Chapitre 6

Eléments caractéristiques, réponses ou textes modaux

Les typologies et visualisations du chapitre précédent donnent des panoramas globaux des tableaux lexicaux, que ceux-ci soient agrégés ou non, qu'ils correspondent aux parties naturelles d'un corpus (chapitres, oeuvres d'un même auteur, articles, discours datés, etc.) ou qu'ils soient constitués à partir de regroupements artificiels (regroupement de réponses libres par catégories, de documents par thèmes, etc.).

Les représentations spatiales fournies par l'analyse des correspondances gagnent à être complétées par quelques paramètres d'inspiration plus probabiliste : les *spécificités* ou *formes caractéristiques*, définies et illustrées dans ce chapitre. Il existe en effet des outils statistiques qui permettent de décrire chacune des classes d'une partition en exhibant tout à la fois les unités qu'elle contient en grand nombre par rapport aux autres classes et, au contraire, les unités qu'elle contient en très petit nombre. Dans le cas des unités de base, qu'il s'agisse de formes, de segments ou de lemmes, on parlera d'éléments caractéristiques ou de spécificités (6.1).

De façon plus générale, et inévitablement plus complexe, on peut aussi mettre en évidence, pour chacune des parties du corpus, des associations caractéristiques composées de plusieurs unités. Ainsi, des regroupements de réponses peuvent être caractérisés par quelques réponses dites caractéristiques (ou encore *réponses modales*). Ces réponses, brièvement, contiennent un maximum de formes spécifiques d'un recueil particulier...

Elles permettent de replacer les formes graphiques dans leur contexte et de résumer chaque recueil. Dans le cas de recherche documentaire, il peut s'agir de documents modaux (caractérisant un thème particulier) et on parlera alors de réponses modales, qu'il s'agisse effectivement de réponses à des questions ouvertes, ou de documents (6.2). Dans le cas de textes plus longs, on pourra

mettre en évidence des unités de contexte de longueur variable (phrases, paragraphes, etc.).[1]

6.1 Formes caractéristiques, spécificités.

L'analyse des correspondances crée une typologie portant à la fois sur l'ensemble des parties du corpus et sur l'ensemble des formes attestées dans celui-ci. Cependant, il peut être utile de compléter ces analyses globales par des calculs probabilistes effectués séparément à partir de chacune des sous-fréquence du tableau lexical.

Différentes méthodes de calcul permettent d'aboutir à ce type de paramètres. Nous avons choisi d'exposer ci-dessous le calcul des éléments caractéristiques ou spécificités, adaptation aux matériaux textuels de tests statistiques classiques[2].

6.1.1 Le calcul des spécificités

Nous noterons, exprimées en nombre d'occurrences de formes simples, les quantités :

- k_{ij} - sous-fréquence de la forme numéro i dans la partie j du corpus;
- $k_{i.}$ - fréquence de la forme numéro i dans l'ensemble du corpus;
- $k_{.j}$ - longueur de la partie numéro j;
- $k_{..}$ - longueur totale du corpus (ou simplement k).

A partir des trois derniers nombres contenus dans le tableau ci-dessus, le calcul de spécificités permet de porter une appréciation sur la sous-fréquence k_{ij} sans recourir à la notion de proportionnalité dont on a expliqué qu'elle était, dans la plupart des cas, mal adaptée à l'étude quantitative des textes, surtout lorsque ceux-ci sont de longueurs très différentes.

Le modèle probabiliste

On commence par imaginer une *population* d'objets d'effectif total k. Parmi tous ces objets, on en suppose $k_{i.}$ portant une "marque" qui les distingue des autres : boules de couleur particulière ou, dans le cas qui nous intéresse,

[1] On trouvera dans Salem (1993) des applications de ce dernier type.
[2] Cf. Lafon (1980) pour un exposé et d'autres exemples sur la méthode des spécificités.

objets correspondant aux occurrences d'une même forme graphique de fréquence totale $k_{i.}$. Les objets restants de notre population sont tous confondus en un même sous-ensemble et considérés comme "non-marqués". Le nombre des objets non marqués est donc égal à $k - k_{i.}$.
Prélevons maintenant, en pratiquant des tirages aléatoires sans remise, un *échantillon* contenant exactement $k_{.j}$ objets dans notre population. A l'issue de ce tirage, le nombre des objets marqués que contient l'échantillon est noté k_{ij}.

Les nombres k, $k_{i.}$, $k_{.j}$ définis plus haut constituent les paramètres du modèle.

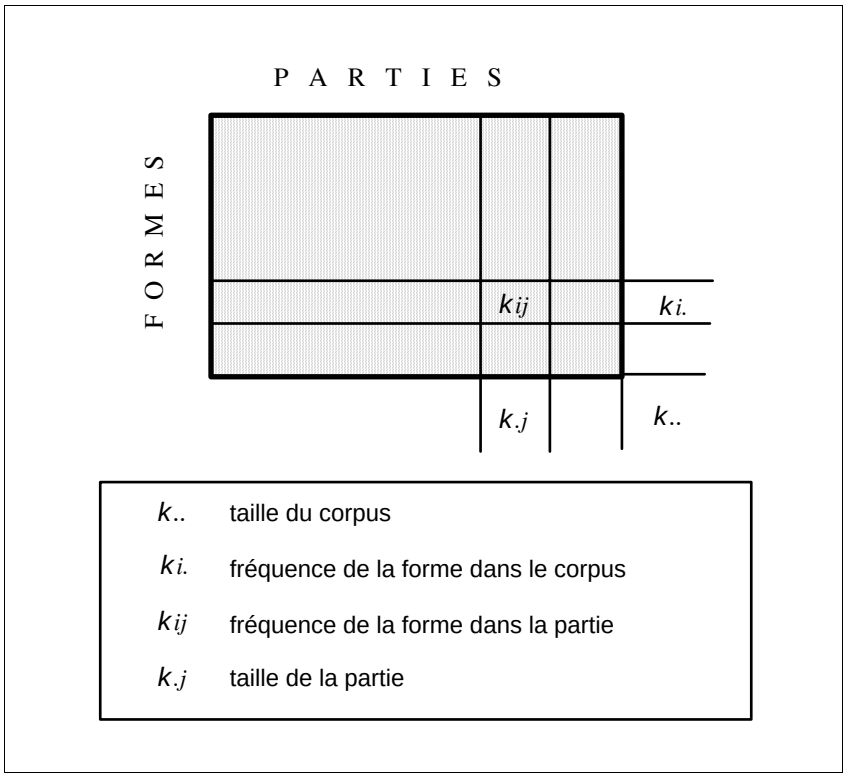

Figure 6.1
Les 4 paramètres du calcul des spécificités

Pour porter un jugement sur le résultat k_{ij}, il nous faut le situer parmi des comptages de même nature qui correspondent à l'ensemble de tous les échantillons, composés de $k_{.j}$ objets, qu'il est possible de prélever à partir de la population de départ.

Pour chaque échantillon de taille $k_{.j}$, le nombre k_{ij} des objets marqués peut prendre une valeur obligatoirement comprise entre 0 et $k_{i.}$, nombre total des objets marqués. Inversement, toujours pour des nombres k, $k_{.j}$ et $k_{i.}$ fixés, et pour chaque nombre n compris entre 0 et $k_{i.}$, il est possible de recenser le nombre $N(n)$ des échantillons de longueur $k_{.j}$ pour lesquels k_{ij} est strictement égal à n.
Si l'on divise chacun des nombres $N(n)$ par le nombre total des échantillons de longueur $k_{.j}$, on obtient une distribution de probabilité (de paramètres : k, $k_{i.}$, $k_{.j}$) sur l'ensemble des nombres compris entre 0 et $k_{i.}$. La somme de ces nombres est égale à l'unité. La loi de probabilité correspondant à un tirage sans remise dans l'hypothèse d'indépendance est la loi hypergéométrique. Cette loi est voisine de la loi multinomiale lorsque les prélèvements sont petits par rapport à la population (on peut alors assimiler tirages sans remise et tirages avec remise).

Nous noterons :

$$Prob\ (k, k_{i.}, k_{.j}, n\ )$$

la probabilité ainsi calculée de voir apparaître exactement n objets marqués lors d'un tirage sans remise d'un échantillon de longueur $k_{.j}$ parmi une population d'effectif total k, sachant que l'ensemble de la population compte $k_{i.}$ objets marqués.

On peut voir figure 6.2 l'exemple d'une telle distribution calculée pour les paramètres :

k	=	160 000	taille du corpus.
$k_{.j}$	=	20 000	taille de la partie j
$k_{i.}$	=	36	fréquence de la forme i.

Comme on le voit, le *mode* de cette distribution (valeur la plus probable) est égal à 4. Les probabilités décroissent rapidement à mesure que l'on s'éloigne de cette fréquence.

Nous pouvons maintenant utiliser la distribution de probabilité ainsi construite à partir des paramètres k, $k_{.j}$ et $k_{i.}$ pour porter un jugement sur la fréquence absolue n_0 observée lors du tirage de notre échantillon. Pour cela nous commencerons par situer n_0 par rapport au mode de la distribution.

Si cette valeur est très proche du mode, nous ne pourrons pas dire grand chose à propos du résultat observé.

Si en revanche elle lui est nettement supérieure, nous calculerons la quantité $P_{sup}(n_0)$ qui est la probabilité de voir apparaître, toujours sous les

hypothèses retenues plus haut, un nombre d'objets marqués égal ou supérieur à n_0 parmi les $k_{.j}$ objets prélevés au hasard.

La probabilité $P_{sup}(n_0)$ est égale à la somme de toutes les probabilités correspondant à des valeurs de n égales ou supérieures à n_0. C'est donc la somme des probabilités $Prob\ (k, k_{i.}, k_{.j}, n\ )$ pour les valeurs de n comprises entre n_0 et $k_{i.}$.
Si n_0 est inférieur au mode, nous calculerons de la même manière la quantité $P_{inf}(n_0)$ qui est la probabilité de voir apparaître, toujours sous les mêmes hypothèses, un nombre d'objets marqués égal n ou inférieur à n_0.

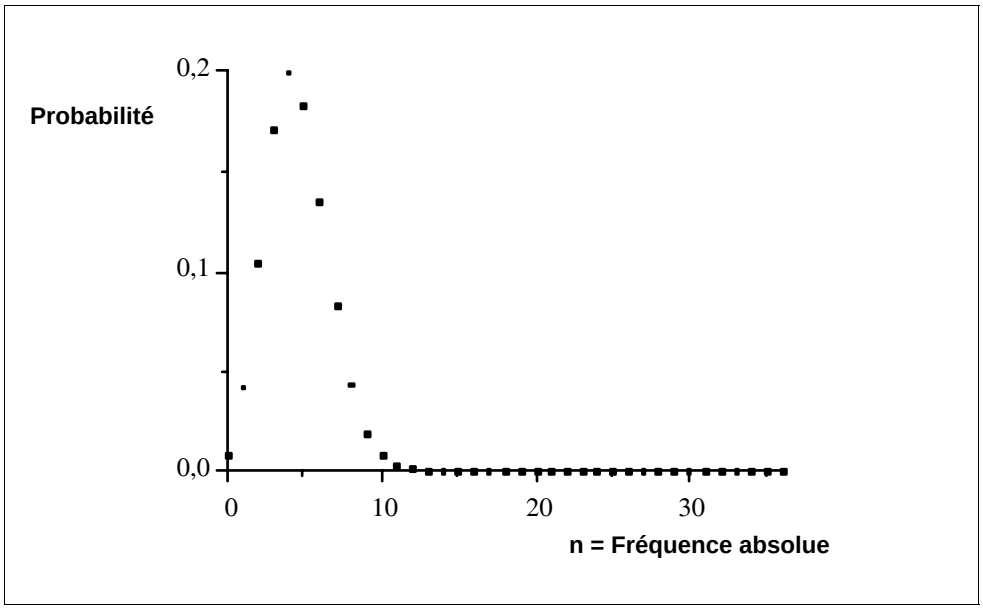

Figure 6.2

Distribution des probabilités hypergéométriques pour les paramètres k =160 000, $k_{.j}$ = 20 000, $k_{i.}$ = 36.

Cette dernière probabilité est égale, quant à elle, à la somme des probabilités $Prob\ (k, k_{i.}, k_{.j}, n\ )$ pour les valeurs de k comprises entre 0 et n_0.

Si la probabilité $P_{inf}\ (n_0)$ est très faible pour un échantillon, on en conclura que cet échantillon compte, par rapport à l'ensemble des échantillons de même type, "anormalement" peu d'objets marqués.

Inversement, si la probabilité $P_{sup}\ (n_0)$ est très faible pour cet échantillon, on conclura que l'échantillon compte un nombre "anormalement" élevé d'objets marqués.

Pratique du calcul des spécificités

Pour chacune des cases du tableau, contenant le nombre k_{ij} des occurrences de la forme numéro i dans la partie numéro j, nous allons apprécier cette sous-fréquence par rapport aux nombres k, $k_{i.}$, $k_{.j}$, en utilisant le modèle décrit plus haut.
Pour sélectionner les probabilités que l'on considérera comme faibles, et par conséquent dignes d'attention, on fixe arbitrairement au début de l'expérience un *seuil* de probabilité qui ne changera pas au cours du calcul sur l'ensemble des sous-fréquences contenues dans le tableau. On note cette valeur par la lettre ***s***.

On effectue ensuite le même type de calcul pour chacune des sous-fréquences k_{ij}. Les paramètres du modèle étant fixés, on peut alors construire la distribution de probabilités sur les nombres compris entre 0 et $k_{i.}$. En fonction de la position de k_{ij} par rapport au mode de la distribution, on calcule ensuite selon la méthode décrite plus haut, les probabilités : $P_{inf}(k_{ij})$ ou $P_{sup}(k_{ij})$.

Formes spécifiques négatives, banales, spécifiques (caractéristiques)

Si la première de ces probabilités se révèle plus faible que le seuil retenu, on en conclura que la sous-fréquence observée k_{ij} est relativement faible, c'est-à-dire que la forme i est plutôt mal représentée dans la partie j par rapport à ce que le modèle hypergéométrique laissait prévoir. Dans ce cas on dira que la forme numéro i est *spécifique négative* pour la partie numéro j, ce qu'on notera *S*-.

Si c'est, au contraire, la seconde de ces probabilités qui se révèle en dessous du seuil ***s***, on en conclura que la sous-fréquence k_{ij} est relativement élevée, c'est-à-dire que la forme i est plutôt abondante dans la partie j. On dira que la forme i est une forme *spécifique positive* pour la partie considérée. On notera *S*+ cette éventualité. On parle aussi dans ce cas de *forme caractéristique.*

Il se peut également qu'aucune des deux probabilités $P_{sup}(k_{ij})$ et $P_{inf}(k_{ij})$ ne se révèle inférieure au seuil ***s*.** On dira dans ce cas que la forme i est *banale* pour la partie j. Les formes banales pour une partie sont signalées par la lettre *b*.

Si une forme ne présente aucune spécificité (i.e. si elle est banale pour chacune des parties du corpus), on dira que cette forme appartient au *vocabulaire de base* du corpus.

Les diagnostics ainsi obtenus par la méthode des spécificités sur chacune des sous-fréquences figurant dans le tableau analysé peuvent être ensuite réordonnés pour permettre différents types de lecture.

La première de ces lectures s'effectue à partir de listes dans lesquelles les diagnostics de spécificité sont classés en premier lieu d'après les formes qu'ils affectent. On peut alors obtenir une caractérisation de chacune des formes du corpus par ses spécificités positives ou négatives dans les différentes parties du corpus.

A l'inverse, on peut pratiquer un autre type de lecture de ces résultats en ordonnant d'abord les diagnostics de spécificité en fonction des parties du corpus.

Ce second type de regroupement permet de rassembler, pour chaque partie, les formes qu'elle sur-emploie (les formes *S*+ de cette partie) ainsi que les formes qu'elle sous-emploie (les formes *S*-). Cette seconde lecture des résultats de la méthode permet de caractériser chacune des parties du corpus tout à la fois par les formes qu'elle utilise plus particulièrement et par celles qu'elle a tendance à éviter.

Dans chacun des états on ordonne les diagnostics de spécificité par ordre de probabilité croissante, c'est-à-dire du plus spécifique au plus banal.

Fréquences minimales, seuil d'absence spécifique

Les formes de faible fréquence, et parmi elles les hapax, constituent un cas limite. En effet, dans le cas d'une forme de fréquence égale à 1, les paramètres : k (longueur du corpus), $k_{i.}$ (fréquence totale de la forme), étant fixés, le diagnostic porté sur la sous-fréquence observée (laquelle ne peut être égale, pour ce qui la concerne, qu'à 0 ou 1) dépend alors essentiellement de $k_{.j}$, la longueur de la partie j. Dans ce cas les résultats du calcul perdent de leur intérêt. Dans la pratique, on se borne donc à calculer les spécificités pour les seules formes dont la fréquence dépasse un certain seuil que l'on appellera la *fréquence minimale retenue* et que l'on note $freq_{min.}$

Pour un seuil de probabilité donné, il existe pour chaque partie j, de longueur $k.j$, une fréquence nj telle que : toute forme de fréquence égale ou supérieure à nj dans le corpus et absente de la partie est spécifique négative pour cette partie. C'est le *seuil d'absence spécifique* (négative) pour la partie j.

6.1.2 Un exemple de calcul des spécificités

Reprenons l'exemple des réponses à la question ***Enfants*** introduit au début de ce chapitre. On peut voir (tableau 6.1, qui reprend la ligne correspondante du tableau 5.3) la ventilation de la forme *chômage* dans les neuf parties constituées à partir du croisement des 3 classes d'âge avec les 3 catégories de diplômes définies plus haut (cf. chapitre 5). Cette forme possède 285 occurrences dans le corpus.

Tableau 6.1

Ventilation de la forme *chômage* dans les réponses à la question *Enfants*

forme : *chômage* (285 occurrences)

Diplôme	A	A	A	B	B	B	S	S	S
Age	-30	-50	+50	-30	-50	+50	-30	-50	+50
Fréquence	35	55	93	25	15	10	19	18	15

Partant de cette ventilation, et compte tenu des longueurs respectives de chacune des neuf parties, le calcul des spécificités fournit (tableau 6.2) au seuil de 0.05, quatre diagnostics (positifs ou négatifs).

Tableau 6.2

Spécificité, au seuil de 0.05, sur la ventilation de la forme *chômage* dans les réponses à la question *Enfants*.

	partie	*fréq totale*	*fréq partie*	*Probabilité*
A -30	285	35	*S+*	0.0006
A +50	285	93	*S+*	0.0168
S -30	285	19	*S-*	0.0159
S -50	285	18	*S-*	0.0023

Comme on le constate, les diagnostics fournis par la méthode ne sont pas tous au même niveau. Certaines des probabilités calculées sont beaucoup plus faibles que les autres. Cependant, si l'on ne retient de ces résultats que le caractère qualitatif spécifique/non-spécifique, on peut résumer, comme au tableau 6.3, l'emploi de la forme par chacune des parties.

Tableau 6.3

Spécificités de la forme *chômage*

Diplôme	A	A	A	B	B	B	S	S	S
Age	-30	-50	+50	-30	-50	+50	-30	-50	+50
Fréquence	35	55	93	25	15	10	19	18	15
Spécificité	*S+*	*b*	*S+*	*b*	*b*	*b*	*S-*	*S-*	*b*

Certaines des formes ne trouvent de spécificité dans aucune des parties du corpus.
Le tableau 6.4 donne quelques-unes de ces formes les plus fréquentes, que l'on a appelé formes banales.

Tableau 6.4

Quelques formes banales.

Diplôme	A	A	A	B	B	B	S	S	S	
Age	-30	-50	+50	-30	-50	+50	-30	-50	+50	
										Tot.
enfants	10	7	15	24	9	20	47	9	7	148
emploi	7	9	5	19	6	5	21	4	3	79
insécurité	3	3	5	14	3	8	18	3	5	62
femme	2	9	18	4	4	5	5	9	4	60

Les résultats de spécificités obtenus à partir des ventilations des formes dans les parties du corpus peuvent être, on l'a vu, ordonnés de différentes manières. On peut rassembler comme nous l'avons fait plus haut les diagnostics relatifs à une même forme graphique en les ordonnant d'après les parties qu'ils affectent. Inversement, on peut établir, pour chacune des parties, la liste des formes qui trouvent des spécificités (positives et négatives) dans cette partie.

Sur le tableau, 6.5 on peut voir, toujours au seuil de 0.05, la liste des formes spécifiques de la partie A-30 (jeunes peu diplômés).

Le tableau s'ouvre sur les formes spécifiques positives triées de la plus spécifique (i.e. celle à laquelle se trouve affectée la plus petite probabilité) à la plus "banale". Viennent ensuite les formes spécifiques négatives triées selon le même critère.

Tableau 6.5

Les diagnostics de spécificité, au seuil de 0.05,

pour la partie A-30 dans les réponses à la question *Enfants*.

partie : A-30 (Aucun diplôme - moins de 30 ans)

forme	*fréq.*	*fréq. totale*		*Probabilité*
chômage	285	35	*S+*	0.0006
le	474	50	*S+*	0.0016
avoir	7	12	*S+*	0.0060
avenir	318	32	*S+*	0.0194
très	18	4	*S+*	0.0313
argent	196	20	*S+*	0.0493
et	204	4	*S-*	0.0013
peut	54	0	*S-*	0.0212
que	78	1	*S-*	0.0259
situation	176	6	*S-*	0.0376

Il est également possible de calculer des spécificités à partir de décomptes dans les parties, ou groupes de parties d'un corpus, des occurrences de toute autre unité constituée à partir de regroupements, ou de subdivisions opérées sur la base des décomptes en formes graphiques. C'est à ce stade que l'on pourra si on le désire regrouper le pluriel et le singulier d'une même forme, les différentes flexions d'un même verbe, afin de comparer les diagnostics obtenus sur ces regroupements à ceux obtenus lors de l'analyse des formes simples.

6.1.3 Liste des formes spécifiques ou caractéristiques

Aux formes caractéristiques sont attachées, dans les listages produits par les logiciels, des *valeurs-test,* qui sont d'autant plus grandes que les probabilités ci-dessus sont petites. Ces valeurs-tests mesurent l'écart existant entre la fréquence relative d'une forme dans une classe avec sa fréquence relative globale calculée sur l'ensemble des réponses ou individus.[1]

Cet écart est normé de façon à pouvoir être considéré comme une réalisation de variable normale centrée réduite, dans l'hypothèse de répartition aléatoire de la forme étudiée dans les classes. Dans cette hypothèse, la valeur-test a *95* chances sur *100* d'être comprise entre les valeurs *-1.96* et + *1.96* .[2] Mais ce

[1] Des logiciels comme SPAD.T et Lexico1 (brièvement décrits en annexe) produisent, pour chaque partition des individus, un bilan exhaustif des formes et segments caractéristiques (spécificités positives et négatives) correspondant à chaque classe (ou partie).

[2] Les valeurs-test ont déjà été évoquées, dans un cadre différent, au chapitre 4, à propos de la description automatique des classes d'une partition par le tableau 4.5.

calcul reposant sur une approximation normale de la loi hypergéométrique n'est utilisé que lorsque les effectifs concernés ne sont pas trop faibles.

Le tableau 6.6 présente pour les quatre catégories "extrêmes" (correspondant aux deux classes d'âge et aux deux niveaux de diplômes extrêmes) les 5 premières spécificités positives et les cinq dernières spécificités négatives.

Les valeurs-test servent de critère de classement des formes à l'intérieur de chaque classe (ou partie). On retrouve les cinq premières formes du tableau 6.5 pour la classe "aucun diplôme-moins de 30 ans", à une interversion près des formes *très* et *avenir*. Cette interversion est due à des bases de calcul différentes.

Tableau 6.6

Formes graphiques caractéristiques (question *Enfants*) pour les 4 classes extrêmes de la partition "âge-diplôme"

	LIBELLE DE LA FORME	POURCENTAGE INTERNE	POURCENTAGE GLOBAL	FREQUENCE INTERNE	FREQUENCE GLOBALE	V.TEST	PROB.
TEXTE NUMERO 1 (A—30 = Moins de 30 ans, Aucun diplôme ou cep)							
1	chomage	3.90	2.25	35.	285.	3.092	.001
2	le	5.57	3.74	50.	474.	2.756	.003
3	avoir	1.34	.61	12.	77.	2.427	.008
4	avenir	3.57	2.51	32.	318.	1.912	.028
5	tres	.45	.14	4.	18.	1.819	.034
5	plus	.22	.69	2.	87.	—1.660	.048
4	situation	.67	1.39	6.	176.	—1.884	.030
3	que	.11	.62	1.	78.	—2.012	.022
2	peut	.00	.43	0.	54.	—2.092	.018
1	et	.45	1.61	4.	204.	—3.122	.001
TEXTE NUMERO 3 S—30 = (Moins de 30 ans, bacc ou université)							
1	financieres	2.62	1.37	34.	173.	3.619	.000
2	couple	1.46	.75	19.	95.	2.718	.003
3	vis	.46	.13	6.	16.	2.674	.004
4	responsabilites	.54	.17	7.	21.	2.672	.004
5	que	1.23	.62	16.	78.	2.565	.005
5	finances	.00	.22	0.	28.	—1.665	.048
4	travail	.69	1.20	9.	152.	—1.718	.043
3	manque	.69	1.26	9.	160.	—1.917	.028
2	sante	.31	.75	4.	95.	—1.920	.027
1	chomage	1.46	2.25	19.	285.	—2.010	.022
TEXTE NUMERO 7 A+50 = (Plus de 50, aucun diplôme ou cep)							
1	ne	2.48	1.52	85.	193.	5.038	.000
2	pas	3.56	2.57	122.	325.	4.123	.000
3	ils	1.05	.62	36.	79.	3.428	.000
4	veulent	.64	.33	22.	42.	3.335	.000
5	sais	.41	.20	14.	25.	2.866	.002
5	problemes	.50	.85	17.	108.	—2.665	.004
4	financiers	.32	.68	11.	86.	—3.051	.001
3	l	4.15	5.32	142.	674.	—3.627	.000
2	avenir	1.55	2.51	53.	318.	—4.348	.000
1	financieres	.55	1.37	19.	173.	—5.108	.000
TEXTE NUMERO 9 S+50 = (Plus de 50ans , Bacc ou université)							
1	sante	1.78	.75	15.	95.	2.969	.001
2	difficultes	1.54	.65	13.	82.	2.759	.003
3	materielles	1.18	.45	10.	57.	2.645	.004
4	assurer	.59	.14	5.	18.	2.549	.005
5	les	5.09	3.49	43.	442.	2.406	.008
5	n	.24	.74	2.	94.	—1.697	.045
4	en	.36	.92	3.	116.	—1.703	.044
3	trop	.00	.41	0.	52.	—1.933	.027
2	si	.00	.44	0.	56.	—2.051	.020
1	y	.00	.44	0.	56.	—2.051	.020

Les spécificités du tableau 6.5 sont calculées sur l'ensemble du corpus, conformément à une tradition bien établie en lexicométrie. Le tableau 6.6, comme les analyses des correspondances du chapitre précédent, est calculé d'après la table de contingence esquissée au tableau 5.3, c'est à dire après intervention du seuil de fréquence des formes.

Dans ce second cas, le choix de ce seuil de fréquence de sélection des formes, qui modifie le nombre total des occurrences du texte partiel ainsi que le nombre d'occurrences pour chaque partie du texte, aura une influence sur les probabilités calculées. Les calculs du tableau 6.6 utilisent de plus une approximation normale de la loi hypergéométrique. On voit cependant que les résultats sont robustes au regard de ces options de calculs.

Le tableau 6.6 a été établi avec un seuil minimal de 14 occurrences par forme, ce qui correspond à 154 formes, et à 12 661 occurrences, sur 15 457.

Le tableau 6.7 reprend le premier texte (Jeunes sans diplômes) dans le cas d'une sélection avec un seuil minimal de 21 occurrences (cas des tableaux et graphiques du chapitre 5), d'où un texte de 12 051 occurrences.

Tableau 6.7

Formes graphiques caractéristiques (question *Enfants*)
Réponses de la classe 1 de la partition "âge-diplôme"
Spécificités positives, seuil de fréq. = 20

	forme	pourcentage interne	pourcentage global	frequence interne	frequence globale	v.test	prob.
TEXTE NUMERO 1	A–30 = (Moins de 30 ans, Aucun diplôme ou cep)						
1	chomage	4.08	2.36	35.	285.	3.069	.001
2	le	5.83	3.93	50.	474.	2.727	.003
3	avoir	1.40	.64	12.	77.	2.414	.008
4	avenir	3.73	2.64	32.	318.	1.887	.030
5	argent	2.33	1.63	20.	196.	1.509	.066
5	plus	.23	.72	2.	87.	–1.671	.047
4	situation	.70	1.46	6.	176.	–1.901	.029
3	que	.12	.65	1.	78.	–2.023	.022
2	peut	.00	.45	0.	54.	–2.100	.018
1	et	.47	1.69	4.	204.	–3.140	.001

On constate que la sélection des formes conservées après seuillage est inchangée (la forme *très* ne fait plus partie du corpus partiel, puisqu'elle n'a que 18 occurrences dans le corpus) mais que les fréquences et les indices de probabilité sont parfois légèrement modifiées.
D'une façon générale, il est recommandé, et cela est conforme à la pratique de la lexicométrie, de calculer les formes caractéristiques sur l'ensemble du

corpus, avant toute sélection par seuil. Toutefois, lorsqu'il s'agit de compléter une visualisation factorielle qui, elle, intervient nécessairement après un seuil, il est loisible de travailler sur le corpus partiel, afin d'assurer une compatibilité des tableaux lexicaux utilisés.

Formes caractéristiques et visualisation factorielle

On peut confronter les résultats du tableau 6.6 aux proximités observables sur la figure 5.1 du chapitre 5 ; il va sans dire que les résultats sont plus complémentaires que concurrents. L'approximation plane de la figure 5.1 fait perdre une certaine information, mais décèle un ordre et des régularités invisibles sur le tableau 6.6.

Une représentation utile consiste à ne garder sur la figure 5.1 que les points les plus caractéristiques, et à les joindre aux points-catégories qu'ils caractérisent.

Les points anti-caractéristiques (spécificités négatives) sont beaucoup plus difficiles à identifier sur les graphiques factoriels. Ainsi, par exemple, la forme *chômage* est la forme la plus *anti-caractéristique* de la classe 3 (S-30 = jeunes instruits), alors que ce n'est pas, loin s'en faut, la forme la plus éloignée (sur la figure 5.1 du chapitre 5) du point S-30.

Attention aux comparaisons multiples !

Le calcul simultané de plusieurs valeurs-test ou de plusieurs seuils de probabilités se heurte à l'écueil des *comparaisons multiples,* bien connu des statisticiens.

Supposons que les parties de texte soient parfaitement homogènes et donc que l'hypothèse d'indépendance entre les formes et les parties soit réalisée. Les valeurs-test attachées aux spécificités, pour une partie donnée, sont alors toutes des réalisations de variables aléatoires normales centrées réduites indépendantes. Dans ces conditions, *en moyenne*, sur 100 spécificités calculées, 5 seront en dehors de l'intervalle [-1.96 , +1.96], et 5 dépasseront la valeur 1.65 (test unilatéral). Le seuil de 5% n'a de sens en fait que pour un seul test, et non pour des tests multiples.

Autrement dit, l'utilisateur non averti trouvera presque toujours "de quoi s'étonner" au seuil de 5%... On résout de façon pragmatique cette difficulté en choisissant un seuil plus sévère. On doit garder à l'esprit les ordres de grandeurs suivants : le seuil 1% correspond à une valeur supérieure à 2.33, et le seuil 1 pour 1 000 à une valeur supérieure à 3.09. Avec ces seuils, beaucoup des spécificités calculées au tableau 6.6 restent significatives.

On se référera à nouveau au tableau 6.6 lors de la recherche des "réponses modales" (ou réponses caractéristiques) de chaque classe. Les réponses modales, étudiées au paragraphe suivant, permettront en effet d'atténuer l'impression de morcellement que peut susciter l'examen des formes isolées (tableau 6.6) en situant ces formes caractéristiques dans leurs contextes.

6.2 Les réponses modales

Une technique simple à mettre en œuvre va permettre de répondre simultanément à plusieurs des objections qui ont été soulevées concernant le caractère fragmentaire et désarticulé de toute étude limitée aux formes isolées de leur contexte immédiat : il s'agit de la sélection automatique des *réponses modales* (appelées encore *phrases caractéristiques* ou *documents caractéristiques*, pour d'autres types d'applications).

Les réponses modales ne seront pas des réponses artificielles dressant un portrait de chaque groupement, mais des réponses authentiques, effectivement fournies par les personnes interrogées, choisies en raison de leur caractère représentatif, pour une catégorie donnée d'individus.

6.2.1 La sélection des réponses modales

Il existe deux modes de sélection de ces réponses modales : un premier mode est fondé sur des calculs de distances selon des critères géométriques simples (sélection suivant la distance du chi-2). Un second mode s'appuie sur les calculs de spécificité.

Sélection suivant la distance du chi-2

Le principe de cette sélection est schématiquement le suivant : une fois décodées par l'ordinateur (phase de numérisation), les réponses libres de k individus (k=1 000, pour fixer les idées) donnent lieu à la construction d'un tableau rectangulaire **T** ayant k lignes (individus ou réponses) et autant de colonnes que de formes graphiques sélectionnées (V= 400, par exemple).

Une réponse est alors une ligne du tableau, donc un vecteur à 400 composantes. Si cette réponse est formée de 25 formes différentes, seulement 25 de ces composantes seront différentes de zéro.

Un groupement de réponses (un discours artificiel) est un ensemble de vecteurs-lignes, et le profil lexical moyen de ce groupement est obtenu en calculant la moyenne des vecteurs-lignes de cet ensemble. Si ce regroupement s'opère selon les modalités d'une question fermée dont les réponses sont codées dans un tableau **Z**, on a vu au paragraphe 5.1 que le tableau lexical agrégé **C** se calcule par la formule:

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

On peut donc calculer des distances entre des réponses et les regroupements de ces réponses. Réponses (lignes de **T**) et regroupements de réponses (colonnes de **C** , ou lignes de **C**', transposée de **C**) sont tous représentés par des vecteurs d'un même espace.

Ces distances expriment l'écart entre le profil d'une réponse et le profil moyen de la classe à laquelle cette réponse appartient. La distance choisie entre ces profils de fréquences sera la distance du chi-2, en raison de ses propriétés distributionnelles soulignées au chapitre 3.

La distance entre un point-ligne i de **T** et un point-colonne *m* de **C** est alors donnée par la formule :

$$d^2(i,m) = \sum_{j=1}^{p} (t../t.j)\,(t_{ij}/t_{i\cdot} - c_{jm}/c_{\cdot m})^2$$

avec les notations usuelles :

- $t..$ désigne la somme globale des éléments du tableau **T**, c'est-à-dire le nombre total d'occurrences.
- $t_{\cdot j}$ désigne la somme des éléments de la colonne j de **T** (nombre d'occurrences de la forme j).
- $t_{i\cdot}$ désigne la somme des éléments de la ligne i de **T** (longueur de la réponse i).
- $c_{\cdot m}$ la somme des éléments de la colonne m de **C** (nombre total d'occurrences de la classe ou du groupement m).

On peut, pour chaque regroupement, classer ces distances par ordre croissant, et donc sélectionner les réponses les plus représentatives au sens du profil lexical, qui correspondront aux plus petites distances. Ce mode de sélection des réponses modales est appelé dans ce qui suit : sélection selon le critère du chi-2.

Sélection par les formes caractéristiques

Un autre mode de calcul de réponses modales est le suivant : une réponse modale d'un groupement est une réponse qui contient, autant que faire se peut, les formes les plus caractéristiques de ce groupement (spécificités). Pour chaque groupement, ces formes sont classées par ordre de significativité (cf. tableau 6.6 par exemple) : le rang d'une forme dans ce classement est d'autant plus petit qu'elle est caractéristique.

Une formule empirique simple consiste à associer à chaque réponse le rang moyen des formes qu'elle contient : si ce rang moyen est petit, cela signifie que la réponse ne contient que des formes très caractéristiques du groupement.

Mieux que les rangs moyens, on peut prendre les valeurs-test correspondantes (au plus petit rang moyen correspond alors la plus grande valeur-test moyenne). On conçoit facilement que certains arbitrages soient nécessaires : entre une réponse courte formée de peu de formes très caractéristiques et une réponse plus longue, qui contient nécessairement des formes moins caractéristiques...

Ce mode de calcul (critère des formes caractéristiques) a la particularité de favoriser les réponses courtes, alors que le critère du chi-2 a au contraire tendance à favoriser les réponses longues.

Quel que soit le mode de calcul, on imprimera en fait plusieurs réponses caractéristiques pour chaque groupement, car il est extrêmement improbable qu'il existe parmi les réponses originales une réponse qui résume à elle seule toutes les particularités d'une catégorie. Dans les listages de sortie, on pourra trouver, par exemple, six réponses par classe (trois réponses pour chacun des deux modes de calcul...).

6.2.2 Mise en oeuvre et exemples

Pour une question ouverte (ici encore, la question ***Enfants*** nous servira d'exemple de référence), et pour une partition de la population (la partition en catégories *âge - diplôme*), on obtient donc, sans codification préalable ni médiation :

- Une visualisation des proximités entre formes et catégories (chapitre 5).

- Les spécificités ou formes caractéristiques de chaque catégorie (paragraphe 6.1 de ce chapitre).
- Les réponses modales de chaque catégorie.

Reprise de l'exemple du chapitre 5

On donnera tout d'abord un exemple de réponses modales correspondant aux quatre classes extrêmes de la partition en neuf classes : les moins de 30 ans sans diplôme, les moins de 30 ans de niveau d'instruction "Bac et plus", les plus de 50 ans sans diplôme, les plus de 50 ans "Bac et plus". On abandonne donc simultanément la classe d'âge et le niveau de diplôme intermédiaires.

Pour chacune de ces quatre classes, on donnera les trois premières réponses modales, pour chacun des deux critères de sélection. Les tableaux 6.8a et 6.8b présentent, pour chacune des quatre classes extrêmes (pour l'âge et le niveau d'instruction) les réponses sélectionnées d'après le critère des formes caractéristiques

Les réponses sont très laconiques : puisque la forme *chômage* est la plus caractéristique de la première classe (tableau 6.6), et qu'il existe des réponses formées de ce simple mot, il est naturel qu'il s'agisse, avec ce critère, de la réponse la plus caractéristique. Le critère de sélection correspondant est alors égal à la valeur test 3.09.

L'article *le* fait légèrement baisser la valeur du critère. On peut vérifier (tableau 6.6) que la valeur de ce critère pour la réponse *le chômage* est la moyenne arithmétique des critères des deux formes qu'elle contient.

Le caractère presque caricatural des réponses sélectionnées par ce premier critère est frappant si l'on compare les tableaux 6.8 et 6.9. L'utilisateur consultera d'abord les premières réponses, qui constituent un résumé, un squelette sémantique, que la lecture des réponses choisies selon le second critère permettra de nuancer et compléter.

Remarques

Cet exemple appelle plusieurs remarques, qui concernent aussi bien l'application elle-même que la méthode.

a) La première est un constat élémentaire, mais qu'il faut faire à ce stade : on a obtenu effectivement un résultat ayant une certaine cohérence, une sorte de résumé de l'information initiale, sans codification ni pré-interprétation du texte.

b) Les réponses ci-dessus sont des réponses originales intégrales, choisies parmi 2 000 réponses saisies... elles permettent de rétablir les formes dans leur contexte, lorsqu'il existe. Elles pourraient d'ailleurs être positionnées sur la figure 6.1, en occupant chacune le centre de gravité des formes graphiques qu'elles contiennent. Mais cela encombrerait une représentation déjà très chargée.

Tableau 6.8 a

Critère des formes caractéristiques
Réponses modales pour la question *Enfants* (Catégories *moins de 30 ans*)

```
CRITERE DE          REPONSES OU INDIVIDUS CARACTERISTIQUES
CLASSEMENT
_____________________________________________________________
TEXTE NUMERO  1  (A—30  = Moins de 30 ans, Aucun diplôme ou cep)
_____________________________________________________________
   3.092 ——    1   chomage
   3.092 ——    2   chomage
   3.092 ——    3   chomage
   3.092 ——    4   chomage
   2.924 ——    5   le chomage
_____________________________________________________________
   TEXTE NUMERO   3  (S—30 = (Moins de 30 ans, bacc/université)
_____________________________________________________________
   2.361 ——    1   raisons financieres
   2.361 ——    2   raisons financieres
   2.361 ——    3   raisons financieres
   2.361 ——    4   raisons financieres
   2.361 ——    5   raisons financieres
```

Tableau 6.8 b

Critère des formes caractéristiques
Réponses modales pour la question *Enfants*
(Catégories *plus de 50 ans*)

```
CRITERE DE          REPONSES OU INDIVIDUS CARACTERISTIQUES
CLASSEMENT
_____________________________________________________________
   TEXTE NUMERO  7   A+50  = (Plus de 50, aucun diplôme ou cep)
_____________________________________________________________
   3.544 ——    1   je ne sais pas
   3.544 ——    2   je ne sais pas
   3.053 ——    3   ne sait pas
   3.053 ——    4   ne sait pas
   3.053 ——    5   ne sait pas
_____________________________________________________________
   TEXTE NUMERO  9   S+50  =  (Plus de 50ans , Bacc/université)
_____________________________________________________________
   2.969 ——    1   sante
   1.890 ——    2   egoisme
   1.843 ——    3   les revenus
   1.823 ——    4   l'egoisme,difficultes materielles
   1.721 ——    5   les difficultes financieres
```

Tableau 6.9

Critère de la distance du chi-2.
Réponses modales pour la question *Enfants*

```
CRITERE DE            REPONSES OU INDIVIDUS CARACTERISTIQUES
CLASSEMENT
-------------------------------------------------------------------
TEXTE NUMERO  1  A-30  = moins de 30 ans, aucun diplôme ou cep)
-------------------------------------------------------------------
 .861 --  1   manque d'argent,l'avenir,le chomage
 .876 --  2   le chomage,le manque d'argent de toutes manieres
 .887 --  3   la peur de l'avenir le chomage
 .887 --  4   la peur de l'avenir,le chomage

-------------------------------------------------------------------
TEXTE NUMERO  3  S-30  = moins de 30 ans, bacc ou université
-------------------------------------------------------------------
 .924 --  1   l'instabilite des relations du couple,la situation
               financiere encore plus et la peur d'un avenir    precaire
             pour les enfants
 .925 --  2     raisons financieres,avenir de l'enfant
 .935 --  3     problemes financiers,l'engagement,le contrat a long    terme
             que cela suppose
 .936 --  4   je ne suis pas peut etre les obligations que cela  impliquent
             et les depenses en plus

-------------------------------------------------------------------
TEXTE NUMERO  7  A+50  = plus de 50, aucun diplôme ou cep
-------------------------------------------------------------------
 .865 --  1   je ne sais pas si on en a pas envie ce n'est pas la
               peine,il y a des gens qui n'aiment pas les enfants
 .909 --  2   je ne peux pas vous dire.. le genre de caractere de      la
             personne
 .914 --  3   la mesentente c'est le seul probleme un couple qui
               ne s'entend pas ils vont pas avoir d'enfants parce
               qu'apres c'est des problemes plus graves
 .918 --  4   les difficultees financieres,il y a des gens qui   n'aiment pas
             les enfants,il vaut mieux ne pas avoir       d'enfants que de
             les rendre malheureux

-------------------------------------------------------------------
TEXTE NUMERO  9  S+50  =  plus de 50ans , bacc/université
-------------------------------------------------------------------
 .894 --  1   l'instabilite economique et sociale du pays et plus
               concretement les difficultes financieres de chacun
               d'entre nous,la base du pouvoir d'achat et l'    insecurite
             de l'emploi
 .908 --  2   une question de logement, ensuite les ressources   financieres,
             ensuite  la situation geographique des       epoux apres ca
             peut etre un certain egoisme,des gens  qui limitent les
             naissances dans le desir de mieux     elever les enfants qu'ils
             auront acceptes
 .913 --  3   la sante de l'epouse,les conditions materielles de
               vie et les causes ideologiques,crainte du futur le
               chomage,la guerre
 .921 --  4   les conditions de vie et les difficultes d'embauche
```

c) Les différences d'expression sont assez considérables, mais ne semblent pas un obstacle à l'analyse de l'information, bien au contraire : Peut-on vraiment considérer, par exemple, que *manque d'argent* invoqué par les jeunes sans diplôme, est tout à fait synonyme des *raisons financières*, mentionnées par les jeunes diplômés ? Bien entendu, un post-codage aurait classé ces deux réponses sous une même rubrique.

Mais le fait que la différence de formes soit liée à une différence de situations socio-économiques pose à nouveau le problème du *contenu* de chacune de ces réponses. N'y a-t-il pas quasi-impossibilité dans un cas et difficulté dans l'autre ? Comme c'est souvent le cas lors de l'utilisation de techniques exploratoires, on est conduit inéluctablement à critiquer l'information de base et même les concepts qui président au recueil de cette information.

d) Il peut arriver, bien que cela ne soit pas le cas ici, que certaines formes rares apparaissent dans les réponses caractéristiques. Il faut noter que pour les deux critères, la sélection des réponses ne se fait que sur les formes retenues, après intervention du seuil de sélection des formes selon leur fréquence (ce seuil vaut 13 dans notre exemple). Mais l'édition des réponses respecte leur libellé original. Il peut donc arriver, dans des cas plutôt exceptionnels, qu'une réponse ne contienne que des formes dont la fréquence est inférieure au seuil, à l'exception d'une seule, qui soit un mot-outil spécifique de la catégorie. Avec le critère de fréquence lexicale (mais pas avec le critère du chi-2), cette réponse pourra apparaître comme caractéristique.

Un exemple en est donné par une des réponses modales de la classe 9 (personnes instruites de plus de 50 ans) : parmi les formes spécifiques de cette classe figurait la conjonction *et*, qui traduit d'ailleurs le fait que les personnes de cette classe sont également les plus disertes (la forme *et* ne figure pas dans le tableau 6.6 car elle occupe le rang 6, avec une valeur-test de 2.3). On obtient comme réponse modale, à un niveau voisin de celles citées plus haut : "*Elles aiment mieux batifoler et rigoler ensemble*", réponse qui ne doit son bon classement qu'à la présence de la forme *et* et à l'absence des autres formes fréquentes du corpus. Le critère de fréquence lexicale n'est donc à utiliser qu'avec circonspection. Le seuil de fréquence doit être fixé suffisamment bas[1].

[1] Notons, toujours dans le cadre empirique de cet exemple, que le fait d'abaisser le seuil de 13 à 6 fait effectivement disparaître cette réponse modale "parasite", mais modifie très peu l'ensemble des autres réponses modales. Une seule nouvelle réponse modale apparaît, (toujours avec le critère de la fréquence lexicale, et toujours pour la classe 4):" *maladies héréditaires* "(*maladies*: fréquence 9, *héréditaires*, fréquence 7).

6.2.3 Autres exemples

Pour donner une idée à la fois de la diversité des réponses modales pour d'autres questions ouvertes et d'autres catégories, et des interprétations auxquelles elles conduisent, on donnera ci-dessous quelques exemples complémentaires (il s'agit de sélections réalisées uniquement selon le critère du chi-2). Ces exemples sont extraits de l'enquête sur les conditions de vie et aspirations des Français (cf. chapitre 3, paragraphe 3.2.2).

Le premier exemple (a) concerne des *inquiétudes générales*, étudiées selon la catégorie socioprofessionnelle des répondants.

Le second exemple (b) concerne un exemple d'explicitation de réponse, du type *"pourquoi ?"*, à une question fermée (cf. chapitre 1, paragraphe 1.4.3).

Le troisième et dernier exemple reprend la question plus méthodologique posée à l'issue de chaque entrevue de cette même enquête (question déjà étudiée au chapitre 5, paragraphe 5.2).

a) Inquiétudes par catégorie socioprofessionnelle

Les réponses modales concernent la question *"Qu'est-ce qui vous inquiète le plus en ce qui concerne votre avenir ?"* pour quelques catégories socio-professionnelles :

Agriculteurs :

— *De pouvoir continuer notre métier jusqu'à la fin de la vie; l'avenir des jeunes dans l'agriculture, problèmes du foncier et de l'équipement lourd en agriculture.*
— *Les problèmes d'énergie, le fuel est trop cher, on en a besoin pour les tracteurs.*
— *Il faudrait que les prix augmentent à la production, sinon l'exploitation ne sera plus rentable.*

Ouvriers :

— *Garantie de l'emploi jusqu'à la retraite, toucher une retraite convenable.*
— *Ne pas disposer de suffisamment d'années de vie pour profiter de la retraite.*
— *L'emploi et le chômage.*

Ménagères, femmes au foyer :

— *Je m'inquiète plus pour l'avenir de mes enfants que pour le mien.*
— *Rien pour moi, mais pour mes enfants: leur orientation, leur formation, leur travail.*
— *Pas de travail pour les jeunes.*

Cadres moyens et supérieurs :

— *Cette crise économique nous entraîne vers une catastrophe, la crise politique vers la guerre.*
— *L'impact de la crise sur mon activité professionnelle par réduction des crédits et du recrutement.*
— *Montée du chômage, elle aura peut-être des conséquences très graves sur le plan psychologique.*

Etudiants :

— *La peur de ne pas trouver un travail intéressant, le ronronnement progressif, la passivité, la démission collective devant les grands problèmes de notre temps.*

Retraités :

— *Tout ce que je demande, c'est de pouvoir vivre comme maintenant, et surtout de garder une bonne santé, le reste, je ne m'en inquiète pas.*

On obtient donc un résumé assez vivant des réponses et de leur dispersion dans différentes catégories sans avoir à lire l'ensemble des deux mille réponses qui correspondent en fait à un texte de près de cent pages.

Bien évidemment, la partition en catégories socioprofessionnelles ne permet pas d'épuiser toutes les formes de réponses : ainsi, si l'on distingue de l'ensemble des "retraités" la catégorie des "retraités avec une seule personne au foyer", on obtient pour cette catégorie les réponses modales suivantes :

— *J'ai peur de tomber malade et d'être seule, le reste ne me fait pas peur.*
— *Etre malade et être seule, ne pas pouvoir rester dans ma maison jusqu'au bout.*
— *La solitude, la santé.*

Ces différences de contenu et de tonalité entre les retraités en général et les retraités vivant seuls illustrent bien l'importance du choix des regroupements à opérer sur les réponses. Aurait-on pu saisir par un post-codage la véhémence des étudiants, l'angoisse des personnes âgées seules, l'attitude oblative des ménagères, le pragmatisme à court et moyen terme des agriculteurs, la globalité des points de vue des cadres ?

Probablement, si le codage est réalisé par un sociologue chevronné, et si l'exploitation statistique des données codées permet de prendre en compte les cooccurrences de ces items ; mais la procédure proposée ici nous paraît moins coûteuse et plus réaliste.

b) Opinions sur le mariage pour une classe d'âge, par sexe

La question ouverte est maintenant la question "*Pouvez-vous dire pourquoi ?"*, à la suite de la question fermée suivante, qui concerne l'institution du mariage (cf. Tabard, 1974) :

Parmi ces opinions, laquelle se rapproche le plus de la vôtre?
Le mariage est:

- *Une union indissoluble.*
- *Une union qui peut être dissoute dans les cas graves.*
- *Une union qui peut être dissoute par simple accord des deux parties.*

On se limite ici à l'examen de deux classes d'une partition croisant le genre et l'âge des répondants (6 000 individus correspondant à trois phases consécutives de l'enquête précitée, de 1978 à 1980) : les hommes de plus de 60 ans et les femmes de plus de 60 ans.

Ces classes d'âges sont toutes deux "traditionalistes" : à la question fermée préliminaire, 48% des hommes et 44% des femmes ont répondu qu'ils considéraient le mariage comme une union indissoluble, alors que le pourcentage moyen obtenu par cette réponse dans l'ensemble de la population (résidents métropolitains âgés de 18 ans et plus) est de 28%.

Les réponses modales des hommes de plus de 60 ans sont :

— *Quand on se marie, on ne fait pas n'importe quoi.*

— *On se marie pour toujours.*

— *Si on se marie, c'est pour toujours, autrement, il vaut mieux rester célibataire.*

— *Je suis catholique, dans ma religion, on ne divorce pas.*

— *Quand on est marié, on doit rester ensemble.*

Les réponses modales des femmes de la même classe d'âge sont plus longues et plus nuancées :

— *De mon temps, c'était comme ça, bien sûr, ce n'est pas interdit de divorcer, mais il vaut mieux passer bien des choses et rester ensemble.*

— *Parce que je suis catholique pratiquante, et que quand on est marié, on l'est.*

— *Plutôt que d'être malheureux toute sa vie, il vaut mieux divorcer, mais avant d'en arriver là, il vaut mieux supporter beaucoup de choses.*

— *Si le mari est saoul du matin au soir, violent, il vaut mieux se séparer, les enfants sont trop malheureux.*

— *En principe, je suis contre le divorce, sauf dans les cas graves (alcoolisme, mauvais traitements infligés aux enfants et au conjoint).*

On relève une assez nette différence d'expression et d'argumentation entre ces deux catégories de personnes âgées.

La forme *malheureux* est trois fois plus fréquente chez les femmes de cette classe d'âge que dans l'ensemble de la population, la forme *supporter* y est quatre fois plus fréquente. Plus circonstanciées, les réponses des femmes sont en retrait par rapport à celles des hommes.

Il est très probable qu'un post-codage "a priori" aurait désincarné les réponses et donc gommé ces différences qui tiennent à la fois à la tonalité globale de la réponse, à la prolixité du répondant, aux nuances de l'argumentation.

C'est en ce sens que l'exploitation des questions ouvertes constitue un prolongement du questionnaire susceptible de ménager des surprises, mais surtout fournissant les moyens de juger la pertinence du libellé des questions et parfois même des hypothèses sous-jacentes au questionnement.

c) Exemple à partir de noyaux factuels

On reprend l'exemple des 22 noyaux factuels donné au chapitre précédent, section 5.2. Sans donner une liste complète de réponses modales pour l'ensemble des 22 classes, on trouvera, à titre d'exemple, des réponses modales relatives à sept d'entre elles :

Classe 17 (hommes ouvriers sans enfant)
Classe 18 (hommes ouvriers avec enfants)
Classe 9 (femmes au foyer de niveau de vie modeste avec enfants)
Classe 10 (femmes au foyer aisées, avec enfants)
Classe 7 (femmes au foyer sans enfant)
Classe 21 (femmes actives de condition modeste avec enfants)
Classe 11 (femmes actives aisées avec enfants)

Examinons tout d'abord les deux catégories d'*hommes ouvriers* :

Classe 17 (hommes ouvriers sans enfants)

— *Non, le questionnaire est très complet.*
— *Je ne connais pas l'utilité de votre questionnaire, mais je trouve qu'il est pas mal fait.*

— *Questionnaire ridicule qui ne peut servir à rien si ce n'est être utilisé par la police ou les impôts.*

Classe 18 (hommes ouvriers avec enfants)

— *Pourquoi faire des différences dans les familles entre deux et trois enfants au niveau des allocations. Ce n'est pas notre faute si nous n'avons pas pu avoir plus de deux enfants, et c'est injuste de nous pénaliser.*
— *Il y a trop d'injustice au niveau de la répartition des allocations familiales, pour un enfant, on devrait avoir la même chose.*
— *Pour les allocations familiales, trop d'écart entre deux et trois enfants, chaque enfant devrait avoir une somme égale, ce sont tous les mêmes Français.*
— *Pourquoi, avec un petit salaire comme le nôtre avons-nous si peu d'allocations familiales, alors que nous avons deux enfants en étude et ceux-ci étant grands nous coûtent plus cher que des enfants en bas-âge.*

Ces dernières interventions en forme de doléances, souvent longues, contrastent avec les réponses libres des ouvriers sans enfants au foyer. On voit que les différences de forme et de contenu des réponses de ces deux groupes pourtant voisins de personnes enquêtées sont importantes.

La partition en noyaux factuels nous donne trois classes de *femmes au foyer* (classes 7, 9, et 10, selon la figure 5.2 du chapitre 5), dont les réponses seront très diversifiées selon la présence ou l'absence d'enfants au foyer, et également selon le niveau de vie.

Pour la classe 9 (*ménagères de niveau de vie modeste, jeunes, avec enfants*), il y a peu de critiques de l'enquête mais une avalanche de remarques et de revendications, comme le suggèrent les réponses modales suivantes.

Classe 9 (femmes au foyer de niveau de vie modeste avec enfants)

— *On aurait pu aborder le sujet de la femme au foyer pour la prendre plus en considération, on ne parle que des femmes qui travaillent*
— *J'aurais aimé que l'on demande aux mères au foyer, et même aux femmes en général si elles accepteraient mieux de rester chez elles si elles touchaient un certain salaire.*
— *La femme au foyer travaille et n'a pas de salaire ni de retraite en rapport avec son travail, elle est handicapée par rapport à la femme qui travaille*

Pour la classe 10, l'attitude est tout à fait différente, comme le laisse prévoir la position du point correspondant sur la figure 5.3 du chapitre 5, plus à gauche que celui de la classe 9, et donc plus proche des remarques et critiques sur l'enquête.

Classe 10 (femmes au foyer aisées, avec enfants)

— *Les réponses pour certaines questions sont trop rigides, pas assez nuancées.*
— *Le questionnaire tourne en rond sur les questions de salaire, on répète plusieurs fois les mêmes questions.*

Enfin, les réponses de la classe 7 sont assez variées, comme peut le laisser prévoir la position du point correspondant sur la figure 5.3.

Classe 7 (femmes au foyer sans enfant)

— *Le libellé des questions limite trop les réponses et empêche de dire ce qu'on pense.*
— *C'est vraiment parce que l'enquêtrice était sympathique que j'ai répondu jusqu'au bout à votre questionnaire, qui, lui, n'est pas très passionnant.*
— *Sur la sécurité sociale, c'est inadmissible de faire attendre si longtemps les gens sans argent pour liquider un dossier.*

Il est intéressant de constater que les réponses des femmes actives urbaines (classes 11, 20, 21) diffèrent assez peu des réponses des femmes au foyer, le principal critère discriminant étant, comme nous l'avons signalé plus haut, la présence d'enfants au foyer, avec également un rôle important dévolu au niveau de vie.

Classe 21 (femmes actives de condition modeste avec enfants)

— *N'a pas été posée la question du salaire de la femme au foyer, suffisant pour éviter que la femme au foyer ne soit obligée de travailler.*
— *Le questionnaire ne parle pas du tout du travail à mi-temps.*
— *On ferait bien de donner un salaire à la femme au foyer afin de permettre qu'elle reste plus chez elle avec ses enfants, ce qui permettrait d'avoir plus d'enfants et de libérer des emplois.*

Alors que les centres d'intérêt et la tonalité générale des réponses fournies par les femmes de la classe 11 sont fort différents :

Classe 11 (femmes actives aisées avec enfants)

— *Trop long, trop détaillé, je n'aime pas parler de ma vie privée.*
— *Non, sinon j'en reviens à la formation, à diplômes égaux, les hommes sont plus favorisés que les femmes, il y a aussi le problème de la régionalisation dans ce domaine, et cela, on en tient absolument pas compte.*

Ces derniers exemples de classes et de réponses montrent que les partitions en noyaux factuels peuvent mettre en évidence une certaine combinatoire de situations qui permettent de détecter, grâce aux réponses modales, des

familles de préoccupations fort différentes. L'intérêt heuristique de la méthode des noyaux factuels tient à la mise en évidence de croisements ou d'interactions qui aurait pu échapper à une sélection faite "a priori" par l'utilisateur.

D'une manière générale, la procédure de sélection des réponses modales apporte des nuances et des compléments importants à l'information fournie par les réponses aux questions fermées. La personne en charge d'analyser l'enquête a sous les yeux la diversité et la complexité de l'individu "répondant", habituellement dissimulées sous des pourcentages qui ne sont neutres qu'en apparence. L'interprétation des réponses donnée par les répondants eux-mêmes mérite aussi d'être prise en considération.