAD* où cette préposition fait fonction de préfixe: @man (sous-atelier) et §§ST (sous-décalage); en général, les emplois de sous, préfixe séparé, sont fréquents dans les exposés; mais peuvent faire l'objet d'erreurs d'orthographe. À l'axe 4, {la} apporte 10% de son inertie: les occurrences de {la} nous semblent toutefois rentrer dans l'usage commun de cette forme. On a écarté {sous} du lexique qui comprendra donc 92 formes; et on a écrété les fréquences maxima de toutes les formes par le programme 'rangreter'; (cf. [TEXTES GRECS 2], *CAD*, Vol.XVIII, n°2, p.148; [POLITIQUE FRANCE], *ibid.*, n°3, p.298).

De façon précise, on se réfère au tableau des profils des lignes (textes). Pour chaque colonne (mot) de ce tableau, sont diminuées les valeurs les plus fortes: les valeurs de rang 1 à rs (ici, rs=5) sont, par transformation linéaire, réduites à s'étaler entre les valeurs initiales de rang rs-1 et rs. On reconstitue alors un tableau de dénombrements {k(i,j)}, en multipliant, par le total k(i) initial, chaque ligne de profil modifiée.

### 1.3 Les résultats présentés

Sur le tableau  $254 \times 92$ , on effectue l'analyse des correspondances (cf. §2.1) et la CAH des ensembles I (textes) et J (mots) (cf. §2.2).

Nous publions, de chacun des plans (1,2) et (1,3), quatre diagrammes: ensemble I, textes désignés chacun par la première lettre de son sigle; (donc \*, pour tous les articles de CSH); ensemble J, le lexique &; ensembles iq et jq des centres des classes des partitions retenues pour chacun des ensembles I et J; ensemble @ des 95 fragments de h@b, adjoints en supplément à l'analyse, chacun marqué par le caractère °.

Pour chacune des CAH, on publie une partition avec arbre et tableau du contenu des classes; et la CAH des i est étiquetée, par Vacor, en terme de contributions apportées par les classes jq (de mots) à la définition des classes iq (de textes).

Dans l'espace engendré par les axes factoriels, chacun des 95 fragments de h@b est affecté au texte le plus proche de CAD+CSH (cf. §2.3).

Pour plus de précision on effectue une analyse générale où figurent en principal, à la fois, les 254 articles de CAD+CSH et les 95 fragments de h@b; soit 349 textes en principal. D'après les contributions des mots à l'analyse factorielle et à la CAH des textes, on considère diverses variantes de cette analyse (cf. §3).

## 2 Croisement du corpus principal (CAD + CSH) avec le lexique $\Delta$ & de mots outil

Comme on l'a expliqué au §1.2, l'analyse se prête à de multiples variantes: mais nous ne rendrons compte ici que d'une seule:

lexique: 92 formes de mots outil:  $\Delta$  & - {sous}

textes : 254 articles: (191 CAD + 63 CSH)

tableau écrété par le programme 'rangreter', cf. §1.2, in fine.

### 2.1 Analyse de correspondance

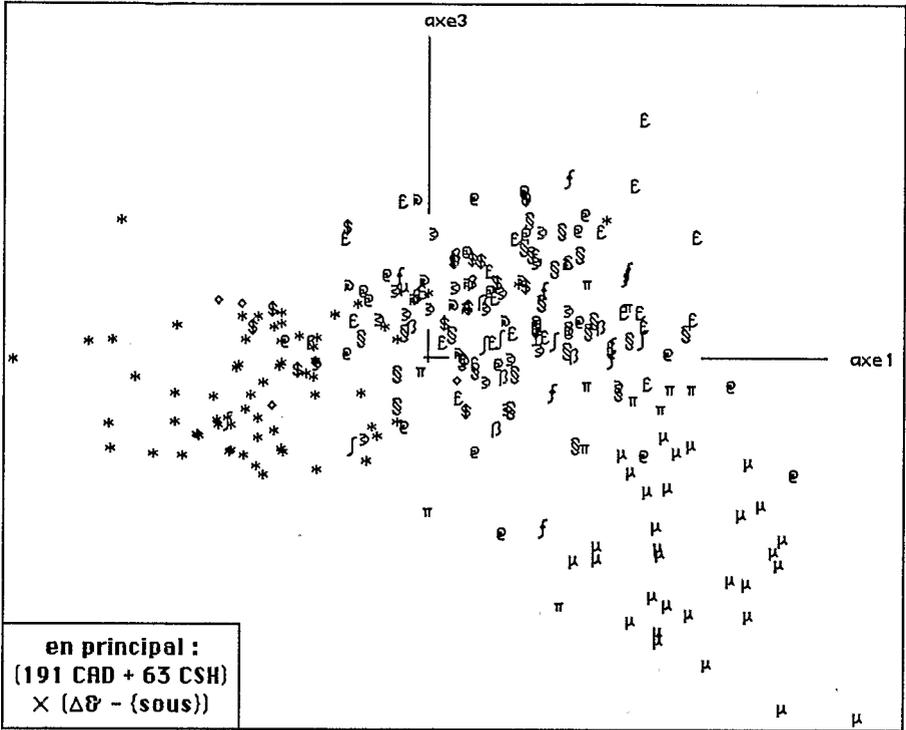
#### 2.1.0 Le tableau des valeurs propres

254 textes (191 CAD+ 63 CSH)  $\times$  92 mots ( $\Delta$  & - {sous})

trace :	1.637e-1									
rang :	1	2	3	4	5	6	7	8	9	10
lambda :	272	183	81	60	55	46	42	39	33	31 e-4
taux :	1662	1119	495	366	335	279	259	241	202	192 e-4
cumul :	1662	2782	3276	3642	3977	4256	4515	4756	4958	5150 e-4

Du 1-er rang au 6-ème, le taux d'inertie décroît fortement; à partir du 6-ème rang, la décroissance est lente et régulière; avec toutefois une dénivellation un peu plus marquée entre 8-ème et 9-ème, ce qui nous a incité à considérer des classifications effectuées dans l'espace engendré par les 8 premiers axes factoriels (cf. §2.2.1).





dans CAD. Vers (F2<0) se mêlent divers symboles: mais prédominent © (ou Ⓓ), cartographie, ® (ou Ⓔ), données par régions, et §, médecine.

Dans le plan (1,3), CSH se concentre au voisinage du demi-axe (F1<0). Les sigles μ occupent le quadrant (F1>0; F3<0). Nous ne tenterons pas de caractériser les symboles qui s'accumulent dans le quadrant (F1>0; F3>0).

μ: Mathématique ou Physique: ex. μSta, stabilité des sous-espaces principaux; μRel, relativisation du continuum spatio-temporel;...

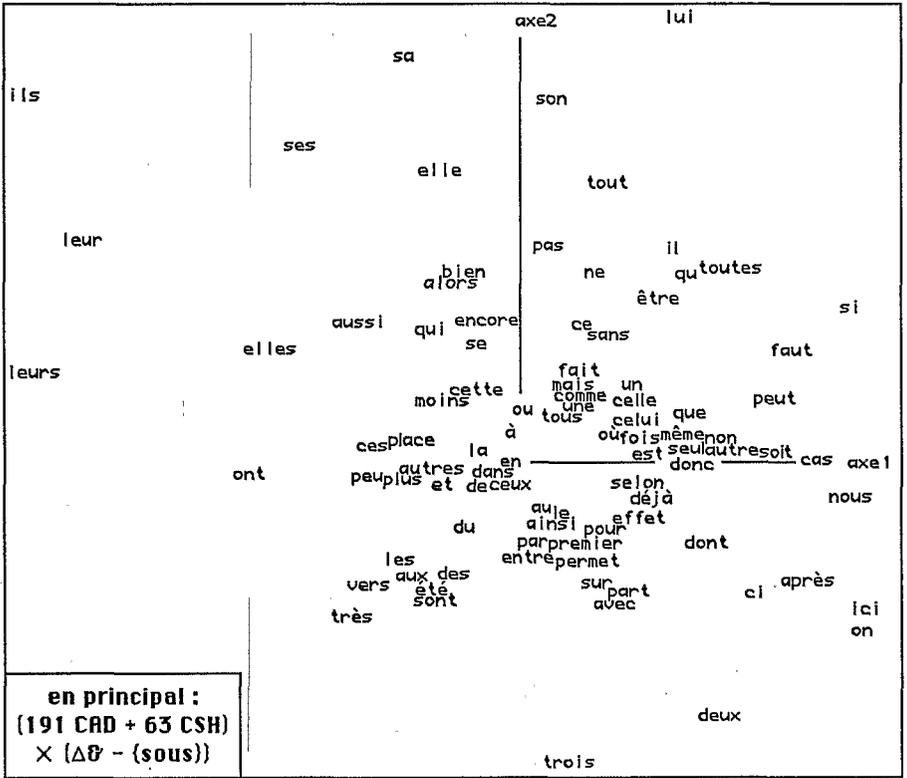
π: Programmation: ex. πCum, programme d'extension d'un tableau par des cumuls de lignes et colonnes;...

f: Philosophie: ex. fqal, qualité et quantité en philosophie et en a. des données; fOrd, l'ordinateur au service du logicien;...

£: Linguistique: ex. £prl, analyse factorielle de la parole; £etr, reconnaissance de l'auteur d'un texte d'après les caractères utilisés; £ñOr, typologie de textes espagnols du siècle d'or;...

\$: Économie: ex. \$Lé1, décomposition d'une matrice de Léontief (1-er article); \$\$sé2, séries chronologiques de l'histoire monétaire (2-ème article);...

β: Bourse (subdivision de \$): ex. βHKg, la bourse de Hong Kong en février-mars 1987; βFm, cours de 26 titres à Paris de Déc 1986 à Juin 1990;...

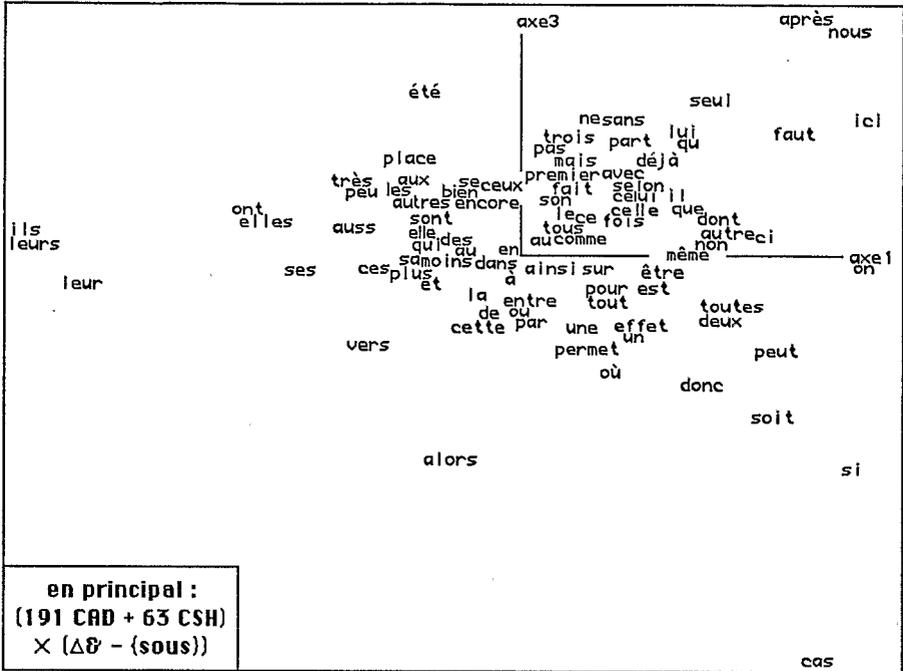


2.1.2 Le nuage J des formes de mots

Sur l'axe 1, {on nous si}, (F1>0), associé à CAD, s'oppose à {ils leur leurs elles}, (F1<0), associés à CSH. Du côté (F2>0), les plus fortes contributions à l'axe 2 viennent de {il ils pas son}; du côté négatif, elles viennent de {des les deux on}.

On remarquera la dispersion sur l'axe 1 des formes pronominales des troisièmes personnes: {ils elles elle il lui}. De même, {dont} et {qui} s'opposent dans le plan (1,2): qui est dans le quadrant (F1<0; F2>0); {dont} est dans (F1>0;F2<0).

Dans le quadrant (F1>0; F3<0) des textes mathématiques, μ, se détachent {cas si soit donc peut}: assurément, ces mots sont des outils ordinaires de la démonstration; mais, d'une part, le lexique a été choisi de telle sorte qu'on peut être assuré que l'usage en est fréquent, non seulement dans CAD, mais également, dans CSH, hors de tout contexte mathématique; et, d'autre part, les



exposés mathématiques de *CAD* se bornent généralement à des schémas de preuve, sans entrer dans les détails techniques.

## 2.2 Classification ascendante hiérarchique

Les éléments sont agrégés de telle sorte que soit minima l'inertie intérieure aux classes; et que, corrélativement, soit maxima l'inertie du nuage des centres de classes. Les résultats obtenus dépendent de l'espace ambiant choisi: ici, avec  $\text{card}J=92$ , l'espace des profils est de dimension 91: une CAH d'après 91 facteurs tient compte de l'ensemble des affinités entre mots et textes du corpus, sans filtrage; au contraire, avec un moindre nombre de facteurs, on se restreint aux dimensions générales qu'on peut attendre de retrouver dans tout corpus analogue.

On'a effectué des CAH d'après 91 facteurs (100% de l'inertie) et d'après 8 facteurs (47,5%); le nombre 8 étant choisi d'après la suite des valeurs propres (ainsi qu'on l'a expliqué au §2.1.0). Dans la suite, sont présentées en détail, avec graphiques dans les plans (1,2) et (1,3) et tableaux de contributions mutuelles, les CAH fondées sur 91 facteurs. On montre, au §2.2.1, sur l'exemple de I, qu'il n'y a pas de différence essentielle entre CAH fondées sur 8 ou 91 facteurs.

c | CAH(8) : Partition en 16 classes : Sigles des textes de la classe c

---

"488	Øfbr	ØOxy	£Rep	SECG	ØLoi	£prl	£Aur	@Src	@stg	*38	§rét	@Ths	µnot	§Pra
"470	Øtki	@man	§SeH	@str	@rch	@chê	*52							
"475	§Lé1	§Arb	§Lé2	@Grc	*09	@Pét	*11							

---

"475		*28	*27	*60											
"486		*03	*65	*62	*14	*43	*63	*61							
"478		*36	Øtk1	*59	*35	*45	*04	*22	*40	*39	*42	*32	*41	*20	*44
"489		*29	*15	*12	*24	ØImm	*37	@ngr	*21	*16	*33	*34	*64	*19	*05
		*23	*10	*26	*25	*30	*53	*56	*07						

---

"492	§cdr	£Lat	£ñOr	£frL	£Dia	@Imm	@vgn	§Jap	§Frn	£Par	§cib	fPrê	*06	*48
"476		*13	*55	*51	*18	*01	*08	*50	*57					
"481		*54	*17	*02										
"480	fusu	fAËc	fOmn	*49	*58									

---

"490	@µVi	§§Nr	£Arb	πPCc	@Cao	@pub	£Gns	§Esp	§Cmc	£SMF	@Jrd	§@ct	@§Tr	@Sal
	@Étu	Øvir	@Mrf	@Mrt	@com	§for	@Elz	@ODA	@Par	@§Qb	@Caf	§&DA	§gel	Øtk2
	§Se4	£Chn	§sST	§gên	§HKg	@Chô	£etr	@stt	§ana	§@Br	§&hX	@Arb	@Vin	£@F@
	§Sé1	@Pnt	@pGr	§Sé2	@prf	§@Bq	@lPr	£HxG	§§pr	@rch	§hml	@fum	@Cal	§cmp
	@Thr	§pél	@µCt	@Alg	£Tr2									
"477	£hxL	@Alg	£mot	§SIC	@Mor	£NT2	£ibr	@&Ps	§&nx	§hm2	£NT1	§HLT	µVAC	@µar
	§§âg	@Exm	§MzT	§2sé	§usi	§eff	§ADN							
"479	πInd	@Prd	π@dg	£Tr1	πprs	§d&d	πCum	§hm3	πCré					

---

"483	£Mt2	£Mt1	fqal	fAri	§voi	§Mrt	@§SF	fLgS	£exp	£sty	@&ps	fDia	§Imp	Ødia
"487	πCor	π@rt	£vox	@µch	@Goo	£cry	£prl	fOrd	µEnt	@µCh	µcah			

---

"491	πpop	µbar	§Tnd	µidé	µprs	πlin	µInB	µMlt	µrec	µniv	µQs1	µSim	µBl3	µBl2
	@Hst	µEPR	µméD	µnq0	µblD	µdéC	µinh	µfus	µbl1	µnqT	µInM	µvoi	µPrt	µRel
	µCv1	µCv2	µbrd	µGut	µΔel									

---

"488		499	503	506
"470	_____	_____	_____	_____
"475	_____	496	501	_____
"486	_____	_____	_____	_____
"478	_____	495	_____	_____
"489	_____	_____	_____	_____
"492	_____	498	502	_____
"476	_____	_____	_____	_____
"481	_____	494	_____	_____
"480	_____	_____	_____	_____
"490	_____	500	_____	505
"477	_____	497	_____	_____
"479	_____	_____	_____	_____
"483	_____	493	_____	504
"487	_____	_____	_____	_____
"491	_____	_____	_____	_____

2.2.1 Classification de l'ensemble I des 254 articles

Nous considérons simultanément deux partitions: les classes de la partition, CAH(91), fondées sur 91 facteurs sont désignées par la lettre i suivie d'un nombre: {i485 i488 i492 ...}; celles de la partition, CAH(8), fondée sur 8 facteurs sont notées {"488 "470 "475 ...}.

c | CAH(91) : Partition en 15 classes : Sigles des textes de la classe c

---

i485	Øfbr	ØOxy	ƒprl	ƒDia	ƒfrL	ƒRep	ƒAur	@ngr	SSeH	*52	@rch	@str	@stg	@Src
i488	ƒtk1	*38	@man	\$Lé2	\$Lé1	@Pét	\$Arb	@Grc	*04	*09	*11			

---

i492	*20	*32	*44	*62	*41	*31	*33	ƒtki	*35	*45	*22	*36	*59	*40
	@chê	*42	*39	*07	*26	*53	*56	*29	*23	*05	*34	*37	*15	*24
		*10	*19											

---

i477	*63	*61	*43	*65	*14	*03								
i480	*28	*60	*27	*64	*16									

---

i209	*18													
i484	*50	*51	*06	*08	*01	*13	*30	ƒImm	*25	*55	ƒhxL	*57	*21	*12
i490	*58	*17	*02	*54										

---

i420	ƒAÉc	ƒMt2	ƒMt1											
i472	ƒPrê	ƒMrt	ƒpêl	ƒFrn	ƒcib	ƒHxG	ƒLat	ƒAOr						
i493	ƒPar	ƒOrd	ƒusu	ƒOmn	*48	ƒAri	@SSF	ƒLgS	ƒqal	*49	ƒdia	ƒvox	ƒvoi	ƒexp
	ƒesty	@Goo	ƒmot	@&ps	ƒImp	ƒprl	ƒdia							

---

i489	ƒpop	ƒTnd	ƒdêc	ƒBl3	ƒinh	ƒbar	ƒMlt	ƒEPR	ƒVAC	@μCh	@μch	ƒcah	ƒEnt	ƒbl1
	ƒbl0	ƒbl2	ƒInM	@Hst	ƒnq0	ƒnqT	ƒfus	ƒInB	ƒniv	ƒSim	ƒGut	ƒQs1	ƒmêd	ƒbrd
	ƒΔel	ƒCv2	ƒCv1	ƒPrt	ƒvoi	ƒRel								

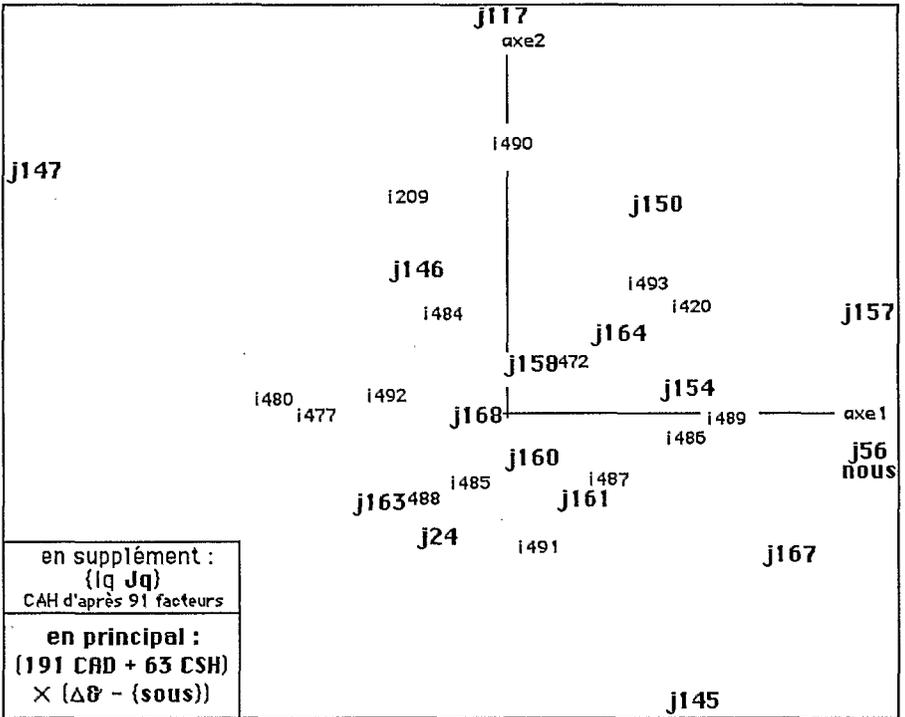
---

i491	SECG	ØLoi	ƒrét	ƒsST	ƒ\$Nr	@Ths	@Mrf	@Mrt	πInd	@Prd	@SQb	ƒChn	@Cao	ƒEsp
	@Par	@μCt	@Caf	@Alg	πPCc	@μVi	@ODA	@Elz	@étu	ƒCmc	ƒSMF	ƒfor	@\$Tr	@Sal
	ƒ&DA	@Vin	ƒ&nX	ƒ\$âg	ƒ\$é2	@Thr	ƒcmp	Øvir	ƒ@F@	@com	ƒfum			
i487	@Alg	@vgn	@Imm	@Mor	@pGr	@prf	ƒcdr	@Cal	ƒgên	ƒ@Bq	ƒHKg	@1Pr	ƒJap	ƒ\$é1
	@Arb	ƒnot	ƒPra	@pub	ƒ@Br	ƒ@ct	@Jrd	ƒGns	@rch	@Pnt	ƒArb	ƒeff	ƒ\$pr	ƒshml
	ƒrec	@μar	ƒcry	π@dg	@stt	ƒTr2	ƒNT1	ƒASIC	ƒNT2	ƒibr	@&Ps	ƒprS	ƒ\$e4	ƒshm2
	ƒtk2	ƒana	ƒ&hX	ƒetr	ƒpêd	πlin	@Chô	ƒSgl	ƒS2é	ƒHLT	ƒADN	ƒusi	ƒMzT	@Exm
i486	π@rt	πCor	ƒd&d	ƒhm3	πCum	ƒTr1	ƒprS	πCré						

---

i485	_____	499	_____	503	_____									
i488	_____													
i492	_____	500	_____											
i477	_____	496	_____											
i480	_____													
i209	_____	494	_____	502	_____	504	_____						506	_____
i484	_____													
i490	_____													
i420	_____	498	_____											
i472	_____	495	_____											
i493	_____													
i489	_____								505	_____				
i491	_____	501	_____											
i487	_____	497	_____											
i486	_____													

On remarque d'abord, dans CAH(91), une branche i500 = {i492 i477 i480}, qui contient près des 2/3 des articles de CSH, presque sans mélange de CAD; lui correspond, dans CAH(8), "501 = {"475 "486 "478 "489}. Ces classes s'écartent de l'origine dans la direction (F1<0). À cette branche de CSH s'agrèe une branche, i499 ou "499, qui comprend des articles divers de CAD avec 4 ou 5 de CSH. Le reste de CSH constitue, dans CAH(91), la branche i502; tandis que dans CAH(8) on a deux subdivisions {"476 "480}



qui, au sein de "502, adhèrent à des subdivisions de CAD. De ce point de vue, la classification CAH(91) apparaît plus claire que CAH(8); mais celle-ci a le mérite d'offrir, au sommet de la hiérarchie, une dichotomie qui oppose à "506 (où rentrent toutes les subdivisions citées jusqu'ici) une branche "505 comprenant exclusivement des articles de CAD; branche dont le contenu se retrouve partagé entre i505 et i498.

§: Médecine (§ à l'imitation du caducée): §hm1, variations de l'état hémodynamique et du taux sérique d'un produit (1-er article); §2sé, efficacité de 2 sédatifs;...

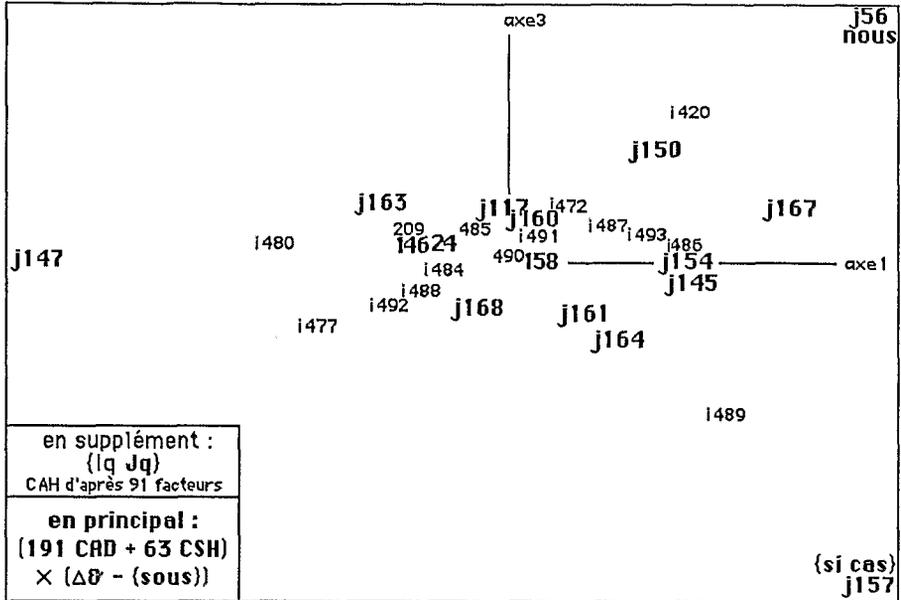
‡: SIDA (subdivision de §): ‡Imm, État du syst immunitaire et histoire clinique des patients infectés par VIH;...

©: données par Régions ou par pays: ©Cao, marché mondial du cacao de 1976/1977 à 1986/1987; ©étu, Étudiants à Paris VI par cycle d'étude et nationalité: 1975-90;...

⊙: Cartographie (subdivision de ⊙): ⊙Alg, Carte de pronostics électoraux pour l'Algérie; ⊙Sal, salaires du secteur privé dans les dép français: 1976-87;...

@: Enquêtes, Sociologie, Psychologie, Ergonomie, descriptions diverses: @Jrd, Enquête scolaire effectuée en Jordanie; @&Ps, validité des échelles en psychologie et psychiatrie et corrélations psychosociales;...

∅: Sciences naturelles (dans peu de cas): ∅vir, typologie de virus de plantes d'après leur protéine de capsid;



Les articles mathématiques de CAD, signalés au §2.1.1 comme occupant le quadrant ( $F1 > 0$ ;  $F3 < 0$ ), s'agrègent dans  $i489 \approx "491$  (cette dernière étant plus pure). La subdivision  $i486 \approx "479$  comprend principalement des notices de programmes: on y trouve aussi  $\text{ftr1}$ , article étiqueté comme linguistique, mais consacré, pour une bonne part, aux programmes d'élaboration des textes et de tri des formes.

La branche  $i504$  se subdivise en  $i502$  (CSH, déjà citée: { $i209$   $i484$   $i490$ }) et  $i498$  (CAD: { $i420$   $i472$   $i493$ }); dont le thème est: Bourse linguistique philosophie. Corrélativement, il nous paraît que les articles de CSH renfermés dans  $i502$  ont un caractère littéraire ou spéculatif marqué: \*18, le Gandhisme en écho; \*50, Force des langues vernaculaires...; \*51, la jeunesse face au multilinguisme... Les 6 subdivisions de  $i504$  sont, dans la partition retenue, les seules dont le centre ait un facteur  $F2$  nettement positif. Aucune n'a un  $F3$  nettement négatif (comme c'est le cas pour  $i489 \approx$  mathématiques).

Certains articles de CAD participent à la fois de plusieurs genres dont les symboles sont alors réunis dans leur sigle; voici quelques cas:

$\pi@rt$ , Programme de représentation cartographique des résultats d'une analyse multidimensionnelle; notice d'utilisation;

$\text{§\$Nr}$ , Consommation de produits pharmaceutiques prescrits ou non, dans la région Nord-Pas-de-Calais;

$@\text{Jsd}$ , Attitudes des adultes de 11 pays vis-à-vis des malades du SIDA.

c | CAH(91) : Partition en 16 classes : Mots de la classe c

---

j158	à fait comme ce tout en
j146	se aussi ses alors encore bien ces qui
j168	et cette la où dans vers de
-----	
j161	permet par pour effet soit entre sur
j160	au place mais du déjà ainsi ceux tous premier autres le
-----	
j24	des
j163	été ont très peu moins plus sont aux les
-----	
j147	ils elles leur leurs
=====	
j56	nous
j145	trois deux
j167	après avec dont autre part ici on
-----	
j157	si cas
j154	non donc est que selon fois même celle celui ci
j164	ou une un peut être
-----	
j117	sa elle lui son
j150	ne pas faut sans seul toute qu il

---

j158	_____	169175	_____	178	_____	181	_____
j146	_____		_____		_____		_____
j168	_____		_____		_____		_____
j161	_____	172	_____		_____		_____
j160	_____		_____		_____		_____
j24	_____	171	_____	180	_____		_____
j163	_____		_____		_____		_____
j147	_____		_____		_____		_____
j56	_____	177	_____	182	_____		_____
j145	_____	173	_____		_____		_____
j167	_____		_____		_____		_____
j157	_____	176	_____	179	_____		_____
j154	_____		_____		_____		_____
j164	_____		_____		_____		_____
j117	_____	170	_____		_____		_____
j150	_____		_____		_____		_____

### 2.2.2 Classification de l'ensemble J des 92 mots et étiquetage de la classification des 254 textes

On considère la partition en 16 classes définie par les 15 nœuds les plus hauts de la CAH fondée sur la totalité des 91 facteurs. Au sommet de la hiérarchie, J se sépare en deux branches, j181 et j182, dont les centres s'opposent suivant l'axe 1.

Les subdivisions retenues de j181, ont leur centre du côté ( $F1 < 0$ ): la seule exception nette est j161 ( $F1 > 0$ ); corrélativement, les subdivisions retenues de j182 ont leur centre du côté ( $F1 > 0$ ), à l'exception de j117. La branche j182 se subdivise en  $j177 = \{j56 j145 j167\}$ , ( $F2 < 0$ ); et  $j179 = \{j157 j154 j164 j117 j150\}$ , ( $F2 > 0$ ).

i485:163+++	145++			
i488:168+			150--	
i492:168+	160+	167-		
i477:168+	147+++	167-	150--	
i480:163+++	147+++	167-	154-	
i209:146++	147++++			
i484:	147++			
i490:			117+++	150++
i420:158+	56++	167++		150++
i472:	56++		117++	
i493:163--	56+		154+	150++
i489:163--		167+	157+++	154+
i491:		167+		
i487:		167++		
i486:	56++	167+		

*Étiquetage des classes de textes  
par les classes de mots*

Parmi ces dernières subdivisions, j157 = {cas si} se signale par sa position excentrique ( $F1 > 0$ ;  $F3 < 0$ ), associée aux textes mathématiques de i489.

Une classe de textes est définie par un profil d'association avec les mots; et peut, en bref, être caractérisée par des liens, positifs ou négatifs, avec des classes de mots. Nous publions un tableau des principaux de ces liens.

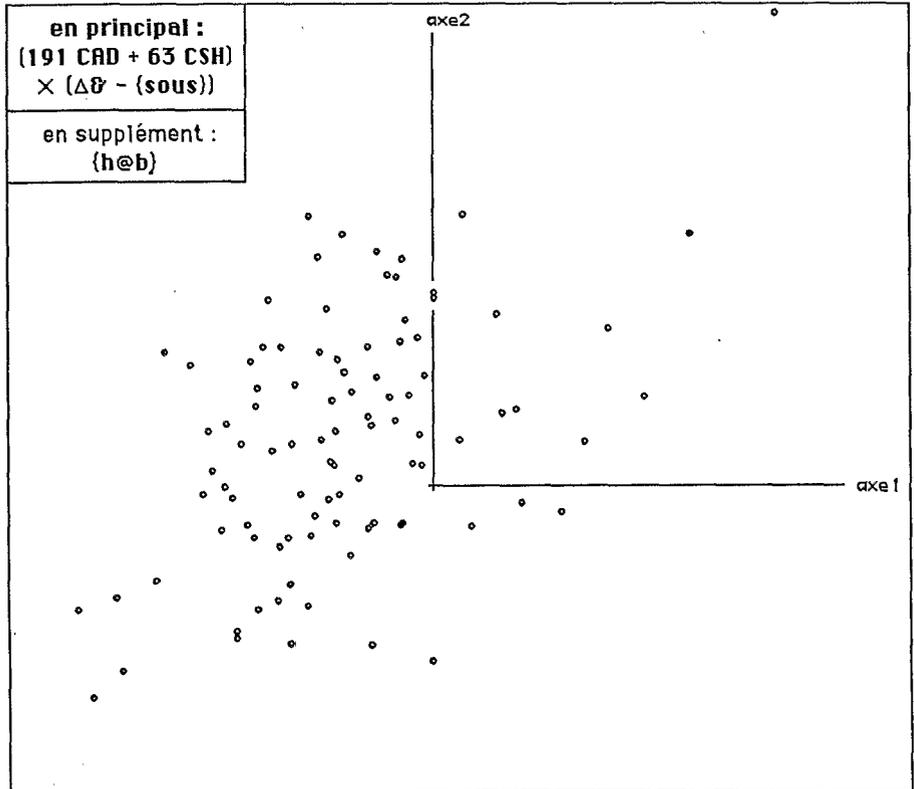
Il apparaît que les subdivisions de j181 sont liées positivement à des classes de textes de *CSH*; tandis que les subdivisions de j182 sont liées aux textes de *CAD*. Toutefois, la classe i480 de *CSH*, située à l'origine dans le plan (1,3) et écartée sur ( $F2 > 0$ ), est fortement liée à j117 = {sa elle lui son}, subdivision de j182.

On remarque dans j163 le groupe {très peu moins plus} de formes exprimant une dénivellation, positive ou négative. La classe i480 de *CSH* a une forte affinité avec j163; c'est le contraire pour i489, mathématiques, et i493 linguistique, philosophie.

Les pronoms {nous on}, qui peuvent être substitués l'un à l'autre dans bien des contextes des articles du corpus (on a vu  $\approx$  nous avons vu), sont dans j177 et à l'extrémité positive de l'axe 1, associés à *CAD*; cependant {on}, avec  $F2 \ll 0$ , s'oppose à i493, ( $\& f$ ); tandis que {nous}, avec  $F3 \gg 0$ , s'oppose à i489, ( $\mu$ ).

Les négations {non} {ne pas} sont dans j179, ( $F1 > 0$ ;  $F2 > 0$ ); mais {ne pas}, dans j150, se signalent par ( $F2 > 0$ ;  $F3 > 0$ ); tandis que {non} s'écarte sur ( $F1 > 0$ ), avec la subdivision j154 (qui comprend également les démonstratifs {celle celui ci}).

Dans les classes de mots on croit reconnaître d'une part, la rhétorique de l'auteur qui argumente pour rallier le lecteur à son point de vue; d'autre part, les outils pour des références précises sans lesquelles chaque pronom propose une énigme.

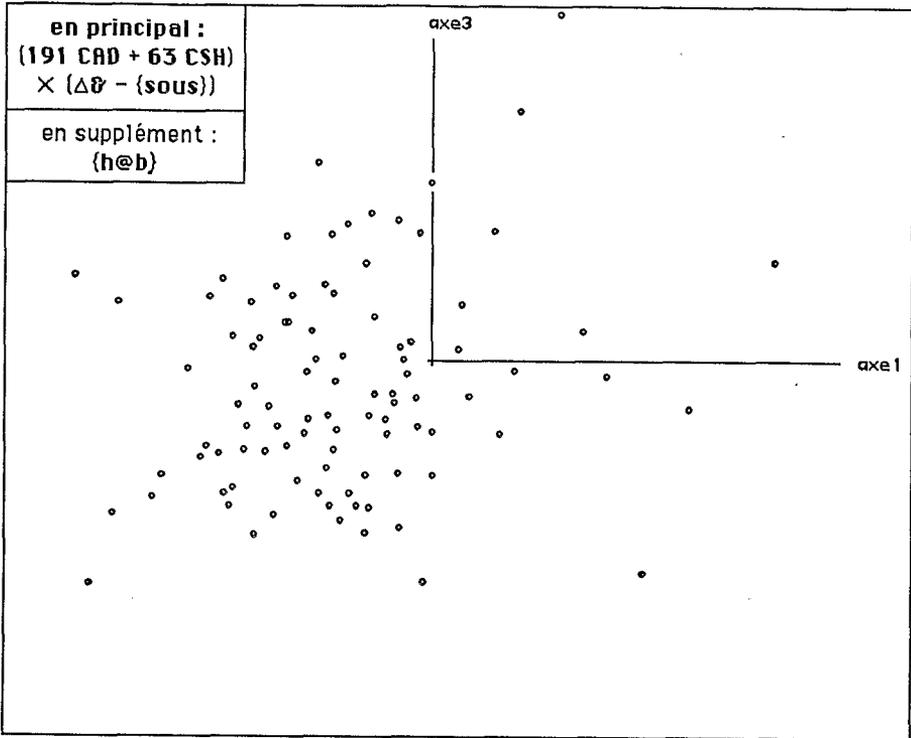


### 2.3 Place des fragments de thèse dans l'espace des articles

Ainsi qu'on l'a dit au §1.1, de la thèse h@b de M. BLANCHET, on a retenu 95 fragments ou §. On a, dans ces fragments, dénombré les occurrences des mots du lexique  $\Delta\&$ ; ce qui permet d'adjoindre les fragments en supplément au nuage N(I) des articles. Le nuage N(h@b) des fragments est représenté sur des graphiques séparés, mais avec la même échelle et le même cadre que pour le nuage N(I) considéré au §2.1.

Dans le plan (1,2), il apparaît que N(h@b) recouvre le sous-nuage N(CSH); mais en déborde pour empiéter sur le domaine où prédomine N(CAD). Dans le plan (1,3), dans la direction de l'axe 3, de part et d'autre du demi-axe ( $F1 < 0$ ), l'étalement est plus grand pour N(h@b) que pour N(CSH).

Afin d'apprécier d'une autre manière la disposition de N(h@b) relativement aux deux sous-nuages d'articles, on a affecté chacun des fragments de la thèse à l'article le plus proche, soit dans l'espace des profils,



rapporté à 91 axes, soit dans l'espace engendré par les axes 1 à 8. Dans le premier cas, 46 fragments sur 95 sont affectés à des articles de *CAD*; dans le deuxième cas, 37 seulement. Il se confirme que la thèse, bien que rentrant dans le domaine de *CSH*, empiète sur *CAD*.

Cependant, parce que, dans le corpus, les articles de *CAD* sont trois fois plus nombreux que ceux de *CSH*, on présume que des proximités contingentes, inscrites sur les axes de rang élevé, ont favorisé l'affectation de h@b à *CAD*; particulièrement avec 91 axes, cas où les affectations à *CAD* sont le plus nombreuses. Afin d'éliminer de tels effets, on a affecté les fragments de la thèse à un petit nombre de centres: les centres de gravité des 16 classes de la partition de I construite dans l'espace engendré par les axes 1 à 8 (cf. §2.2.1): 37 fragments se trouvent affectés à l'une des classes {"488 "470 "492} où *CAD* prédomine nettement sur *CSH*.

Reste à préciser, quant à l'usage des mots outil, la zone intermédiaire où h@b déborde de *CSH* dans *CAD*: pour cela on effectue une analyse où sont en principal fragments de thèse et articles des deux revues.

### 3 Analyse du corpus des 254 articles avec le corpus des 95 fragments de thèse

349 textes (191 CAD+ 63 CSH+ 95 h@b) × 92 mots ( $\Delta\& - \{\text{sous}\}$ )

trace :	1.861e-1									
rang :	1	2	3	4	5	6	7	8	9	10
lambda :	275	189	88	79	62	49	47	44	40	37 e-4
taux :	1477	1016	473	423	335	261	252	236	213	197 e-4
cumul :	1477	2493	2966	3389	3724	3985	4238	4473	4686	4883 e-4

#### 3.0 Analyse préliminaire et choix du lexique

On analyse d'abord le tableau principal croisant les 349 textes avec le lexique ( $\Delta\& - \{\text{sous}\}$ ); et on effectue des CAH sur I et J dans l'espace engendré par les axes 1 à 8. Nous publions une partition des textes en 12 classes; avec l'arbre étiqueté en termes de facteurs.

On retrouve, pour les 254 articles, une distribution générale déjà vue au §2.2.1, pour CAH(91). Au sommet de la hiérarchie, se détache une branche 691 qui compte 48 articles de CSH (\*); s'y joignent 39 fragments de h@b (dont les sigles commencent par le caractère '°'). Le reste, 696, se subdivise en 694 et 695: cette dernière branche se compose quasi exclusivement d'articles de CAD, avec une subdivision, 678, de textes mathématiques.

Dans 694, on remarque la subdivision 682, composée exclusivement de 44 fragments de h@b; 682 s'agrège à 681 et 685, où prédomine CAD, notamment en  $\{\text{£ f } \beta\}$ . Le reste de 694 forme les subdivisions 680 et 676, où prédomine CSH.

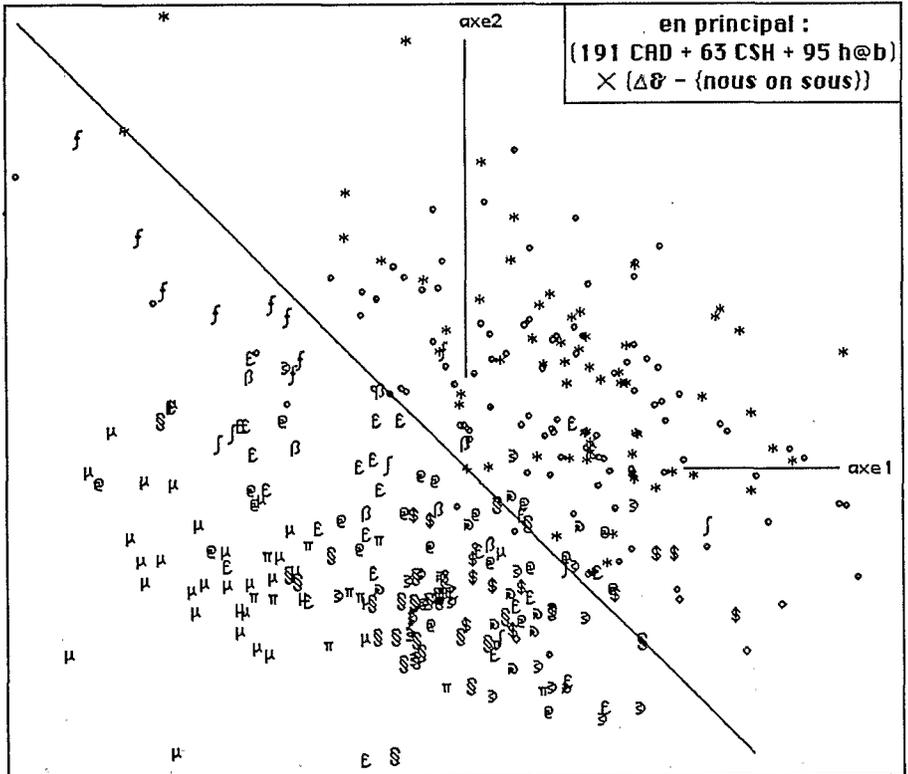
Les subdivisions 682 (44 '°') et 681 (CAD; £...) se signalent par une fréquence maxima du pronom 'nous': on présume donc que ce pronom délimite une grande partie de h@b, et la met en rapport avec CAD. De plus, 682 se détache sur (F3<0), associée à {de la}, qui contribuent pour 28% à l'inertie de l'axe 3.

Dès qu'on élimine {nous on} du lexique, l'agrégation à CAD d'une partie de h@b disparaît. On a une nette dichotomie entre CAD et (CSH + h@b); et, dans la branche de la CAH qui comprend '\*' et '°', une classe de 69 '°', où les seuls intrus sont 2 CAD.

Si on restreint le lexique aux 88 formes de ( $\Delta\& - \{\text{sous nous on de la}\}$ ), la dichotomie subsiste; mais la séparation entre CSH et h@b est moins nette: on a toutefois une classe composée de 48 '°', 1CAD et 2 '\*'.

Dans la suite, on considérera exclusivement l'analyse fondée sur le tableau principal  $349 \times (\Delta\& - \{\text{sous nous on}\})$ ; avec des CAH dans l'espace engendré par les huit premiers axes factoriels.





### 3.1 Analyse de correspondance: 349 textes × 90 mots

349 textes (191 CAD+ 63 CSH+ 95 h@b) × 90 mots (Δ& - {sous nous on})

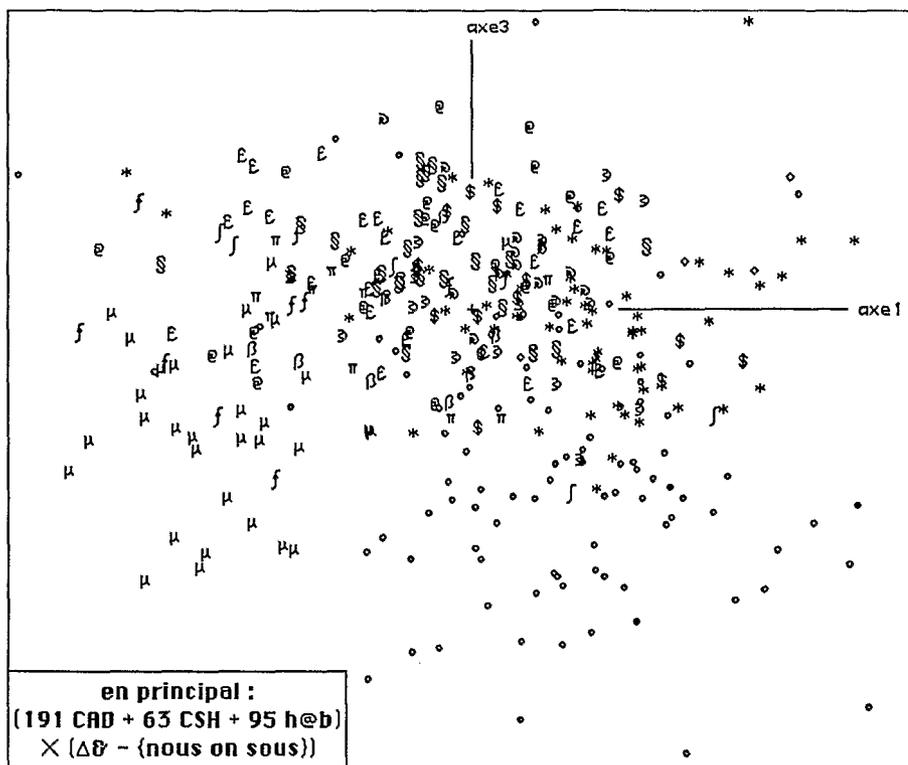
trace : 1.775e-1

rang :	1	2	3	4	5	6	7	8	9	10	
lambda :	237	176	89	72	55	48	46	42	38	36	e-4
taux :	1336	990	503	406	311	268	259	235	214	201	e-4
cumul :	1336	2327	2830	3236	3547	3815	4075	4309	4524	4725	e-4

#### 3.1.1 Le nuage I des textes

Dans le plan (1,2) comme dans le plan (1,3), chaque texte est marqué par le caractère initial de son sigle: i.e. une étoile '\*' pour tous les 63 articles de CSH; un symbole du thème pour les 191 articles de CAD, 'o' pour les fragments de la thèse h@b.

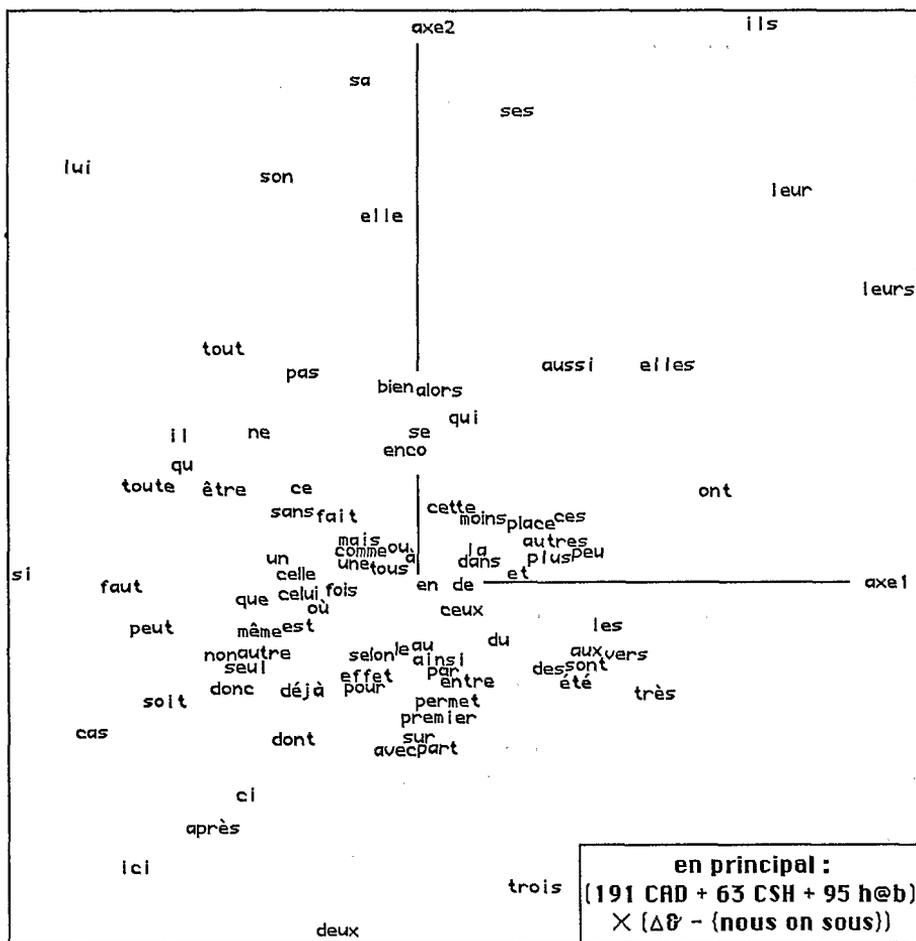
Dans le plan (1,2), se séparent, approximativement, (CSH + h@b) situé au-dessus de la deuxième bissectrice des axes, et (CAD) situé au-dessous; du côté (F1 < 0), on distingue un groupe d'articles philosophiques (f: F2 > 0); et un amas dense d'articles de mathématiques (μ: F2 < 0).



Du plan (1,3), on remarque particulièrement la bande inférieure ( $F3 < 0$ ): avec, à gauche ( $F1 < 0$ ), un amas de 'μ'; puis, s'étalant jusqu'à l'extrémité ( $F1 > 0$ ), un sous-nuage presque pur de '°', fragment de la thèse h@b.

Nous avons dit que, dans le plan (1,2), la deuxième bissectrice sépare (CAD) de (CSH + h@b). Cependant si l'on met à part quelques points très excentriques dans le quadrant ( $F1 < 0$ ;  $F2 > 0$ ), le nuage N(I) a, comme on pouvait l'attendre d'après la décroissance des valeurs propres, la forme d'une ellipse dont le grand axe est l'axe 1. La deuxième bissectrice n'est donc pas un axe de symétrie pour N(I).

À une telle configuration générale est lié le fait qu'à l'extrémité positive de l'axe 1, on a une accumulation de textes de (CSH + h@b) {\*, °}; et du côté ( $F1 < 0$ ) une accumulation d'articles de (CAD). Cette distribution sur l'axe 1 se voit le plus clairement dans le plan (1,3): avec, à gauche, des caractères divers, parmi lesquels {μ} prédomine; et, à droite, {\*} dans le quadrant ( $F1 > 0$ ;  $F3 > 0$ ); {°} dans { $F1 > 0$ ;  $F3 < 0$ }.

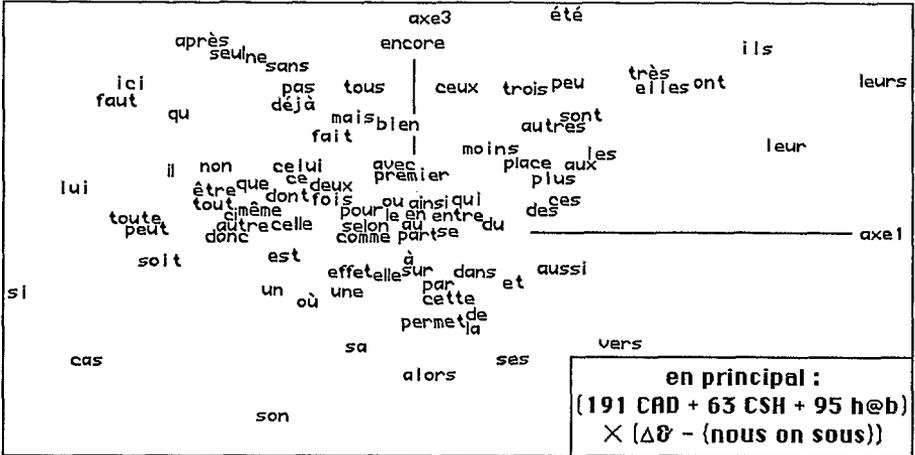


### 3.1.2 Le nuage J des mots

Dans le quadrant ( $F1 > 0$ ;  $F2 > 0$ ) du plan (1,2) les mots {ils leur ses leurs elles}, outils des troisièmes personnes, se détachent sur une zone d'où CAD est absent.

Dans le quadrant opposé, ( $F1 < 0$ ;  $F2 < 0$ ), {cas soit ci après ici}, apparaissent associés aux articles mathématiques de CAD.

Dans la bande ( $F3 < 0$ ) du plan (1,3), on a, à gauche, ( $F1 < 0$ ), les mots {peut si soit cas} associés à  $\mu$ ; et, vers ( $F1 > 0$ ), {vers aussi ses et dans de la...}, sur la zone réservée à h@b.

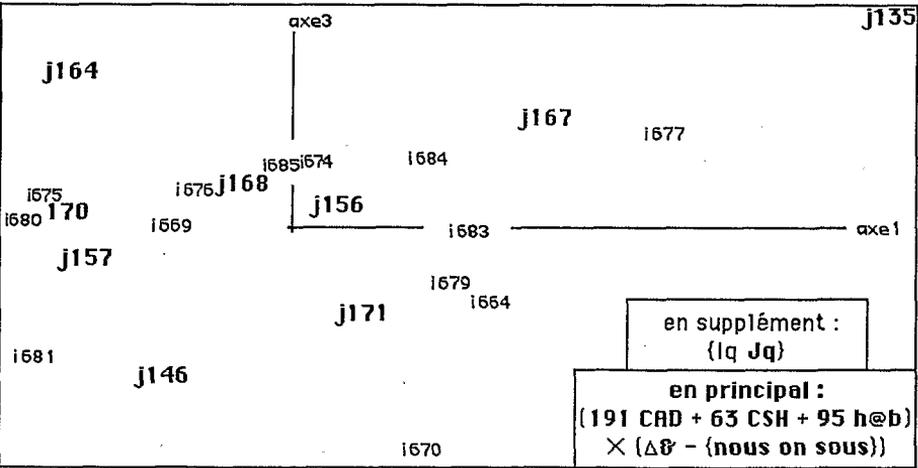


Le détail de la répartition des mots sera mieux compris, d'après les classifications de I et J, sur les nuages des centres de classe.

### 3.2 Classification ascendante hiérarchique

#### 3.2.1 Classification de l'ensemble I des 349 articles et fragments

La définition des classes est assez claire pour dispenser d'en donner un long commentaire. Comme on pouvait l'attendre d'après l'examen de I dans le plan (1,2), les deux branches 695 et 696, en lesquelles se divise I, sont à peu près séparées par la 2-ème bissectrice du plan (1,2); et leurs centres s'opposent suivant la 1-ère bissectrice. On a déjà dit que 695 s'identifie, à peu près, à (CSH + h@b); avec les (2/3) de h@b constituant la classe 670.



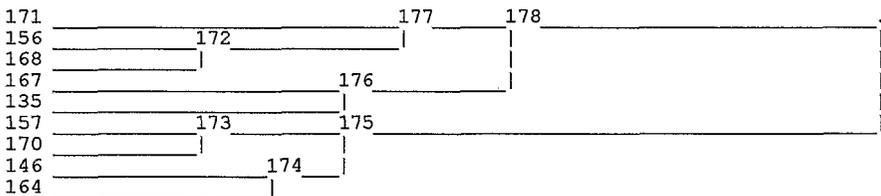
c	CAH(8) Partition en 13 classes : Sigles des textes de la classe c
i670	°A01 °H08 °H04 °E11 °H02 °I16 °G01 °C01 °B05 °E01 °I09 °G10 °I18 °D09 °H06 °H07 °H01 °C02 °H05 °I08 °B03 °C04 °F03 °I06 °I07 °D08 °B04 °G02 °F01 °D07 °D06 °E09 °G09 °I19 °I13 °E05 °E04 °E08 °I12 °B02 °E10 °G06 °B01 °G05 °I10 °E06 °D01 }tki °H11 °H09 °H12 βJap °G03 °D05 °I17 °G11 °F07 °H10 °J09 °I04 °H03 *15 °F06 °E03 °G08 °I11 °F02 °F08 °C03 °F05 °D04
i664	\$Lé1 *09 \$Arb \$Lé2 ©Grc ©Pét *11 °F04 *44 *31 @rch *20 *41
i683	*65 °E02 *03 °D02 °G07 *62 *14 °G04 *43 °I15 *53 °J03 °I02 *34 *33 ffrL *56 *07 *26 *01 *25 *29 *05 *23 *30
i679	\$sST \$rét @Imm @vgn *21 *52 °E07 °J07 °J12 *24 *10 *19 *16 }tk1 *59 °H13 °H14 °J02 *36 *39 *04 °I01 £Dia *45 °J01 *22 *35 *42 *32 *40
i677	*63 *61 *27 *64 *60 *28
i675	*54 °J04 *49 *58
i685	*12 *06 *48 £Par °J08 βcib fPrê @Alg £hxL °K01 *55 °D03 *18 *13 *51 *08 *57 *50 °J10 £sty *17 *02
i684	∅Oxy ∅fbr *38 @Ths @chê @pub μnot \$Pra ∅loi \$ECG °I14 \$Esp ©\$Tr £Rep £Arb \$SNr @μVi °I03 @Jrd @pGr βpôl ©Pnt ∅vir ©ODA ©Elz ©Mrf ©Mrt £prl @man \$SeH @str @stg @Src
i674	©Alg ©Caf £Gns ©Par ©\$Qb ©Cao πPCc ©étu ©com ©Sal \$for \$Cmc £SMF @μCt £Tr2 \$ect @rch \$cmp @Thr \$MzT @fum £Chn }tk2 \$gel ©Chô }ana ©Arb \$gên \$QBr \$shX \$Se4 \$hml \$Spr \$usi £etr @stt \$&DA \$&nx \$&âg \$ADN Seff \$2sé @Exm
i680	£vox μeah @μCh μVAC @Goo @μch μEnt £prl \$Imp }exp }dia fAri £MT2 £MT1 fOmn fdia fusu fAÉc fOrd °J05 °J06 βvoi ©SSF fLGs fgal
i676	£ñOr \$cdr £Lat @prf βFrn μrec βMrt \$Sé1 @Vin £©F© \$Sé2 ©1Pr ©Cal \$©Bq £HxG \$hm2 ©μar £NT1 βSIC @Mor £NT2 £cry £mot £ibr @&Ps @&ps
i669	*37 £Aur @Prd βHKg π©dg °I05 }Imm @ngr π©rt πCor \$HLT πInd £Tr1 \$hm3 \$d&d πCum πprs πCré
i681	@Hst μEPR μnq0 °J11 μΔel μméd μCv1 μCv2 μGut μRel μvoi μbrd μPrt πpop \$Tnd πlin μprs μInβ μMlt πméd μQs1 μSim μdéc μblδ μbar μniv μinh μbl2 μBl3 μbl1 μInM μnqT μfus
i670	F3--- 690 693 F1++ 695
i664	_____ 695: vers (F1>0; F2>0)
i683	_____ 686 689 _____
i679	_____
i677	F1+++ _____
i675	F1-- _____ 691 F2++++ _____
i685	_____
i684	_____ 688 _____ F2-- _____ 696
i674	_____ 696: vers (F1<0; F2<0)
i680	F1--- _____ 692 _____ 694 F1-- _____
i676	_____ 687 _____
i669	_____
i681	F1--- F2- F3- _____

Dans 696, on a distingué 6 subdivisions de CAD: celles-ci n'ont pas toutes une interprétation nette; mais on a 681 ≈ mathématiques; et, dans les subdivisions de 692, une concentration de linguistique et de philosophie; avec, dans 669, prédominance des notices de programmes.

i670 :	171+++		146++	164--
i664 :	156++			164--
i683 :		135++	157-	
i679 :	171++			
i677 :		167+++	135++++	
i675 :		167--		146+++++ 164+++
i685 :			135+++	146+ 164++
i684 :		167+++		146-
i674 :	171-	168++++		146-
i680 :		167--	157+ 170+	164++
i676 :	168+++			164++
i669 :			157++++	
i681 :	167--		157+ 170++++	

*Étiquetage des classes de textes  
par les classes de mots*

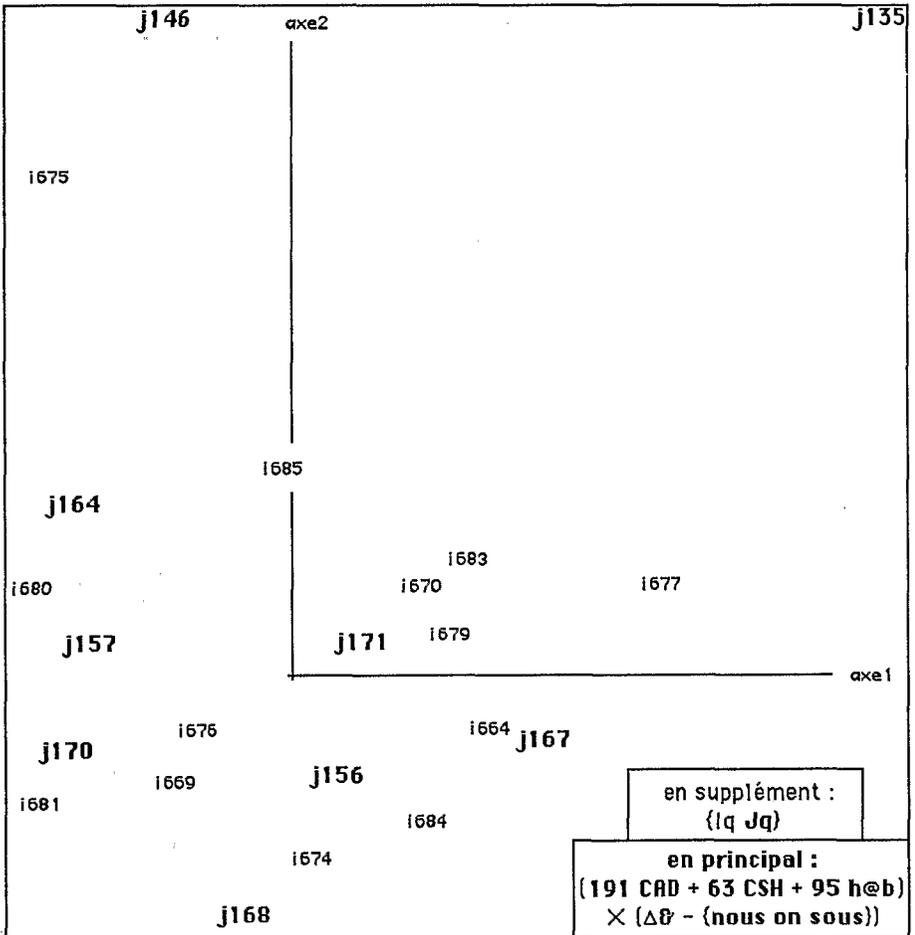
c	CAH(8) Partition en 9 classes : Mots de la classe c
171	à une en permet par de alors cette aussi qui dans place se la et
156	premier le du vers
168	trois deux ici après ainsi entre au effet selon sur part avec ci dont
167	des ces peu ceux moins plus elles ont été très autres sont aux les
135	ils leurs leur
157	être soit peut ou un
170	où autre celui même que pour non donc est si cas
146	elle lui son sa ses
164	tous bien encore celle fois comme mais fait ce déjà sans seul faut pas ne qu toute tout il



**3.2.2 Classification du lexique J = Δ& - {nous on sous}**

À la différence de l'ensemble I des textes, l'ensemble J des mots se partage en deux branches, 178 et 175, dont les centres s'opposent suivant l'axe 1. On note, dans le plan (1,2), très à l'écart vers (F2>0), les deux subdivisions j135 {ils leurs leur} et j146 {elle lui son sa ses} dont la première rentre dans 178 et la seconde dans 175; j135 renvoie à la 3-ème personne du pluriel, et j146, à la 3-ème personne du singulier; or {elles} (féminin pluriel) est dans 167, subdivision de 178; tandis que {il} (masculin singulier) est dans 164, subdivision de 175.

Plusieurs des classes déjà vues au §2.2.2, se retrouvent, peu modifiées, dans la présente partition.



Pour saisir le sens de l'ensemble des affinités entre mots et textes mises en évidence par les traitements statistiques, il convient d'examiner, avec l'aide du tableau d'étiquetage des classes de textes par les classes de mots, les graphiques présentant, sur les plans (1 × 2) et (1 × 3) issus de l'analyse du tableau 349 × 90, les projections des centres des classes des deux partitions retenues: classes de textes (initiale i, caractère maigre sur les graphiques) d'une part; classes de mots (initiale j, caractère gras sur les graphiques) d'autre part.

Considérons, à titre d'exemple, l'opposition, sur l'axe 2, entre des textes littéraires, culturels ou idéologiques, et des textes plus techniques (mathématiques, physique, programmation).

On a du côté ( $F2 > 0$ ) les classes de textes i675 et i685, expliquées par les mots des classes j146 et j164:

**j146**={elle lui son sa ses}, prépondérants dans i675,

**j164**={tous bien encore celle fois comme mais fait ce déjà sans seul faut pas ne qu' toute tout il};

mots auxquels il faut encore ajouter, pour la classe i685,

**j135**={ils leurs leur}, prépondérants dans cette dernière;

et de l'autre côté de l'axe 2, ( $F2 < 0$ ), la classe i681, expliquée par les mots:

**j170**={où autre celui même que pour non donc est si cas},

et, dans une moindre mesure,

**j157**= {être soit peut ou un}.

Cette dernière classe de textes, i681, comprend en majorité des articles de mathématiques et de physique (dont le sigle commence par  $\mu$ ), des notices de programmes (de sigle  $\pi...$ ); et, en outre:

@Hst: [MOD. HISTOG.]: Inadéquation de la description d'un histogramme par moyenne et écart type;

§Tnd: [TENDINITES]: "Efficacité comparée de deux gels dans les tendinites superficielles"; article qui, n'étant bien corrélé qu'avec l'axe 2 très négatif, n'est pas représentatif de la classe i681; et que nous éliminons;

°J11: extrait de thèse, mal représenté dans l'espace des 8 premiers axes (comme le signale le listage d'analyse des correspondances; non publié); ce qui interdit d'en interpréter la place dans la CAH.

Aux deux derniers textes près, §Tnd et °J11, tous les articles de la classe i681 ont donc des thèmes s'apparentant aux mathématiques.

Considérons maintenant la composition des deux premières classes citées, i675 et i685, qui s'opposent à i681 sur l'axe 2. La classe

i675={\*54 °J04 \*49 \*58},

comprend quatre textes, de caractère littéraire:

\*54: "L'émigration dans la littérature égyptienne contemporaine"

°J04: extrait de thèse: où l'on cite largement des réponses à la question ; "Pour quelles raisons êtes-vous venus habiter à Marne-la-Vallée?"; plus précisément, les réponses émanant d'habitants qui, n'ayant pas véritablement choisi d'habiter dans cette ville nouvelle, font état d'un 'Choix imposé'.

\*49: 'Compte rendu du livre: vers une révolution francophone'

\*58: "Emancipation des jeunes et migrations interafricaines", où l'on rapporte successivement, en style parlé, les propos d'un jeune homme puis d'une jeune femme qui décrivent les péripéties de leur vie.

La classe i685 est beaucoup plus nombreuse:

i685={\*12 \*06 \*48 £Par °J08 ßCib fPrê @Alg £hxL °K01 \*55 °D03 \*18 \*13 \*51 \*08 \*57 \*50 °J10 £sty \*17 \*02};

de cette classe, nous éliminons quelques textes que le listage d'analyse des correspondances signale comme étant mal représentés en projection sur l'espace engendré par les 8 premiers axes: {°J08 ßcib °K01}. Les textes restants s'apparentent au genre culturel ou idéologique:

- \*12: Réforme agraire et fiction municipale au Mexique
- \*06: Comment être pauvre et citoyen: l'Eglise au Brésil
- \*48: La contribution des langues au développement : domaine de la santé
- £Par: Le Parménide de Platon: répartition des vocables
- fPrê: [ARIST. PRET]: Aristote et le prêt à intérêt
- @Alg: [QUEST. ALGERIE]: Questionnaire sur l'Avenir politique de l'Algérie
- £hxL: [HEXA. DACTYL. LATIN]: composition métrique de l'hexamètre dactylique latin
- \*55: Campagnes et villes de l'Inde: deux cultures antagonistes?
- °D03: L'environnement est un processus actif et continu (fragment de thèse: généralités se référant à divers ouvrages et concernant les interactions des composantes de l'environnement entre elles)
- \*18: Gandhisme: traditions de résistance et résistance au colonialisme
- \*13: Retour en arrière: les haciendas dans leur gloire: Veracruz
- \*51: La jeunesse face au multilinguisme en Casamance (Sénégal)
- \*08: Du salarié au petit producteur libre: expériences de réforme agraire
- \*57: Légende à revoir: l'ouvrier du bâtiment brésilien sans feu ni lieu
- \*50: Langues vernaculaires et exogamie linguistique: le cas du Vaupés colombien
- °J10: "Les habitants envisagent-ils de déménager?" (fragment de thèse analysant l'ensemble des réponses à cette question; mais où les propos cités occupent relativement peu de place)
- £sty: [PROGRAMME]: Programme de recherches en stylométrie (où il s'agit, non de programmation mais de deux pages de considérations générales sur la façon d'aborder la stylométrie)
- \*17: La lecture des traces dans des sociétés sans écriture
- \*02: Emploi dans le bâtiment et insertion des migrants à Bogota, Colombie

Tous ces textes, rappelons-le, sont caractérisés par l'abondance des mots: {il leur leurs} tandis que les textes plus littéraires de la classe i675 sont caractérisés particulièrement par les mots de la classe **j146**:

**j146**= {elle lui son sa ses}.

Le lecteur pourra poursuivre, par lui-même, l'interprétation détaillée de la CAH; lui manqueront, toutefois, les listages numériques, non publiés dans le présent article, et auxquels nous avons parfois eu recours.

#### 4 Sommaire et perspectives

Les études de vocabulaire [CAD XII-XVII] parus dans CAD en 1993 portaient sur 191 articles, tous issus de CAD. On aborde ici la stylistique générale des publications scientifiques en adjoignant à ces 191 articles de CAD, des textes relevant d'autres méthodes scientifiques, à savoir: 63 articles publiés par l'ORSTOM dans les *Cahiers des Sciences Humaines (CSH)* avec, en complément, 95 fragments d'une thèse de psychologie qui traite d'Urbanisme, en rendant compte d'une enquête sur l'habitat, effectuée dans la ville nouvelle de Marne-la-Vallée.

L'étude est fondée sur l'emploi qui est fait, dans les textes, de mots outil; en acceptant comme tel tout mot qui n'évoque pas directement le thème de l'article. Les études antérieures citées ci-dessus ont montré que, malgré cette précaution, le thème et le genre interfèrent avec le style.

Les mots outils sont ici choisis d'après les seuls articles de CAD et de CSH; les fragments de thèse, bien que n'ayant pas participé à l'établissement de la liste des mots outil, seront analysés sur ces mêmes mots. Quant au choix lui-même, il se fait très naturellement en prenant l'intersection des deux ensembles constitués par les 200 formes les plus fréquentes (sans distinction de mot plein ou vide) des sous-corpus issus respectivement de CAD (191 articles) et de CSH (63 articles); il suffit alors d'éliminer, de cette intersection, les formes (peu nombreuses) qui n'ont pas toujours valeur d'outil dans leurs emplois recensés dans les deux sous-corpus de CAD et de CSH. On a ainsi retenu 92 mots outil.

Le tableau de base, considéré au §2, croise les 254 articles, issus de CAD et de CSH, avec les 92 mots outil retenus. On l'écrète par le programme 'rangreter'.

D'une façon générale, les textes sont classés par CAH, d'après les facteurs issus d'analyse de correspondance préalable. Toutefois, les fragments de la thèse ont, au §2, un traitement particulier, puisqu'ils sont adjoints en éléments supplémentaires au tableau principal  $254 \times 92$ : la CAH étant faite sur

les 254 articles principaux, on recherche, pour chaque fragment de thèse, décrit par ses facteurs, l'article principal dont il est le plus proche; et, de même, après avoir choisi une partition en 16 classes de la CAH, on recherche le centre de classe dont le fragment de thèse est le plus proche.

Dans un second temps (§3), on prend le corpus entier, y compris les fragments de thèse, en principal. Mais alors, on est amené à ôter deux mots du lexique de mots outil. Analyse de correspondance et CAH portent donc sur un tableau principal  $349 \times 90$ .

Dans les deux cas, les articles issus de *CAD* d'une part, de *CSH* d'autre part, constituent deux sous-nuages qui ne s'interpénètrent pas, sauf sur une frange limitrophe des deux sous-nuages; et, dans l'analyse du tableau  $349 \times 90$ , il apparaît que les fragments de thèse ne se mêlent qu'imparfaitement aux articles de *CSH*; dont ils sont, toutefois, plus proches que du corpus de *CAD*.

Pour la classification, on se place, au §3, non dans l'espace à 89 dimensions, qui est l'espace ambiant au nuage, mais dans l'espace plus restreint engendré par les 8 premiers axes, espace où le nuage se projette de façon plus dense. Il faut toutefois signaler que la qualité de représentation des fragments de thèse sur ces 8 premiers axes est souvent médiocre, voire très mauvaise; car la thèse n'a pas été prise en compte pour le choix du lexique (sinon pour ôter deux mots). Lors de l'interprétation des résultats, il convient d'éliminer les textes (fragments de thèse ou articles) mal représentés.

Les résultats de cette étude confirment ceux des études précédentes citées ci-dessus: le style d'un article est marqué, voire imposé, par le genre, sinon le thème, du sujet dont il traite. Nous croyons l'avoir montré par le commentaire détaillé qui termine le §3.

Ce premier pas vers une étude générale est donc encourageant et devrait se poursuivre par la recherche des caractères stylistiques d'œuvres, de plus en plus nombreuses et diverses, conformément à ce qui est proposé dans [PROGRAMME], in *CAD* Vol. XIII, 1988.

La quasi totalité des textes publiés aujourd'hui ont été saisis et mis en forme sur support informatique. Il faut souhaiter que la loi impose désormais le dépôt de ces fichiers; dûment normalisés suivant une convention internationale, dont l'UNESCO devrait assumer le patronnage. À l'analyse statistique des textes, s'ouvriraient ainsi des perspectives illimitées...