LES CAHIERS DE L'ANALYSE DES DONNÉES

A. AÏT HAMLAT

Analyse des répétitions et indexation automatique des documents

Les cahiers de l'analyse des données, tome 9, n° 2 (1984), p. 173-204

<http://www.numdam.org/item?id=CAD_1984__9_2_173_0>

Les Cahiers de l'Analyse des Données
Vol. IX - 1984 - nº 2 - p. 173-204

ANALYSE DES RÉPÉTITIONS ET INDEXATION AUTOMATIQUE DES DOCUMENTS

[IND. DOC.]

par A. Aït Hamlat (1)

1 *Du problème aux données* : En considérant brièvement le rôle d'un système d'information documentaire (§ 1.1), on découvre la nécessité d'une extraction automatique du langage représentatif d'un corpus. Jusqu'à présent on a principalement tenté de fonder cette extraction sur des modèles probabilistes (§ 1.2). Ayant eu accès à un corpus de 266 compte-rendus de visite issus de chercheurs du groupe pétrolier Elf-Aquitaine (§ 1.3), nous avons pu explorer une autre voie (*) en soumettant à l'analyse des données divers tableaux issus de ce corpus (§§ 2 et 3).

1.1 *Systèmes d'information documentaire* : L'information cognitive (connaissances scientifiques techniques...) donne lieu à deux formes d'organisation

- les bases de données documentaires et/ou textuelles
- les banques de données (numériques)

Nous nous intéressons ici aux bases de données documentaires et/ou textuelles sous l'aspect plus général des systèmes d'information documentaire .

Un service de documentation doit généralement conduire une série d'opérations que l'on appelle la chaîne de traitement de l'information documentaire et dont voici les grandes étapes :

a) collecte des documents à enregistrer.

b) indexation : caractérisation des sujets dont traite le document par juxtaposition des concepts indépendants les uns des autres et pouvant être combinés au moment de la recherche. Ces concepts sont extraits d'un thésaurus (ensemble organisé de descripteurs, sélectionnés pour leur aptitude à représenter avec le maximum d'efficacité le contenu des documents d'un champ donné).

(*) Pour des applications antérieures de l'a. des données en documentation (cf. G. Seguin Thèse U.C. Bernard 1978 ; et D. Mullet, Marquis Thèse 1981.

(1) Docteur de 3° cycle en statistique ; stagiaire de recherche CNRS.

0339-3097/84/02 173 32/$ 5.20/© Gauthier-Villars

c) condensation ou rédaction de résumés.

d) recherche des références des documents qui répondent à une question plus ou moins précise d'un utilisateur.

e) diffusion de l'information recueillie aux utilisateurs d'un centre soit de façon systématique, (bulletins signalétiques, catalogues, etc.), soit de façon sélective, chaque personne étant informée régulièrement des nouveautés apparues dans son domaine d'activité.

L'intégration de nouveaux documents dans un centre (étape c) de la chaîne de traitement nécessite une analyse minutieuse de leur contenu et reste une opération complexe. Le développement de procédures permettant une aide à l'indexation automatique est donc intéressant.

Sans entrer dans le détail nous proposons un schéma de la chaîne de traitement dans un système d'information documentaire.

1.2 Extraction automatique du langage représentatif d'un corpus suivant un modèle probabiliste : En bref, on considère comme fixé l'ensemble D des documents du corpus. Relativement à un lot particulier de n documents, le calcul des probabilités permettra de déterminer si oui ou non le mot M est caractéristique du lot (i.e. Si la proportion dans ce lot de documents contenant le mot M est significativement supérieure à celle calculée pour un lot de même effectif n issu de D par tirages successifs aléatoires indépendants.

De façon précise on fait choix d'une fonction de pertinence et après fixation d'un seuil, on détermine les mots retenus comme les plus pertinents, pour caractériser le lot n au sein de D.

Plus généralement, le schéma probabiliste a été le fondement méthodologique de la plupart des travaux de statistiques dans le domaine lexical. Le modèle binomial avec la loi normale et la loi de Poisson imposées dans cette discipline par Muller, Guiraud,... en sont les expressions les plus connues.

Parmi les travaux de cette nature, très peu se sont appliqués en fait à caractériser les concepts contenus dans le corpus analysé, la détermination du vocabulaire représentatif d'un corpus était établie en vue essentiellement de la comparaison avec un autre corpus et n'avait pas spécifiquement pour objectif la caractérisation du contenu sémantique de la collection considérée. C'est au contraire avec cet objectif en vue que nous avons entrepris de soumettre à l'analyse des données l'intéressant corpus auquel nous avons eu accès.

1.3 Présentation du corpus

1.3.1 Origine des données : le corpus concerne l'activité technologique des chercheurs du groupe Elf Aquitaine dont voici un aperçu : exploration et production pétrolière dont la complexité technologique requiert des recherches fondamentales, notamment universitaires, sur les produits tensio-actifs et les micros-émulsions, etc. ; gestion de l'énergie ; amélioration des rendements d'usage des produits pétroliers, en particulier par des techniques électroniques de régulation ; stockage de l'énergie, énergies nouvelles, sans oublier qu'en matière de communication, d'innovation et de traitement de l'information au sein de l'entreprise, le groupe consacre beaucoup d'efforts.

1.3.2 *Description préalable des concepts du corpus* : De façon précise, le corpus réunit 268 compte-rendus de visite, se rapportant à la nature et aux résultats des activités internes et des contacts extérieurs de chercheurs du Groupe, compte-rendus produits par les chercheurs eux-mêmes.

Bien que les documents soumis à l'étude recouvrent un domaine assez vaste de la connaissance scientifique, il a été arrêté après lecture de ces documents et compte tenu des nouveaux arrivants, une liste hiérarchisée de thèmes qui servira de référence dans la suite du travail et sera utile pour la présentation des résultats. Cette liste est constituée de dix sept grands thèmes pour chacun desquels nous avons distingué les sous-thèmes correspondants, eux-mêmes subdivisés en sous-parties lorsqu'il y a lieu.

Nous donnons ci-dessous la liste des thèmes, avec à titre d'exemple, le détail du thème V (Chimie et Agrochimie).

1 - EXPLORATION-PRODUCTION
2 - RAFFINAGE-DISTRIBUTION
3 - GESTION DE L'ENERGIE
4 - BIOTECHNOLOGIES
5A- CHIMIE
- 5.a - Matériaux composites
- 5.b - Synthèse
- 5.c - Techniques de séparation, purification
- 5.c 1 - Chromatographie
- 5.c 2 - Résines
- 5.d - Habitat (isolation, peintures...)
- 5.e - Thioorganique, tiochimie
- 5.f - Chimie fine
- 5.g - Chimie lourde

5 B- AGROCHIMIE
6 - ENVIRONNEMENT
7 - ACCUMULATEURS
8 - BUREAUTIQUE (voir 15)
9 - UTILISATIONS DE BIOMASSE EN VUE DE BESOINS ENERGETIQUES
10 - CAPTEUR
11 - COMBUSTION
12 - COMMUNICATION (voir 13)
13 - COOPERATION INDUSTRIELLE
14 - ECONOMIE
15 - INFORMATIQUE
16 - INNOVATION
17 - SCIENCES SOCIALES

1.3.3 *Elaboration des données qui nous ont été soumises* :Avant tout traitement, les textes se présentent sous la forme de documents écrits de longueur variable, ordinairement d'une vingtaine de lignes environ. Ces textes sont ensuite enregistrés sur support électronique en vue de leur traitement.

Nous décrirons brièvement les étapes de ce traitement effectué par une chaîne de programmes mise au point au C.E.A. par l'équipe d'Andreewski et Fluhr.

La saisie se fait au moyen d'un programme conversationnel permettant un certain nombre de corrections et d'éditions. Le texte qui n'a encore subi aucun traitement linguistique est alors découpé par un automate en chaînes de caractères candidates à être des mots. La difficulté réside dans la diversité des séparateurs dont aucun ne détermine à coup sûr la fin du mot, notamment du fait des sigles tels que U.N.E.S.C.O. .

On a recours à l'analyse morphologique pour reconnaître les différentes unités lexicales (mots) dont se composent les documents ; détecter les éventuelles erreurs typographiques ; associer un certain nombre d'informations (valeurs grammaticales hors contexte, genre, nombre,...) utiles pour la suite.

Des expressions comme "en majeure partie", "mettre en oeuvre",... sont regroupées en une seule entité. Cela est réalisé à l'aide d'un dictionnaire d'expressions idiomatiques d'environ 1.300 entrées. Le reste des mots composés est reconnu par des algorithmes. Certaines ambiguïtés du langage relèvent de catégories grammaticales différentes. Par exemple CAR peut être nom ou conjonction, MARCHE substantif ou participe, etc. . De telles ambiguïtés sont résolues automatiquement par un programme d'analyse syntaxique à apprentissage sans aucune intervention manuelle. En sortie, la phrase doit avoir une structure correcte (deux substantifs ne peuvent se suivre).

Les mots sans aucun caractère informatif sont éliminés sur des critères grammaticaux (mots rentrant dans des catégories grammaticales vides : pronoms, prépositions, articles : le, la, les, l', de, des, aux, ce, ces, nous, vous, ou, par, en effet, à moins de, et, or,... ; les mots à éliminer figurent dans un dictionnaire...). Ne sont conservés (dans la mesure du possible) que les mots pleins, autrement dit : les substantifs, verbes, adjectifs,...

A chaque mot, on associe sa forme standard en distinguant les homonymes et tenant compte des catégories grammaticales. Toutefois, cette chaîne de traitements ne fonctionne pas parfaitement : le système a laissé subsister quelques ambiguïtés, etc. (CF; e.g. § 2.1).

1.3.4 *Structure des données retenues* : La normalisation peut faire perdre une information importante : le fichier brut obtenu à l'issue de ces traitements est une suite de mots (par ordre alphabétique) avec un pointeur vers le (ou les) documents où il apparaît ; la description d'un document se réduit finalement à une suite d'occurrences de mots.

Compte tenu des diverses éliminations, lemmatisations et réductions, le fichier comporte 8 100 mots, totalisant 50.998 occurrences. Nous donnons sur une page la liste des 273 mots les plus fréquents chacun précédé de sa fréquence ; pour les fréquences ($\leq$ 33) on s'est borné à récapituler sur deux colonnes les nombres de mots afférents à chacune. On voit qu'il y a 3 518 *hapax* (mots employés une seule fois dans le corpus) ; 1 272 mots employés deux fois... ; que le mot le plus fréquent est *étude* (f = 312), suivi de *produit* (234) et *charbon* (229). Un lecture rapide de la liste permet de préciser par des mots les thèmes énumérés au § 1.3.2).

LISTE DES 273 PREMIERS MOTS DE LA LISTE

PRECEDES DE LEUR FREQUENCE

312 :	ETUDE			
234 :	PRODUIT			
229 :	CHARBON			
212 :	RECHERCHE			
193 :	TECHNIQUE			
191 :	PROCEDE			
186 :	PROJET			
180 ·	SYSTEME			
178 :	POINT			
174 :	SOLAIRE			
172 .	DONNER			
166 :	ETE			
162	MILLIMETRE	PROBLEME		
158 :	SOCIETE			
147 :	METTRE	PERMETTRE		
146 .	AN			
145 :	COMBUSTION	INDUSTRIEL	INFORMATION	
144 :	ENERGIE	ENERGIE	UTILISER	
143 ·	DEVELOPPEMENT	MARCHE	MARCHE	
140	ESSAI			
138 .	ELF	PROGRAMME		
135 :	DEUX			
131	EST			
130	UNITE			
126 :	LABORATOIRE			
123 :	CHAUFFAGE	PREMIER		
121 ·	ACTION			
118 .	APPLICATION			
115 :	GAZ			
111 :	H.2.0.			
109	PARTIR	PRODUCTION		
108 .	PAC	FAIT	FUEL	
104 .	CHAUDIERE	GRAND		
103	TYPE			
102	COUT	COUT	INTERESSER	OBTENIR
101 .	IMPORTANT			
99 .	DOMAINE	MATERIEL		
96 :	PRESENTER	TRAITEMENT		
94 :	PROPOSER			
93 :	CONTACT	LIT		
92 :	NOUVEAU			
91 :	BOIS			
90 :	GROUPE	PERSONNE	VOIR	10
89 .	INSTALLATION	PILOTE		
86 ·	MISER			
84 :	CHALEUR	MOYEN	TRAVAIL	
83 :	ETUDIER	PAS		
82 :	100			
81 :	ACTUELLEMENT			
80	ACTIVITE			
79	COLLABORATION	DEVELOPPER	INTERET	STOCKAGE
	THERMIQUE	TRAVAILLER		
78 .	ENTREPRISE	FABRICATION	PETIT	
77 :	OBJECTIF	TEMPERATURE	VISITE	
76 :	FRANCE	BASE		
75 :	A.N.	AIR		
74 .	CONTRAT	POMPE	VISITER	
73	UTILISATION			
71 ·	POSSIBLE	PRIX	RESULTAT	
70 :	U.S.A.	SOLUTION		
69 :	NIVEAU	REUNION		
68 :	ENVISAGER			
67	CENTRE	EFFECTUER		
66 :	RAPPORT			
65 :	REALISER			
64	COUR	COURS		
62 ·	FLUIDISER	000		
61 :	POSSIBILITE	TECHNOLOGIE		
60 :	ALLER	BANQUE	CONNAITRE	MOIS
	PREVOIR			
59 :	D.	APPAREIL	CAS	CHARGE
	MOTEUR	SERVICE		
58 :	CRES	BRULER	INTERESSANT	NECESSAIRE
57 .	EGALEMENT	PHASE	PRINCIPAL	20
56 ·	POSER			
55 :	ETRE	ILE	MELANGE	TEMPS
54 :	15			
53 ·	TERME			
52	EQUIPE	CONCERNER	ENSEMBLE	ENVIRON
51	BON	BUREAU	CHIMIQUE	FAIBLE
	NOUVEL			
50 .	BREVET	HUILE	PAYS	
49	ELECTRIQUE	FOURNIR	FRANCAIS	QUALITE
	SCIENTIFIQUE			
48	ACTUEL	DEFINIR	PETROLIER	REACTEUR
	SOUHAITER	SUITE	TROIS	30
47 :	ECONOMIE	AGIR	PRESENT	
46 :		INFORMATIQUE	JOUR	PENSER
45 .	ANALYSE	FIN	FONCTIONNEMENT	LOGICIEL
	MESURE	PROFESSEUR	SUJET	USINE
44		GAZEIFICATION	LUIRE	PARAITRE
	REALISATION	TEST	TROUVER	
43 :	NI ; VOIE	COMPRENDRE	CONSTITUER	DEMANDER
	DISPOSER	EXISTER	FOND	PRODUIRE
42 :	10F	ACCORD	FONCTION	HAUT
	METHODE	VAPEUR	40	
41 :	ECONOMIQUE	COKE	REGULATION	RENDEMENT
40 :	ELEMENT	ELEVER	CERTAIN	CONSTRUCTION
	DIFFERENT	FOUR	INDUSTRIE	KG
	PUISSANCE	TRAITER	1981	
39 .		DISCUSSION	FABRIQUER	GENERAL
	METHANOL	MAIS	MAISON	PILOTER
	PLACE	PROPOSITION	TAILLE	200
38 :		AMERICAIN	CALCULATEUR	CENTRER
	COMBUSTIBLE	CONDITION	DEMONSTRATION	DERNIER
	EMPLOYER	FLUIDE	GESTION	JAPONAIS
	MAINTENIR		OPERATION	
37	AND	I.F.P.	AIDE	ANNEE
	CONDUIRE	EXPERIENCE	MATIERE	NOTER
	TESTER			
36 :	ETAT	DELAPALME	ENGINEERING	MENER
	PRESENTATION	RENCONTRER	SUR	SURFACE
	TUBE			
35 :	ECHANTILLON	I.B.M.	AFFAIRE	APPORTER
	BESOIN	CAPTEUR	INGENIEUR	INVESTISSEMENT
	LOURD	OBJET		
34 :	IDIC	M2	CELLULE	CONSTRUIRE
	EXPLOITATION	EXTERIEUR	PERSONNEL	PORTER
	PRESSION	SYNTHESE	UNIVERSITE	

FRÉQUENCE, NOMBRE

n	V_n
1	3.518
2	1.272
3	688
4	461
5	301
6	238
7	195
8	168
9	122
10	93
11	81
12	67
13	59
14	66
15	47
16	40
17	41
18	35
19	45
20	34
21	36
22	21
23	24
24	20
25	12
26	13
27	19
28	24
29	18
30	18
31	10
32	16
33	13
.	.

2 *A la découverte du corpus*

2.0 Panorama des analyses : Au terme de notre étude nous sommes parvenu à la conclusion que l'analyse des répétitions des mots au sein des documents est le procécé le plus efficace que nous connaissions pour extraire automatiquement le langage représentatif d'un corpus (§ 3). Cependant les très nombreuses analyses effectuées sans recourir à ce procédé, nous ont apporté des résultats assez intéressants pour que nous en rendions compte brièvement dans ce § 2.

Pour énumérer les analyses il est commode de poser les notations que voici :

DOC : l'ensemble des 268 documents, (éventuellement restreint à 250 au § 3).

MOT : un ensemble de mots, variable suivant les analyses : on précisera au besoin par un chiffre le sous-ensemble considéré.

T : un ensemble de tranches du vocabulaire total des 8 100 mots (cf. § 1.3.4) définies par la fréquence ; on désignera par t une tranche et par Mt l'ensemble des mots qu'elle comprend (e. g. la 1-ère tranche M1 comprend les 17 mots de fréquence comprise en 146 et 312).

Ceci posé, on rend compte au § 2.1 d'une première analyse de correspondance DOC ×MOT , portant sur 250 mots. Au § 2.2 on considère le tableau DOC × T où k(doc,t) désigne le nombre total des occurrences des mots de la tranche de fréquence t dans le document "doc". Au § 2.3 on tente l'analyse d'un tableau mixte DOC(T ∪ M1 ∪ M2), où , en bref, on adjoint au tableau DOC × T, les 59 colonnes du tableau DOC × MOT, afférentes aux mots les plus fréquents (tranches 1 et 2). Au § 2.4 on étudie la répartition de l'inertie des points du nuage N(MOT) au sein des différentes tranches de fréquence M1, M2,....

2.1 Première analyse DOC × MOT (268 × 250) : Faute d'un meilleur critère, à la recherche duquel est précisément consacré l'essentiel de notre travail, on a retenu ici les 250 mots les plus fréquents, représentant 18.629 occurrences; (Une analyse globale 268 × 8000 (4500 en ôtant les hapax) aurait été difficile à faire ! Disons pour servir aux chercheurs tentés par une semblable expérience qu'il conviendrait de construire le tableau DOC × DOC suivant

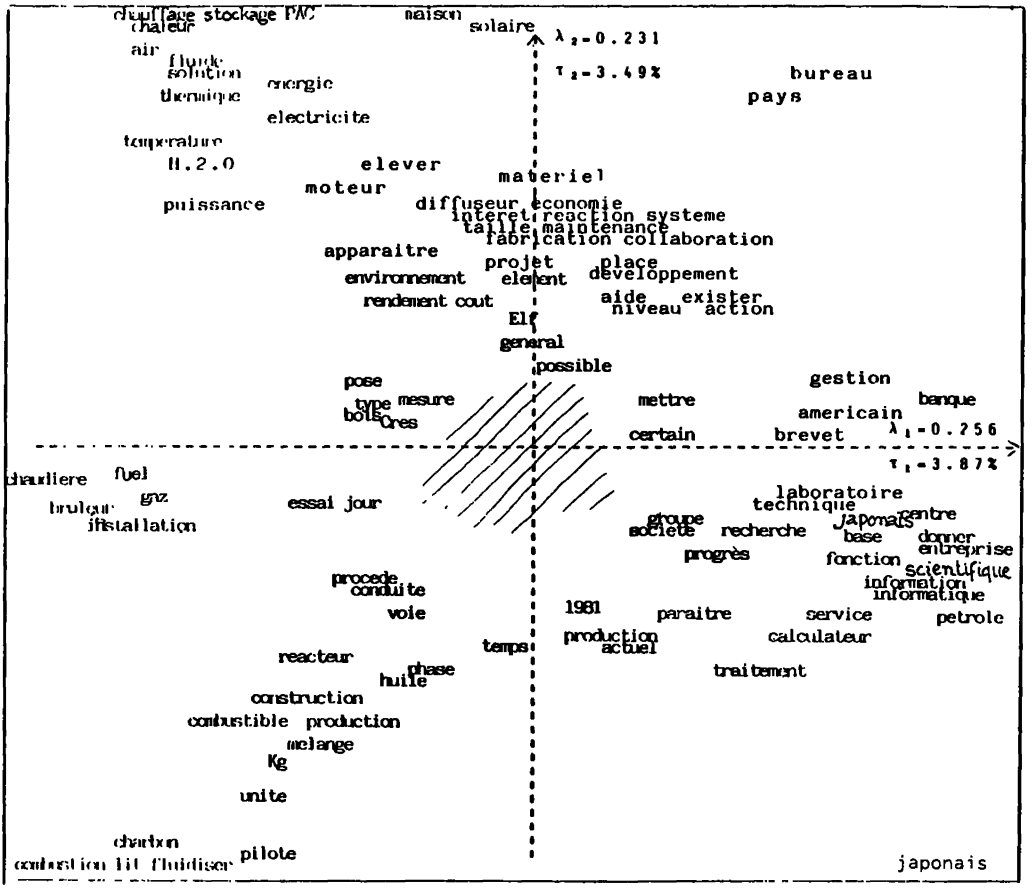
$$k(doc,doc') = \Sigma\{k(doc,m)\,k(doc',m)/k(m) \mid m \in MOT\}$$

où k(doc,m) désigne le nombre d'occurrences du mot m dans doc ; et k(m) le nombre total d'occurrences de m dans le corpus. Ce tableau, analogue d'un tableau de BURT, donne les mêmes facteurs normalisés que DOC × MOT, avec des valeurs propres au carré. On a les facteurs sur MOT en adjoignant une colonne supplémentaire par mot. Le coût de la construction du tableau DOC × DOC est acceptable vue la structure du fichier, cf. § 1.3.4, organisé par mots ; il reste une analyse 268 × 268 ; et un retour au fichier pour inclure les éléments supplémentaires. On comprendra que nous nous soyons borné à un tableau 268 × 250 !)

Les valeurs propres obtenues et les taux sont :

$$\lambda_1 = 0.256,\ \lambda_2 = 231,\ \lambda_3 = 0.190,\ \lambda_4 = 0.160,...$$

$$\tau_1 = 3.9\%,\ \tau_2 = 3.5\%,\tau_3 = 2.9\%,\ \tau_4 = 2,4\%$$



§2.1 CORRESPONDANCE MOTS-DOCUMENTS

plan 1-2, 250 mots les plus frequents .

La considération des listes de contributions et corrélations sur le premier axe a permis de sélectionner de chaque côté (cf. plan 1×2)

F1 > 0	CTR	COR	F1 < 0	CTR	COR
INFORMATION	750	320	CHARBON	410	170
ENTREPRISE	410	220	COMBUSTION	400	240
BANQUE	480	280	CHAUDIERE	230	130
DONNER	330	170	LIT	380	200
SCIENTIFIQUE	280	300	FLUIDISER	280	190
JAPONAIS	300	250	GAZ	170	120

Parmi ces mots, on soulignera que (BANQUE et DONNER) viennent en fait du syntagme BANQUE DE DONNEES, de même (LIT et FLUIDISER) doivent être pris comme LIT FLUIDISE : ces deux exemples nous amènent à critiquer le prétraitement décrit au § 1.3.3 : d'une part , la

reconnaissance des locutions n'a pas été parfaite ; d'autre part, la réduction de la forme "DONNEES" à la base verbale DONNER semble inopportune. En revanche l'analyse des correspondances apparaît capable de reconstituer l'information.

En bref cet axe oppose les thèmes "Recherche " et "Information", situés du côté positif, au thème "Energie", qui figure du côté négatif en tant que moyen et source de production. La présence du vocable JAPONAIS avec l'information s'explique par le fait que le corpus contient des textes fort longs se rapportant à l'information au Japon.

Le second axe distingue l'énergie classique du côté négatif avec CHARBON, COMBUSTION ; et l'énergie nouvelle du côté positif avec SOLAIRE, MAISON, PAC.

Le quatrième axe reprend le thème de l'information déjà mis en évidence par le plan 1 × 2 avec la séparation de l'information en tant qu'application de l'informatique (base de données, logiciel, calculateur...) et de l'information au Japon.

Ainsi, comme on peut l'observer sur le plan 1 × 2 (figuré ici avec autant de mots que la place le permet), les groupements obtenus à l'issue de cette analyse décrivent de façon concrète les documents par la détermination es thèmes auxquels ils se rapportent.

On peut remarquer que les mots qui ont bien contribué dans le plan 1 × 2 ont leur fréquence totale d'apparition dans le corpus,assez bien répartie,sur l'intervalle de fréquence total 37<–>312 des deux cent cinquante (250) mots considérés ici ; avec cependant une légère concentration du côté des fréquences supérieures (>) à 100. Cette remarque sera confirmée au § 2.4.

2.2 Technicité et tranches de fréquence du vocabulaire : Divers essais de découpage en tranches du vocabulaire ont donné lieu à des analyses de tableau DOC × T. plusieurs principes nous ont guidé dans le choix du découpage. D'abord si l'on suppose que l'apparition des mots courants résulte d'un processus poissonien alors 50 occurrences ne diffèrent pas statistiquement de $50 \pm \sqrt{50}$: de ce point de vue, il n'y a pas lieu de ranger dans des classes différentes les mots dont la fréquence varie de 43 à 57. Même si le modèle poissonien est inacceptable (cf. § 1.2) l'ordre de grandeur est à retenir. Ensuite à une tranche de fréquence il correspond un ensemble de mots rentrant dans cette tranche ; et finalement un ensemble d'occurrences; par exemple les mots dont la fréquence varie de 93 à 145 totalisent 4957 occurrences : ce nombre est le poids de la tranche. On fera en sorte que les tranches aient toutes des poids équivalents. Dans la thèse sont rapportés les résultats (d'ailleurs concordants) obtenus avec des ensembles T1 et T2 de 12 et 8 tranches respectivement.

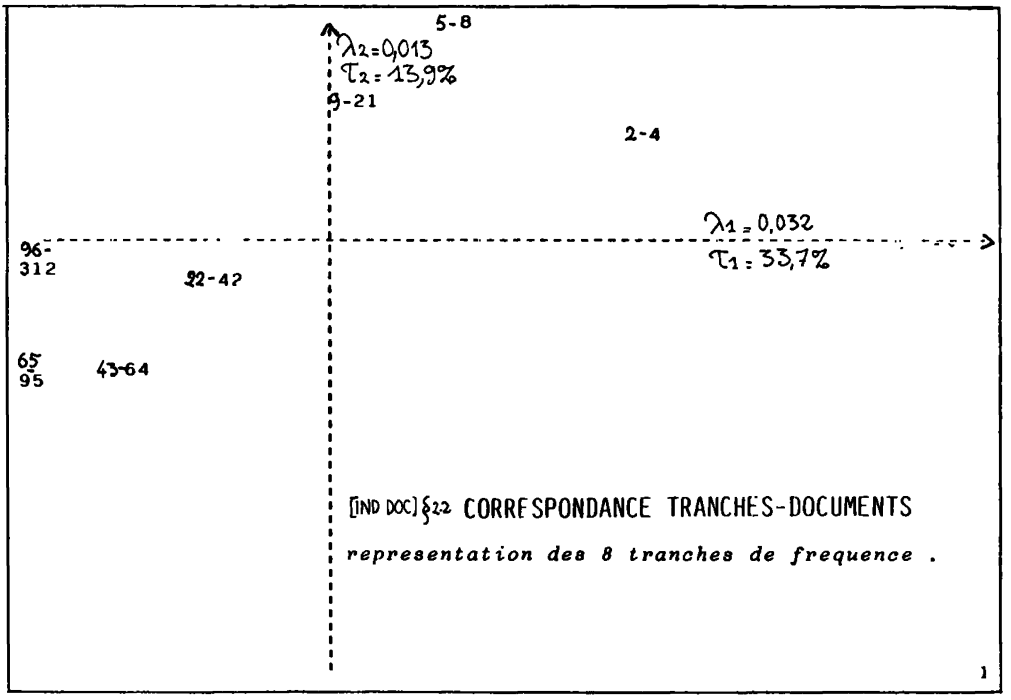
Voici quelques commentaires sur le 1-er axe issu du tableau DOC × T2 (où, rappelons-le, k(doc,t) = nombre total des occurrences dans doc des mots rentrant dans la tranche de fréquence t).

Ce premier axe ($\lambda_1 = 0{,}032$; $\tau_1 = 33{,}6\%$) oppose les fortes aux basses fréquences et grâce à la représentation simultanée de l'analyse des correspondances on peut voir, associée aux fortes fréquences une classe de documents : D072, D125,D126, D173, D182, D200, D204, D248, et D266 qui s'oppose à une autre associée aux faibles fréquences : D023, D068, D074, D152, D121, D123, D140, D159, D160, D163, D165, D185, D186, D226, D227, D269.

Un retour aux textes montre qu'en fait la première classe a trait aux thèmes suivants : CHAUFFAGE, COMBUSTION, INFORMATION , POMPE A CHALEUR (PAC), BOIS et LIT FLUIDISE alors que la seconde classe se rapporte à des sujets tels que MOTEUR, TURBINE, CAPTEUR, LUBRIFIANTS, BIOTECHNOLOGIES, GENETIQUE, COKACE.

On comprend pourquoi ces classes de documents se départagent ainsi, par l'emploi des fréquences. Les documents de la première classe se rapportent chacun à un concept d'ordre général même s'il s'agit de thèmes précis : CHAUFFACE, COMBUSTION, etc. d'où utilisation de vocables relativement peu techniques. Au contraire dans la seconde classe, on a beaucoup de descriptions techniques telles que: description d'une TURBINE DE MOTEUR, homologation d'un LUBRIFIANT, mise au point d'une nouvelle technologie et cela avec l'emploi de mots très spécifiques donc peu fréquents.

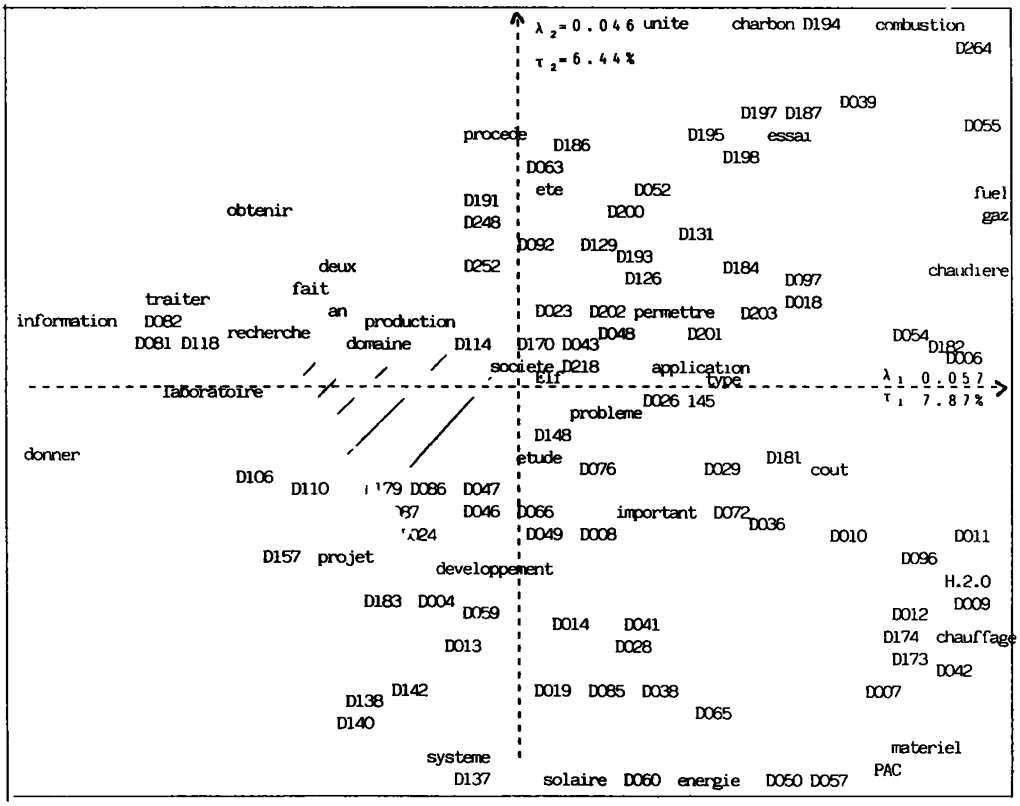
En somme, comme l'annonçait le titre du § 2.2, ce 1-er axe est un axe de technicité. Plus généralement des analyses analogues pourraient donner une typologie stylistique d'un corpus.



Ce résultat est intéressant en ce qu'il montre une certaine cohérence dans l'emploi de l'ensemble du vocabulaire : mais comme on pouvait l'attendre, faute d'avoir pris les mots isolément, on n'a aucune typologie thématique des documents. C'est ce qui nous a incité à tenter une analyse mixte combinant des informations traitées aux §§ 2.1 et 2.2.

2.3 *Analyse mixte DOC×(T ∪ M1 ∪ M2)* : Nous avons tenté de nombreuses variantes d'analyse mixte, avec en colonne, comme descripteurs de documents, à la fois des mots (comme au § 2.1) et des tranches (comme au § 2.2). Le graphique donné ici est issu d'une analyse, où l'on a retenu les 10 premières tranches de la partition T2 en éliminant les deux dernières, celles des hapax et celle des mots de fréquence 2 & 3), et les 59 mots les plus fréquents (totalisant 14.655 occurrences). A ceci près qu'il y a moins de mots, ce plan s'interprète comme celui illustrant le § 2.1. Comme on a porté les points Dxxx désignant les documents, on peut voir groupés sur l'axe 1 négatif les trois documents D082, D081, D118 relatifs au traitement de l'information au Japon (§ 2.1). Mais le mot JAPONAIS, dont la fréquence est seulement 38 (cf. tableau du § 1.3.4), est absent. Les variables "tranches", peu écartées du centre (donc non représentées ; à l'exception d'une seule notée "145" qui comprend les mots de fréquence 93 à 145), n'apportent pas de contribution notable.

Ceci nous a incité à faire d'autres analyses en multipliant les colonnes afférentes aux tranches par un coefficient choisi pour équilibrer les contributions de ces colonnes et celles des mots retenus (cf. A Hamrouni et Y. Grelet ; in C.A.D. Vol II n° 3, 1977 ; ou le volume ENS2, V n° 5).



§2.3 CORRESPONDANCE TRANCHES ET MOTS - DOCUMENTS

plan 1-2, 10 premieres tranches augmentees des 59 mots les plus frequents

2.4 Répartition de l'inertie au sein des diverses tranches de fréquence : Il est classique en analyse des correspondances d'alléger un tableau en éliminant les éléments qui apportent de faibles contributions. Ici, sur un tableau 268 × 4500 le seul calcul des inerties afférentes à chaque mot requiert un programme particulier. C'est pourquoi on a d'abord calculé l'inertie des mots dans les analyses partielles DOC × Mt, où Mt est l'ensemble des mots rentrant dans une tranche de fréquence (cf. § 2.2). Ces analyses ont permis de segmenter chaque tranche suivant l'inertie : on donne au § 2.4.1 un aperçu des résultats obtenus. Sur cette segmentation on a voulu reprendre une analyse analogue à celle du 2.2 mais plus fine : mais pour des raisons expliquées au § 2.4.2 on n'a pas obtenu la typologie stylistique espérée. Enfin au § 2.4.3 on a étudié la distribution conjointe de l'inertie et la fréquence pour 273 mots pris ensemble ce qui permet de distinguer un sous-ensemble du vocabulaire représentatif des thèmes du corpus (cf. § 2.4.4 analyse factorielle).

2.4.1 Etude séparée des classes de fréquence : Pour chaque tranche MOT1, MOT2 ,..., MOTt, d'un découpage du vocabulaire suivant la fréquence, on a un tableau de correspondance DOC × MOTt ; avec sur l'ensemble MOTt une colonne INR (donnant en millièmes les $f_m \| f^m_{DOC} - f_{DOC} \|^2$) : d'après cette colonne on peut construire un histogramme de INR sur MOTt. Nous donnons ici des histogrammes en 3 classes, afférents aux 4 premières tranches avec pour les deux tranches de plus forte fréquence un commentaire des résultats.

Tranche 1 (fréquence totale FT de 146 à 312). A l'intérieur de cette tranche qui contient les vocables les plus fréquents, on a défini trois classes 11, 12, 13. La classe 11, d'inertie maxima comprend les 2 mots CHARBON (FT = 229) ; SOLAIRE (FT = 174) où l'on reconnait des thèmes. On peut donc admettre que les mots de cette classe sont pertinents. La classe 12 d'inertie moyenne se compose des 4 mots: SYSTEME (FT = 180) ; PRODUCTION (FT = 109) ; SOCIETE (FT = 158) ; DONNER (FT = 172) qui n'ont plus de valeur de thèmes mais ont une certaine relation avec ceux-ci. La classe 13 d'inertie minima est plus volumineuse que les deux précédentes. On y trouve les mots : ETUDE, POINT, PROBABLEMENT, etc. ; dont le premier (ETUDE), avec FT = 312 est le plus usité du corpus. Ces mots n'ont plus aucun rapport avec les principaux concepts du corpus mais s'avèrent nécessaires lorsqu'on désire construire des phrases ayant un sens littéral clair. Ils sont sémantiquement comparables aux mots outils de la langue.

Tranche 2 (fréquence totale FT de 93 à 145). Ici encore on a défini trois classes suivant l'inertie. La classe 21 d'inertie maxima compte 4 mots : TRAITER, PROGRAMME, INFORMATION, PAC. On reconnaît le thème général de "l'information" avec en vue son traitement, tandis que PAC (pompe à chaleur) évoque l'effort de reconversion de l'énergie classique en énergies nouvelles à faible coût et facilement disponible. La classe 22 d'inertie moyenne, véhicule le thème "combustion" car elle contient FUEL, COMBUSTION ; avec LABORATOIRE, CONTACT, ESSAI, ELF, mots liés au thème de "recherche" au niveau du groupe. La classe 23, d'inertie minima, rappelle la classe 3 de la tranche 1 ; mais au sein d'une liste de mots outils on note tout de même ici la présence des vocables pertinents CHAUDIERE, CHAUFFAGE. Quand on descend dans la liste des fréquences, les mots ne sont plus statistiquement stables ; leur répartition dans les diverses classes a un caractère aléatoire ; le test lié à l'inertie ne filtre plus les mots pertinents. Cependant, l'efficacité du critère de répartition des mots pertinents selon le test de l'inertie s'améliorerait avec l'accroissement du nombre de documents du corpus, qui deviendrait exactement représentatif de l'ensemble des thèmes que peut recouvrir le fonds documentaire.

§2.4.1 HISTOGRAMMES DES CLASSES D'INERTIE

TRANCHE 1

INTERVALLE DE FREQUENCE [145-312]

* ETUDE			
* AN			
* PROCEDE			
* METTRE			
* PERMETTRE			
* POINT			
* PROBABLEMENT			
*			
* ETE			
* MILLIMETRE	* SYSTEME		
* TECHNIQUE	* PRODUCTION		
* PROJET	* SOCIETE	* CHARBON	
* RECHERCHE	* DONNER	* SOLAIRE	
* 13	* 12	* 11	
* 13	* 4	* 2	

TRANCHE 2

INTERVALLE DE FREQUENCE [93-145]

* INDUSTRIE			
* MARCHE			
* TYPE			
* PARTICIPATION			
* DEUX			
* OBTENIR			
* DOMAINE			
* PREMIER			
* INTERET			
* PRODUIRE			
* DEVELOPPER			
* MATERIEL			
* APPLICATION			
* ACTION			
* GAZ	* FUEL		
* H.2.0	* LABORATOIRE		
* ENERGIE	* CONTACT		
* CHAUDIERE	* COMBUSTION	* TRAITER	
* CHAUFFAGE	* ESSAI	* PROGRAMME	
* PROPOSER	* ELF	* $PAC	
* UNITE	* LIT	* INFORMATION	
* 23	* 22	* 21	
* 21	* 7	* 4	

TRANCHE 3

INTERVALLE DE FREQUENCE [61-92]

* OBJECTIF			
* RESIDU			
* THERMIQUE			
* ACTION			
* GROUPE			
* TEMPERATURE			
* COLLABORATION			
* VISITER			
* NOUVEL			
* PRIX			
* VISITE			
* TRAVAIL			
* $FRANCE	* TECHNIQUE		
* CENTRE	* FABRICATION		
* AN	* POMPE		
* CHALEUR	* BASE		
* EFFET	* SOLUTION		
* AIR	* FLUIDE		
* CONTRACT	* PILOTE	* STOCKAGE	
* REUNION	* ENTRER	* BOIS	
*S-CLASSE	* 33	* 32	* 31
* NB MOTS	* 20	* 8	* 2

TRANCHE 4

INTERVALLE DE FREQUENCE [43-60]

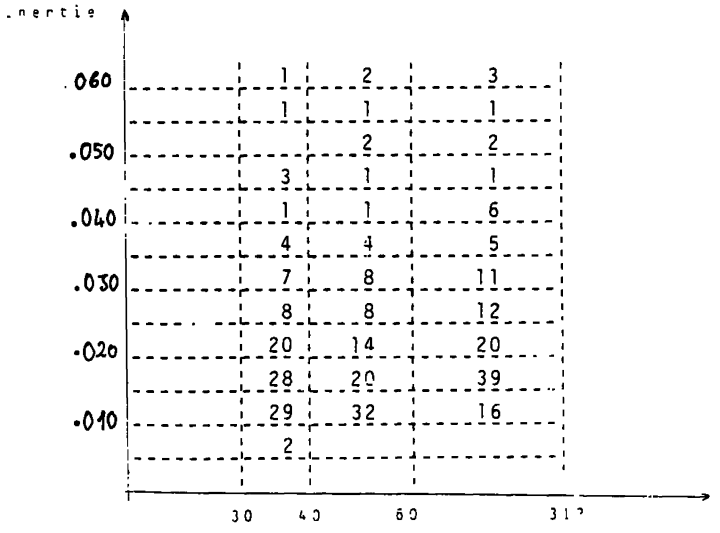
* .			
* .			
* .			
* .			
* .			
* .			
* CHIMIQUE			
* NOUVEAU			
* THERMIQUE			
* REALISATION			
* FOND	* LOGICIEL		
* PETROLIER	* BRULEUR		
* BUREAU	* PAYS		
* MELANGE	* APPARAITRE		
* $CRES	* MOTEUR		
* PROF	* BANQUE		
* ECONOMIE	* INFORMATIQUE	* NICKEL	
*S-CLASSE	* 43	* 42	* 41
* NB MOTS	* 74	* 7	* 1

2.4.2 *Essais d'analyse DOC × ST* : Au § 2.4.1 chaque tranche du vocabulaire a été partagée en 3 suivant l'inertie : par exemple de l'ensemble T des 4 premières tranches on a fait un ensemble ST de 12 sous-tranches. L'usage de ces tranches ne correspond-il pas à des caractères stylistiques, plus précis que celui vu au § 2.1 ? Pour le voir, on a analysé le tableau DOC × ST (où k(doc,st) = nombre total d'occurrences dans doc de mots de la sous-tranche st).

L'analyse a donné un étalement intéressant des documents ; mais parce que, ainsi qu'on l'a vu au § 2.4.1, plusieurs sous-tranches ont une nette signification thématique, cet étalement est sémantique, non stylistique. Il pourrait en être autrement sur une expérience à plus grande échelle, ou en écartant les sous-tranches de faible effectif (précisément celles qui ont une forte valeur sémantique). Ne peut-on pas croire en effet que, par exemple, les mots outils rares, marquent un autre style que les mots outils fréquents ?

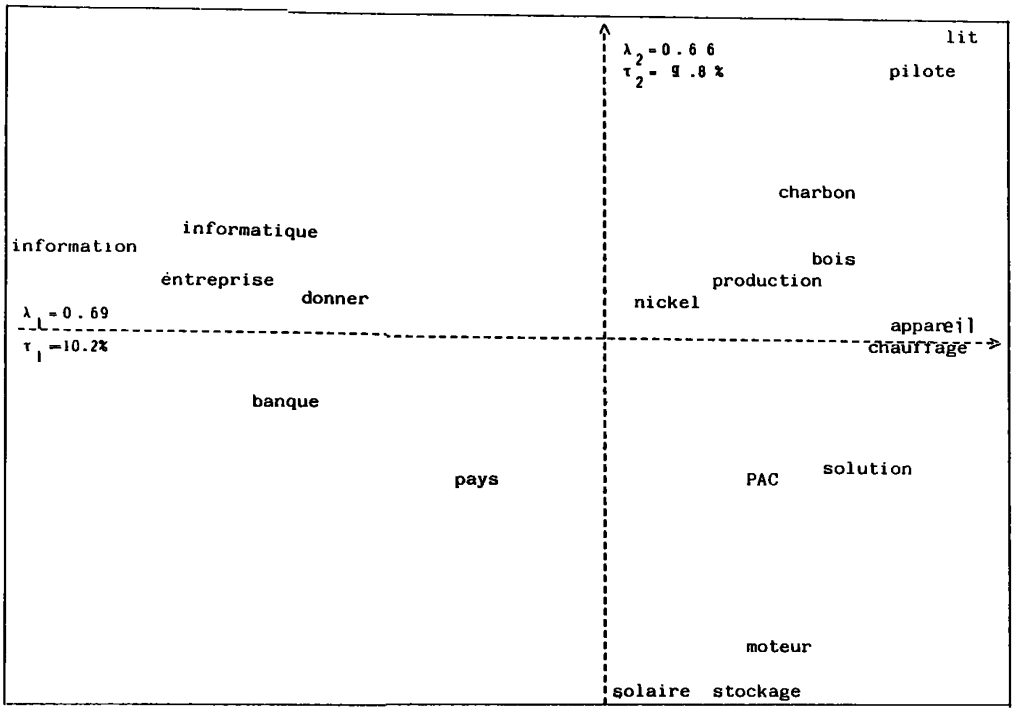
2.4.3 *Histogramme double ou fréquence et inertie* : L'analyse des correspondances nous incite à adopter le critère d'inertie ; mais il est clair que ce critère ne protège pas contre l'excentricité de mots peu fréquents. Il a paru bon de vérifier qu'au moins pour les mots de fréquence notable, l'inertie n'a rien à voir avec la fréquence et que, par conséquent on peut choisir les mots à valeur sémantique d'après le seul critère de l'inertie, la fréquence fournissant seulement un seuil inférieur.

L'histogramme ci-joint donne la répartition des valeurs de l'inertie pour les trois tranches de fréquence supérieure (les inerties INR × trace = $f_j \| f^j_I - f_I \|^2$ ont été calculées par un programme spécial sur le tableau DOC × MOT (268 × 313). On voit qu'il existe des inerties très fortes ou fortes en proportions semblables dans chaque tranche. Ce qui confirme que l'inertie à elle seule peut servir à choisir un sous-ensemble de mots en ne gardant la fréquence que pour s'arrêter quand elle devient trop faible (plus précisément, le critère d'arrêt devrait être non la fréquence définie comme nombre d'occurrences dans le corpus, mais la fréquence définie comme nombre de documents où le mot se rencontre, avec ou sans répétitions ; cf. § 3.1).



HISTOGRAMME DOUBLE : variation de l'inertie en fonction de la fréquence sur l'ensemble des 313 mots les plus fréquents.

2.4.4 *Choix d'un vocabulaire pertinent d'après l'inertie* : On a pu d'après leur inertie importante choisir 29 mots constituant un ensemble MOT1, qu'on a croisé avec l'ensemble des DOC les contenant. Après avoir mis en supplémentaire deux de ces mots (PAChaleur et NICKEL) on a obtenu trois facteurs intéressants. Le plan 1 × 2 diffère peu de ceux déjà présentés (aux §§ 2.1 et 2.3) ; l'axe 3 associe économie d'énergie à recherche relative aux moteurs ; apparaissent ensuite stockage d'énergie ; bois avec lit fluidisé... Résultats intéressants mais que l'on retrouve sous une forme plus complète et bien ordonnée après sélection des mots selon le critère de la répétition (§ 3.3).



§2.4.4 DOCUMENTS X MOT1

Plan 1-2 : répartition des mots (19) dont l'inertie est maxima ; (PAC et NICKEL sont en éléments supplémentaires).

2.5 *Conclusion* : Au § 2, les pratiques lexicométriques en usage : constitution d'index hiérarchique, découpage du vocabulaire en tranches de fréquence, etc., ayant pour objectif la caractérisation du vocabulaire pertinent d'un corpus ont été reprises et améliorées grâce à la mise en place de tris selon le critère d'inertie, dans chaque tranche. Ces opérations sont à l'origine de la constitution d'un ensemble de termes significatifs, ensemble qui sera toutefois élargi et précisé au § 3.

3 *Etude du corpus fondée sur le décompte des répétitions des mots:*

En bref, les mots pertinents se signalent par le fait qu'ils sont généralement l'objet de répétitions au sein des documents où on les rencontre (§ 3.1). Une analyse de la correspondance entre ces mots et l'ensemble des documents fournit d'emblée des résultats intéressants (§ 3.2). Mais seule la classification automatique permet d'apprécier dans quelle mesure on a obtenu une typologie utile pour organiser et retrouver l'information (§3.3).

3.1 *Sélection des mots pertinents d'après leur taux de répétition* :

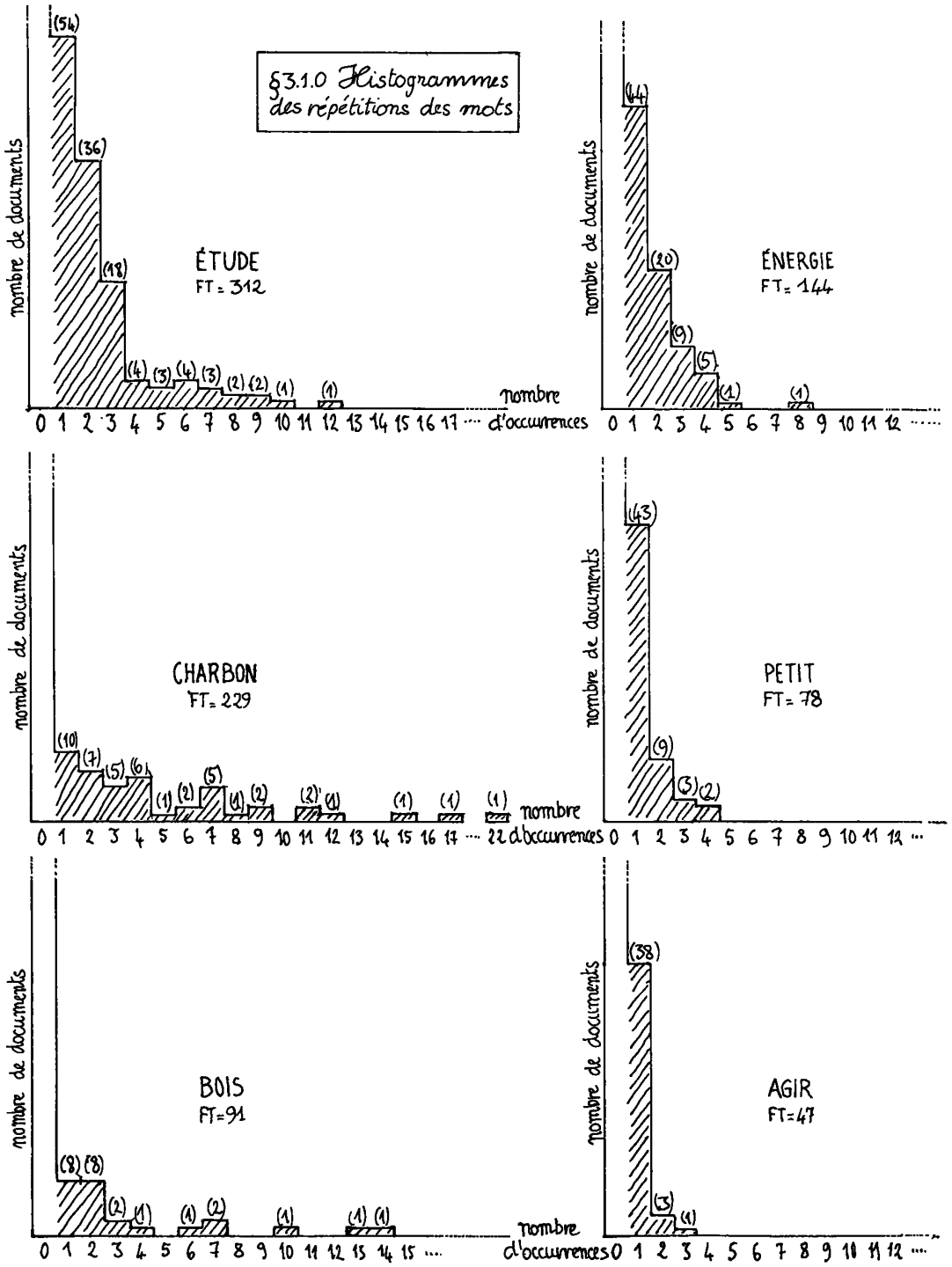
En examinant quelques histogrammes on voit entre les taux de répétitions de quelques mots, des différences qui semblent en rapport avec la charge sémantique (§ 3.1.0). L'ensemble de l'information contenue dans les histogrammes, constitue un tableau de correspondance dont l'analyse précise les divers modes de répétition (§ 3.1.1) et suggère une partition du vocabulaire en trois classes : mots pertinents (§ 3.1.2.1) ; concepts généraux (§ 3.1.2.2) ; mots outils (§ 3.1.2.3).

3.1.0 *Examen des histogrammes de répétition de six mots* : Le nombre des occurrences d'un mot dans un document de notre corpus peut varier de 0 (absence de mot dans le document) à 36 (le mot INFORMATION se rencontre 36 fois dans le document 204, le plus long du corpus, consacré à l'information au Japon). De façon précise on peut pour chaque mot construire un histogramme, avec en abscisse le nombre d'occurrences (0, 1, 2,..., n,..., 36) et en ordonnée, dans le créneau n , le nombre des documents où le mot est employé exactement n fois . Comme on pouvait l'attendre après que les articles, prépositions, conjonctions, formes du verbe être, etc., aient été écartés du corpus par un prétraitement (cf. § 1.3.3) le créneau zéro (nombre des documents d'où le mot est absent) est toujours le plus haut : le mot étude lui même dont le nombre total d'occurrences est 312 est absent de plus de la moitié des documents (de 140 à 268) : pour cette raison, ce créneau n'a pas été figuré dans toute sa hauteur. Les créneaux suivants (nombre de documents où le mot apparaît une fois, deux fois...) sont plutôt de hauteur décroissante : mais d'une part les fluctuations d'échantillonnage rendent cette décroissance irrégulière, d'autre part le profil général de décroissance diffère sensiblement d'un mot à l'autre.

Comparons les 4 mots ETUDE, ENERGIE, PETIT, AGIR : le créneau 1 a presque la même hauteur dans les 4 cas (54, 44, 43, 38) : cependant, le nombre total d'occurrences varie à peu près en progression géométrique de raison 1/2 (FT = 312, 144, 78, 47) ; car tandis que pour ETUDE l'histogramme décroît assez lentement jusqu'à un dernier créneau qui signale un document où le mot est répété 12 fois, la chute est brutale pour PETIT, et dans aucun document le mot n'est présent plus de 4 fois ; ENERGIE, étant intermédiaire. Quant à AGIR il figure 2 fois dans 3 documents, 3 fois dans un seul ; et ne dépasse dans aucun document le nombre 3.

Tout à l'opposé les deux mots CHARBON et BOIS ne présentent pas de créneau (hormis le zéro) dont la hauteur dépasse 10 : pourtant leurs FT respectives sont 229 et 91. Mais d'une part le nombre maximum d'occurrences trouvées dans un seul document est 22 pour CHARBON et 14 pour BOIS ; d'autre part les créneaux 1 et 2 sont de hauteur comparable, avec au-delà une décroissance lente ; particulièrement pour CHARBON dont le nombre moyen d'occurrences (par document où il est présent) est : $229/31 \approx 7$.

Du point de vue du sens, les 6 mots présentés ici sont bien différents entre eux : CHARBON et BOIS sont des objets bien déterminés qui , compte-tenu du corpus (cf. § 1.3.2) peuvent être le thème unique


§3.1.0 Histogrammes des répétitions des mots
ÉTUDE
FT = 312
nombre de documents
nombre d'occurrences
ÉNERGIE
FT = 144
CHARBON
FT = 229
PETIT
FT = 78
BOIS
FT = 91
AGIR
FT = 47

d'un document, ou un thème principal : auquel cas le mot en cause sera forcément répété : car aucune virtuosité de style ne permet de l'éviter. Au contraire PETIT est un adjectif dont la répétition ne s'impose que dans des contextes très particuliers : son emploi ne dépend pas du thème du document, et corrélativement la présence du mot n'a aucune valeur comme indicateur du contenu. Le verbe AGIR est , lui aussi, sémantiquement neutre (en fait il figure dans la locution "il s'agit de"). Quant à ENERGIE et ETUDE on ne peut les considérer comme des mots vides ; au sein d'un corpus plus vaste ENERGIE pourrait être discriminant ; mais dans les documents issus de la compagnie ELF (cf. § 1.3.1) le concept est partout et l'apparition du mot ne signifie à peu près rien ; enfin ETUDE définit un univers encore plus vaste qu'ENERGIE ; il est de plus caractéristique du rôle des ingénieurs en mission qui ont rédigé les comptes-rendus : il n'y a rien à conclure du fait que certains usent de ce mot avec prédilection.

3.1.1 Analyse MOT × N : représentation de l'ensemble N : Pris individuellement, l'histogramme apporte une information précise et suggestive ; la comparaison d'un petit nombre d'histogrammes entre eux est possible, même peu précise ; présenter sous forme d'histogrammes un bilan des répétitions des 237 mots les plus fréquents du corpus apporte une aide médiocre à la synthèse. On considère donc, classiquement, ces histogrammes comme les lignes d'un tableau de correspondance. De façon précise on pose :

MOT : ensemble des 237 mots les plus fréquents du fichier.

N : ensemble des 37 entiers de 0 à 36.

$k(mot,n) = Card\{doc \mid doc \in DOC \ ; \ k(doc,mot) = n\}$;

i.e. k(mot,n) est le nombre des documents du corpus contenant exactement n fois le mot.

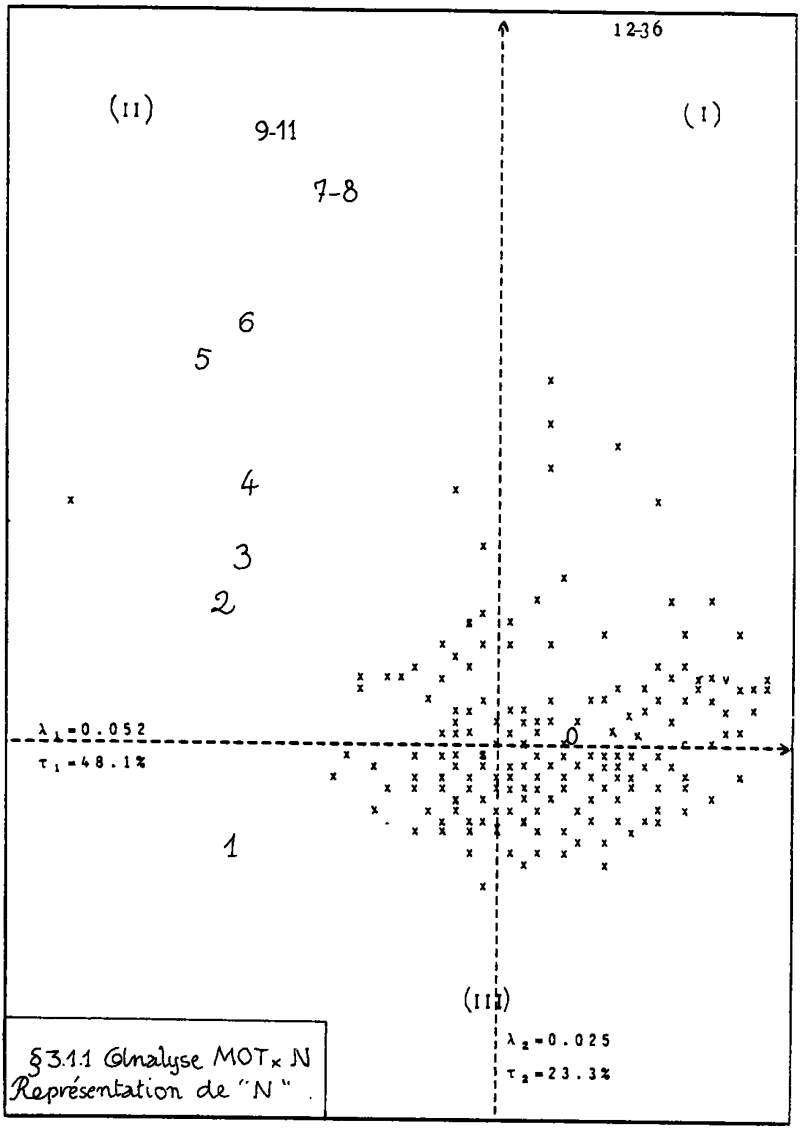
En fait les valeurs les plus élevées de n se rencontrent rarement et donc aléatoirement : pour obtenir des résultats stables, il convient de cumuler les colonnes de rang élevé. En vertu du principe d'équivalence distributionnelle, le choix précis des cumuls effectués importe peu, ainsi qu'on l'a vérifé par des essais. Pour le présent exposé, on a retenu les résultats issus d'un tableau dont l'ensemble des colonnes est :

$$N = \{0, 1, 2, 3, 4, 5, 6, 7-8, 9-11, 12-36\} \ ;$$

e.g. la dernière colonne cumule toutes les grappes de 12 à 36 occurrences d'un même mot dans un seul document.

L'analyse du tableau MOT × N fournit dans le plan (1,2) 71% de l'inertie : au-delà nous n'avons rien d'interprétable, encore moins d'utilisable pour le traitement des documents. Nous donnerons d'abord l'interprétation du nuage N, sur une figure où les mots sont placés comme de simples points ; puis sur trois figures partielles agrandies nous considérerons en détail le vocabulaire (cf. § 3.1.2).

Sur l'axe 1, le nombre n = 0 (caractère d'absence du mot) s'oppose à tous les autres n (à l'exception du cumul 12 - 36 très éloigné sur l'axe 2 mais attiré dans le quadrant F2 > 0, F1 > 0 par le mot INFORMATION, employé 36 fois dans le document 204) : d'ailleurs avec COR1(0) = 998, le point 0 est presque exactement situé sur de demi-axe F1 > 0. Le premier axe oppose donc l'absence (F1 > 0) à la présence (F1 < 0) quelle que soit par ailleurs le taux de répétition : on pourra vérifier cette interprétation sur le nuage des mots (en tenant compte toutefois de ce que F1(mot) n'est pas exactement déterminé par le nombre de doc où se trouve le mot, puisque les abscisses



F1(n) varient pour n ≥ 1).

Sur l'axe 2, les modalités de présence de 1 (une seule fois) à 12 - 36 (de 12 à 36 fois) se rangent dans leur ordre naturel : corrélativement un mot aura un facteur F2 d'autant plus élevé que dans les documents où il se trouve, il apparaît avec plus de répétitions. Plus précisément la formule barycentrique donne :

$$F2(mot) = \lambda_2^{-1/2} \Sigma\{k(mot,n)\ F2(n) \mid n = 1 \ldots 256\}$$

dans cette formule on a omis le terme en n = 0 , puisque F2(0) ≈ 0 ;

mais il faut prendre garde que le dénominateur k(mot) = 256 = Card DOC, n'est pas seulement la somme des k(mot,n) pour n ≥ 1, mais comprend aussi k(mot,0) : en sorte qu'à profil égal sur N-{0}, un mot sera d'autant plus proche de l'origine sur l'axe 2, qu'il est présent dans moins de documents. Cependant le signe de F2(mot) a un sens indépendant de cela : les emplois isolés comptés dans k(mot,1) attirent les mots vers le bas (F2 < 0) ; tous les autres emplois (par 2 occurrences ou plus dans un document) l'entraînent vers le haut (F2 > 0) et d'autant plus que le taux de répétition n est plus élevé.

3.1.2 *Analyse MOT × N : représentation de l'ensemble des mots* : Compte tenu de l'interprétation donnée à l'axe 2, on ne s'étonnera pas que l'opposition entre les deux demi-plans F2 < 0 et F2 > 0 soit très claire en ce qui regarde le sens et la fonction des mots. L'axe 1, quant à lui représente à peu près la fréquence : celle-ci n'a pas d'interprétation sémantique tranchée, et même s'il est clair que les mots les plus fréquents ne peuvent caractériser le thème d'un document avec précision, rien ne suggère *a priori* que le zéro de l'axe 1 marque une frontière. Cependant il apparaît satisfaisant de partager le plan 1 × 2 en trois domaines :

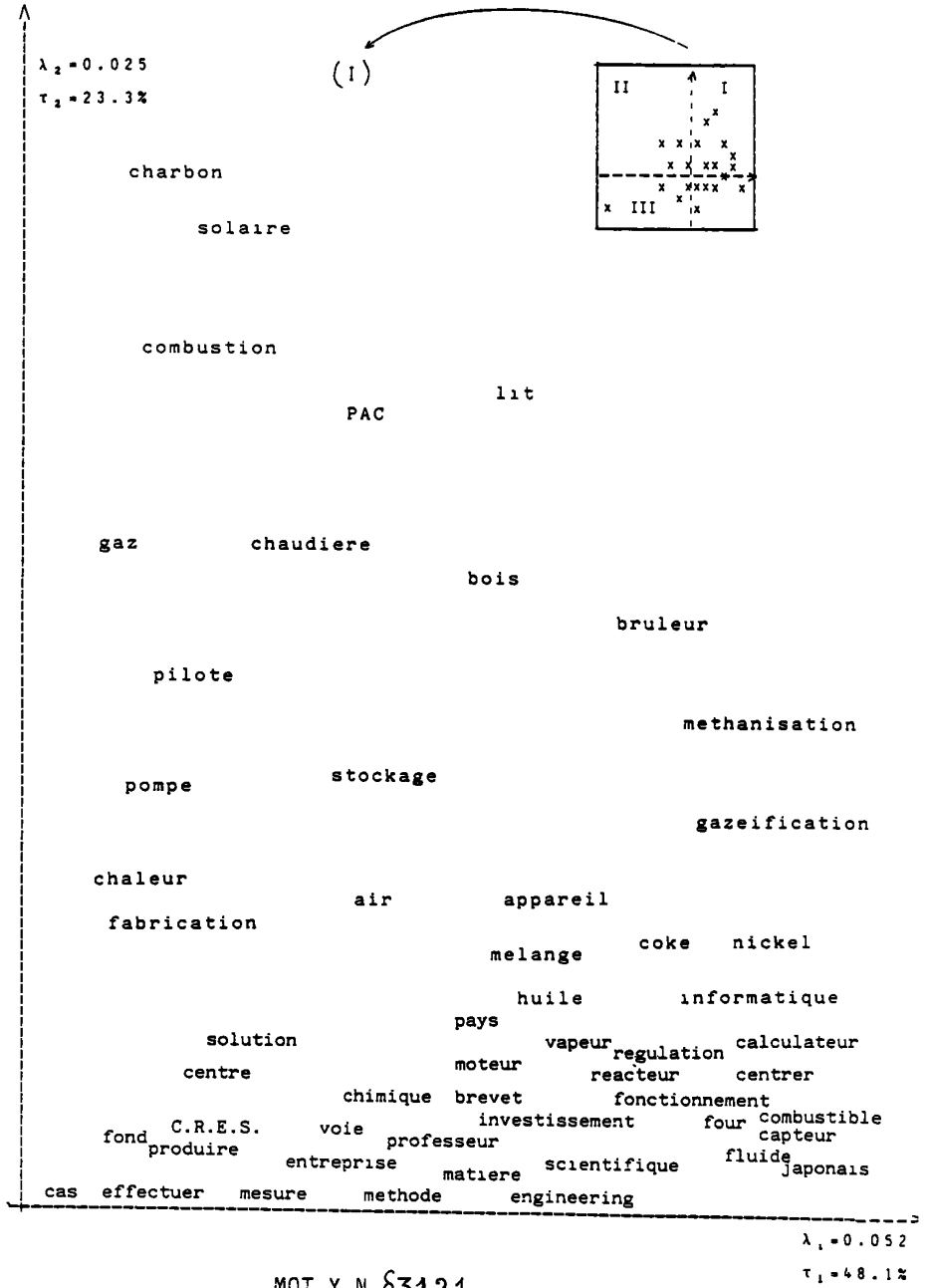
I le quadrant F1 > 0 ; F2 > 0 ;
II le quadrant F1 < 0 ; F2 > 0 ;
III le demi-plan F2 < 0 .

Il nous a seulement paru opportun d'adjoindre au vocabulaire des mots pertinents que renferme le domaine (I), quelques "frontaliers" ayant valeur thématique (cf § 3.2).

3.1.2.1 *Le quadrant (I) des mots pertinents représentatifs de la diversité du corpus* : Ce quadrant ne contient aucun élément de N : mais il est intermédiaire entre "0" et les valeurs élevées de n : on y trouve des mots qui, parmi ceux retenus, sont de fréquence relativement basse et s'emploient volontiers en grappe. Deux de ceux-ci CHARBON et BOIS nous sont déjà connus : ils ont valeur de thème ; les mots qui les accompagnent, particulièrement si l'on s'éloigne quelque peu de l'axe F1 (i.e. si F2 est nettement positif), sont de même pour la plupart susceptibles de caractériser le contenu d'un document. Il y a peut-être des exceptions : mais l'effectif relativement faible du corpus ne permet pas de mettre en oeuvre de nouveaux filtres... Nous dirons qu'en général la concentration des occurrences d'un mot dans peu de documents, en conférent une certaine spécificité à celui-ci, lui octroie un apport informatif considérable ; et que l'analyse des répétitions a mis en évidence cet ensemble de caractères.

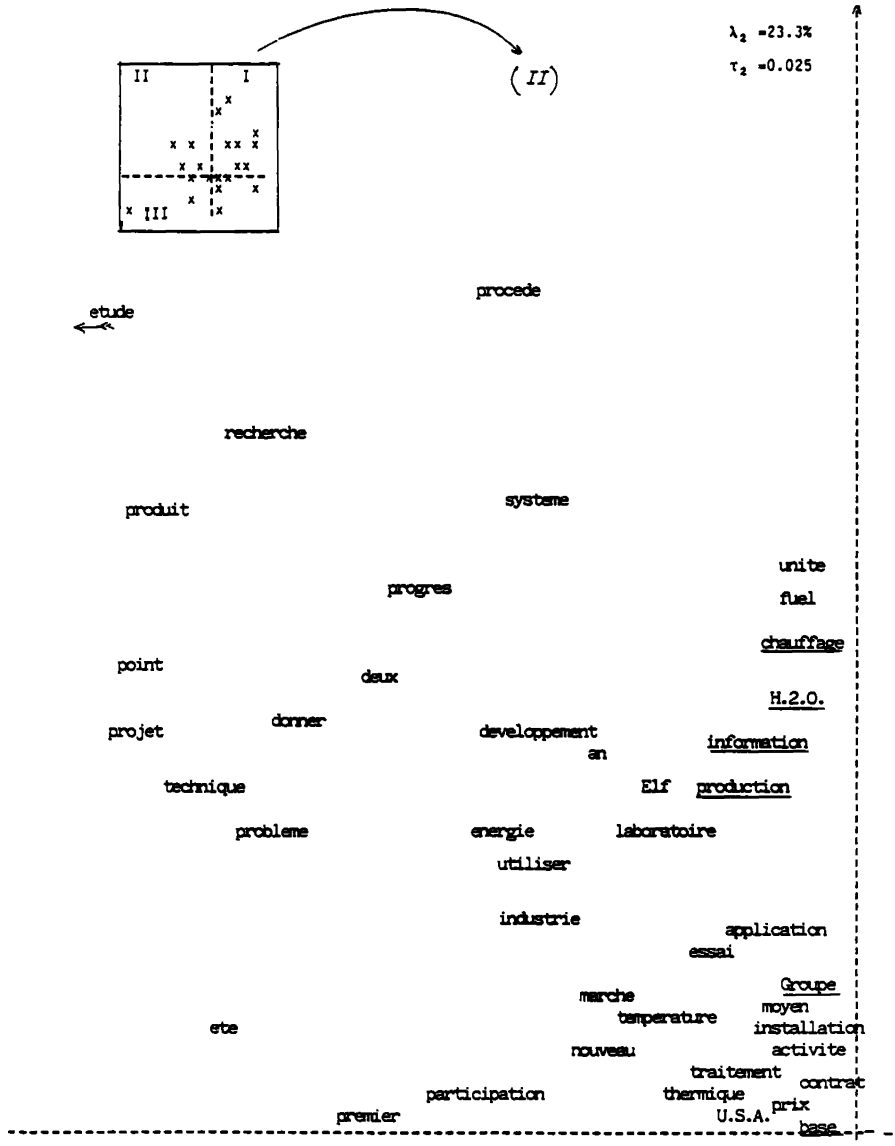
3.1.2.2 *Le quadrant II des concepts généraux de l'ensemble du corpus:*

Ce quadrant contient des mots fréquents, objets d'assez nombreuses répétitions : deux de ceux-ci, ETUDE et ENERGIE ont fait l'objet d'un commentaire fondé sur leurs histogrammes (cf § 3.1.0). Les mots qui les accompagnent ayant des profils voisins relèvent de la même tendance : ils véhiculent des concepts trop généraux pour caractériser un document donné. Toutefois ainsi qu'on l'a annoncé, la frontière entre (I) et (II) ne s'impose pas à l'interprétation : c'est pourquoi on a dans II souligné quelques mots, proches du demi-axe (F2>0) et susceptibles d'être adjoints à (I) pour constituer un vocabulaire représentatif de la diversité du contenu des documents.



MOT X N §31.21

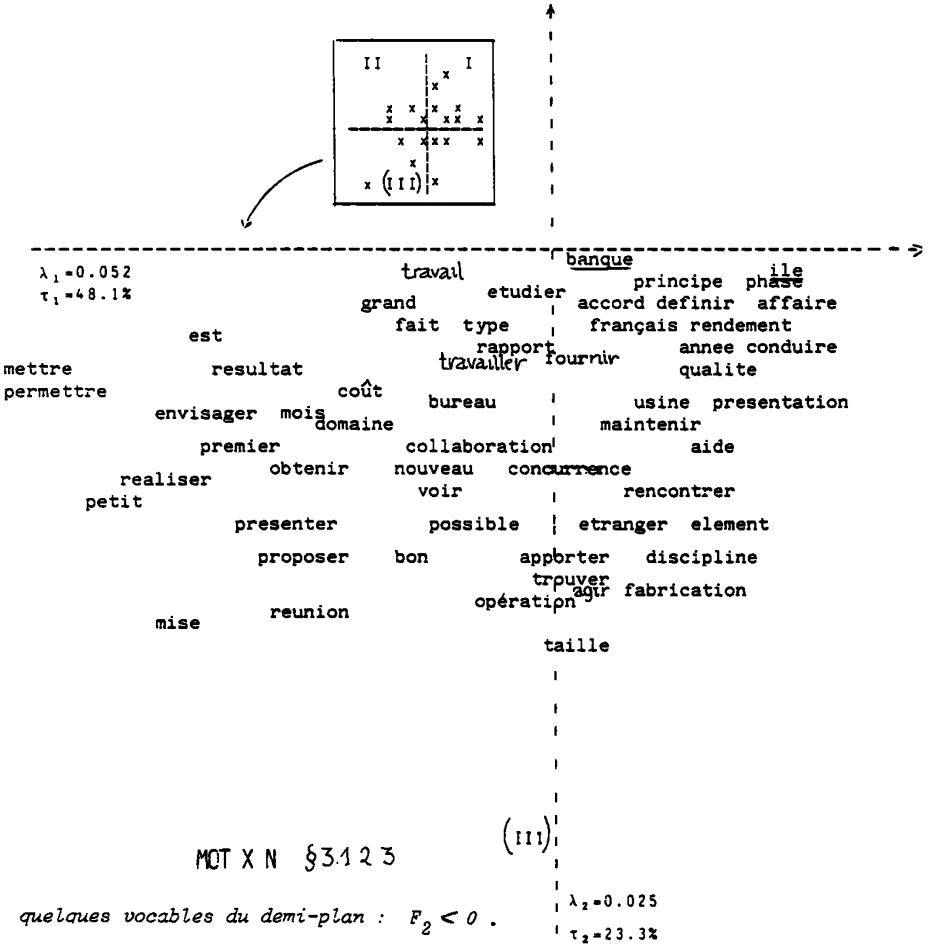
vocables du quadrant I .



MOT X N §3.1.2.2

vocables du quadrant II .

3.1.2.3 *Le demi plan (III) de mots outils de l'exposé* : Il s'agit de mots de fréquence moyenne, fortement associés au caractère n = 1, i.e. à l'absence de répétition. On a déjà vu PETIT et AGIR ; les autres mots de (III) sont comme ceux-ci de simples outils de l'exposé, dont la rédaction use occasionnellement ; en général il évitera de les répéter. Certains de ces mots sont plutôt rares : qu'un document en offre des répétitions nombreuses ne pourra être qu'une exception ; et les plus fréquents eux-mêmes auront un histogramme dominé par le pic n = 1. Toutefois on a trouvé dans (III) deux mots susceptibles de nous intéresser par leur contenu : ce sont ACTION et BANQUE (de données principalement !)

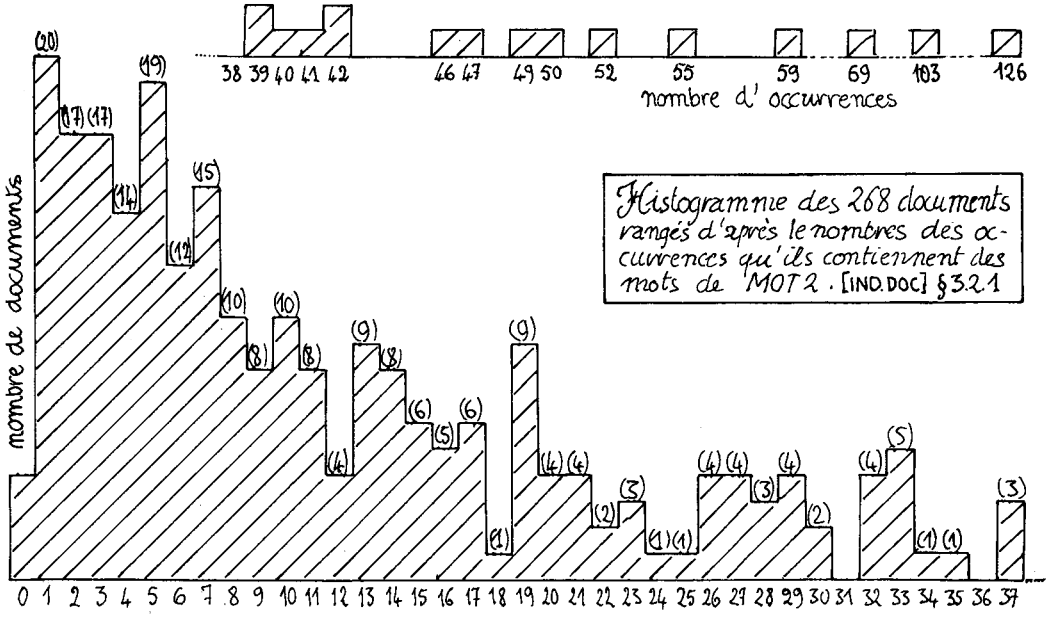


MOT X N §3.1.2.3

quelques vocables du demi-plan : $F_2 < 0$.

3.2 *Analyse factorielle croisant les documents avec un vocabulaire représentatif*

3.2.1 *Les ensembles en correspondance* : Ainsi qu'on l'a expliqué au § 3.1.2 l'ensemble des mots pertinents a été délimité d'après l'analyse des répétitions : il s'agit essentiellement des mots du quadrant (I) (mots de fréquence moyenne survenant volontiers en grappes) complétés de quelques mots "frontaliers" soulignés sur les graphiques des régions (II) et (III). Toutefois des contraintes techniques nous ont fait perdre les mots de fréquence < 34 : soit Réacteur, Japonais, combustible, capteur, calculateur, gazéification, méthanisation. Finalement on a un ensemble de 51 mots notés MOT2 (après MOT1, ensemble de 29 mots constitué suivant le critère de l'inertie : cf. § 2.4). Ce choix fait, on doit éliminer 4 documents où ne figure aucun des mots retenus : reste 264 textes, dont l'ensemble sera encore noté DOC. L'histogramme ci-joint montre que parmi ceux-ci, il y en a 68 qui contiennent moins de 5 occurrences de mots de MOT2 : la place de ces documents dans les analyses factorielles et les CAH ne pourra être que peu significative.



3.2.2 *Résultats de l'analyse DOC × MOT2* : L'analyse donne sur chacun des 6 premiers axes des associations intéressantes entre certains mots et des groupes de documents. Pour aider à l'interprétation nous soumettons au lecteur, en plus des mots remarqués sur les axes de 1 à 4, les documents symbolisés par leur contenu selon la nomenclature introduite eu § 1.3.2. La signification des numéros de cette nomenclature est rappelée en marge.

Axe 1 = λ_1 = 0,65 ; τ_1 = 6,8% : on retiendra seulement l'extrémité positive de l'axe.

F1 < 0 : INFORMATION ; ENTREPRISE ; SCIENTIFIQUE ; INFORMATIQUE

12 = Communication
14 = Econ.(d'énergie)
15 = Inf. et b. de données
1 = Explorat. Product.

Documents : (12);(14a);(15a);(12);(15)
(15);(12,15);(12,15);(1,12).
(par exemple le dernier document cité concerne le thème 1 : Expl. production ; et 12, communication).

Axe 2 : $\lambda_2 = 0,58$; $\tau_2 = 6,1\%$

F2 > 0 : NICKEL ; mot unique COR2(NI) = 928 ; CTR2(NI) = 840.

7 = Accumulateurs
5f = Chimie fine

Documents : (12);(7);(5f);(7,12);(7,12).

Axe 3 : $\lambda_3 = 0,53$; $\tau_3 = 5,6\%$; on considère les deux extrémités de l'axe.

F3 < 0 : SOLAIRE, mot unique associé aux 6 documents suivants

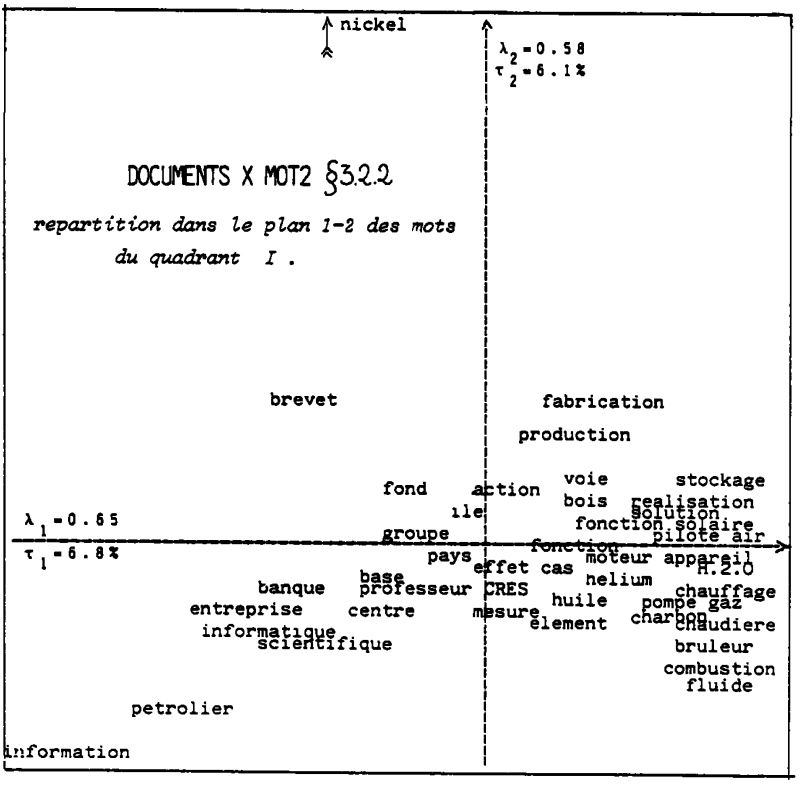
13 = Coop. Industrielle
16 = Energies Nouvelles
3c = Thermique habitat
3d = P.A. Chaleur
3c = Solaire photo volt.
3f = Géothermie
3g = Echangeurs

Documents : (3c,3d,3e,3g);(4a,11a,13);
(3c,3d,3e,13);(3d,3e,16a,);
(3d,3g,3f);(3d,16a)

F3 > 0 : COMBUSTION, CHARBON, LIT, FLUIDE associés aux

11 = Combustion
2d = Huiles lourdes
3a : Stockage d'énerg.
4a : Biotech. (pétrole)
6a : Trait. eau (poll.)

Documents : (4a,11);(2d,4a,11,13);(11);(11,16);
(3g,11);(11);(11);(3a,6a,6b,11)



Axe 4 : $\lambda 4 = 0,45$; $\tau 4 = 4,8\%$

F4 > 0 : BOIS , mot unique associé aux

9 = Utilis. de la Biomasse 5b = Agrochimie 1e = Gaz 2b = Carburants 3b = Bois (autre que biomasse) 5g = chimie lourde	Documents : (11),(1a,2b,5g,9);(9);(3b);(9); (9);(9);(9);(9);(9);(5b,9); (3b,13) où la fréquence du n° 9 (Biomasse) ne surprendra aucunement

3.3 *Classifications sur le vocabulaire et les documents* : Le tableau DOC × MOT2, dont l'analyse est rapportée au § 3.2, est à la base de classifications effectuées sur chacun des deux ensembles en correspondance. Il s'agit dans tous les cas de classification ascendante hiérarchique (CAH), avec pour critère la maximisation de l'inertie interclasse. Des classifications ont été faites directement sur le tableau de correspondance : e.g. l'ensemble MOT2 étant identifié au nuage (MOT2) des profils sur DOC des colonnes du tableau. Toutefois on a obtenu les meilleurs résultats en se plaçant dans l'espace engendré par les 10 premiers axes factoriels : ce qui revient à utiliser l'analyse factorielle pour filtrer l'information que contient le tableau. Le chiffre 10 a été choisi après essais d'un nombre moindre de Dimensions. Etant dans R^{10}, on a assez naturellement trouvé avantage à retenir de chaque CAH une partition en 11 classes, 11 étant le nombre des sommets d'un simplexe dans l'espace à 10 dimensions (comme 3 pour un triangle du plan...). Ce choix est d'ailleurs imposé par l'histogramme des indices de niveau (cf. § 3.3.2). Nous considérerons successivement : la CAH sur le vocabulaire MOT2 (§ 3.3.1) dont l'interprétation détaillée est aisée ; puis la CAH sur l'ensemble DOC (§ 3.3.2) dont il est difficile de donner plus que les grandes lignes sans publier les documents eux-mêmes, dont cependant le listage constitue à lui seul un volume. Enfin le croisement des partitions retenues à partir des deux CAH, utilisées pour interpréter la classification des documents, suggère de plus un procédé pour classer rapidement les documents nouveaux, et retrouver tout ce qui dans le corpus concerne un thème (§ 3.3.3).

3.3.1 *Classification sur l'ensemble MOT2* : Le détail de la classification et de l'interprétation est donné sur deux pages qui se font face. La partition retenue présente pour nous l'intérêt d'avoir isolé les principaux concepts déjà connus du corpus, non sans mettre en évidence l'importance de deux nouvelles rubriques : BOIS et STOCKAGE. Quant à la structure de l'arbre on notera d'abord que l'élément NICKEL se sépare de l'ensemble à un niveau élevé (premier branchement) ; puis l'emboitement des classes dans le reste de l'ensemble se fait comme suit :

Noeud 100 : n(98) ∪ n(99) = n(100) ; τ = 11,3%

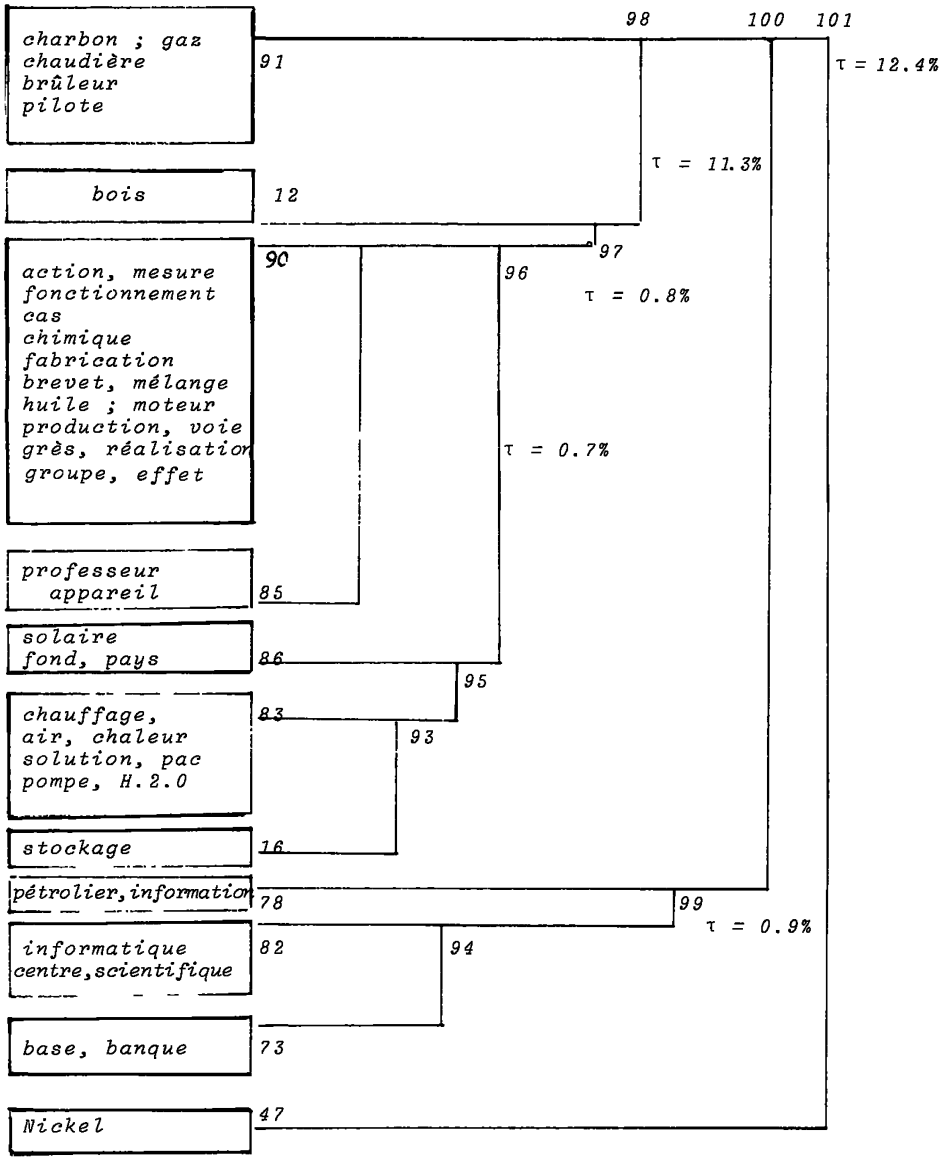
A ce niveau s'inscrit la séparation entre ce qui a trait à l'INFORMATION, COMMUNICATION,... classe 99 et le reste des thèmes COMBUSTION, BOIS, GENIE CHIMIQUE, SOLAIRE... compris dans la classe 98 , nous signalons également le débranchement.

Noeud 98 : n(91) ∪ n(97) = n(98) ; τ = 9,4%

Au niveau de ce noeud le thème COMBUSTION (classe 91) se sépare de BOIS, GENIE CHIMIQUE, SOLAIRE... (classe 97).

Noeud 97 : n(12) ∪ n(96) = n(97) ; τ = 8,7%

A ce niveau-là nous avons encore séparation entre une classe réduite à un élément le BOIS (classe 12) et GENIE CHIMIQUE, SOLAIRE , CHAUFFAGE... (classe 96). Cette séparation mérite notre attention car contrairement à la notion classique de BOIS liée à CHAUFFAGE, BOIS en est ici dissocié, du fait de sa valorisation dans diverses applications dans le cadre de la biomasse : gazéification, liquéfaction du bois.



§ 4.5.2 C.A.H. SUR LES VOCABLES DE MOT2
(sur tableau de 10 facteurs)

Représentation des 11 classes de la partition, τ est la part d'inertie du noeud

91 : Combustion est le thème général de cette classe ; la présence du mot PILOTE, rappelle l'importance des essais dans ce domaine.

12 : Réduite au seul mot BOIS, cette classe rappelle l'importance particulière de cette matière dans le corpus.

90 : FABRICATION, FONCTIONNEMENT, PRODUCTION, caractériseraient le mieux cet ensemble de mots divers : HUILE,MELANGE, MOTEUR, s'introduisent comme applications ; REALISATION appelle dépôt ou acquisition de BREVET par le GROUPE ou le laboratoire (CRES).

85 : Recherche et coopération avec les laboratoires (universitaires).

86 : SOLAIRE, définit indéniablement le thème de la classe ; avec référence aux PAYS chauds, ou sous-développés, que le FOND de développement international aide à développer des projets souvent liés au SOLAIRE.

83 : CHAUFFAGE et PAC (pompe à chaleur) : deux thèmes dont le premier fait appel au second ce qui explique leur association.

16 : STOCKAGE, apparaît ici isolé ; ce qui attire notre attention sur ses emplois dans les documents : S. souterrain d'énergie, de chaleur, etc., apparaît comme une nouvelle rubrique dont les spécialistes ont confirmé l'intérêt.

78 : INFORMATION, est un thème général, dont les deux classes suivantes précisent la structure au sein de la classe 99.

82 : INFORMATIQUE, thème qui dans le corpus comprend l'informatisation de l'entreprise, le calcul scientifique, les modèles prévisionnels.

73 : BANQUES et BASES de données : Thème qui préoccupe les grandes entreprises françaises désireuses de constituer leurs propres réserves d'informations scientifique et économique.

47 : NICKEL, dans notre corpus, est considéré comme constituant de nouveaux types d'accumulateurs, auxquels sont consacrés d'importantes recherches spécialisées.

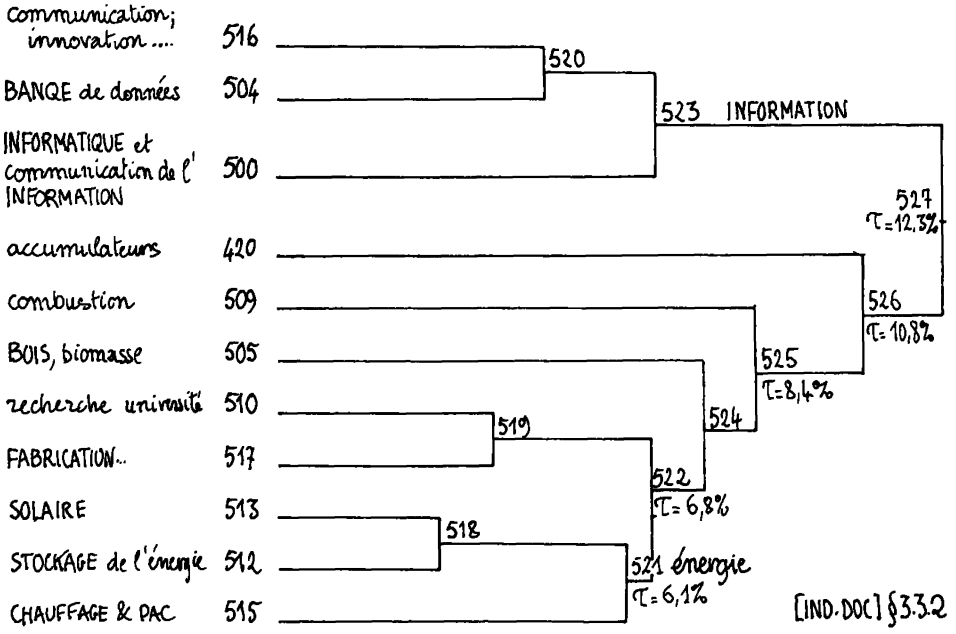
3.3.2 *Classification sur l'ensemble DOC* : L'ensemble DOC est restreint aux 264 documents présentant au moins une occurrence des mots de MOT2 ; comme on l'a dit au § 3.2.1, seuls environs 200 documents contiennent au moins 5 occurrences, ce qui semble un minimum au-dessous duquel on ne peut garantir une bonne reconnaissance du thème.

La classification retenue ici (cf. § 3.3 en tête), est faite dans l'espace engendré par les 10 premiers axes issus de l'analyse du tableau DOC×MOT2 (cf. § 3.2). L'histogramme ci-joint suggère fortement de considérer la partition en 11 classes définies par les 10 noeuds les plus hauts de cette CAH. L'interprétation a confirmé l'intérêt de cette partition.

J	I(J)	A(J)	B(J)	T(J)	T(Q)	HISTOGRAMME DES INDICES DE NIVEAU DE LA HIERARCHIE
527	523	523	526	123	123	**********
526	460	420	525	108	231	**********
525	356	509	524	84	314	**********
524	343	505	522	81	395	**********
523	311	520	500	73	468	**********
522	288	519	521	68	536	**********
521	262	518	515	61	597	**********
520	213	516	504	50	647	**********
519	201	510	517	47	694	**********
518	191	513	512	45	739	**********
517	95	514	492	22	761	**********
516	73	511	491	17	779	**********
515	59	498	496	14	792	**********
514	44	508	502	10	803	**********
513	43	507	439	10	813	**********
512	38	329	495	9	822	**********

[IND. DOC] §3.3.2

Pour interpréter les classes, nous avons d'abord (comme au § 3.3.2) assimilé chaque document à la prescription abrégée de son contenu faite suivant la nomenclature introduite au § 1.3.2. Pour 8 classes sur 11 (font exception : 516, 510, 517) ce procédé donne un thème indubitable. Mais souvent il reste dans une classe des documents qui semblent étrangers à son thème. Parfois (cf. § 3.2.1) ce sont des textes où figurent seulement 1 ou 2 occurrences de MOT2 ;



parfois en retournant au document initial, on découvre que l'indexation suivant la nomenclature du § 1.3.2 était peu satisfaisante ; d'ailleurs le tableau DOC ×MOT2, permet de voir sur quel mot particulier ambigu (e.g. STOCKAGE d'énergie ou non ; BASE de données ou B. de fabrication...) la CAH a pu faire une fausse affectation. Enfin particulièrement pour les classes 516, 510, 517, on a consulté le tableau croisant les partitions en 11 classes de DOC et MOT2 (cf. § 3.3.3). Finalement on a retenu les interprétations qui suivent :

Classe 516 : 34 documents : 6 contiennent moins de 4 occurrences de MOT2 ; les autres associent généralement INFORMATION, INFORMATIQUE et BANQUE de données, à un thème industriel ou scientifique particulier. Si l'on considère les subdivisions de 516, on peut distinguer 4 classes [491], communication et innovation ; [481] , logiciels (e.g. en géophysique) ; [474] ≈ [491] ; [476] chimie, thermodynamique et biotechnologie.

Classe 504 : 23 documents : 19 concernent les banques et bases de données ; 1 la bureaucratique ; 3 autres abordent des thèmes divers, mais l'un de ceux-ci ne comporte qu'une seule occurrence des mots retenus.

Classe 500 : 7 documents : 5 concernent l'INFORMATIQUE et la communication de l'INFORMATION ; 2 ne contiennent chacun que deux occurrences de MOT2.

Classe 420 : 2 documents relatifs aux accumulateurs et associés à NICKEL.

Classe 509 : 18 documents : 15 relatifs à la combustion ; plus 3 autres dont 2 abordent des thèmes connexes(pyrolyse...).

Classe 505 : 14 documents, tous relatifs à la biomasse : valorisation et récupération du BOIS, délignification, combustible.

Classe 510 : 27 documents : malgré la diversité des thèmes abordés, la classe peut être caractérisée par la recherche, généralement en collaboration avec des laboratoires dirigés par un PROFESSEUR, en vue de FABRICATION, FONCTIONNEMENT, PRODUCTION.

Classe 517 : 66 documents : comme la 510, la classe 517 est associée à FABRICATION, FONCTIONNEMENT et PRODUCTION ; mais ici le thème principal est la combustion.

Classe 513 : 19 documents : 15 relatifs à l'énergie SOLAIRE ; 2, consacrés à la chromatographie, abordent le thème des énergies nouvelles ; 2 ne comportent chacun que deux occurrences de MOT2.

Classe 512 : 16 documents : gestion et STOCKAGE de l'énergie (SOLAIRE, photovoltaïque, etc.) ; document parasite, qui ne contient que trois occurrences de MOT2, est attiré ici par le mot STOCKAGE. (mis dans un autre contexte que S. de l'énergie).

Classe 515 : 38 documents : CHAUFFAGE, Pompe A Chaleur, combustion (des asphaltes ; par brûleur spécial : par gazogène...) ; 1 seul document parasite relatif au traitement des eaux usées , attiré par le mot POMPE.

Finalement, on a obtenu 8 classes associées à des thèmes définis dont l'importance est certaine (même si comme pour "accumulateurs" les documents sont peu nombreux). La classe 516, si elle n'a pas de thème unique, rentre nettement dans le domaine de l'INFORMATION et de l'INFORMATIQUE : classe 513. Seules les classes 510 et

517 (réunies en la classe 519) sont définies par une préoccupation générale (FABRICATION, PRODUCTION...) plutôt que par un thème (bien que la combustion prédomine dans 517. Compte tenu de ce que beaucoup de compte-rendus abordent des sujets multiples cette classification est satisfaisante.

Quant aux documents ne contenant que peu d'occurrences de MOT2, certains sont très brefs et quasi inclassables. D'autres ont un thème précis, mais insufisamment représenté dans notre corpus de 268 textes pour avoir fourni des mots au vocabulaire retenu : ils pourront apparaître sur un corpus complété par des documents nouveaux. Par exemple le mot microordinateur n'est pas dans MOT2.

3.3.3 *Croisements entre classes de textes et classes de mots et affectation des documents nouveaux* : Notons respectivement CDOC et CMOT les partitions en 11 classes retenues pour DOC et MOT2. On définit entre CDOC et CMOT une correspondance par la formule :

$$k(cd,cm) = \Sigma\{k(doc,mot) \mid doc \in cd \ ; \ mot \in cm\} \ ;$$

en d'autres termes k(cd,cm) est le nombre total des occurrences des mots de la classe cm dans les documents de la classe cd. On obtient le tableau ci-joint, déjà utilisé, nous l'avons dit, pour aider à l'interprétation des classes de documents proposés au § 3.3.2.

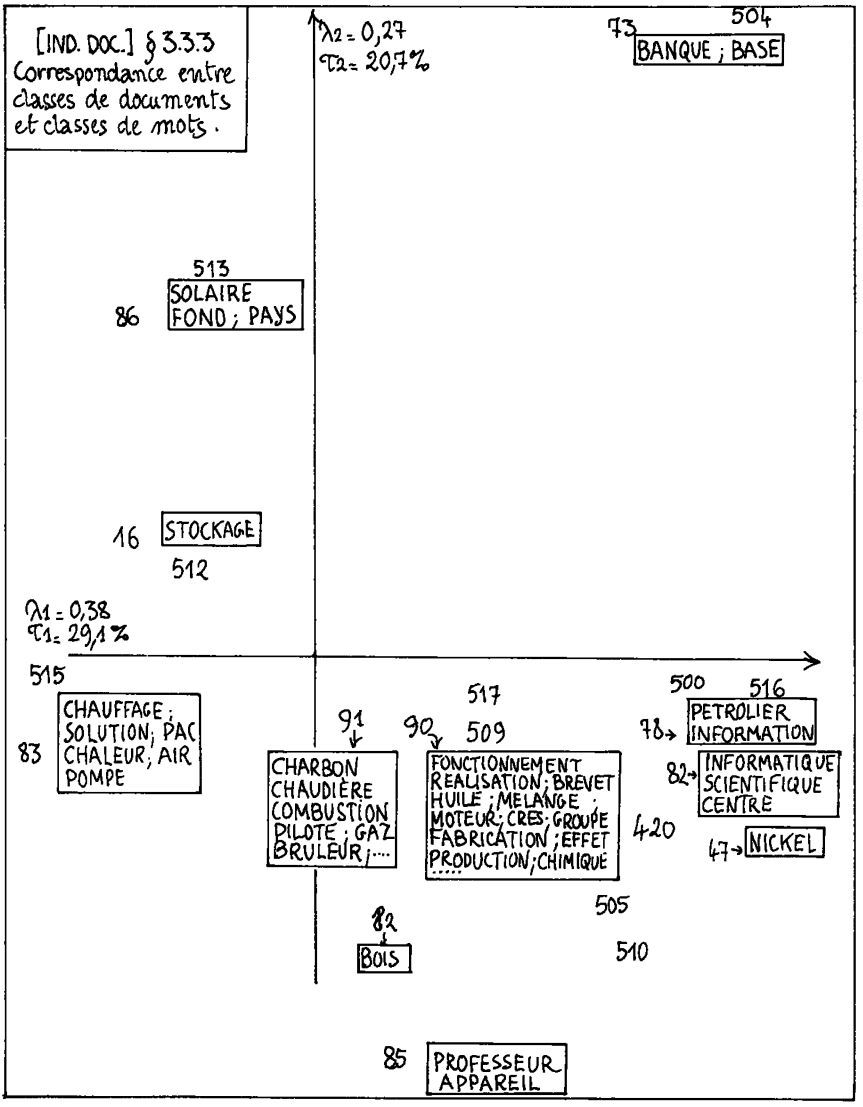
[IND.DOC] §3.3.3	combustion	BOIS	FABRICATION FONCTIONNEMENT	recherche universitaire	SOLAIRE	CHAUFFAGE; PAC	STOCKAGE d'én.	INFORMATION	INFORMATIQUE	BANQUE de d.	NICKEL
CDOC \ CMOT	91	12	90	85	86	83	16	78	82	73	47
516	2	0	72	8	9	1	1	54	62	22	0
504	0	0	30	0	0	0	0	5	3	57	0
500	0	0	4	0	1	0	0	15	28	3	0
420	0	0	13	0	1	1	0	0	0	0	30
509	409	2	74	2	6	12	2	5	8	2	0
505	22	73	47	5	6	3	2	5	4	0	0
510	9	2	99	64	1	18	0	8	6	6	0
517	256	0	553	8	39	108	10	82	113	27	13
513	6	1	50	0	139	64	3	1	9	13	0
512	6	6	64	1	32	39	38	1	1	0	0
515	985	7	102	17	33	399	3	11	6	6	185

Le tableau peut se lire directement. Toutefois, comme les lignes et colonnes ont des poids inégaux, certaines de leurs affinités apparaissent mieux après analyse factorielle. Une première analyse a montré sur un axe l'association entre cd = 420 et cm = 47, i.e. deux documents relatifs aux accumulateurs avec NICKEL. Afin d'avoir sur le plan 1 × 2 une image aussi riche que possible, on a mis en supplémentaire les classes cd = 420, 516 et cm = 82, 47, 91. Le graphique ci-joint confirme l'ensemble des associations déjà notées au § 3.3.2.

Enfin, puisque le cumul suivant les classes de CMOT paraît suffire pour interpréter la CAH des documents, on s'est demandé si ce même cumul ne fournirait pas un moyen simple pour indexer tout document automatiquement, avec ou sans machine. On a donc construit le tableau DOC × CMOT :

$$k(doc,cm) = \Sigma\{(doc,mot) \mid mot \in cm\} ;$$

dans ce tableau, il correspond à tout document de DOC (et plus généralement à tout document nouveau) une ligne de 11 nombres k(doc,cm) donnant le cumul par classes des occurrences dans DOC des mots de MOT2. Il apparaît que par simple lecture de cette ligne on peut déterminer de quoi parle le document.



En effet, après s'être fixé un seuil s (déterminé par l'expérimentation) on peut déterminer quelles sont les classes de mots cm pour lesquelles k(doc,cm) ≥ s ; on en conclura que les thèmes définis par ces classes sont spécifiques du document considéré. L'intérêt de cette méthode est que, à la différence de la CAH, elle permet éventuellement de reconnaître plusieurs thèmes propres à un seul texte. Pour des documents très courts comme le sont les nôtres, le choix du seuil s est embarrassant car un seuil très faible est peu sûr ; mais un seuil élevé élimine la plupart des documents. En prenant s = 10 nous avons vérifié qu'on ne commettait pas d'erreur d'affectation ; mais la CAH a montré qu'on pouvait descendre bien au-dessous.

Corrélativement, en lisant une colonne cm du tableau DOC × CMOT, on peut retrouver rapidement tous les documents rentrant totalement ou partiellement dans le thème "cm". Dans la pratique, on pourra utiliser un seuil variable, d'abord élevé pour retrouver un petit nombre de documents pertinents ; et plus bas ensuite, si les informations retrouvées sont insuffisantes en sorte qu'il faut poursuivre l'exploration du corpus.

Avec un ensemble de textes très étendu, on peut concevoir une hiérarchie de traitements : placer d'abord le document dans une branche d'après un vocabulaire général ; puis le situer au sein de la branche choisie d'après le vocabulaire spécifique de celle-ci. Les vocabulaires utilisés devraient être mis périodiquement à jour, en reprenant l'analyse MOT × N du § 3.1 sur la tranche des documents assez récents pour être utiles au chercheur.