Chapitre 5

Typologies, visualisations

Comment appliquer dans des situations réelles les techniques multidimensionnelles définies sur des exemples d'école aux chapitres précédents ? La complexité de l'information et la multiplicité des points de vue possibles ne permettent pas de proposer une voie unique menant du problème à une solution définitive. Dans ce chapitre, nous allons au contraire tenter de reconnaître divers chemins permettant de reculer quelque peu le moment de la nécessaire intervention interprétative de l'utilisateur. En somme, on ambitionne d'étendre le domaine de l'analyse contrôlable et reproductible, ces deux vocables peu élégants étant peut-être moins polémiques que les qualificatifs *objectif* et *automatique*...

Il s'agira ici d'aider à la lecture d'une série de textes, qu'il s'agisse de textes littéraires, de documents, ou de réponses à des questions ouvertes regroupées en textes artificiels (regroupement par classes d'âge, par profession, par niveau d'instruction ou tout autre critère pouvant présenter de l'intérêt vis-à-vis du phénomène étudié).

Quels sont les textes les plus semblables en ce qui concerne le vocabulaire et la fréquence des formes utilisées (autrement dit, quels sont les textes dont les profils lexicaux sont similaires) ? Quelles sont les formes qui caractérisent chaque texte, par leur présence ou leur absence ? On reconnaît là les questions auxquelles permet de répondre l'analyse des correspondances du *tableau lexical entier* (tableau croisant formes graphiques et textes). Ce type de traitement fera l'objet du paragraphe 5.1 intitulé : *Analyse des correspondances sur tableau lexical*.

Pour le cas des réponses libres dans les enquêtes, l'approche que nous proposons présuppose que les réponses ont été préalablement regroupées. Or les critères de regroupement les plus pertinents ne sont pas connus a priori. Il n'est d'autre part pas toujours possible, compte tenu du nombre de variables pouvant servir de critère de regroupement, et donc du nombre encore plus

grand de croisements de ces variables, d'essayer toutes les combinaisons possibles.

On proposera en fait trois stratégies d'analyse lorsqu'il n'existe pas de variable privilégiée pour regrouper les textes :

a) L'utilisation d'une partition de synthèse. Une technique de classification automatique déjà esquissée au chapitre 4 qui permet de résumer en une seule partition de synthèse les différentes caractéristiques des personnes interrogées. Cette technique, dite de "partition en noyaux factuels" est exposée sur un exemple au paragraphe 5.2 intitulé : *Les noyaux factuels*.

b) Une analyse directe des réponses non regroupées. Si les réponses paraissent suffisamment riches pour être traitées isolément, une analyse directe du tableau lexical entier croisant formes graphiques et réponses peut être opérée. Une telle analyse produira une typologie des réponses, en général assez grossière, et, de façon duale, une typologie portant sur les formes. Les réponses émanent d'individus, dont les caractéristiques sont connues par leurs réponses aux autres questions posées lors de l'enquête. Il sera donc possible d'illustrer ces typologies par des caractéristiques qui auront le statut de variables supplémentaires ou illustratives. Ce traitement direct des réponses pourra conduire à la réalisation d'un post-codage partiellement automatisé. Les problèmes et les éléments de solutions relatifs à cette approche directe seront abordés au paragraphe 5.3, intitulé : *Analyse directe des réponses ou documents.*

c) Une analyse juxtaposant des tables de contingence. Dans ce cas, pour un même vocabulaire V sélectionné, on construit autant de tableaux lexicaux qu'il y a de variables nominales, et l'on soumet à l'analyse la juxtaposition de ces tableaux. Cette méthode, d'interprétation plus délicate, est esquissée à partir d'un exemple au paragraphe 5.4, intitulé : *Analyse des correspondances à partir d'une juxtaposition de tableaux lexicaux*.

Enfin, toutes ces approches qui font intervenir des tableaux impliquant des formes graphiques peuvent aussi être réalisées à partir des unités statistiques que sont les segments répétés. Le paragraphe 5.5, intitulé *Analyse des correspondances à partir du tableau des segments répétés*, reprendra l'optique du premier paragraphe en substituant aux formes les segments. Ces divers traitements seront présentés à partir d'exemples. La plupart d'entre eux se réfèrent à un même corpus de réponses à une question ouverte.

5.1 Analyse des correspondances sur tableau lexical

On a vu au cours des chapitres précédents comment les réponses pouvaient être numérisées de façon complètement "transparente" pour l'utilisateur. Le résultat de cette numérisation peut prendre deux formes différentes, matérialisées par deux tableaux **R** et **T**.

5.1.1 Les tableaux lexicaux de base

Le tableau **R** a autant de lignes qu'il existe de réponses. On note en général ce nombre de lignes par k, nombre d'individus (il peut y avoir des réponses vides, mais il est commode de réserver une ligne pour chaque individu interrogé, de façon à assurer une correspondance aisée avec les autres informations disponibles sur ces individus).

Le tableau **R** a d'autre part un nombre de colonnes égal à la longueur de la plus longue réponse (nombre d'occurrences dans cette réponse). Pour un individu i, la ligne i du tableau **R** contient les adresses des formes graphiques qui composent sa réponse, en respectant l'ordre et les éventuelles répétitions de ces formes. Ces adresses renvoient au vocabulaire propre à la réponse. Le tableau **R** permet donc de restituer intégralement les réponses originales.

En pratique, le tableau **R** n'est pas rectangulaire, car chacune de ses lignes a une longueur variable. Les nombres entiers qui composent le tableau **R** ne peuvent dépasser *V*, étendue du vocabulaire.

Le tableau **T** a le même nombre de lignes que le tableau **R**, mais il possède autant de colonnes qu'il y a de formes graphiques utilisées par l'ensemble des individus, c'est-à-dire *V* colonnes. A l'intersection de la ligne i et de la colonne j de **T** figure le nombre de fois où la forme j a été utilisée par l'individu i dans sa réponse. Il s'agit donc d'une table de contingence (réponses ∞ formes), c'est-à-dire d'un tableau lexical entier.

Le tableau **T** peut être aisément construit à partir du tableau **R**, mais la réciproque n'est pas vraie : l'information relative à l'ordre des formes dans chaque réponse est perdue dans le tableau **T.**

En fait, le tableau **R** est beaucoup plus compact que le tableau **T** : ainsi, une réponse contenant 20 occurrences (pour un lexique de 1 000 formes) correspond à une ligne de longueur 20 du tableau **R** et à une ligne de longueur 1 000 du tableau **T** (cette dernière ligne comprenant au moins 980 zéros...). Les traitements statistiques et algorithmiques qui mettront en jeu le

tableau **T** seront en réalité programmés à l'aide du tableau **R**, moins gourmand en mémoire d'ordinateur.

5.1.2 Les tableaux lexicaux agrégés

Dans la plupart des applications, les réponses isolées sont trop pauvres pour faire l'objet d'un traitement statistique direct : il est nécessaire de travailler sur des regroupements de réponses.

Reprenant les notations du chapitre 3, on désignera par **Z** le tableau disjonctif complet à k lignes et p colonnes décrivant les réponses de k individus à une question fermée comportant p modalités de réponse possibles, ces réponses étant mutuellement exclusives. Autrement dit, une ligne de **Z** ne comportera ici qu'un seul "1", et *(p-1)* "0".

Contrairement au tableau de l'exemple du chapitre 3 (tableau 3.9), le tableau **Z** ne concerne qu'une seule question, et comporte donc un bloc unique. Chaque question fermée de ce type permet de définir une partition des individus interrogés.

Le tableau **C**, obtenu par le produit matriciel:

$$\mathbf{C} = \mathbf{T}' \mathbf{Z}$$

est un tableau à V lignes (V est, rappelons-le, le nombre total de formes distinctes) et p colonnes (p est le nombre de modalités de réponses à la question fermée considérée) dont le terme général c_{ij} n'est autre que le nombre de fois où la forme i a été utilisée par l'ensemble des individus ayant choisi la réponse j.

Un exemple de tableau **C**, de dimension modeste, est fourni par le tableau 3.1 du chapitre 3, avec $V = 14$, et $p = 5$. La question fermée était dans ce cas : *Quel est le diplôme d'enseignement général le plus élevé que vous avez obtenu ?*

Il est donc aisé, pour toute question fermée dont les réponses sont codées dans un tableau $\mathbf{Z}_q$, de calculer le tableau lexical agrégé $\mathbf{C}_q$ par la formule :

$$\mathbf{C}_q = \mathbf{T}' \mathbf{Z}_q$$

et donc de comparer les profils lexicaux de différentes catégories de population. Chaque tableau $\mathbf{C}_q$ donne un point de vue différent (le point de vue de la question fermée q) sur la dispersion des profils lexicaux des réponses à la question ouverte étudiée.

5.1.3 Seuil de fréquence pour les formes

Ces comparaisons de profils lexicaux n'ont de sens, d'un point de vue statistique, que si les formes apparaissent avec une certaine fréquence : les hapax ou même les formes rares seront écartés de la phase de comparaisons de fréquences. Ceci a pour effet de réduire considérablement la taille du vocabulaire effectivement pris en compte.

Pour une question ouverte posée à 2 000 personnes, compte tenu de la structure de la gamme des fréquences du vocabulaire, une sélection des formes apparaissant au moins 10 fois peut, dans bien des cas, ramener la valeur de V de 1 000 (pour fixer les idées) à 100...

5.1.4 Présentation de l'exemple

On reprendra, en vraie grandeur, l'exemple de réponses à une question ouverte qui a été présenté dans une version simplifiée au chapitre 3 et que nous avons appelé : la question ***Enfants***.

Numérisation du texte

Le bilan de la première phase de numérisation du texte réalisée par le programme de traitement est le suivant :

Sur 2 000 réponses, on relève 15 457 occurrences, avec 1 305 formes graphiques distinctes. Si l'on sélectionne les formes apparaissant au moins 20 fois, il reste 12 051 occurrences de ces formes, qui sont au nombre de 117.

Le tableau 5.1 donne la liste alphabétique de ces formes les plus fréquentes ; celles-ci sont ensuite classées par ordre de fréquences décroissantes dans le tableau 5.2. Cette "mise à plat" de l'information de base peut paraître surprenante. En fait, il ne s'agit que d'une étape intermédiaire, et les traitements ultérieurs vont redonner forme à cet inventaire.

5.1.5 Construction du tableau lexical agrégé

Les réponses à la question ***Enfants,*** dont on peut trouver quelques exemples au paragraphe 2.2 du chapitre 2 sont en général assez lapidaires. Dans ces conditions, deux réponses qui semblent au premier abord traduire des préoccupations similaires, telles par exemple : *manque d'argent* et *problèmes financiers* peuvent n'avoir aucune forme graphique en commun, et donc ne pas être reconnues comme proches.

Tableau 5.1

Vocabulaire des réponses à la question *Enfants* (Ordre alphabétique)

NUM.	MOTS EMPLOYÉS	FREQUENCES	NUM.	MOTS EMPLOYES	FREQUENCES
1	a	218	60	l	674
2	actuelle	83	61	la	666
3	argent	196	62	le	474
4	au	44	63	les	442
5	aucune	33	64	leur	71
6	aussi	22	65	liberte	52
7	avec	21	66	logement	52
8	avenir	318	67	maladie	38
9	avoir	77	68	manque	160
10	beaucoup	23	69	materielle	27
11	bien	21	70	materielles	57
12	c	98	71	monde	21
13	ca	41	72	moyens	44
14	ce	34	73	n	94
15	cela	23	74	ne	193
16	chomage	285	75	on	84
17	conditions	48	76	ont	24
18	conjoncture	22	77	ou	58
19	couple	95	78	par	23
20	crainte	24	79	parce	22
21	crise	25	80	pas	325
22	d	332	81	peur	160
23	dans	67	82	peut	54
24	de	915	83	plus	87
25	des	208	84	pour	161
26	deux	26	85	pouvoir	41
27	difficile	28	86	probleme	37
28	difficultes	82	87	problemes	108
29	du	155	88	professionnelle	22
30	economique	54	89	qu	51
31	egoisme	109	90	quand	28
32	elever	51	91	que	78
33	emploi	79	92	question	48
34	en	116	93	questions	23
35	enfant	148	94	qui	76
36	enfants	148	95	raison	41
37	entente	22	96	raisons	176
38	est	190	97	responsabilite	21
39	et	204	98	responsabilites	21
40	etre	69	99	ressources	49
41	faire	22	100	s	55
42	fait	25	101	sais	25
43	faut	35	102	sante	95
44	femme	60	103	se	36
45	finances	28	104	si	56
46	financier	24	105	situation	176
47	financiere	91	106	son	28
48	financieres	173	107	sont	44
49	financiers	86	108	sur	29
50	gens	25	109	surtout	34
51	guerre	27	110	tout	41
52	il	105	111	travail	152
53	ils	79	112	trop	52
54	incertain	45	113	un	172
55	incertitude	29	114	une	120
56	independance	23	115	veulent	42
57	insecurite	62	116	vie	179
58	instabilite	21	117	y	56
59	je	62			

Il s'agit là d'un problème important sur lequel on reviendra à plusieurs reprises dans ce chapitre et dans ceux qui suivent. En bref, tout dépend du degré de finesse que l'on assigne à l'analyse, et aussi de la taille de l'échantillon, indissolublement liée à cette exigence.

Tableau 5.2

Vocabulaire de la question *Enfants* (Ordre lexicométrique)

NUM.	MOTS	EMPLOYES	FREQUENCES	NUM.	MOTS	EMPLOYES	FREQUENCES
33	de		915	42	economique		54
76	l		674	82	logement		52
77	la		666	81	liberte		52
78	le		474	145	trop		52
79	les		442	109	qu		51
31	d		332	45	elever		51
98	pas		325	119	ressources		49
11	avenir		318	22	conditions		48
21	chomage		285	112	question		48
1	a		218	69	incertain		45
34	des		208	6	au		44
53	et		204	90	moyens		44
3	argent		196	135	sont		44
92	ne		193	148	veulent		42
52	est		190	115	raison		41
150	vie		179	16	ca		41
131	situation		176	140	tout		41
116	raisons		176	105	pouvoir		41
63	financieres		173	84	maladie		38
146	un		172	106	probleme		37
104	pour		161	128	se		36
85	manque		160	58	faut		35
100	peur		160	18	ce		34
41	du		155	138	surtout		34
141	travail		152	7	aucune		33
49	enfant		148	137	sur		29
50	enfants		148	70	incertitude		29
147	une		120	60	finances		28
48	en		116	37	difficile		28
44	egoisme		109	110	quand		28
107	problemes		108	134	son		28
67	il		105	66	guerre		27
15	c		98	86	materielle		27
127	sante		95	36	deux		26
26	couple		95	30	crise		25
91	n		94	56	fait		25
62	financiere		91	124	sais		25
103	plus		87	65	gens		25
64	financiers		86	29	crainte		24
93	on		84	61	financier		24
2	actuelle		83	94	ont		24
39	difficultes		82	13	beaucoup		23
68	ils		79	113	questions		23
47	emploi		79	71	independance		23
111	que		78	96	par		23
12	avoir		77	19	cela		23
114	qui		76	108	professionnelle		22
80	leur		71	51	entente		22
54	etre		69	55	faire		22
32	dans		67	24	conjoncture		22
74	je		62	97	parce		22
72	insecurite		62	8	aussi		22
59	femme		60	73	instabilite		21
95	ou		58	89	monde		21
87	materielles		57	14	bien		21
154	y		56	118	responsabilites		21
130	si		56	117	responsabilite		21
123	s		55	10	avec		21
101	peut		54				

Une analyse de contenu rapide sur petit échantillon se heurtera à des problèmes de dispersion des formes relatives à un même lemme, et aussi aux problèmes de synonymies. Il n'en est pas de même si l'on s'astreint à traiter des ensembles importants de réponses. Dans ce cas, les fréquences permettent en quelque sorte de consolider les emplois de tel terme ou

locution, et parfois d'identifier des catégories de locuteurs qui les utilisent de manière privilégiée.

Il convient donc, dans un premier temps, de trouver des regroupements d'individus pertinents vis-à-vis du phénomène étudié. On reviendra sur la façon de choisir un tel regroupement. Dans le cas particulier de l'exemple traité, les individus sont regroupés en 9 classes, qui diffèrent soit par l'âge, soit par le niveau d'instruction générale. Cette partition en 9 groupes est obtenue en croisant deux partitions élémentaires qui contiennent chacune 3 classes[1].

Tableau 5.3
Tête de liste du tableau lexical entier croisant les 117 formes de fréquence supérieure à 20 avec la partition *âge-diplôme* en 9 groupes

	A -30	B -30	S -30	A -50	B -50	S -50	A +50	B +50	S +50
a	14	17	25	34	10	29	61	16	12
actuelle	7	6	7	18	8	5	20	8	4
argent	20	18	22	38	11	17	57	4	9
au	1	2	6	6	3	7	14	0	5
aucune	0	4	2	7	2	0	12	5	1
aussi	3	2	1	4	4	3	3	0	2
avec	2	2	2	4	1	4	6	0	0
avenir	32	40	43	58	21	38	53	17	16
avoir	12	5	11	8	2	5	25	7	2
beaucoup	0	1	1	8	2	1	10	0	0
bien	3	2	2	4	1	2	6	1	0
. . . . .									

Le tableau 5.3 est en tout point analogue au tableau 3.1 du chapitre 3, à ceci près qu'il comporte l'intégralité des formes graphiques de fréquences supérieures à 20 (117 au lieu de 14) et une partition des individus en 9 classes intégrant l'âge des répondants (au lieu d'une partition en 5 classes par niveau d'instruction).

[1] Première partition : trois classes d'âge : moins de 30 ans (notée -30), entre 30 et 50 ans (-50), plus de 50 ans (+50). Seconde partition : trois classes de niveaux d'instruction: Aucun diplôme ou Certificat d'études primaires (notée A), BEPC ou équivalent (notée B), Bac et études Supérieures (notée S).

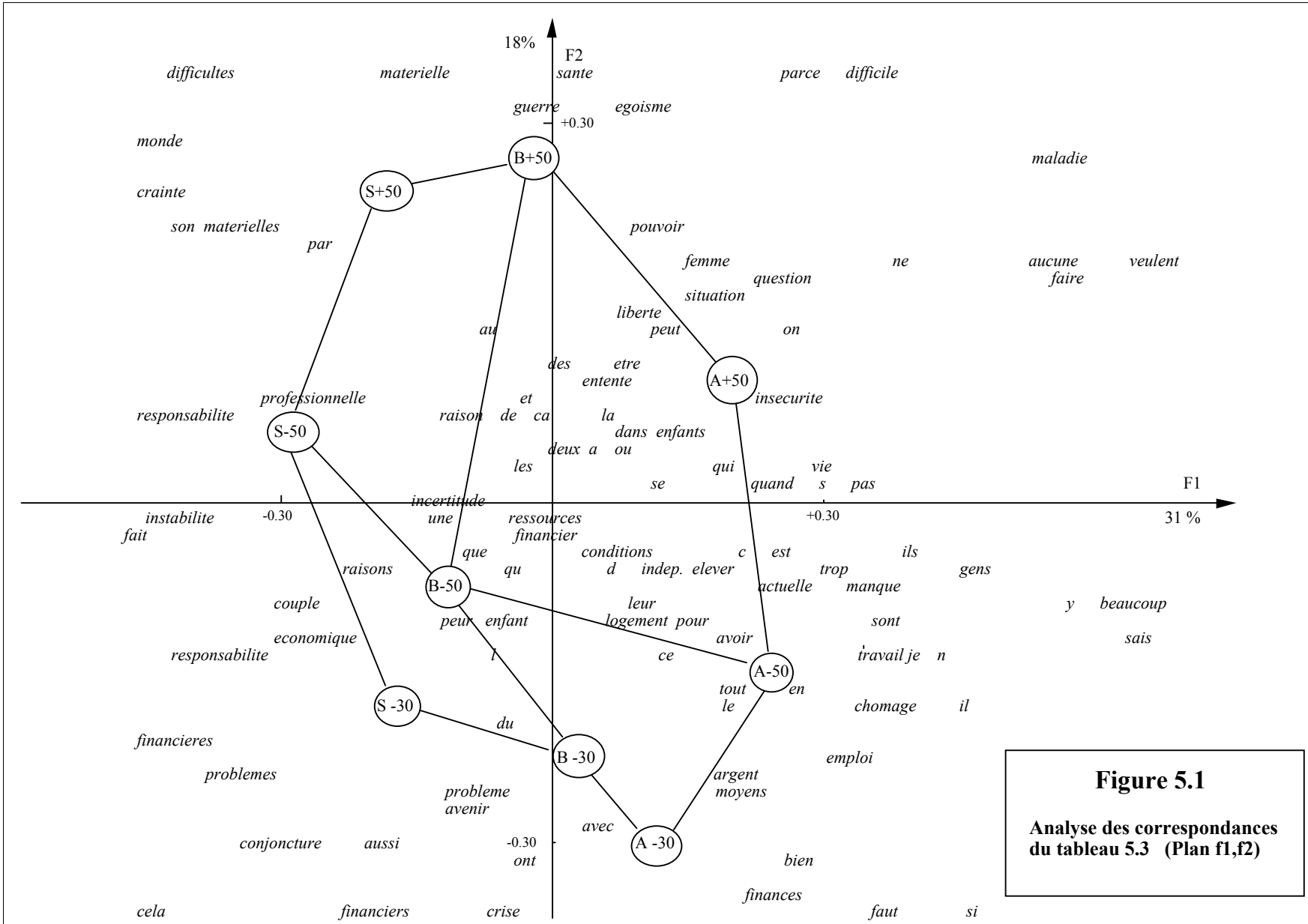

Figure 5.1

Analyse des correspondances du tableau 5.3 (Plan f1,f2)

Pour lire efficacement l'information contenue dans ce tableau, il faut calculer les tableaux de profils-lignes et de profils-colonnes, et représenter les distances entre formes d'une part, entre catégories d'âge et d'instruction, d'autre part. C'est précisément la vocation de l'analyse des correspondances que de procéder à cette double description.

5.1.6 Analyse et interprétation du tableau lexical

La figure 5.1 représente le premier plan factoriel (c'est-à-dire le plan des deux premiers facteurs) de l'analyse des correspondances du tableau 5.3. Les deux premières valeurs propres valent respectivement 0.035 et 0.021, et correspondent à 31% et 18% de la trace (ou encore inertie totale, ou variance totale).

La disposition des points-colonnes est d'une régularité assez étonnante : à partir d'une information purement lexicale (les éléments des profils-colonnes), on retrouve le caractère composite de la partition des individus en 9 classes. A âge égal, les individus sont d'autant plus instruits qu'ils sont situés vers la gauche du graphique ; à niveau d'instruction égal, ils sont d'autant plus âgés qu'ils occupent une position élevée sur l'axe vertical.

Ainsi donc, ces vecteurs décrivant pour chaque catégorie la fréquence de 117 formes (choisies à partir d'un simple critère de fréquence) permettent de reconstituer simultanément la gradation des âges, et celle du niveau de diplôme à l'intérieur de chaque classe d'âge.

L'examen des positions des points représentant les formes reste assez frustrant, à ce niveau d'analyse limitée aux seules formes, en raison de l'absence de contexte.

On devine des *je ne sais pas, je ne vois pas* du côté des personnes peu instruites et/ou âgées. Les formes *monde*, *responsabilités*, *conjoncture* sont invoquées par les personnes les plus instruites. On remarque que la forme *avenir* est résolument du côté des plus jeunes, alors que *égoïsme,* mais aussi *santé*, *maladie* sont du côté des réponses fournies par les personnes les plus âgées.

On note d'ailleurs que les trois classes "jeunes" sont plus resserrées que celles de leurs aînés, phénomène que l'on retrouve souvent dans un cadre socio-économique plus général : à l'intérieur de la classe des personnes âgées, on trouve en effet un éventail de situations (niveau de vie, niveau d'instruction) plus large que chez les jeunes.

Interprétation

On peut faire quelques remarques à ce stade:

a) L'indexation automatique des formes et les calculs de fréquences qu'elle permet ignorent délibérément de nombreuses informations de type sémantique ou syntaxique qui sont accessibles à tout lecteur des textes considérés. La synonymie n'est pas prise en compte, pas plus que l'homonymie.

   La pratique de ce type d'analyse sur des échantillons importants (ici : 2 000 réponses) montre que ces objections peuvent être assez facilement levées dans le cas de discours artificiels construits par juxtaposition de réponses et pour lesquels on recherche surtout des éléments de répétition.

   Dans ce contexte statistique, les dépouillements en formes graphiques se révèlent souvent plus intéressants que les dépouillements en lemmes. Les formes *problème* au singulier et au pluriel occupent des positions similaires sur la figure 5.1 (partie inférieure gauche), ce qui prouve qu'il n'était pas gênant de distinguer les deux formes. Le seuil de fréquence choisi (formes apparaissant plus de 20 fois) a éliminé la forme *difficultés* qui apparaît exactement 20 fois, et n'a laissé que le singulier. L'analyse avec un seuil inférieur, ou le positionnement du pluriel de ce mot en élément supplémentaire montre que les deux formes sont au contraire très distantes, (en bas à droite au singulier, en haut à gauche au pluriel) ce qui n'est pas moins intéressant, car cette opposition renvoie à des contextes d'utilisation très différents (brièvement : mentions de *difficultés matérielles* pour les personnes plutôt âgées et instruites, contre *difficulté de trouver du travail* pour des répondants sensiblement plus jeunes et moins instruits).

   A propos d'autres exemples, on a pu noter que la présence de flexions distinctes d'un même verbe (comme: *peuvent, peut, pouvoir, puisse*) et de synonymes aide à confirmer l'interprétation de certaines zones du plan factoriel : le regroupement de formes graphiques différentes possédant des affinités au plan sémantique est au contraire un critère de validation supplémentaire des résultats empiriques.

b) La prise en compte aveugle de mots-outil (comme : *de, les, quand, que, pour, ...*) n'alourdit en rien l'analyse. Ces formes n'apparaîtront dans une analyse factorielle lexicale que si leur répartition n'est pas uniforme dans les textes, autrement dit que si elles caractérisent électivement

certains discours : dans ce cas, elles méritent effectivement d'être situées par rapport aux autres formes.[1]

c) L'ordre des formes graphiques dans les réponses n'est pas pris en compte : chaque discours est pour les programmes de calcul un "sac de mots", dont seul le profil de fréquences est actuellement exploité. Cette objection est sérieuse, encore qu'un profil de fréquences se révèle à l'expérience beaucoup plus riche en information qu'on ne l'imagine a priori.

Si dans l'absolu, un profil lexical, c'est-à-dire ici une suite de 117 sous-fréquences, n'a pas grand sens, la confrontation de plusieurs profils lexicaux est au contraire riche d'enseignements dans l'optique comparative choisie ici.

Toujours dans une optique fréquentielle, la recherche des *segments répétés* permet de prendre en compte les occurrences d'unités plus riches au plan sémantique que les formes isolées. La sélection des *réponses modales* qui sera évoquée plus bas répond également à plusieurs des objections précédentes en mettant en évidence les contextes les plus fréquents de certaines de ces formes.

Stabilité vis-à-vis d'une lemmatisation interne

Les remarques a) et b) du paragraphe précédent posent la question de la stabilité des résultats en cas de suppression d'une "liste de mots-outils" et/ou de remplacement systématique des formes graphiques par les lemmes auxquelles on peut les rattacher[2].

Le temps d'une expérience, une procédure de lemmatisation interne au sous-corpus utilisé après intervention du seuil permettra de vérifier la stabilité des structures obtenues.

Il s'agit en quelque sorte de vérifier si la structure observée[3] (disposition relative des neuf points catégories sur la figure 5.1) n'est pas un artefact dû à la présence de formes grammaticales particulières, auquel cas les catégories se distingueraient avant tout par la forme de leur discours, et non par le contenu des réponses dont on peut supposer qu'il est plutôt lié aux mots pleins.
Dans la liste de formes du tableau 5.1, on a supprimé les formes suivantes :

---

[1] Une étude sur l'image de l'institution du mariage (Lebart, 1982b) a montré, par exemple, que l'adverbe *quand* est surtout utilisé lors de réponses traditionalistes (*quand on se marie, c'est pour la vie*, ou encore : *quand on se marie, c'est pour avoir des enfants).*

[2] Nous verrons au chapitre 7 une étude en grandeur réelle sur ce même sujet.

[3] Le mot anglais *pattern* serait encore plus précis.

a, au, c, ca, ce, cela, d, dans, de, des, du, en, et, l, la, le, les, n, ne, on, ou, par, parce, quand, que, qu, qui, s, se, si, sur, un, une, y

De plus, on a rassemblé en de mêmes unités les ensembles de formes suivants (la première forme de chaque ligne subsiste et remplace la ou les suivantes).

enfant	*enfants*
avoir	*ont*
finances	*financier, financière, financiers , financières*
problème	*problèmes*
matérielle	*matérielles*
raison	*raisons*
question	*questions*
responsabilité	*responsabilités*
être	*est , sont*
pouvoir	*peut*
faire	*fait*

Il s'agit en quelque sorte d'une lemmatisation interne à l'ensemble des formes sélectionné, (après constitution d'un sous-corpus par seuillage sur la fréquence des formes) et non d'une vraie lemmatisation comme celle présentée au chapitre 2 (paragraphe 2.8).

Il est clair que lorsque la lemmatisation intervient sur la totalité du texte original, elle influe également sur la constitution de l'ensemble des formes que l'on sélectionne par seuillage. En particulier, les verbes, beaucoup plus dispersés en français que les noms et les adjectifs, sont défavorisés par une sélection à partir des fréquences de formes, puisque de nombreuses flexions d'un même verbe peuvent apparaître dans le texte de départ avec une fréquence inférieure au seuil de fréquence retenu.

Néanmoins, l'optique prise pour cette expérience est celle d'un calcul de perturbation et de stabilité interne au sous-corpus sélectionné. Le vocabulaire est ici réduit de 117 formes à 68 "pseudo-lemmes". Pour cet exemple particulier, le graphique 5.1 n'est pas profondément modifié par cette diminution considérable du nombre de formes.

On a pu vérifier que les points-catégories sont disposés d'une façon qui respecte l'ordre des classes d'âge et des niveaux de diplôme. Ce qui est une présomption supplémentaire (et non une preuve) du lien entre les catégories et le contenu des réponses.

5.2 Les noyaux factuels

On a vu qu'il était souvent nécessaire de regrouper les réponses pour pouvoir procéder à des analyses de type statistique. Les profils lexicaux d'agrégats de réponses ont plus de régularité et s'interprètent plus facilement que les réponses isolées. Les exemples précédents ont montré comment ce regroupement a priori pouvait être réalisé à partir des variables retenues en fonction d'un choix d'hypothèses de départ. Mais ceci suppose une bonne connaissance a priori du phénomène étudié, situation qui n'est en général pas réalisée dans les études dites exploratoires.

La technique dite des *noyaux factuels* ou des *situations-types* brièvement exposée au chapitre 4 (exemple du paragraphe 4.3.2 relatif aux techniques de classification) va permettre de donner des éléments de réponse à ce problème.

Etant donnée une liste de descripteurs ou de variables caractérisant les individus, on se pose maintenant le problème de répartir ces individus en groupes les plus homogènes possible vis-à-vis des caractéristiques retenues... sans en privilégier aucune a priori. Comme on l'a dit au chapitre 4, il s'agit en quelque sorte d'approcher à l'intérieur des classes le *toutes choses égales par ailleurs,* si difficile à réaliser en sciences sociales. C'est précisément le type d'opération que permet d'obtenir, autant que faire se peut, un algorithme de classification, selon les modalités définies lors de l'exemple qui clôt le chapitre 4.

Un exemple

L'exemple qui suit concerne une question ouverte posée à l'issue des interviews de l'enquête déjà citée sur les conditions de vie et aspirations des Français (en 1980 et 1981), et pour laquelle on dispose de 4 000 réponses libres. Le questionnaire abordait successivement les thèmes suivants : Famille et politique familiale, Environnement, Niveau de vie, Vie au travail, quelques thèmes généraux (Science, Justice), Vie sociale, Loisirs (cf. Chapitre 2, paragraphe 2.2).

Le libellé de cette question ouverte était le suivant :

Vous venez d'être interrogés longuement sur vos conditions de vie et votre environnement. Peut-être auriez-vous aimé donner votre avis sur certains points non prévus dans ce questionnaire rigide. Avez-vous des remarques à formuler ?

On désire avoir une vue d'ensemble des réponses, mais la variété des thèmes abordés dans cette enquête, et donc l'étendue du contenu potentiel des

réponses ne permettent pas de savoir a priori quels peuvent être les critères les plus pertinents de regroupements des réponses.

A partir d'une batterie de 13 descripteurs, dont la liste figure ci-dessous, on regroupe les 4 000 individus enquêtés en 22 classes (dont une classe "résiduelle", représentant moins de 3% de l'échantillon, qui contient des questionnaires incomplets, ou un ensemble hétérogène de combinaisons de caractéristiques peu fréquentes).

Les deux dernières variables sélectionnées constituaient, à l'époque de l'enquête, d'assez bons indicateurs d'équipement pouvant remédier à certaines lacunes ou à certains biais de la variable "revenu".

Liste des descripteurs *(variables actives de la classification)*

	Nombre de modalités
Catégorie socio-professionnelle	15
Statut matrimonial	6
Diplôme d'enseignement général le plus élevé	8
Situation actuelle de la personne interrogée	8
Présence ou non d'enfants < 16 ans au foyer	2
Statut d'occupation du logement	5
Croisement sexe-activité	4
Taille d'agglomération	5
Nombre de personnes vivant actuellement au foyer	5
Croisement sexe-âge de l'enquêté	8
Revenu global du foyer	7
Possession d'un lave-vaisselle	2
Possession de plusieurs voitures	2

La figure 5.2 représente la coupure de l'arbre hiérarchique correspondant aux 22 classes finalement retenues et donne une description sommaire de la composition de chacune de ces classes.

La figure 5.3 donne une esquisse du plan factoriel que l'on obtient en soumettant à l'analyse des correspondances le tableau lexical entier croisant, en lignes les 150 formes graphiques les plus fréquentes, et en colonnes les 22 classes (ou noyaux factuels). Seules quelques formes ont été représentées sur cette figure. Le caractère composite et systématique de la partition met en évidence des particularités ou des oppositions qui auraient pu être omises à partir de regroupements plus sommaires. Les grands traits de la description fournie par cette figure auraient peut-être été saisis par des regroupements question par question.

Ainsi les personnes instruites manifestent des opinions critiques vis-à-vis du questionnaire, alors que les personnes plus âgées évoquent, en moyenne, des problèmes qui les concernent particulièrement (*retraite, impôts*, etc).

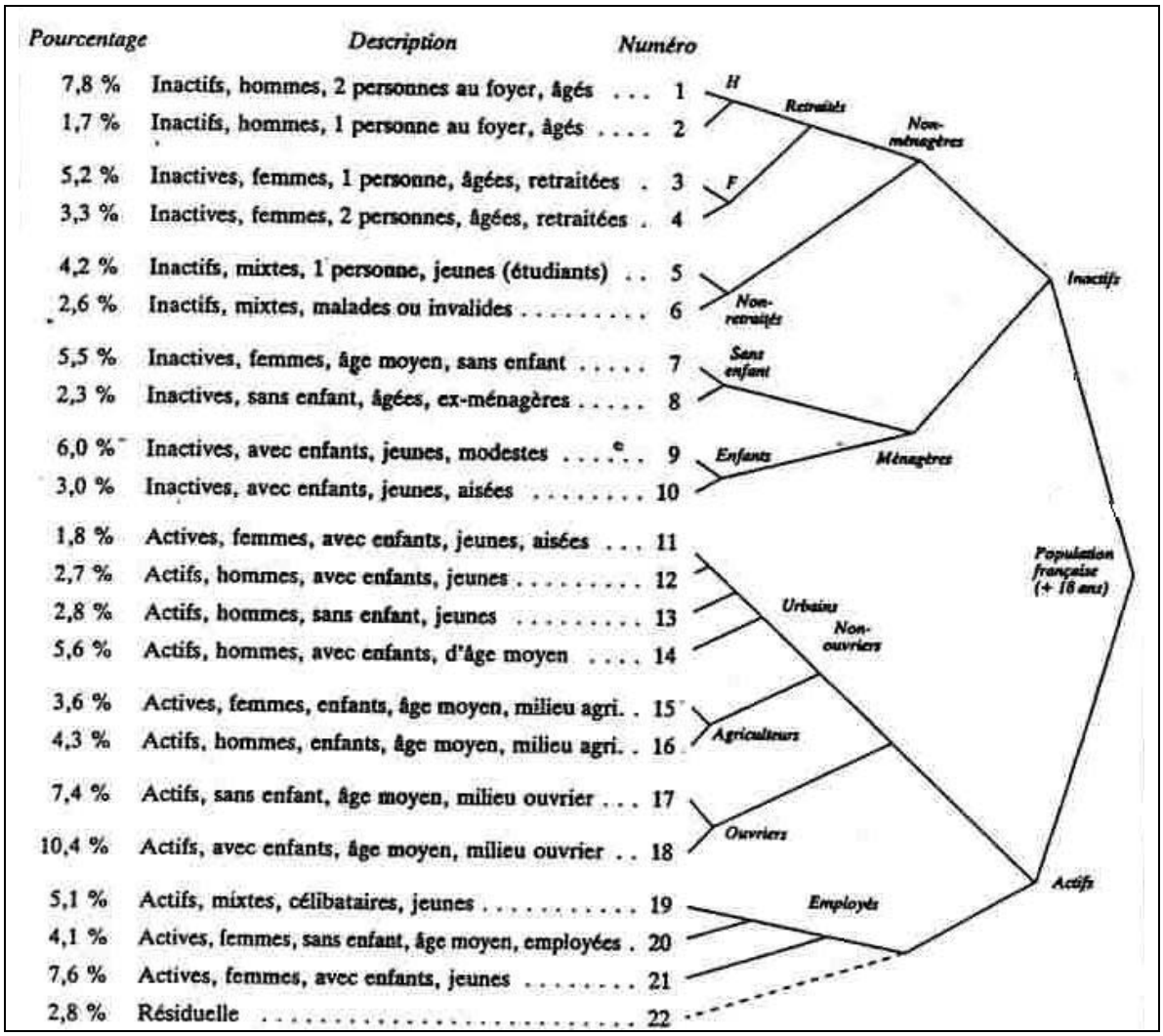

Figure 5.2

Partition en 22 noyaux factuels d'un échantillon de 4 000 personnes

Mais des nuances plus difficiles à saisir peuvent apparaître. On note ainsi, par exemple, une différence considérable de profils lexicaux entre les ouvriers de sexe masculin selon qu'ils sont sans enfant (classe 17), ou avec enfants (classe 18).

Ces derniers,visiblement intéressés et motivés, posent des problèmes non abordés dans le questionnaire, ou prolongent des débats amorcés lors de l'interview, alors que les premiers se contentent de quelques remarques assez laconiques sur le questionnaire.

Ce sont les regroupements des 22 classes observables sur la figure 5.3 qui donnent l'idée des variables sous-jacentes. Les noyaux 1, 2, 3, 4, 8 en bas à

droite sont tous composés de personnes âgées, alors que les noyaux 12 et 13 sont formés de jeunes hommes actifs.

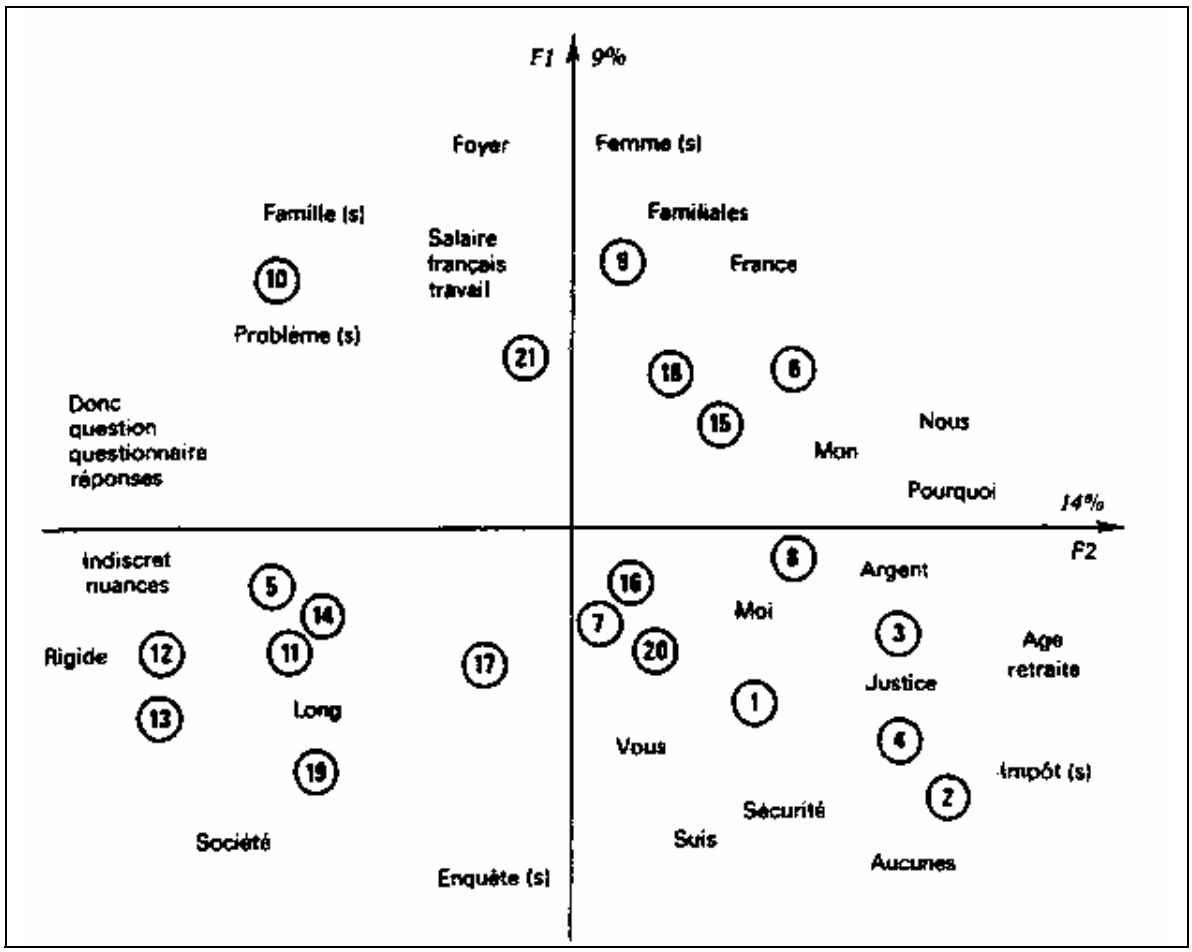

Figure 5.3

Proximités entre les 22 noyaux factuels et quelques formes (Question: *remarques à l'issue de l'interview"*)

Les classes 9, 10 et 21, constituées de femmes, évoquent souvent les lacunes du questionnaire sur certains problèmes de politique familiale. A propos de cette même question ouverte, la technique des réponses modales présentée au chapitre 6 va permettre de situer dans leurs contextes les formes représentées sur cette figure, et d'enrichir considérablement l'interprétation de ces classes.

5.3 Analyse directe des réponses ou documents

Une part importante de notre travail a été consacrée jusqu'ici à l'analyse des profils de fréquences de pour des textes relativement importants du point de vue de leur longueur.
Pour obtenir de tels textes à partir des fichiers de réponses libres, nous avons dû procéder à des regroupements a priori de ces réponses.

Mais certaines des méthodes d'analyse statistique exploratoire présentées aux chapitres 3 et 4 peuvent être appliquées avec profit aux réponses individuelles. Ce traitement direct des données individuelles est recommandé dans les deux cas suivants :

a) Lorsque les réponses sont suffisamment riches du point de vue lexical pour que les profils de fréquence puissent être comparés avec profit, ce qui peut être le cas dans les interviews prolongées, dans le domaine psychologique ou médical par exemple, ou encore lors du traitement d'unités textuelles présentant un certain volume dans le domaine de l'étude des textes socio-politiques.

b) Lorsque l'on veut procéder à un travail préliminaire de description, de regroupement et de mise en ordre.

Il est clair que le premier cas se situe dans le cadre des méthodes préconisées aux paragraphes précédents : une description directe des réponses est maintenant possible. Elle ne ferme d'ailleurs pas la porte définitivement à des regroupements ultérieurs, si cela peut aider l'interprétation ou permettre d'éprouver certaines hypothèses.

Le second cas est sensiblement différent : la notion de profil n'a plus le même sens. En termes statistiques, la variance inter-individus n'a plus le même statut que la variance inter-catégories. Les réponses se distinguent alors plus par la présence ou l'absence de formes que par de véritables variations entre profils de fréquences.

5.3.1 Comment interpréter les distances ?

On commencera par prendre un exemple simple de réponses libres à une question sur la sécurité routière, dont le libellé est le suivant :

Après une question fermée préliminaire :

A votre avis, est-il possible de diminuer fortement le nombre des tués et des blessés dans les accidents de la route ? (réponses: *oui/non*), on demande à ceux qui ont répondu *oui* (environ 80% des enquêtés en 1982) : *Que faut-il faire pour cela ?*
On peut alors trouver des réponses du type[1] :

— *développer l'usage des transports en commun*

ou encore :

— *inciter les gens à se servir le plus possible du train, du bus.*

[1] Cf. Boscher, (1984).

qui sont deux réponses ayant respectivement 7 et 13 occurrences de formes, sans aucune forme graphique commune, et dont les contenus sont pourtant assez voisins.

Inversement, les deux réponses suivantes à la même question :

— *respecter les limitations de vitesse*
— *faire respecter les limitations de vitesse*

ne se distinguent que par une seule forme graphique et ont cependant des contenus sensiblement différents.

Cet exemple correspond à une situation réelle, mais plutôt exceptionnelle. Lors des traitements de tableaux lexicaux agrégés, les réponses de ce type sont en général noyées dans des classes dont le profil lexical moyen possède, lui, une certaine régularité. Quoi qu'il en soit, ces exemples montrent que les distances entre réponses individuelles ne pourront pas être interprétées facilement.

Notre préoccupation n'est pas cependant de procéder à un décodage exhaustif de l'information, mais d'utiliser les redondances et répétitions, lorsqu'elles existent, pour simplifier cette information.

Il est clair qu'une analyse directe, dans ces conditions, pourra au moins regrouper les réponses identiques ou similaires, en laissant dans un premier temps "non-classées" les réponses que distingue l'originalité de leur forme. Il s'agit en somme d'une aide au post-codage, opération manuelle évoquée au chapitre 1, paragraphe 1.4.4.

5.3.2 Analyse du tableau clairsemé T

Cette analyse directe des réponses revient à soumettre à une analyse descriptive le tableau **T** défini plus haut (paragraphe 5.1), dont les k lignes sont les réponses et les V colonnes correspondent à un ensemble de formes.[1] Ceci appelle plusieurs remarques.

a) La proximité entre deux formes, c'est-à-dire entre deux colonnes du tableau **T** sera d'autant plus grande que ces formes apparaîtront souvent dans les mêmes réponses (et non plus seulement dans les mêmes textes ou agrégats de réponses), ce qui permettra de mieux appréhender les

[1] Le traitement d'un tableau aussi grand impliquera en général la mise en oeuvre d'algorithmes de calcul particuliers, utilisant le tableau réduit **R** au lieu du tableau **T,** et évitant le calcul et le stockage d'une matrice à diagonaliser d'ordre (V,V) (cf. par exemple, Lebart, 1982a).

voisinages syntagmatiques. L'analyse directe rendra mieux compte des contextes.

b) Les formules de transition du chapitre 3 (paragraphe 3.2) donnent aux représentations factorielles issues de l'analyse des correspondances de la matrice **T** une interprétation assez simple : les *k* individus seront approximativement aux centres de gravité des formes qu'ils auront utilisées, et réciproquement, ces formes seront approximativement positionnées aux centres de gravité des individus qui les auront choisies pour s'exprimer (nous disons approximativement pour rappeler la présence d'un coefficient dilatateur β pour chaque axe, qui déplace la position des centres de gravités...).

c) Les *k* lignes du tableau **T** représentent les réponses ou les individus interrogés. Les réponses aux questions fermées du questionnaire des mêmes *k* individus peuvent constituer les colonnes d'un tableau $\mathbf{T}^{+}$, et donc être positionnées comme éléments illustratifs (ou supplémentaires) sur les plans factoriels issus de l'analyse des correspondances de **T**. Ceci permet une visualisation très rapide des proximités entre les mots et les caractéristiques des répondants. En pratique, cela suggère des critères de regroupement.

Les modalités pratiques de ce type d'approche exploratoire seront exposées à partir d'un exemple.

5.3.3 Exemple d'application

Ici encore, on reprendra l'ensemble des 2 000 réponses à la question ouverte ***Enfants***.

Le tableau **T** compte ici *k*= 2 000 lignes et 117 colonnes (liste des formes figurant dans le tableau 5.1 au début de ce chapitre). Notons qu'un tel tableau, appelé *tableau clairsemé*, contient 95% de valeurs nulles, puisque les quelques 12 000 occurrences se répartissent dans plus de 234 000 cases !
La figure 5.4 représente le plan des deux premiers facteurs issus de l'analyse des correspondances du tableau **T**.
Les deux mille points-lignes correspondant à des individus anonymes ne figurent pas sur ce graphique, mais le tableau 5.4 nous donne les coordonnées, sur les mêmes axes factoriels, de certaines caractéristiques des répondants. C'est la figure 5.5 qui donnera, pour ce même plan factoriel, les positions de ces caractéristiques, projetées a posteriori en tant que modalités illustratives.

Tableau 5.4

Positionnement des modalités illustratives dans le plan de l'analyse directe des réponses

MODALITES	*EFF.*	*DISTO*	*COORDONNEES*		*VALEURS-TEST*	
			1	*2*	*1*	*2*
Genre (Sexe)						
homme	963	1.20	.07	.04	6.1	3.8
femme	1037	.83	-.05	-.03	-6.1	-3.8
statut matrimonial						
célibataire	348	5.21	.06	.13	2.8	5.6
marié(e)	1191	.59	-.02	-.04	-3.3	-5.2
cohabitation mar.	127	19.02	.27	.10	6.3	2.3
divorcé(e)	119	15.82	.24	-.02	6.3	-.5
veuf/ve	215	9.05	-.22	.01	-7.6	.3
instruction						
sans diplôme	435	4.62	-.20	-.01	-9.3	-.5
cep	642	1.96	-.09	-.07	-6.4	-5.1
bepc	396	4.01	.13	.02	6.5	1.0
bacc	339	4.73	.18	.09	8.6	4.4
université	180	8.29	.08	.05	2.7	1.7
autres diplômes	8	476.82	-.50	-.02	-2.3	-.1
age en 3 classes						
moins de 30 ans	509	3.02	.08	.15	4.9	9.0
de 30 à 50 ans	705	1.80	.08	-.04	5.9	-2.7
plus de 50 ans	786	1.54	-.12	-.06	-10.2	-5.3

Il convient cependant d'effectuer au préalable une préparation du tableau lexical clairsemé. Certaines réponses ne contiennent aucune des formes sélectionnées au seuil de fréquence que l'on a fixé, et s'éliminent d'elles-mêmes ; d'autres ne contiennent qu'une seule de ces formes : l'analyse des correspondances a alors pour propriété de les isoler sur un axe qui aura une valeur propre voisine de 1, voire égale à 1 si cette forme apparaît exclusivement dans des réponses où elle est isolée. On restreint dans la pratique la typologie aux réponses ayant au moins deux formes sélectionnées.

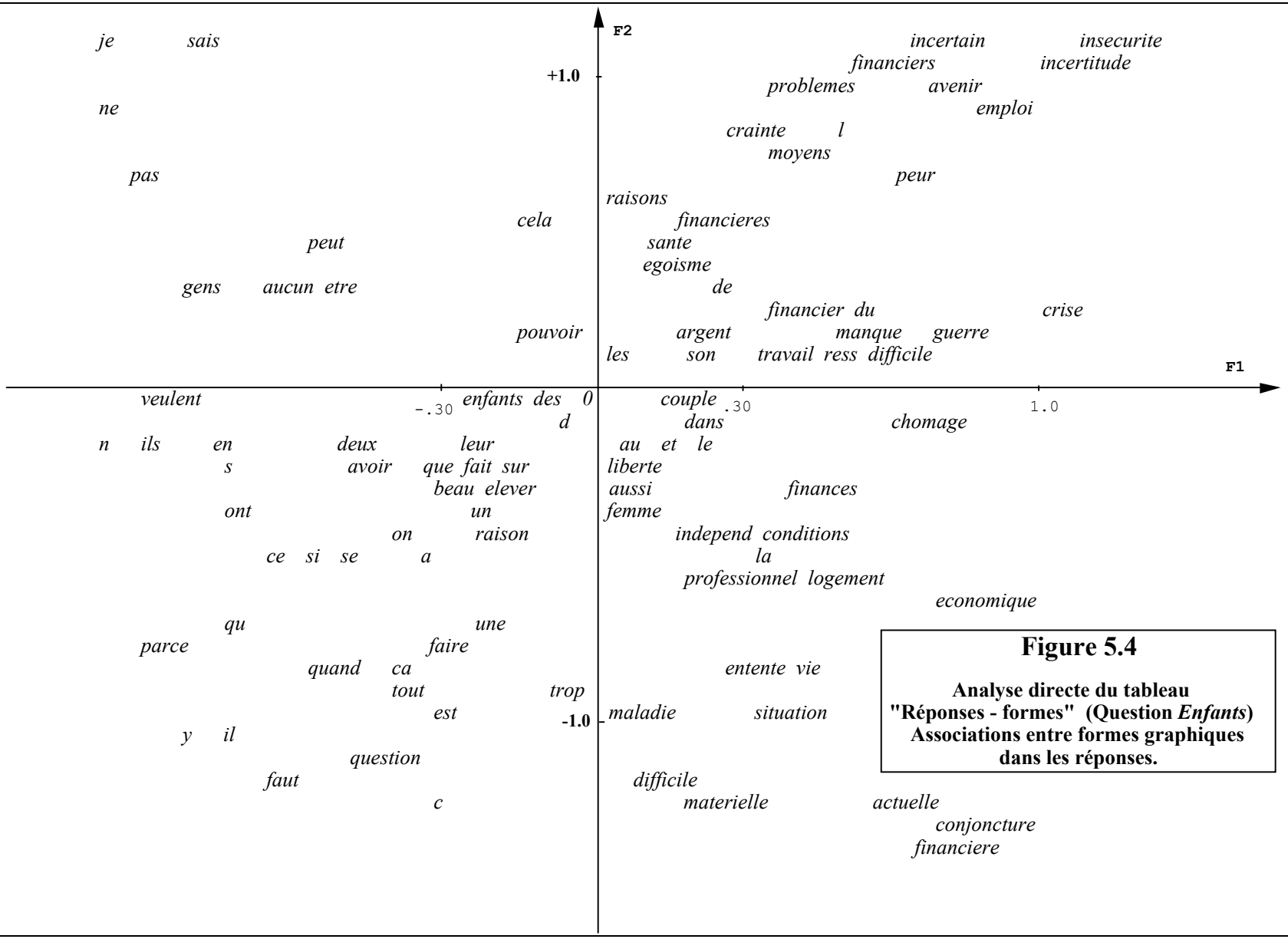

Figure 5.4

Analyse directe du tableau "Réponses - formes" (Question *Enfants*) Associations entre formes graphiques dans les réponses.

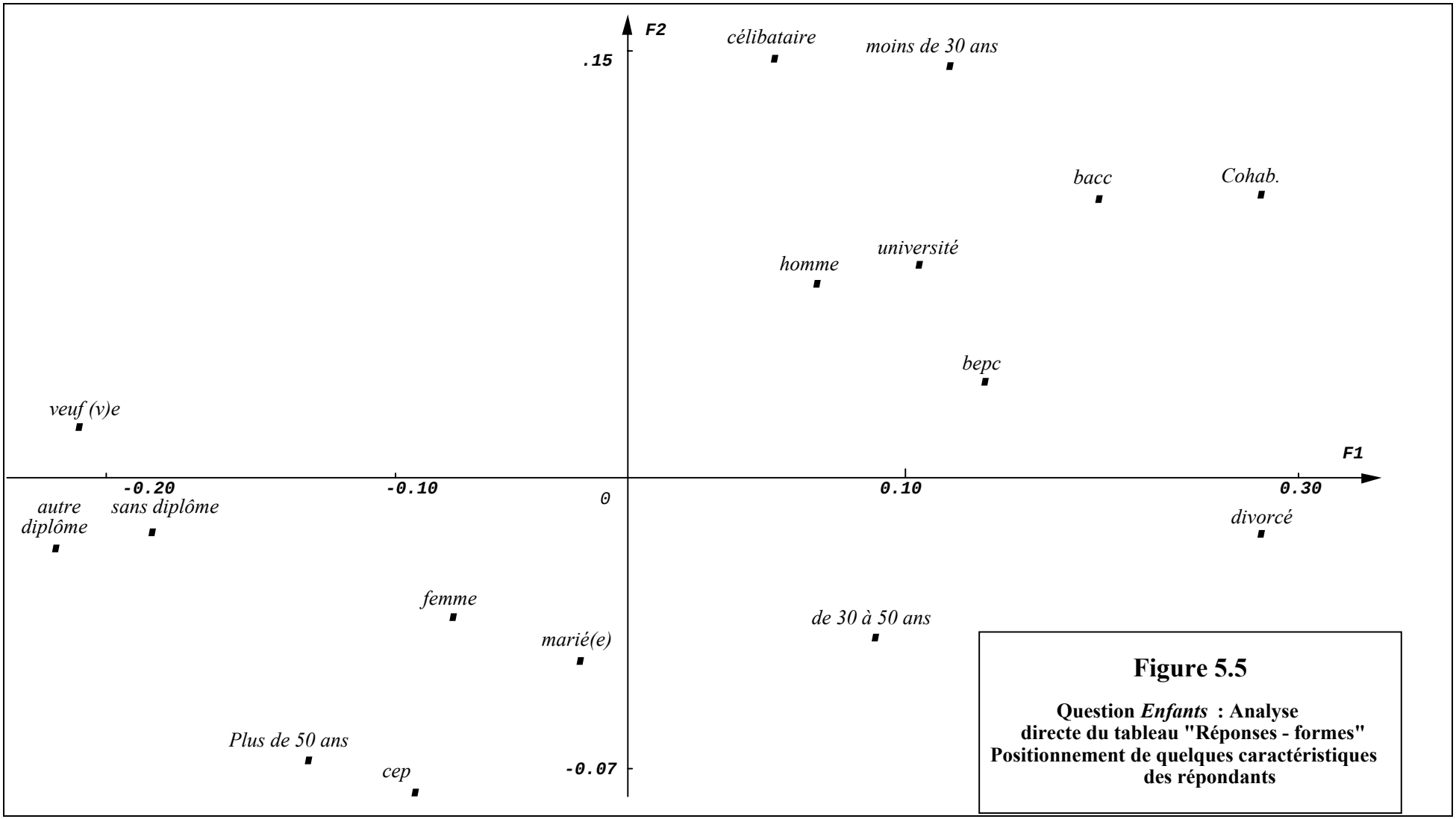

Figure 5.5

Question *Enfants* : Analyse directe du tableau "Réponses - formes" Positionnement de quelques caractéristiques des répondants

Lecture de la figure 5.4

Les proximités entre formes traduisent maintenant des cooccurrences de ces formes dans les réponses (et non plus dans des parties de textes ou dans des agrégats de réponses) ; on ne s'étonnera donc pas de voir grossièrement reconstitués certains éléments de phrases sur le graphique.

On lit ainsi sur la partie supérieure gauche de la figure 5.4 : *je ne sais pas,* ou encore, plus bas *ils ne veulent pas*, ou *ils veulent* et/ou peut-être *ils peuvent*.

On note en effet une certaine concentration de "formes à caractère grammatical" (appelées parfois mots-outil) dans la partie gauche du graphique, qui apparaît comme plus caractéristique des personnes âgées peu instruites, avec, comme conséquence de la structure démographique de la population interrogée, plus de femmes (figure 5.5). Cette concentration semble traduire des difficultés d'expression, des hésitations ou des réticences à répondre, plus fréquentes dans ces dernières catégories.

On note aussi qu'une partie des liaisons observées traduisent simplement le fait que les réponses sont soumises aux règles de la syntaxe, (accords entre substantifs et adjectifs au singuliers ou au pluriel, par exemple), et donc que les proximités entre formes ne dépendent pas seulement du contenu des réponses. Mais ceci n'est que partiellement vrai, car le tableau 5.4 et la figure 5.5 montrent une très forte polarisation socio-démographique de ce plan factoriel, et donc que, comme c'est souvent le cas lors de l'étude quantitative des matériaux textuels, la forme et le fond ne peuvent être facilement dissociés au niveau statistique.

En fait, dans l'analyse d'un tableau clairsemé comme le tableau **T,** les premiers axes factoriels n'épuisent qu'une faible part de l'information contenue dans le tableau de départ.

Il s'agit là d'un résultat général pour ce type de codage. Il faut cependant se garder d'interpréter avec pessimisme les faibles taux d'inertie obtenus sur les premiers axes car l'inertie totale, dans ce cas, ne constitue pas la seule information de référence. Elle contient un "bruit" incompressible (au sens de la théorie du signal), imputable au caractère clairsemé du tableau.

La décroissance des valeurs propres est très faible ($\lambda_1 = 0.418$, $\lambda_2 = 0.310$, $\lambda_3 = 0.291$, $\lambda_4 = 0.287$, etc.

Les pourcentages d'inertie correspondant valent $\tau_1 = 3.23$, $\tau_2 = 2.40$, $\tau_3 = 2.25$, $\tau_4 = 2.22$, ...). Les dix premiers axes ne restituent que 21% de l'inertie totale.

On n'obtient donc pas, dans ce cas, une synthèse visuelle, comme dans celui des tableaux agrégés (cas des figures 5.1, 5.3) mais un "épluchage" progressif de l'information.

On notera que l'échelle de la figure 5.5 est près de dix fois plus petite que celle de la figure 5.4 : autrement dit, portées sur la figure 5.4, ces caractéristiques seraient toutes concentrées près de l'origine des axes.

L'explication de ce phénomène est simple : les réponses du type *je ne sais pas*, par exemple, sont souvent le fait de personnes âgées, mais il s'en faut de beaucoup que toutes les réponses de ces personnes se réduisent à ces quelques formes. D'où une position beaucoup moins excentrée mais néanmoins très significative, du point qui représente cette catégorie.

Il est facile de sélectionner les modalités supplémentaires (ou illustratives) les plus significatives, au sens statistique de ce terme. Prenons l'exemple du graphique (non représenté ici, puisque les répondants sont anonymes) des 2 000 points-individus ou points-lignes sous-jacents à la figure 5.4. On sait que le point représentatif de la catégorie supplémentaire "personnes de plus de 50 ans" est le centre de gravité (dilaté) des points-individus concernés.

Les coordonnées de ce pseudo-centre de gravité figurent sur la dernière ligne du tableau 5.4. Pour plus de clarté, ces centres de gravité ont été portés sur une figure séparée, la figure 5.5. Dans l'hypothèse où ces individus pourraient être considérés comme tirés au hasard parmi les 2 000 personnes enquêtées (hypothèse qui dénoterait ici un vocabulaire indifférencié pour cette catégorie), ce centre de gravité partiel ne doit pas trop s'éloigner du centre de gravité du nuage (origine des axes factoriels).

On peut convertir cette distance au centre de gravité en "valeur-test", qui sera alors la réalisation d'une variable normale centrée réduite (deux dernières colonnes du tableau 5.4). Autrement dit, dans l'hypothèse d'un tirage au hasard, la valeur-test d'une catégorie supplémentaire a 95 chances sur 100 d'être comprise entre -1.96 et +1.96 . Comme on le lit sur le tableau 5.4, la valeur-test du point "plus de 50 ans" sur l'axe horizontal est de -10.2 ! C'est ce qu'il est convenu d'appeler une modalité hautement significative. Toutes les modalités ont d'ailleurs, pour ce qui concerne notre exemple, des valeurs-test significatives sur le premier axe, et presque toutes sur le second.

La consultation séquentielle, pour les couples d'axes factoriels successifs, des graphiques analogues des figures 5.4 et 5.5 permet donc de dégager systématiquement les formes ou groupes de formes choisis par les mêmes répondants, et d'identifier ces répondants par leurs caractéristiques. Les caractéristiques ainsi mises en évidence permettront ensuite de procéder à des regroupements et d'obtenir des visualisations plus synthétiques.

Une application directe des techniques de classification sera également très profitable à ce stade exploratoire. Elle peut alors constituer la première étape d'un *post-codage assisté*, qui consiste à regrouper les réponses similaires de façon à réduire les dimensions du problème.

Mais il sera indispensable de prendre alors en compte l'information de type segmentale à laquelle sera consacré le paragraphe 5.5.

5.4 Analyse des correspondances à partir d'une juxtaposition de tableaux lexicaux

Rappelons les différents points de vue adoptés jusqu'ici pour les analyses de réponses à des questions ouvertes ou plus généralement de textes courts et nombreux à propos desquels existe une information externe.

La première étape (paragraphe 5.1) a permis de procéder à un regroupement a priori, à partir d'une partition de l'ensemble des textes considérée comme privilégiée.

La seconde (paragraphe 5.2) suggère de construire une partition synthétique et polyvalente à partir des variables externes disponibles, en l'absence de variable privilégiée a priori.

La troisième (paragraphe 5.3), toujours à cause de l'absence d'un critère de regroupement privilégié, montre qu'une analyse directe des réponses, malgré le *bruit* considérable qu'elles contiennent, et le peu de signification au plan statistique des distances inter-individuelles, permet de mettre en évidence certains traits structuraux, et de sélectionner pour un éventuel regroupement ultérieur des variables externes qui leur sont probablement liées.

La quatrième étape, abordée dans le présent paragraphe, sera consacrée à l'analyse, non plus d'une table de contingence, mais d'une juxtaposition de tables de contingences. Cette juxtaposition est obtenue de la façon suivante : en lignes figurent toujours les unités statistiques de base (formes, segments, ou lemmes), en colonne figurent les partitions juxtaposées correspondant à différentes variables.

Il ne s'agit pas ici d'une partition de synthèse car il y a simplement juxtaposition et non croisement. Les distances entre formes sont donc des distances moyennes, pour lesquelles chacune des partitions a la même importance.[1]

[1] Cette stratégie d'analyse proposée par Benzécri, a été implémentée dans le logiciel SPAD (1984) (procédure TALEX), puis dans SPADT (1988). Son intérêt dans le cas ou

Il faut donc que ces partitions ne soient pas trop hétérogènes, pour que l'interprétation des proximités entre formes reste possible

Comme dans les sections précédentes 5.1 et 5.3, on présentera un exemple d'application relatif à la même question et la même enquête, de façon à permettre des comparaisons avec dans un cadre maintenant familier pour le lecteur. Le seuil de fréquence sera plus bas (14 au lieu de 20) ce qui donne 154 formes au lieu des 117 sélectionnées dans les tableaux 5.1 et 5.2. En effet, comme les catégories sont juxtaposées et non croisées, les effectifs des regroupements sont importants, et il est alors possible de travailler sur des fréquences plus faibles.

Les questions qui définiront les partitions à juxtaposer seront précisément celles qui figurent en lignes du tableau 5.4 : *genre, statut matrimonial, niveau d'instruction, âge en trois classes*. Bien entendu, il est possible de croiser certaines d'entre elles si l'on remarque lors d'une phase de dépouillement préalable que les interactions correspondantes présentent un intérêt particulier.

Tableau 5.5

Premières lignes de la juxtaposition de tables de contingence
***En ligne* : les 16 premières formes**
***En colonne* : 4 partitions de l'échantillon**

	Genre		Statut Matrimonial				Instruction /Diplômes						Age		
	H	F	CE	MA	CO	DI	VE	SA	CE	BE	BA	UN	AU	–30	–50	+50
a	99	119	39	134	12	9	24	46	63	43	46	19	1	56	73	89
actuelle	32	51	15	54	2	5	7	15	30	22	14	2	0	20	31	32
argent	95	101	45	103	11	15	22	51	64	33	29	17	2	60	66	70
au	19	25	11	26	4	1	2	8	13	5	11	7	0	9	16	19
aucune	20	13	2	26	1	0	4	8	11	11	2	1	0	6	9	18
aussi	12	10	4	13	4	0	1	2	8	6	4	2	0	6	11	5
avec	12	9	3	16	0	0	2	1	11	3	1	5	0	6	9	6
avenir	154	164	66	192	19	23	18	53	90	78	75	22	0	115	117	86
avoir	27	50	17	50	1	0	9	12	33	14	12	6	0	28	15	34
beaucoup	9	14	3	17	0	0	3	5	13	3	0	2	0	2	11	10
bien	7	14	4	11	1	1	4	5	8	4	4	0	0	7	7	7
c	44	54	15	61	2	6	14	23	35	17	14	8	1	24	34	40
ca	23	18	7	24	3	3	4	8	15	4	11	3	0	8	17	16
ce	13	21	4	22	2	2	4	3	16	6	3	6	0	8	14	12
cela	11	12	9	14	0	0	0	2	5	5	4	7	0	11	7	5
chomage	125	160	41	184	14	16	30	72	111	50	40	11	1	79	88	118
.....																

les partitions sont constituées par de nombreuses questions fermées a été souligné par Cibois (1990).

Le tableau 5.5 est un extrait (16 premières lignes sur 154) de la table qui sera soumise à l'analyse des correspondances. Chaque occurrence d'une forme apparaît quatre fois (une fois dans chaque sous-tableau). On voit que ce type d'analyse peut être intéressant lorsque l'on a affaire à des petits échantillons, car les profils-colonnes restent fondés sur des comptages importants.

Enfin, la Figure 5.6 donne le plan principal de visualisation obtenu par l'analyse de ce tableau composite. Le premier axe correspond à 34% de l'inertie, le second à 13%.

Lecture de la figure 5.6

Les résultats présentés ici contiennent moins de détails que les résultats correspondants représentés par la figure 5.1, en ce qui concerne les effets séparés de l'âge et du niveau d'instruction, mais le graphique est cependant très riche : la présence du statut matrimonial et du genre (sexe) est une information intéressante.

Le fait que le niveau d'instruction soit plus détaillé est également appréciable, car les diplômes *Bacc* et *Université*, regroupés précédemment, sont distants sur la figure 5.6, comme d'ailleurs les rubriques *Cep* et *Sans diplôme*.

On remarque par exemple que les cinq formes construites à partir du vocable *finance* attestées dans le corpus (finances, financier, financiers, financières, financière) se regroupent dans le même quadrant, avec la constellation (bacc, célibataire, cohabitation, moins de 30 ans, divorcé(e)), mais également avec les couples singulier/pluriel (*problème, problèmes* et *responsabilité, responsabilités*). Ces formes étaient plus dispersées précédemment.

On retrouve l'expression particulière des personnes âgées et peu instruites, et la plupart des grandes oppositions observables lors des visualisations précédentes.

On peut aussi noter la stabilité de certains traits au travers de différentes enquêtes (ici par exemple le fait que l'emploi de la forme *enfant* au singulier correspond à un niveau d'instruction plus élevé en moyenne que l'emploi du pluriel *enfants*).

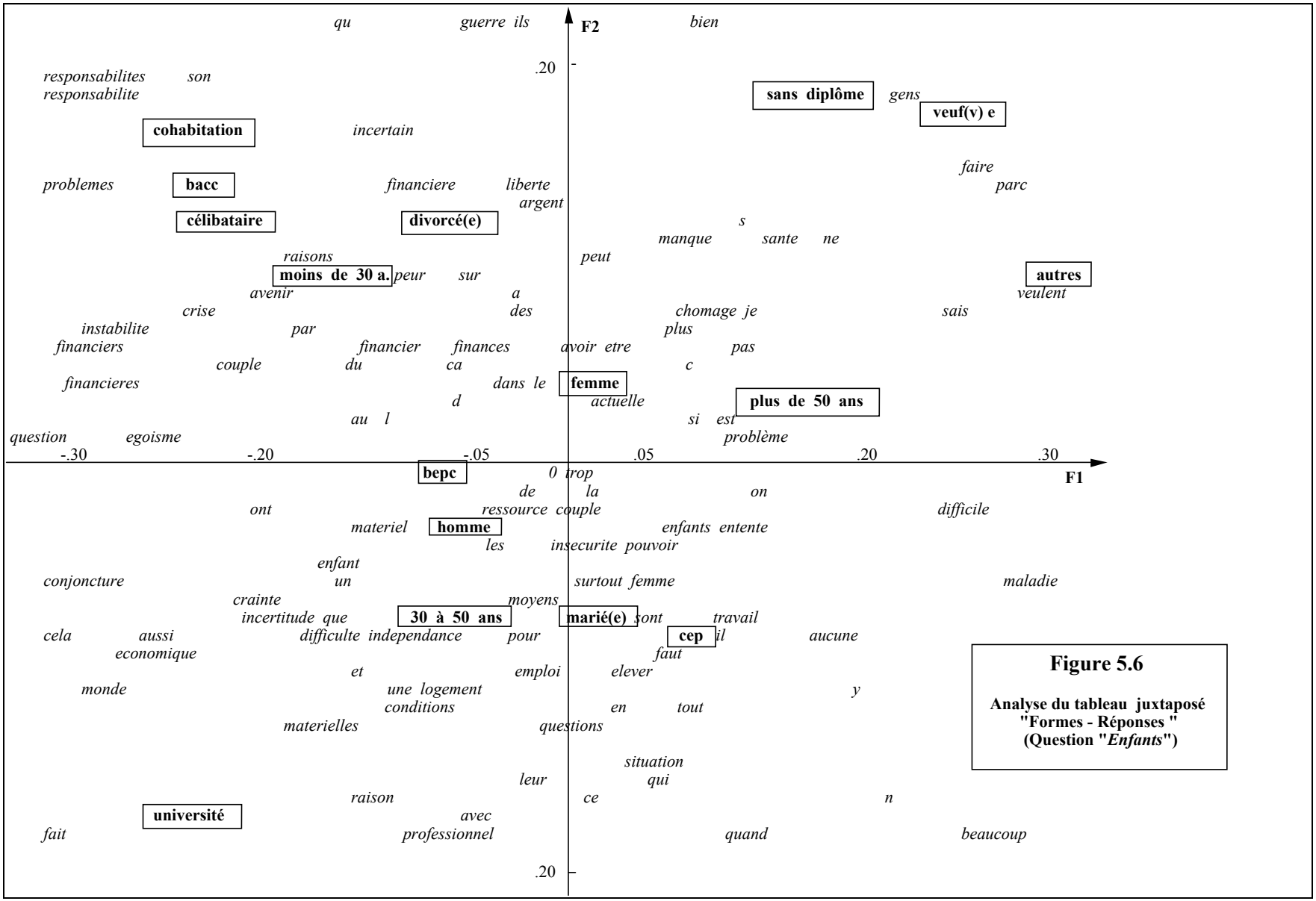

Figure 5.6

Analyse du tableau juxtaposé "Formes - Réponses " (Question "*Enfants*")

5.5 Analyse des correspondances à partir du tableau des segments répétés.

Cette section recoupe en fait les quatre sections précédentes ; les opérations effectuées à partir des formes pouvant être généralisées aux segments. Il est possible d'utiliser directement les données segmentales dans le cas d'une partition privilégiée, qui a fait l'objet du paragraphe 5.1, mais aussi dans le cas d'une partition obtenue à partir des noyaux factuels. (paragraphe 5.2).
Dans le cas d'une analyse directe des réponses, cela reste possible si les réponses sont longues et riches lexicalement (paragraphe 5.3). Enfin, cela est également possible dans le cas de juxtaposition de tableaux de contingence (paragraphe 5.4).

On a choisi de présenter l'utilisation de segments dans les phases de visualisation réalisées à partir du premier exemple traité dans la section 5.1. On reprendra donc la même question ouverte, et la même partition en 9 classes croisant les variables Age et Niveau de diplôme.

Les tableaux segmentaux sont en général assez volumineux. A titre d'exemple, on notera que pour les 2 000 réponses à la question ouverte de référence ***Enfants***, alors que 249 formes graphiques ont plus de 6 occurrences, on relève 330 segments de longueur 2 apparaissant au moins 6 fois, 133 segments de longueur 3, 39 de longueur 4, 18 de longueur 5, un seul de longueur 6 (tous ces segments étant assujettis à apparaître au moins 6 fois). On a donc un tableau de 523 segments distincts dont l'effectif total est égal à 8 670.

Comme les formes, les segments sont évidemment des éléments de description des réponses, et plus généralement, des parties de corpus. Mais la notion de profil segmental n'est pas aussi pure que celle, largement utilisée aux chapitres précédents, de profil lexical. Un tableau lexical entier comme le tableau 5.3 est en effet homogène parce que doublement additif : pour une ligne, la somme des termes est la fréquence globale de la forme, pour une colonne, la somme est la longueur de la partie correspondante du corpus.

La somme des colonnes d'un tableau analogue dont les lignes seraient des segments n'a plus une signification aussi naturelle : les segments ne sont pas indépendants et constituent des éléments d'information largement redondants. Doit-on utiliser des seuils de fréquences différents selon la longueur des segments? Doit-on utiliser simultanément les segments et les formes?

Toutes ces interrogations sont justifiées, mais l'expérience concrète des traitements statistiques d'informations segmentales montre en fait une grande robustesse des résultats vis-à-vis de ces choix. En bref, on ne

bouleverse que rarement le système des distances entre catégories ou parties en complétant les profils lexicaux des catégories ou parties par leurs profils segmentaux. Mais la présence des segments sur les graphiques et les listages va faciliter grandement leurs interprétations en contribuant à lever l'équivoque du contexte immédiat.

Un exemple

Toujours pour la question ***Enfants***, on soumet à l'analyse des correspondances le tableau à 523 lignes et 9 colonnes croisant l'ensemble des 523 segments apparaissant plus de 6 fois avec les 9 modalités de la variable "âge-éducation".
On retrouve alors dans le plan des deux premiers facteurs le réseau régulier de proximités entre les 9 catégories déjà observé sur la figure 5.1, obtenue directement à partir des formes lexicales. Pour alléger la présentation, on publiera ci-dessous une analyse portant sur un extrait des 523 segments précédents obtenu de la façon (automatique...) suivante : on retient les segments de longueur 2 apparaissant plus de 50 fois, ceux de longueur 3 ou plus s'ils apparaissent plus de 6 fois.
On obtient 111 segments, qui suffiront eux-aussi, comme les formes lexicales originales, et comme l'intégralité des 423 segments, à reconstituer le réseau régulier "âge-éducation" dans le plan des deux premiers facteurs.

Le tableau 5.6 donne la liste intégrale de ces 111 segments avec leurs fréquences. Le tableau **T** soumis à l'analyse compte donc 111 lignes et 9 colonnes. L'effectif total du tableau ainsi constitué s'élève cette fois à 2 647 occurrences de segments. Rappelons que les segments participent cette fois à l'analyse en qualité d'éléments actifs.

La figure 5.7 représente le plan principal de visualisation issu de l' analyse des correspondances de la table de contingence d'ordre (111 x 9) croisant en ligne les 111 segments du tableau 5.6 avec les neuf catégories de la variable croisée Sexe-Age utilisée au paragraphe 5.1. Les deux premières valeurs propres valent respectivement 0.13 et 0.07, et correspondent respectivement à 33% et 18% de la trace. Pour les deux premiers facteurs, les pourcentages d'inertie sont donc très semblables aux pourcentages obtenus sur les formes simples[1]. La figure 5.7 qui représente les deux premiers axes factoriels de l'analyse de **T** rappelle évidemment la figure 5.1, à ceci près que la structure est moins régulière, et a subi une légère rotation.

[1] Les valeurs propres sont plus grandes, ce qui était prévisible, car la table de contingence segmentale est beaucoup moins pleine (2 647 occurrences au lieu de 12 051 pour la table de contingence 5.3 qui concerne les occurrences de formes). Le coefficient 12 051/2 647 = 4.55 est approximativement un facteur d'échelle entre les deux ensembles de valeurs propres.

Tableau 5.6

Inventaire partiel des segments répétés *(Question "Enfants")*
Seuil pour les segments de longueur 2 : 50 ; de longueur 3 ou plus : 6

Freq	Long.	Segment	Freq	Long.	Segment
8	3	a pas de	11	3	le chomage le
7	3	avec un enfant	9	3	le chomage les
17	3	conditions de vie	9	3	le fait de
8	3	crainte de l'avenir	31	3	le manque d'argent
10	3	d'avoir un enfant	35	3	le manque de
7	3	dans la famille	10	3	le travail de
12	3	dans le couple	7	4	le manque de moyens
113	2	de l'avenir	12	4	le manque de travail
117	2	de la	7	5	le fait de ne pas
53	2	de travail	9	5	le travail de la femme
9	3	de l'avenir pour	59	2	les enfants
21	3	de la femme	19	3	les conditions de
7	3	de la situation	9	3	les conditions materielles
49	3	de la vie	10	3	les difficultes de
26	3	de ne pas	18	3	les difficultes financieres
8	3	de vie actuelle	15	3	les moyens financiers
9	4	de ne pas pouvoir	7	3	les problemes d'argent
8	4	de plus en plus	9	3	les problemes financiers
55	2	des enfants	12	3	les raisons financieres
7	3	des raisons de	15	4	les conditions de vie
8	3	des raisons financieres	8	5	les difficultes de la vie
11	4	difficultes de la vie	53	2	manque d'argent
11	3	il n'y a	88	2	manque de
22	3	il y a	8	3	manque de liberte
10	4	il n'y a pas	8	3	manque de moyens
7	4	il y a des	13	3	manque de ressources
7	4	il y en a	23	3	manque de travail
8	4	je n'en vois pas	11	3	n'y a pas
20	4	je ne sais pas	7	4	n'y a pas de
17	3	l'avenir de l'enfant	12	3	ne pas avoir
12	3	l'avenir des enfants	13	3	ne pas pouvoir
7	3	l'avenir le chomage	9	3	ne sont pas
13	3	l'insecurite de l'emploi	14	3	ne veulent pas
51	2	la femme	7	3	pas de travail
79	2	la peur	7	3	pas les enfants
87	2	la situation	103	2	peur de
127	2	la vie	65	3	peur de l'avenir
11	3	la conjoncture actuelle	8	3	peur de la
8	3	la crainte de	15	3	peur de ne
9	3	la femme travaille	14	3	peur du chomage
44	3	la peur de	7	4	peur de l'avenir pour
10	3	la peur des	12	4	peur de ne pas
15	3	la peur du	8	5	peur de ne pas pouvoir
8	3	la situation actuelle	22	3	pour les enfants
11	3	la situation economique	51	2	problemes financiers
22	3	la situation financiere	8	3	qui n'est pas
7	3	la situation materielle	7	3	raison de sante
23	3	la vie actuelle	93	2	raisons financieres
7	3	la vie chere	15	3	raisons de sante
17	3	la vie est	9	3	raisons financieres et
32	4	la peur de l'avenir	12	4	travail de la femme
7	4	la peur des responsabilites	52	2	un enfant
152	2	le chomage	7	3	une question de

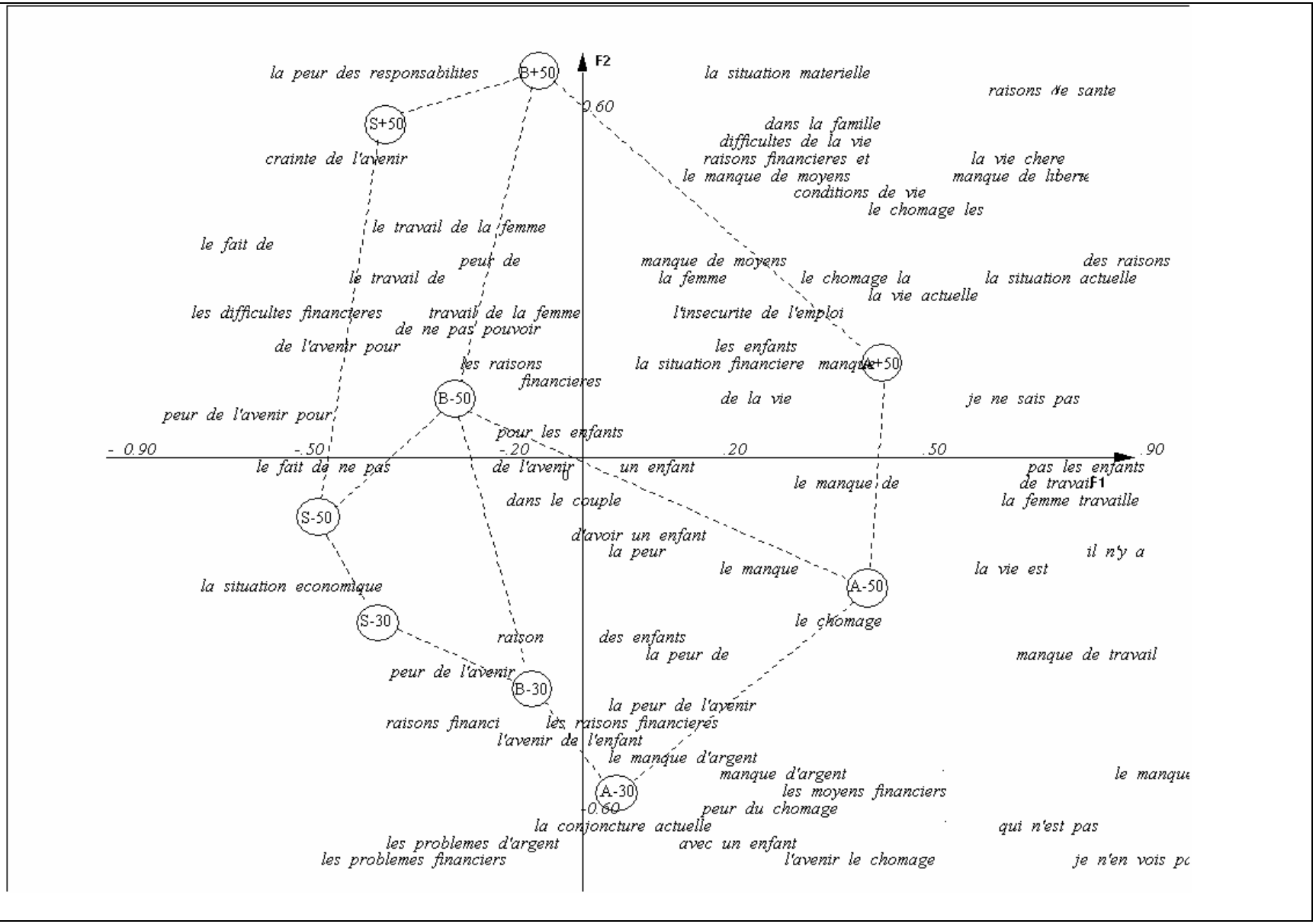

Figure 5.7 : Analyse des correspondances du tableau segmental - Question Enfants. Plan (f_1,f_2)

Rappelons que les coordonnées d'un point-catégorie sont les composantes de son profil segmental. Il n'était donc pas du tout évident a priori de retrouver simultanément l'ordre des classes d'âge et l'ordre des niveaux de diplôme. Il y a donc variation continue des segments utilisés avec l'âge (à niveau de diplôme constant) et avec le niveau de diplôme (à âge constant).

Les positions des points-segments vont permettre de compléter les interprétations que suscitaient les positions des points-formes de la figure 5.1.

La figure 5.7 suscite quelques remarques spécifiques:

a) La première de ces remarques porte sur la similitude, en ce qui concerne les positions des points-catégories, des figures 5.1 et 5.7. Il eût été équivalent, dans ce cas particulier caractérisé par une grande stabilité du "pattern" des catégories, de positionner les points-segments en éléments illustratifs sur la figure 5.1.

b) Le fonctionnement des segments répétés comme "sélecteur automatique de contexte" éclaire incontestablement certains points laissés obscurs par l'analyse directe des formes graphiques.
Examinons par exemple la position de la forme *manque* sur la figure 5.1, près des points *A-50* (aucun diplôme, entre 30 et 50 ans) et *A+50* (aucun diplôme, plus de 50 ans). Il lui correspond les segments *le manque, le manque de, manque de* qui occupent des positions au voisinage des points homologues *A-50* et *A+50* sur la figure 5.7.
Mais les expansions de ces segments (segments plus longs qui les contiennent) sont beaucoup plus dispersées sur cette figure : *le manque de travail* , en bas tout-à-fait à droite, caractérise donc les personnes sans diplôme, *le manque d'argent*, sur l'axe vertical, près du point *A-30* , caractérise plutôt les jeunes, *le manque de moyens*, beaucoup plus haut, entre les points *A+50*. et *B+50* appartient plutôt aux réponses des plus âgés.
Il en va de même pour la forme *peur*, qui est seulement *peur du chômage* chez les plus jeunes peu instruits (*A-30*), et *peur des responsabilités* chez les plus âgés et plus instruits, entre les points *S+50* et *B+50*. La forme synonyme *crainte* n'apparaît, en concurrence avec *peur*, que chez les répondants plus instruits.

c) Comme cela a déjà été signalé plusieurs fois, le contenu des réponses est fortement lié aux termes de l'expression... au point que l'on peut penser que la forme joue tour à tour les rôles d'aide ou d'obstacle dans le déchiffrage de l'information. Quelquefois, une idée qui semble être la même au prime abord est exprimée de façons différentes, avec

pourtant des mots voisins. Ainsi, ce sont surtout les personnes plus âgées qui mentionnent l'activité féminine comme une des raisons pouvant faire hésiter à avoir un enfant. Mais parmi ces personnes, les plus instruites disent : *le travail de la femme* (sous le point *S+50*), alors que les moins instruites disent simplement : *la femme travaille* (entre les points *A+50* et *A-50*). Ces différences de formes (qui traduisent des différences de lieux d'expression, de lieux d'observation), stimulantes dans toutes phases de recherche ou d'exploration, appartiennent à ce que l'on peut appeler l'aspect sociolinguistique du matériau analysé[1]. Il s'agit de nuances qui disparaissent dans une opération de post-codage, ou qui peuvent être masquées par une lemmatisation. Les personnes plus éduquées nominalisent plus facilement et désignent plutôt le phénomène que l'action correspondante[2]. Mais ces différences ne se limitent peut-être pas à la forme, et c'est dans ce sens que ces matériaux bruts, ni codés ni transformés, peuvent stimuler le sociologue : les deux segments renvoient-ils aux mêmes activités ?

d) Une quatrième remarque porte sur l'itinéraire parcouru et sur l'outil dont on dispose en fin de course. A partir d'une saisie aveugle et automatique du texte des réponses et des caractéristiques des répondants, on est en mesure d'obtenir, sans intervention manuelle ni interprétative, des représentations du type de celle présentée à la figure 5.7, plus suggestive que celle de la figure 5.1, elle-même beaucoup plus suggestive que les tableaux de fréquences.
Le seul choix important qui a été effectué est celui de la variable de regroupement (ici la variable "âge-diplôme"), mais il ne s'agit évidemment pas d'un choix unique et définitif, puisque l'analyse peut être refaite pour n'importe quel autre critère. Toutes les remarques faites à ce sujet à propos des analyses de formes s'appliquent. En particulier, une partition en noyaux factuels préliminaire peut aider à trouver le critère ou la combinaison de critères ad hoc.

e) Enfin, il faut insister sur le caractère complémentaire des analyses segmentales : la figure 5.7, volontairement limitée aux seuls segments, ne se substitue pas à la figure 5.1, elle l'enrichit.
Prenons l'exemple de la forme *crise ,* qui apparaît 25 fois dans le corpus (cf. tableau 5.1). Le segment le plus fréquent qui contient cette forme (*la crise*) n'apparaît que 19 fois ; chacun des segments plus

[1] A propos des variations des modes d'expression dans les diverses couches sociales, et plus généralement de la sociologie du langage, on pourra consulter, par exemple, la synthèse de Achard (1993).
[2] Cf. aussi les travaux de Somers (1966), que l'on évoquera à nouveau au chapitre 8.

longs, (*la crise économique, la crise de société, la crise actuelle*), a une fréquence d'apparition inférieure à notre seuil choisi égal à 6, ce qui a pour effet de faire disparaître cette forme de la figure 5.7, ce qui est ici dommageable car les contenus des segments restent proches de ceux de la forme isolée.

Parmi les solutions possibles à ce type de problème, on peut proposer une analyse simultanée des formes et des segments, ou bien adopter une approche dans laquelle soit les formes, soit les segments figurent en éléments supplémentaires.