L'ANALYSE STATISTIQUE DES RÉPONSES LIBRES DANS LES ENQUÊTES SOCIO-ÉCONOMIQUES

par

Ludovic LEBART(*)

RÉSUMÉ. – On utilise des questions ouvertes dans les enquêtes socio-économiques lorsque l'éventail des réponses est trop grand, trop diversifié, ou imparfaitement connu. On propose ici une chaîne de traitements statistiques qui s'appliquent aux réponses libres saisies directement sous forme littérale, sans codage ni intervention manuelle. Outre des statistiques lexicales classiques, la procédure comporte une visualisation graphique des proximités entre les mots utilisés et les catégories auxquelles appartiennent les individus, et une sélection des réponses les plus caractéristiques de ces catégories. Plusieurs exemples d'application réels sont présentés; ils concernent l'enquête annuelle du CREDOC sur les conditions de vie et aspirations des Français.

ABSTRACT. – STATISTICAL ANALYSIS OF FREE ANSWERS IN SOCIO-ECONOMIC SURVEYS. Open ended questions are asked in socio-economic surveys when the sets of response-items are too large, or diversified, or insufficiently known beforehand. A chain of statistical treatments is applied to free answers registered directly in their textual form, without coding or tampering. Besides the traditional lexical statistics, the procedure includes a graphical visualization (using correspondence analysis) of the proximities between points representing the words and the groups to which the individuals belong. The final step consists of a selection of the answers which characterize such groups. Several examples of application are given. They relate to the CREDOC annual survey on the living conditions and aspirations of the French.

SOMMAIRE

(*) Maître de Recherches au CNRS et chercheur au CREDOC, 142, rue du Chevaleret, Paris 13e. L'auteur remercie A. DESROSIÈRES et Y. HOUZEL pour les remarques très pertinentes qui lui ont été faites.

Cet article présente et commente une chaîne de traitements pour l'analyse statistique des réponses aux questions ouvertes dans les enquêtes socio-économiques. A partir des réponses saisies directement sous forme littérale sur un support informatique, on procède à des regroupements et des éditions sélectives de textes, à des calculs de statistiques lexicales (accompagnées de visualisation par analyse des correspondances) et à des sélections de « discours typiques » ou de « réponses modales ».

La chaîne de traitement fonctionne actuellement en routine au CREDOC sur les fichiers du système d'enquêtes annuelles sur les conditions de vie et aspirations des Français. La transparence des différentes étapes et le caractère pratique des résultats ont conduit à multiplier les questions ouvertes dans les phases récentes de ce système, en raison notamment de l'originalité et du pouvoir heuristique de l'information recueillie.

Après avoir brièvement rappelé le statut actuel, les avantages et les inconvénients des réponses libres et de l'information de type textuel dans les enquêtes, évoqué le traitement conventionnel par post-codage manuel, on examinera successivement les trois étapes du traitement : éditions sélectives, analyse factorielle lexicale, mise en évidence des réponses caractéristiques (ou encore : réponses modales). Des exemples d'application « en vraie grandeur » (corpus de 6 000 réponses extrait de l'enquête précitée) illustreront la chaîne des opérations.

1. LES RÉPONSES LIBRES, POURQUOI?

On est conduit à utiliser des questions ouvertes dans les enquêtes de type socio-économique pour des raisons qui peuvent concerner aussi bien le contenu du questionnement que des aspects techniques du recueil d'information.

Classiquement, une question est laissée ouverte lorsque l'éventail des réponses possibles n'est connu que partiellement. Certains items de réponse relativement peu fréquents pourraient en effet être négligés à la suite d'essai sur des échantillons-pilotes nécessairement réduits.

L'éventail des réponses peut au contraire être parfaitement connu, mais être trop vaste (exemple : titres de trois hebdomadaires lus régulièrement ou occasionnellement) pour faire l'objet d'une codification de la part de l'enquêteur au moment de l'interview; il peut être aussi trop complexe et nécessiter des

arbitrages qui peuvent dans certains cas dépasser les possibilités de l'enquêteur. Ainsi, même dans le cas de caractéristiques aussi classiques que la « branche d'activité » de l'entreprise où travaille la personne interrogée, il est souvent préférable de ne prélever qu'une réponse littérale au moment de l'interview.

Les réponses peuvent être également composites (par exemple : activités principales durant le week-end), et dans ce cas une question ouverte peut, dans une certaine mesure, remplacer plusieurs batteries de questions fermées très lourdes à administrer.

Il est également possible que les items de réponse aient un caractère conjoncturel, et ne puissent être prévus lors de la conception du questionnaire, ou que cela n'ait pas sens de les suivre dans le temps. C'est le cas par exemple de la question sur la participation à des actions en faveur de l'environnement, dans les enquêtes annuelles du CREDOC sur les conditions de vie et aspirations des Français. (Des items de réponses tels que « Larzac », « Plogoff » sont trop spécifiques pour s'insérer dans une grille de codification permanente et trop importants pour ne pas être pris en compte).

D'une façon générale, on peut dire que les questions ouvertes fournissent une information riche et étendue, ayant plus le caractère d'une expression que d'une réaction. Cette information a toutefois sa propre spécificité et ne peut en aucune façon être comparée à celle issue des questions fermées, pour lesquelles la personne interrogée est sollicitée (mais parfois aussi lassée, quelquefois contrariée) par les batteries de réponses qui lui sont proposées.

Cette spécificité fait qu'il n'y a pas vraiment concurrence entre les deux types de questionnements : pour repérer et suivre dans le temps un pourcentage, pour classer des items, pour toutes questions à réponse dichotomique ou conduisant à un positionnement sur une échelle sémantique, il est clair que les questions fermées sont irremplaçables... alors que lorsque les sujets abordés sont nouveaux, que la durée de l'interview nécessite une certaine adhésion (obtenue en limitant l'inquisition et en laissant des possibilités d'expression), que l'on recherche autant des idées que des suffrages, les questions ouvertes apparaissent très utiles.

D'un point de vue technique, il faut souligner que la médiation de l'enquêteur est vraisemblablement plus importante dans le cas des questions ouvertes, en particulier parce que celui-ci a la charge de retranscrire les réponses qu'il peut modifier de façon parfois sensible (s'il est, par exemple, contraint de résumer après des hésitations). La médiation de l'enquêteur sera de toute façon plus manifeste, puisque la forme même des réponses (leur longueur en particulier) peut être affectée lors du recueil de l'information.

2. LE POST-CODAGE MANUEL

Le principe de ce prétraitement très classique des réponses libres est le suivant : on construit pour une question ouverte une ou plusieurs questions fer-

mées à partir des réponses relevées sur un échantillon de questionnaires (qui représentent de l'ordre de 10 % de l'échantillon total); puis des chiffreurs répondent alors à ces questions en codifiant les réponses littérales pour l'ensemble de l'échantillon.

Cette procédure largement utilisée n'est cependant pas sans inconvénient :

— La médiation de l'enquêteur n'est évidemment pas éliminée, mais s'y ajoute celle du chiffreur, qui elle, dans tous les cas, nécessite arbitrage, interprétation, et donc « équation personnelle ».

— L'information est nécessairement appauvrie : la qualité de l'expression, la spécificité du vocabulaire, le ton utilisé sont des matériaux d'analyse perdus lors d'un post-codage.

— Les réponses rares, originales, ou complexes sont affectées à des items résiduels qui sont donc très hétérogènes et perdent de ce fait toute valeur opératoire.

— Enfin, les réponses aux questions ouvertes sont d'emblée multidimensionnelles et la codification est une opération trop délicate pour être considérée comme une phase préliminaire et quelque peu subalterne.

Ce sera là d'ailleurs un des enseignements de nos techniques de traitements textuels directs, le regroupement des réponses peut souvent se faire à différents niveaux qui n'apparaissent qu'après des analyses statistiques élaborées utilisant d'autres variables de l'enquête.

Prenons un exemple emprunté à l'enquête précitée (vagues 1978-1979-1980). On dispose de 5 805 réponses libres à la question ouverte :

« *Qu'est-ce-qui vous inquiète le plus en ce moment en ce qui concerne votre avenir?* »

Parmi les mots les plus fréquemment utilisés dans les réponses, citons (avec entre parenthèses le nombre d'occurrences) SANTE (764), EMPLOI (723), RIEN (686), ENFANTS (643), CHOMAGE (573), AVENIR (563), TRAVAIL (526), (sur un total de 27 718 mots dont 3 031 sont distincts).

Limitons-nous aux réponses concernant l'emploi et l'avenir des enfants, et donnons quelques exemples de réponses :

1. « *Le chômage pour mes enfants* ».

2. « *Je m'inquiète plus pour l'avenir de mes enfants (emploi) que pour le mien* ».

3. « *Ma sécurité d'emploi parce que je suis seule et divorcée avec un fils de quinze ans* ».

4. « *L'avenir des enfants, trouveront-ils du travail, auront-ils une formation en rapport avec la vie qui les attend sans passer par le chômage avant de travailler; le coût de la vie (pétrole, électricité)* ».

5. « *Mes enfants, leur situation dans l'avenir, j'aimerais qu'ils ne connaissent pas de guerre, un avenir mieux que nous* ».

6. « Le problème du chômage pour les enfants et pour mon mari ».

7. « De mourir avant d'avoir élevé mes enfants, les laisser à mon mari ».

8. « Le chômage des jeunes, réduction des heures de travail pour que tout le monde travaille ».

9. « L'emploi pour mes enfants, ils feront des études et lorsqu'ils seront titulaires d'un diplôme, trouveront-ils un emploi? ».

On voit sur cet exemple à propos du thème « Avenir des enfants », l'imbrication et les interrelations entre les différents items de réponse. Sur ces neuf réponses seulement (plus de 500 concernent directement ou indirectement l'avenir des enfants), on dispose d'éléments d'information sur :

— une hiérarchie des inquiétudes; la sécurité d'emploi des parents seuls avec enfants; l'adaptation de la formation à la vie professionnelle; le coût de la vie; la crainte de la guerre; la présence des chômeurs multiples dans les familles; le problème des jeunes en général; l'utilité des diplômes...

Bien entendu, la consultation de réponses plus nombreuses et la prise en considération des autres thèmes d'inquiétude montreraient l'enchevêtrement de plusieurs centaines d'items de fréquences très variables (certaines réponses pouvant comporter jusqu'à sept items).

On pourrait utiliser — et on utilise d'ailleurs en pratique — une codification retenant par exemple vingt items principaux, en attachant deux ou trois questions fermées (ayant chacune les vingt items de réponse) à la réponse libre, pour pouvoir enregistrer deux ou trois items par réponse... On a ainsi vingt possibilités pour le premier item cité, vingt pour le second, etc. Mais il devient difficile de faire usage des combinaisons d'items (au nombre de 400 dans le cas des combinaisons deux à deux). On peut noter d'ailleurs l'incohérence selon laquelle un item de base (par exemple « l'évolution des prix agricoles ») peut ne pas figurer dans les vingt items les plus fréquents, et donc être exclu du post-codage, alors que des combinaisons deux à deux beaucoup plus rares (par exemple, « santé » et « trouver un premier emploi ») sont prises en compte. En bref, on peut dire que les réponses libres donnent lieu à une information assez spécifique, difficile à utiliser selon des schémas classiques.

Les réponses aux questions ouvertes sont en fait des éléments statistiques que l'on commence seulement à pouvoir traiter de façon efficace : ce sont des « *vecteurs clairsemés* ». Que l'on imagine un tableau rectangulaire T ayant autant de lignes que d'individus interrogés (6 000 pour notre exemple) et de colonnes que de mots distincts utilisés (3 031 pour la question relative aux inquiétudes); chaque réponse pourra être décrite par une ligne de 0, 1, 2... selon que le mot correspondant est absent, présent une fois, deux fois, etc, dans la réponse. Ce mode de représentation (lui-même insuffisant, car il ne prend pas en compte l'ordre des mots dans la réponse) va permettre néanmoins de travailler dans un « espace des réponses » donnant des outils de description intéressants. Mais avant cela, les possibilités de gestion, d'édition et

de calcul des ordinateurs nous suggèreront des traitements élémentaires permettant de respecter l'information brute, et néanmoins de la rendre plus utilisable.

3. LES ÉDITIONS SÉLECTIVES

On va utiliser les possibilités de classement et d'édition de l'ordinateur pour établir, dans une première phase, une *aide à la lecture des réponses libres.*

La donnée de base du traitement sera un enregistrement en format libre alphanumérique des réponses (à l'intérieur d'une réponse, les séparateurs entre formes lexicales (1) étant constitués par des séquences de longueur quelconques de blancs, virgules, points, apostrophes, parenthèses). L'unité d'enregistrement comporte un numéro d'individu (permettant d'établir une correspondance avec d'autres questions, ouvertes ou fermées) et le texte de la réponse en format libre, dont la longueur totale est limitée, dans les applications courantes, à 750 caractères (10 cartes perforées avec 75 colonnes utiles).

Le premier traitement élémentaire consiste, pour une question ouverte, à *regrouper les réponses selon les classes d'une partition jugée pertinente vis-à-vis du problème étudié.* Ainsi, on peut être amené à étudier les inquiétudes des Français par catégorie socio-professionnelle. On éditera donc les textes initiaux pour les ouvriers spécialisés, les instituteurs, les professions libérales, etc. Le fait d'avoir reclassé les réponses peut faire apparaître des discours plus homogènes, la répétition de certains thèmes apparaissant plus clairement.

Un second traitement possible consiste à positionner les réponses originales sur les cartes factorielles, à l'emplacement des individus concernés, en considérant ces réponses comme de simples *identificateurs.* Dans l'hypothèse où cette procédure est techniquement réalisable, on espère alors voir apparaître des zones privilégiées où se regroupent des mots, des thèmes, des associations de thèmes.

En fait, cette seconde procédure est impraticable lorsque les réponses sont nombreuses et longues. On se ramène alors à la précédente en découpant l'espace factoriel en zones homogènes et disjointes, ce qui revient à établir une nouvelle partition des individus à partir des mêmes variables actives que celles de l'analyse factorielle.

On aboutit dans les deux cas à des « *discours artificiels* » formés de juxtapositions de réponses libres relatives à une même classe. Si l'on dispose de 18 catégories socio-professionnelles par exemple, le tableau T (individus x formes lexicales) évoqué au paragraphe précédent va voir ses 6 000 lignes regroupées en 18 lignes. Le tableau ainsi agrégé (classe x formes) que l'on appellera C, contient, à l'intersection de la ligne i et de la colonne j le nombre de fois où,

(1) « Enfants » et « Enfant » sont deux formes lexicales du même mot « Enfant ».

dans la classe *i*, se rencontre la forme *j*. On a en fait un tableau de contingence apte à être décrit par l'analyse des correspondances. Cette dernière méthode peut en effet donner une représentation graphique des associations entre classes (vis-à-vis de leurs similitudes de profils lexicaux), des associations entre formes ou mots (vis-à-vis de leurs profils de répartition dans les classes), et enfin des associations privilégiées formes-classes (voir paragraphe 4). De tels graphiques constituent une précieuse aide à la lecture des *discours artificiels*. On verra plus bas que cette lecture peut être allégée par la sélection de « réponses modales » (paragraphe 5).

4. ANALYSE FACTORIELLE LEXICALE

Ce paragraphe va décrire une procédure d'analyse des réponses regroupées selon les classes d'une partition des individus enquêtés. L'individu statistique « réponse » sera ignoré lors de cette étape qui s'intéresse seulement aux relations existant entre les classes et les formes lexicales. C'est dire que cette procédure s'applique aussi bien à n'importe quelle série de textes (l'article cité en référence (Lebart, 1981) donne un listage commenté du programme de calcul et un « exemple miniature » d'application à une série de cinq poèmes).

Le programme va reconnaître, puis compter et classer les formes lexicales, calculer la table de contingence *C* croisant les classes et les formes, procéder à l'analyse des correspondances de cette table. Entre la saisie du texte brut et la consultation de graphiques tels que le graphique 2 ci-après, ne s'interpose donc aucune intervention manuelle.

4.1. Principe technique du programme

Le programme ne comporte aucune limite pour la longueur des textes, qui sont lus ligne par ligne - chaque ligne est lue lettre par lettre, et les séquences de lettres sans séparateurs seront les formes lexicales à collationner. La rapidité de la procédure tient au fait que des emplacements en mémoire centrale sont réservés pour les formes selon leur longueur (seules les 16 premières lettres d'une forme sont prises en compte).

La répartition maximale par longueur dans le programme de calcul est la suivante :

TABLEAU I

Nombre de lettres *(en caractères)*	Nombre de formes *distinctes*
1	30
2	90
3	150
4	250
5	350
6	420
7	450
8	420
9	350
10	250
11	150
12	100
13	60
14	50
15	40
16	20
	Total : 3 180

Cet histogramme des formes par longueur est satisfaisant pour les textes de l'ordre de 40 000 mots. A chaque application, le nombre de formes de chaque longueur effectivement rencontrées est édité, et permet de réajuster éventuellement ces paramètres (ce réajustement est nécessaire pour des textes écrits dans d'autres langues que le Français).

Chaque forme lue est comparée aux formes de même longueur précédemment retenues. Ainsi le mot « EDUCATION » est comparé aux mots de neuf lettres déjà trouvés, lettre par lettre. Il y en a moins de 350; le test sur la seule première lettre élimine déjà les 9/10e des mots de cette longueur; on voit donc que le nombre d'opérations est minime. S'il est nouveau, le mot vient enrichir le thésaurus.

REMARQUES :

Etant donné le caractère purement formel de l'opération, les fautes d'orthographe (assez nombreuses dans ce type de recueil), les erreurs de lecture ou de perforation de texte créent des formes lexicales parasites et multiplient d'ailleurs indéfiniment le nombre de formes distinctes rencontrées. Les formes repérées au-delà des effectifs maximaux prévus appartiennent en grande majorité à cette catégorie. Elles sont ignorées.

— Un seuil de fréquence minimal permet de réduire les dimensions de la table de contingence (classes × formes) et d'éliminer les mots très rares (les mots apparaissant moins de 10 fois, par exemple, sont éliminés).

— Il existe des logiciels (en général assez lourds) effectuant des traitements lexicométriques plus complets. L'étape dont il est question ici est légère, transportable, et s'intègre facilement dans une chaîne d'analyses statistiques beaucoup plus élaborées que celles habituellement disponibles.

L'étape suivante est une analyse des correspondances de la table de contingence, avec production de graphiques des plans factoriels, où les mots sont repérés par 4, 8 ou 12 lettres selon la place disponible (voir graphique 2, p. 50).

4.2. Exemple illustratif

On donnera un exemple de réponses libres venant en complément de réponses à une question fermée sur le mariage (vague 78) de l'enquête sur les conditions de vie et aspirations des Français, auprès de 2 000 français de 18 ans et plus.

Libellé de la question :

— « *Parmi ces opinions, quelle est celle qui se rapproche le plus de la vôtre :*

Le mariage est :

1. Une union indissoluble.

2. Une union qui peut être dissoute dans les cas graves.

3. Une union qui peut être dissoute par simple accord des deux parties.

4. Ne sait pas »?

La question ouverte qui suit est simplement :

— « *Pouvez-vous dire pourquoi* »?

La partition utilisée pour regrouper les réponses comporte sept classes; elle a été établie par classification à partir des principales questions d'opinions de l'enquête (questions fermées). Nous ne pouvons donner ici les détails de la construction de cette typologie en familles d'opinions. Le lecteur pourra se reporter aux rapports annuels du système d'enquête.

Le graphique 1 donne la position des centres de classes dans le plan factoriel de l'analyse de la même batterie de questions d'opinions. Elle doit permettre d'identifier de façon sommaire les classes.

On ne publiera pas ici, faute de place, la table de contingence (7 × 187) croisant les sept classes d'opinions et les 187 formes lexicales les plus fréquentes (elles apparaissent au moins 12 fois). On note que 1 748 formes distinctes ont été retenues (à partir d'un texte de 14 894 mots). Il y a donc 1 561 formes qui apparaissent moins de 12 fois...

Le tableau II ci-après donne simplement la liste de ces formes et leurs fréquences absolues et relatives.

Le graphique 2 donne le résultat final de l'analyse factorielle lexicale : les proximités entre formes, entre classes, et entre formes et classes. On a reproduit ici le listage original du graphique afin de souligner le caractère automatique de son obtention à partir du texte brut des réponses (et la mention de l'appartenance des individus aux sept classes).

Les deux premiers axes extraits expliquent respectivement 17 et 9 % de l'inertie totale.

Le graphique 2 (p. 50) est assez difficile à lire, par suite du tronquage de certains mots à quatre ou huit caractères pour des raisons de précision graphique (les mots de moins de trois lettres sont exclus). (Pour éviter trop de superposition de mots, l'échelle verticale a été exagérément dilatée : la figure réelle

est beaucoup plus large que haute, et l'on interprètera surtout les distances sur l'axe horizontal).

On note que la classe 1 occupe une position assez isolée sur la gauche du graphique, avec un vocabulaire très spécifique : plus impersonnel, parfois plus savant, presque juridique : « institution, droit, contrat, parties, société, liberté, couple, conjoints ». On trouve au contraire sur la partie droite : « mon, moi, nos, suis, nous » et des mots tels que « famille, morale, foyer ». On remarque également à gauche trois formes lexicales du verbe pouvoir : « peuvent » (près de l'axe horizontal), « pouvoir » (un peu plus haut) et « peut » (encore plus haut et près de l'axe vertical) et à droite deux formes de devoir : « doit » et « doivent » (vers le haut).

Le fait de travailler sur des formes lexicales et non des mots n'est pas un inconvénient grave lorsque le corpus est grand, comme c'est le cas ici. Au contraire, les formes suggèrent des utilisations particulières des mots, et permettent de mieux conjecturer le sens des phrases.

On retrouve des expressions (noter par exemple la proximité à droite entre « meilleure » et « pire »). Certains regroupements de mots donnent des informations sur le contenu des réponses (« condition », « religieuse », « sacré », « engage », sur la partie droite) mais aussi sur le ton, ou l'éventuelle aisance (« normal », « évident », « logique », dans la moitié gauche).

GRAPHIQUE 1

Positions relatives des sept classes d'opinions

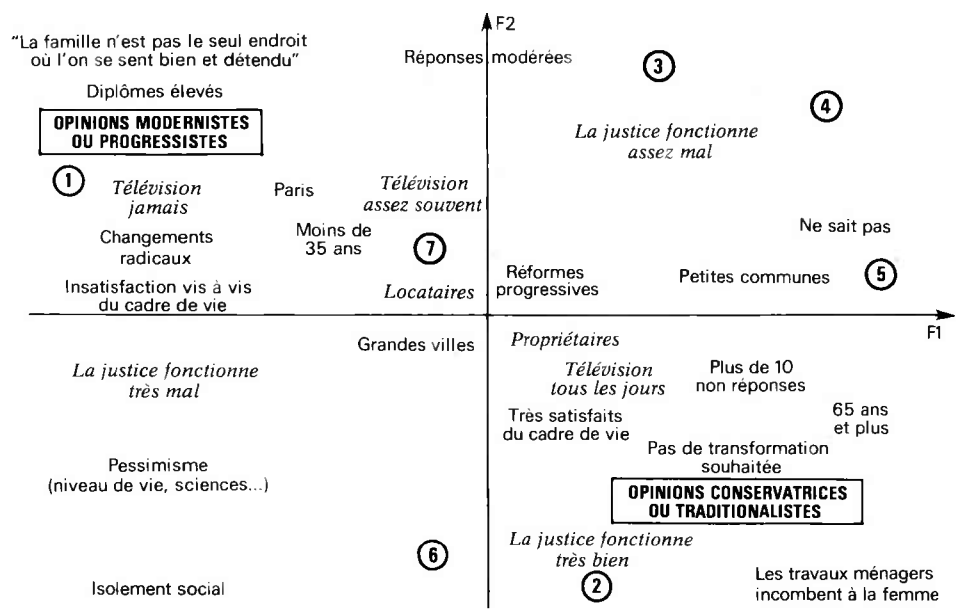


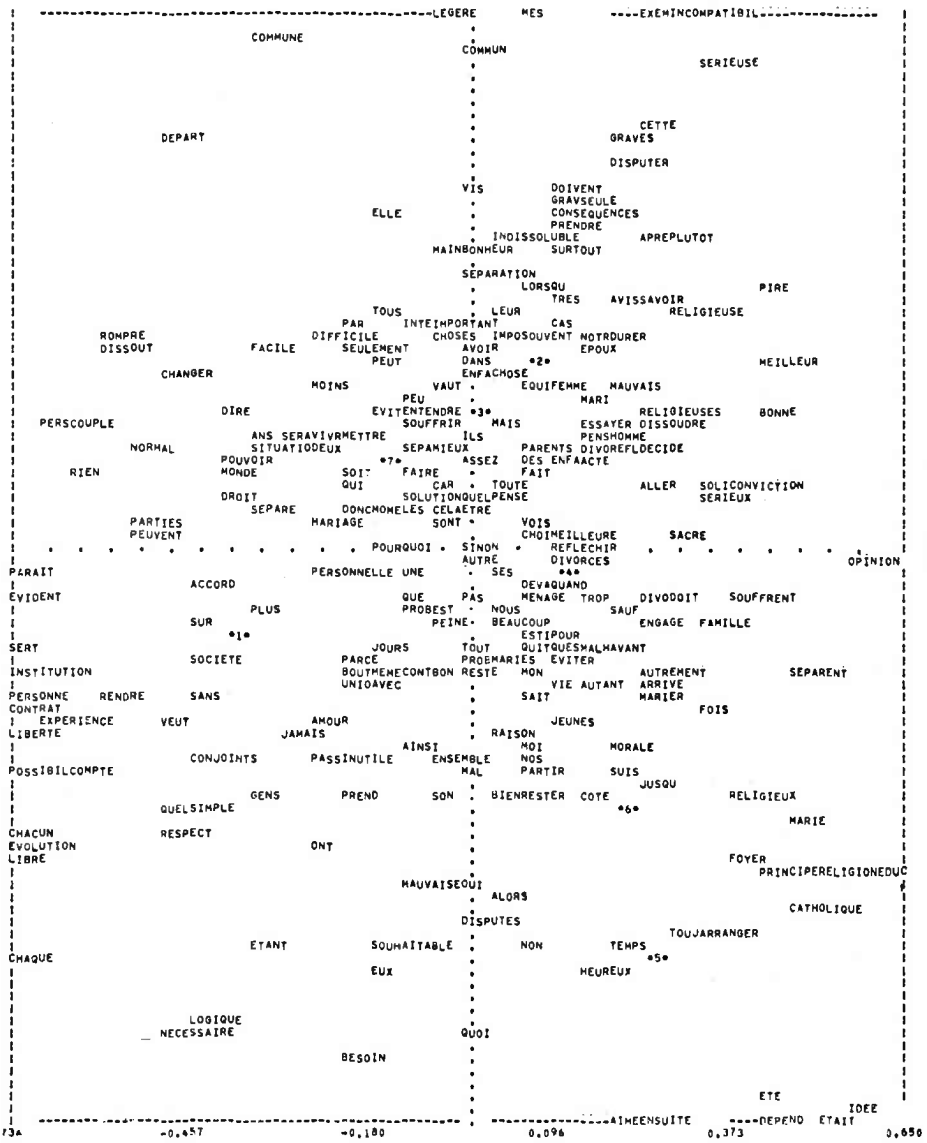
TABLEAU II

Questions sur le mariage
Les 187 mots les plus fréquents

Fréquences	%	Libellés	Fréquences	%	Libellés	Fréquences	%	Libellés
814	5,47	Est	36	0,24	Solution	18	0,12	Inutile
624	4,19	Pas	36	0,24	Quitter	18	0,12	Plutôt
536	3,60	Pour	36	0,24	Réfléchir	18	0,12	Certains
405	2,72	Des	35	0,23	Gens	18	0,12	Aussi
386	2,59	Les	35	0,23	Vraiment	18	0,12	Mauvais
354	2,38	Enfants	35	0,23	Pouvoir	18	0,12	Partir
302	2,03	Que	34	0,23	Ils	18	0,12	Tous
276	1,85	Quand	34	0,23	Séparation	18	0,12	Peuvent
263	1,77	Plus	34	0,23	Entente	17	0,11	Mal
251	1,69	Une	33	0,22	Simple	17	0,11	Problème
239	1,60	Vie	33	0,22	Sans	17	0,11	Alors
230	1,54	Séparer	33	0,22	Eviter	16	0,11	Leur
219	1,47	Mariage	32	0,21	Autre	16	0,11	Supporter
213	1,43	Mieux	31	0,21	Son	16	0,11	Entendent
205	1,38	Faut	31	0,21	Engagement	16	0,11	Lorsque
191	1,28	Peut	30	0,20	Femme	16	0,11	Epoux
189	1,27	Marié	29	0,19	Toute	16	0,11	Devrait
165	1,11	Cas	29	0,19	Chacun	16	0,11	Equilibre
159	1,07	Vaut	28	0,19	Pourquoi	15	0,10	Enfant
149	1,00	Ensemble	28	0,19	Non	15	0,10	Doivent
133	0,89	Entend	28	0,19	Donc	15	0,10	Catholique
124	0,83	Dans	28	0,19	Normal	15	0,10	Ménage
123	0,83	Doit	27	0,18	Parents	15	0,10	Parties
114	0,77	Etre	27	0,18	Moi	15	0,10	Incompatibilité
104	0,70	Parce	26	0,17	Fois	15	0,10	Religion
100	0,67	Qui	26	0,17	Sur	14	0,09	Entendre
94	0,63	Vivre	26	0,17	Rien	14	0,09	Libre
94	0,63	Rester	25	0,17	Mon	14	0,09	Penser
89	0,60	Mais	25	0,17	Contrat	14	0,09	Meilleure
85	0,57	Divorce	25	0,17	Foyer	14	0,09	Homme
78	0,52	Accord	25	0,17	Quelque	14	0,09	Idée
73	0,49	Divorcer	25	0,17	Pense	14	0,09	Jamais
73	0,49	Graves	25	0,17	Contre	14	0,09	Trouve
69	0,46	Très	24	0,16	Important	14	0,09	Difficile
67	0,45	Fait	24	0,16	Après	14	0,09	Légère
66	0,44	Tout	24	0,16	Pire	14	0,09	Quelque
64	0,43	Deux	23	0,15	Autant	14	0,09	Impossible
64	0,43	Bien	23	0,15	Amiable	14	0,09	Moins
58	0,39	Avec	23	0,15	Indissoluble	13	0,09	Religieuses
58	0,39	Faire	23	0,15	Aux	13	0,09	Conjoints
56	0,38	Suis	23	0,15	Sérieux	13	0,09	Institution
56	0,38	Trop	22	0,15	Malheureux	13	0,09	Moment
55	0,37	Grave	22	0,15	Car	13	0,09	Bout
53	0,36	Union	22	0,15	Savoir	13	0,09	Passer
52	0,35	Sont	21	0,14	Personnes	13	0,09	Dire
51	0,34	Cause	21	0,14	Soit	13	0,09	Mauvaise
50	0,34	Famille	21	0,14	Raison	13	0,09	Autrement
49	0,33	Liberté	21	0,14	Marché	13	0,09	Exemple
47	0,32	Avoir	21	0,14	Mari	13	0,09	Lorsqu'
47	0,32	Cela	20	0,13	Séparé	13	0,09	Heureux
45	0,30	Par	20	0,13	Mésentente	13	0,09	Engagé
45	0,30	Chose	20	0,13	Possible	13	0,09	Arriver
45	0,30	Avant	20	0,13	Nous	12	0,08	Prendre
40	0,27	Même	20	0,13	Choses	12	0,08	Dissoudre
40	0,27	Toujours	20	0,13	Raisons	12	0,08	Sonne
40	0,27	Surtout	19	0,13	Temps	12	0,08	Bon
40	0,27	Marier	19	0,13	Question	12	0,08	Envie
40	0,27	Peine	19	0,13	Meilleur	12	0,08	Entre
39	0,26	Problèmes	19	0,13	Personne	12	0,08	Bonheur
38	0,26	Couple	19	0,13	Choix	12	0,08	Dépend
38	0,26	Comme	19	0,13	Cette	12	0,08	Côté
37	0,25	Préférable	19	0,13	Sinon			
37	0,25	Sait	19	0,13	Sacré			

GRAPHIQUE 2

Question ouverte sur le mariage
Reproduction du listage : proximités entre formes lexicales et classes.



Notons également un effet du *psittacisme* fréquemment observé dans les interviews: les mots de la question fermée introductive sont largement repris, puisque l'on trouve, dans le quadrant inférieur gauche : « simple », « accord », « deux », « parties », et dans le quadrant opposé : « indissoluble », et « cas », « grave ».

On notera enfin la position souvent significative des *mots-outils*. Un test statistique simple permet en effet de vérifier si un mot, compte tenu de sa fréquence et de sa distance à l'origine des axes, peut être considéré comme réparti de façon aléatoire dans les classes. Ainsi, le mot « quand » (à droite et en bas de l'origine), le mot « lorsqu' » (plus haut) traduisent souvent des réponses de type traditionnel (« quand on se marie, c'est pour la vie »). Les mots « donc », « car », « parce que », caractérisent, à gauche de l'origine, des réponses peut-être plus argumentées.

On voit que cette visualisation, automatique et peu coûteuse, peut aider à la lecture des sept « discours artificiels » formés de juxtapositions de réponses libres pour chacune des sept classes d'opinions. Il est clair qu'il ne s'agit pas ici d'un traitement qui épuise le sujet, ne serait-ce que parce que le regroupement des réponses est hypothéqué par le choix d'une partition. Il ne s'agit donc que d'un point de vue.

Cependant, ce type d'analyse lexicométrique souffre d'une faiblesse classique (qui serait encore plus importante d'ailleurs s'il s'agissait d'analyses unidimensionnelles) : le sens des réponses peut seulement être conjecturé en l'absence du contexte des mots. La procédure de sélection des réponses modales va permettre de lever les incertitudes relatives aux divers emplois possibles des mots.

5. LES RÉPONSES MODALES

L'ensemble des réponses relatives à une classe a jusqu'à présent été considéré, si l'on veut, comme un « sac » rempli de formes lexicales entassées pêle-mêle, sans tenir compte des associations entre formes à l'intérieur des réponses. L'individu statistique « réponse », volontairement ignoré, va de nouveau être utilisé.

5.1. Principes techniques

Revenons au tableau clairsemé T croisant en lignes les individus (i.e. : les réponses) et en colonnes les formes lexicales. On a vu que l'on pouvait réduire considérablement le nombre de colonnes de T en écartant les formes dont la fréquence absolue est inférieure à un seuil jugé convenable.

La table de contingence C, qui s'obtient par agrégation de lignes de T, a aussi pour colonnes les formes lexicales. On peut donc positionner les lignes de T comme des éléments supplémentaires dans l'analyse de C; autrement dit, on peut faire apparaître les *réponses complètes* sur un graphique tel que celui figurant ci-contre.

L'abscisse x_i de la réponse i (formée de $t_{i.}$ mots) sur l'axe α, sur lequel le mot j a pour abscisse φ^j_α, et auquel correspond la valeur propre λ_α, est donnée par la formule classique :

$$x_i = (1/\sqrt{\lambda_\alpha}) \sum_j (t_{ij}\varphi^j_\alpha / t_{i.})$$

Les points-réponses et les points-classes sont dans le même espace (dont les coordonnées sont les formes), et donc on peut interpréter maintenant des proximités entre *classes* et *réponses*. Bien entendu, la visualisation est quasi impossible car les réponses sont trop nombreuses. On se contentera donc de sélectionner, pour chaque classe, les points-réponses les plus proches, puis d'éditer les réponses complètes correspondantes. La distance « réponse-classe » se calculera directement comme distance du chi-deux entre les lignes des tableaux T et C.

Ces réponses ne sont pas des « réponses moyennes de classe », qui seraient des réponses artificielles construites à partir des mots les plus fréquents dans la classe, mais des « réponses modales », c'est-à-dire des réponses réelles, utilisant le plus de mots caractéristiques de la classe.

En utilisant les notations usuelles pour décrire les marges des tableaux T et C, le carré de la distance de la réponse i à la classe k s'écrit :

$$d^2(i,k) = \sum_j \frac{t_{..}}{t_{.j}} \left(\frac{t_{ij}}{t_{i.}} - \frac{c_{kj}}{c_{k.}} \right)^2$$

On note d'ailleurs que $t_{.j} = c_{.j}$ (fréquence absolue du mot ou de la forme j). Pour chaque valeur de k, ces distances sont calculées et classées, et les réponses correspondant aux 20 plus petites sont éditées (20 est le paramètre retenu en pratique; les mots les plus fréquents peuvent en effet se répartir dans plusieurs réponses modales distinctes).

Autres critères de sélection des réponses modales :

D'autres critères peuvent être utilisés pour exprimer la proximité entre une réponse et une classe. On peut, par exemple, commencer par chercher les mots les plus caractéristiques de chaque classe. Les plans factoriels, comme le graphique 2, en donnent une idée, mais il n'est pas nécessaire de se limiter à ce qui n'est qu'une approximation plane. En comparant la fréquence relative d'un mot dans une classe à sa fréquence relative dans la population, et en tenant compte de la fréquence absolue du mot, on peut obtenir un classement des mots selon leur « pouvoir de caractérisation » à l'intérieur de chaque classe. Une réponse est alors décrite par le rang moyen des mots qui la composent. Les réponses de plus faible rang moyen seront caractéristiques de la classe (on remplace en fait les rangs r par une fonction monotone du type $\log(1+r)$, afin de ne pas pénaliser trop durement les réponses qui contiendraient quelques mots mal classés en même temps que des mots très bien classés).

En pratique, on publiera simultanément, pour chaque classe, les mots caractéristiques, et les réponses modales selon les deux critères précédents, qui ne sont pas toujours identiques; la distance du chi-deux favorisant les réponses assez longues, le critère du rang moyen les réponses courtes. La convergence des deux critères est excellente lorsque les classes ont des profils lexicaux typés.

Exemple 1 :

On comparera les classes 1 et 4 de l'exemple précédent, classes qui s'opposent sur l'axe factoriel 1 (voir graphique 2).

— Les formes les plus fréquentes de la classe 1 sont : (avec entre parenthèses le pourcentage de la forme en général, suivi du pourcentage de la forme dans la classe, évidemment plus élevée).

— Plus (2.4; 3.6), liberté (0.4; 0.9), chacun (0.3; 0.6), personnes (0.2; 0.5), contrat (0.2; 0.5); que (2.8; 3.8), couple (0.4; 0.7), mariage (2.0; 2,8), ..., amiable (0.2; 0.4), gens (0.3; 0.6).

— Pour la classe 4, on a le classement analogue :

— marié (1.7; 2,5), certains (0.2; 0.4), indissoluble (0.2; 0.4), inutile (0.2; 0.4), quand (2.5; 3.2), sérieux (0.2; 0.4), vraiment (0.3; 0.5), graves (0.6; 0.9), question (0.2; 0.3), meilleur (0.2; 0.3) ..., pire (0.2; 0.4).

Parmi les réponses modales de la classe 1, citons :

— « *Parce que le mariage est un contrat* ».

— « *Le mariage est un contrat à l'amiable* ».

— « *Le mariage ne concerne que les deux parties, si elles sont d'accord, elles se séparent* ».

— « *Afin de respecter la liberté de chacun* ».

Parmi les réponses modales de la classe 4, on relève :

— « *Quand on se marie, c'est pour la vie* ».

— « *On est marié, c'est pour le meilleur et pour le pire* ».

— « *Question de morale* ».

— « *Quand on se marie, c'est pour fonder un foyer* ».

On voit sur ces exemples que les listes de mots caractéristiques et les phrases modales donnent un aperçu assez synthétique de la forme et du contenu des réponses de chaque classe. Les réponses modales permettent de doter d'une ossature les éléments foisonnants des graphiques factoriels.

Exemple 2 :

Toujours à propos de la même question ouverte sur le mariage, on va montrer les résultats obtenus en utilisant maintenant une nouvelle partition des individus, établie par simple croisement des caractéristiques de base : sexe et âge. Les données concernent 6 000 individus (trois phases consécutives de l'enquête précitée).

On va se limiter ici à l'examen de deux classes particulières de la partition : les hommes de plus de 60 ans, et les femmes de plus de 60 ans. Classiquement, ces classes d'âge sont traditionalistes : à la question fermée préliminaire sur le mariage, 48 % des hommes et 44 % des femmes de plus de 60 ans ont

répondu qu'ils le considéraient comme une « union indissoluble », contre 28 % pour l'ensemble de la population.

Réponses modales des hommes de plus de 60 ans :

— « *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* ».

Réponses modales des femmes de plus de 60 ans.

— « *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 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 note une différence de ton, d'expression, de sensibilité, d'argumentation entre ces deux catégories de personnes enquêtées. Le mot « malheureux » est presque trois fois plus fréquent chez les femmes de cette classe d'âge que dans l'ensemble de la population, et le mot « supporter » quatre fois plus fréquent. Ces différences n'auraient vraisemblablement pas pu apparaître après un post-codage réalisé sur les 6 000 réponses non regroupées; ici, cette sélection de réponses brutes peut faire penser, par exemple, que le mariage n'est pas vécu de la même façon par les hommes et les femmes de cette génération. Les réponses circonstanciées des femmes de cette classe d'âge ont le mérite de resituer le problème que se pose le réalisateur de l'enquête parmi les problèmes que se posent les personnes enquêtées. Elles constituent un véritable prolongement du questionnaire, dans des directions qui échappent, et c'est une chance, aux prérogatives du « questionneur ».

5.2. Réalisation pratique : Les partitions instrumentales

On a vu qu'il était nécessaire de regrouper les réponses pour pouvoir procéder à des analyses de type statistique, le tableau d'incidence T croisant les formes lexicales et les individus (c'est-à-dire les réponses) étant trop clairsemé pour subir un traitement tel quel. Lors des exemples précédents, les partitions utilisées pour regrouper les réponses étaient dans le premier cas une partition

construite pour mettre en évidence des familles d'opinions, dans l'autre une partition élémentaire établie en croisant les deux caractéristiques de base que sont le sexe et l'âge.

Il peut arriver que, au départ, aucune partition particulière ne s'impose, et que l'on désire néanmoins avoir une vue d'ensemble des réponses à une question ouverte. S'il s'agit d'identifier par les caractéristiques des répondants les différents thèmes abordés, on pourra utiliser une *partition en noyaux factuels,* c'est-à-dire une partition de l'échantillon croisant le maximum de caractéristiques de base, avec cependant la contrainte d'un effectif minimal par classe. De telles partitions s'obtiennent en appliquant une procédure de classification automatique aux individus de l'échantillon décrits par une batterie de ces caractéristiques de base (les individus agrégés ensemble seront ceux qui ont le plus de caractéristiques de base en commun).

En bref, une partition en noyaux factuels est la plus fine des partitions que l'on peut construire avec des combinaisons de variables objectives, tout en étant compatible avec une induction statistique convenable, ce qui interdit les classes d'effectifs trop faibles. On la qualifie d'instrumentale dans la mesure où elle n'est qu'un intermédiaire de calcul et non un résultat final : il s'agit d'un instrument de détection d'interactions entre caractéristiques factuelles, vis-à-vis d'un phénomène particulier (qui est dans le cas présent l'information contenue dans une ou plusieurs réponses libres).

L'analyse factorielle lexicale permet ensuite de regrouper les classes qui ont répondu de la même façon (c'est-à-dire, ayant les mêmes profils lexicaux) et donc de ne garder que les caractéristiques, ou les croisements de caractéristiques discriminants vis-à-vis des thèmes exprimés.

Exemple d'application : remarques et appréciations à l'issue d'une enquête

Lors des deux dernières vagues du système d'enquête sur les conditions de vie et aspirations des Français, le questionnaire se terminait par la question ouverte suivante :

— « *Vous venez d'être interrogé(e) 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?* ».

Pour avoir une vue d'ensemble des réponses libres à cette question, on a utilisé une partition des 4 000 individus concernés en 22 noyaux factuels. Cette partition a été établie à partir d'une batterie de descripteurs objectifs comprenant notamment la catégorie socio-professionnelle, le sexe, l'âge, le statut matrimonial, le revenu, la taille d'agglomération, des caractéristiques d'équipement, etc. Ce choix de 22 classes correspond à une coupure satisfaisante de l'arbre hiérarchique. Il correspond simultanément à une certaine finesse de description, sans introduire trop d'encombrement ou de complexité au niveau de la présentation des résultats.

On trouvera, dans l'encadré ci-après, la liste des 22 noyaux avec leurs fréquences relatives et une description sommaire de leurs caractéristiques dominantes.

Le graphique 3 schématise le plan factoriel issu de l'analyse des correspondances de la table de contingence croisant les 22 noyaux factuels et les 300 formes les plus fréquentes.

On note, par exemple, que les personnes âgées, retraitées ou non, sont assez bien regroupées, dans le quart inférieur droit du graphique (noyaux 1, 2, 3, 4, 8).

Les jeunes actifs des deux sexes et les étudiants (noyaux 5, 11, 12, 13, 19) sont également relativement regroupés à gauche et s'opposent aux précédents, avec toutefois l'exception du noyau 21 (jeunes femmes actives avec enfants, de condition modeste) qui sera intéressant à analyser.

Les deux noyaux d'ouvriers 17 et 18 qui ne diffèrent que par la présence d'enfants au foyer, sont cependant assez éloignés.

CARACTÉRISTIQUES DOMINANTES DES 22 NOYAUX FACTUELS

	(%)	
1.	*8*	Retraités hommes, deux personnes au foyer
2.	*2*	Retraités hommes, seuls
3.	*5*	Retraitées femmes, seules
4.	*3*	Retraitées femmes, deux personnes au foyer
5.	*4*	Etudiants
6.	*3*	Invalides, malades, autres non actifs
7.	*6*	Ménagères sans enfants au foyer, d'âge moyen
8.	*2*	Ménagères de plus de 60 ans
9.	*6*	Ménagères avec enfants, condition modeste
10.	*3*	Ménagères avec enfants, aisées
11.	*2*	Jeunes femmes actives avec enfants (urbaines, aisées)
12.	*3*	Jeunes hommes actifs avec enfants (urbains, aisés)
13.	*3*	Jeunes hommes actifs sans enfants (urbains, aisés)
14.	*6*	Hommes actifs d'âge moyen avec enfants (urbains)
15.	*6*	Hommes actifs (milieu rural)
16.	*4*	Femmes actives (milieu rural)
17.	*7*	Ouvriers hommes (sans enfants au foyer)
18.	*10*	Ouvriers hommes (avec enfants)
19.	*5*	Jeunes actifs des deux sexes, célibataires
20.	*2*	Femmes actives sans enfants (au foyer)
21.	*8*	Jeunes femmes actives avec enfants (employées, condition modeste)
22.	*2*	Classe résiduelle (très hétérogène)

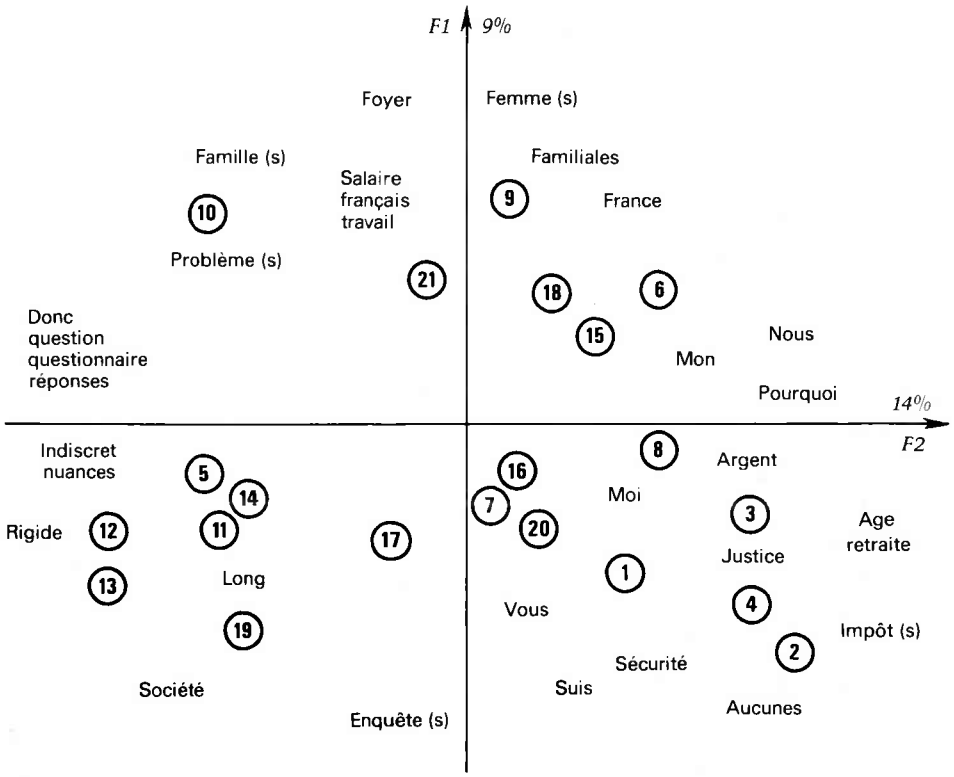
L'examen systématique des réponses modales par grands groupes va situer dans leur contexte les mots du graphique 3, et mettre en évidence l'ancrage factuel des thèmes exprimés :

a) Les retraités (noyaux 1, 2, 3, 4, 8)

Les réponses du type « rien », « aucunes » sont nombreuses, mais beaucoup de commentaires prolongent l'interview : ils concernent principalement l'insuffisance des retraites, les impôts (sur les retraites), la justice, la sécurité.

GRAPHIQUE 3

Représentation schématique de certaines proximités :
« formes lexicales × noyaux factuels »
(question ouverte : « remarques à l'issue de l'interview »).



Exemples de réponses modales :

— « *Non, à part que je voudrais bien que l'on s'occupe des vieux* ».

— « *C'est plutôt des jeunes qu'il faut interroger, à mon âge, on n'attend plus grand chose, on a fait sa vie* ».

— « *La solitude des personnes âgées, et surtout les lourdes charges qu'ils doivent supporter après une vie de travail, l'impôt sur les retraites, c'est triste* ».

— « *Le plus grave concerne la justice qui est trop indulgente pour les criminels. Je suis pour la peine de mort, au moins pour leur faire peur* ».

— « *Je trouve que ce n'est pas bien qu'on vous enlève des impôts sur une retraite, on ne devrait payer d'impôts que sur des retraites élevées, pas sur les petites retraites* ».

b) Les ouvriers (noyaux 17 et 18)

On a noté une différence considérable de profils lexicaux entre les ouvriers avec enfants au foyer et les autres. Les premiers, visiblement intéressés et motivés, posent des problèmes non abordés, ou prolongent des discussions amorcées plus tôt; les seconds se contentent de faire des commentaires assez laconiques sur le questionnaire. Donnons à titre d'exemple des réponses modales dans le cas de la présence d'enfants au foyer :

— « *Pourquoi faire des différences dans les familles entre 2 et 3 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...* ».

Les interventions de ce type abondent... elles sont souvent très longues, et montrent le besoin de s'exprimer sur ce sujet des individus de cette catégorie; elles contrastent avec les remarques de la catégorie pourtant voisine socio-économiquement des ouvriers sans enfants au foyer. Ces derniers déclarent par exemple :

— « *Non, le questionnaire est très complet* ».

— « *Je ne connais pas l'utilité de votre questionnaire, mais je trouve qu'il est pas mal fait* ».

ou au contraire :

— « *Questionnaire ridicule qui ne peut servir à rien si ce n'est d'être utilisé par la police ou les impôts, je le trouve indiscret* ».

On voit quelle peut être la différence de ton et de contenu de ces deux groupes voisins de personnes enquêtées.

c) Ménagères (noyaux 7, 9, 10)

Ici encore, les réponses sont très diversifiées selon la présence ou non d'enfants au foyer, et également selon le statut social.

Pour le noyau 9 (ménagères modestes, jeunes avec enfants), comme dans le cas des ouvriers avec enfants, l'aspect « critique de l'enquête » est négligé, au profit d'une avalanche de remarques et de revendications sur certains points manifestement très préoccupants pour les personnes interrogées.

Exemples de réponses modales :

— « *On aurait pu aborder le sujet de la femme au foyer, pour la prendre davantage 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* ».

— « *Trop de femmes qui travaillent au détriment des hommes, ça ferait du travail pour les hommes, et la femme s'occuperait de ses enfants (moins d'enfants à la rue)* ».

— « *Le manque d'information pour les femmes au foyer, il faudrait faire quelque chose pour ça...* »

Pour le noyau 10 (Ménagères aisées avec enfants), l'attitude est assez différente : le questionnaire est critiqué, dans sa forme notamment :

— « *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, pour le noyau 7 (ménagères sans enfants au foyer), il y a à la fois des remarques sur l'intervieuw et des réflexions sur les thèmes abordés :

— « *Le libellé des questions limite trop les réponses, et empêche de dire réellement ce que l'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 passionnant* ».

— « *Sur la Sécurité sociale, c'est inadmissible de faire attendre si longtemps les gens sans argent pour liquider un dossier* ».

— « *Pour une mère au foyer, les questions concernant le temps libre et les week-ends ne se posent pas de la même manière : la séparation temps libre - temps de travail n'est pas aussi nette* ».

d) Etudiants, jeunes actifs (noyaux 5 et 19)

Ces deux catégories de jeunes ont en commun le fait d'émettre des opinions très critiques vis-à-vis du questionnaire; toutefois, les jeunes actifs sont beaucoup plus nombreux à se plaindre de la longueur du questionnaire.

Exemples de réponses modales :

— « *Ce questionnaire est vraiment trop long, je suis gênée par la formulation de certaines questions fermées, c'est trop directif, pas assez souple* ».

— « *C'est trop long, ça manque de nuance, j'aurais aimé donner mon avis sur la place des dépenses pour les équipements militaires nuisant à l'environnement* ».

— « *Ce questionnaire implique que les gens qui vous répondent jouent le jeu, il n'est pas apolitique, ce questionnaire, j'ai accepté d'y répondre uniquement sur votre bonne mine...* ».

e) Femmes actives urbaines (noyaux 11, 20, 21)

Ici encore, réponses très différentes selon qu'il s'agit de femmes actives sans enfants au foyer (noyau 20), qui, dans l'ensemble, n'ont aucune remarque à ajouter ni de commentaire à faire; de femmes avec enfants de conditions modestes, qui émettent des réflexions ou des revendications nombreuses et circonstanciées (comme les ménagères avec enfants de conditions modestes, cf. ci-dessus); de femmes avec enfants plus instruites et plus aisées, qui sont plus réservées et plus critiques (noyau 11).

Le fait le plus notable est que les interventions des femmes actives de conditions modestes (avec enfants au foyer) rejoignent celles des ménagères en ce qui concerne le travail des femmes :

— « *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* ».

— « *Penser un peu plus à la femme qui travaille, j'aimerais beaucoup que la femme touche un salaire tout en restant à la maison pour s'occuper de ses enfants* ».

Alors que le ton et les observations des femmes actives « aisées » sont différentes :

— « *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* ».

f) Actifs des deux sexes, ruraux (noyaux 15 et 16)

Très peu de remarques proprement dites ou de critiques, mais un retour sur les thèmes jugés intéressants : la justice, le chômage, et évidemment, les problèmes spécifiquement agricoles.

— « *Il faudrait beaucoup plus de justice dans la société* ».

— « *Il n'y a pas assez de justice sur terre, dans l'agriculture, le revenu devrait être plus sûr* ».

— « *Le questionnaire est surtout fait pour des salariés habitant un milieu urbain, il est fait par des citadins, ignorant les conditions de vie rurales, et les conditions de vie des exploitants agricoles...* ».

— « *Développer les questions sur la justice, et poser des questions sur la peine de mort, pour une autre enquête..* ».

g) Hommes actifs urbains, non ouvriers (noyaux 12, 13, 14)

Dans l'ensemble, on trouve beaucoup de critiques et de remarques sur le questionnaire, avec des questions sur l'utilisation qui va être faite des résultats;

Exemples de réponses modales :

— « *Certaines questions sont peu précises, on demande des réponses par tout ou rien, ma pensée n'est pas aussi sommaire, ça sert à faire des moyennes mais c'est tronquer les débats..* ».

— « *Non, c'est assez complet, on a abordé tout ce qu'il y avait d'important, mais on n'est pas assez préparé à l'enquête, il pourrait y avoir un courrier plus complet avant* ».

— « *Non, je trouve que mon opinion a pu s'exprimer suffisamment, et je pense que les gens en général devraient pouvoir s'exprimer davantage* ».

On voit sur cet exemple que l'utilisation d'une partition instrumentale assez fine permet de détecter certains types d'interactions responsables d'attitudes très différentes. Les lacunes, les a priori, voire l'ethnocentrisme du questionnaire sont dénoncés directement par certaines catégories que la partition a permis d'isoler. L'homogénéité et la spécificité des réponses de certains groupes sont un des résultats surprenants de l'analyse de cette question ouverte pourtant très évasive.

CONCLUSION

La chaîne de traitement se termine par un retour à l'information brute, et cela peut rassurer ceux qui craignent à juste titre les diversions ou l'éloignement du réel que peuvent parfois susciter les approches quantitatives.

Ici le programme classe, édite, pose quelques jalons quantitatifs (fréquences lexicales), représente les associations formes-classes selon la procédure transparente, criticable, reproductible de l'analyse des correspondances (sans arbitraire de codage puisqu'il s'agit ici de vrais tableaux de contingence), et enfin sélectionne les réponses modales selon des critères également explicites. Il n'y a donc pas de « boîte noire », ni même de longue phase sans contrôle dans la suite des opérations. On fera cependant quelques brèves remarques pour

conclure car bien évidemment tous les problèmes ne sont pas résolus, et les problèmes nouveaux sont nombreux.

Les exemples ont montré que pour des enquêtes nationales qui ne sont pas de simples monographies, on peut maintenant avoir accès à une information de base vivante et multiforme. Que cette information se manifeste sous forme de réponses parfois chaleureuses, poignantes, virulentes fait partie intégrante de la réalité sociale à l'étude; l'aridité des matériaux n'est évidemment pas un critère de scientificité...

Cependant, le traitement reste essentiellement exploratoire; s'il est plus puissant que les procédures de post-codage au niveau de la description qualitative des résultats, il ne peut encore remplacer cette technique pour donner une image quantifiée de l'importance des grands thèmes cités. On imagine cependant aisément quelle aide ce genre de traitement peut apporter lors d'une recodification de l'information. Une des directions de recherche actuelle concerne précisément l'inclusion des techniques de post-codification dans la chaîne des opérations.

On a pu noter que la (ou les) réponse(s) modale(s) d'une classe n'est pas la réponse de l'individu modal de la classe, lorsque cette notion existe. Les réponses modales se calculent dans l'espace de formes lexicales, alors que les individus modaux (ou centraux) se calculent dans l'espace des variables ou de leurs caractéristiques de base. Les relations entre « réponses typiques » et « réponses d'individus typiques » méritent une étude particulière.

Une autre approche consiste à décrire les individus directement par leurs réponses libres, soit parce que celles-ci sont suffisamment longues (entretien), soit en juxtaposant simultanément plusieurs réponses relatives à plusieurs questions ouvertes, de façon à ce que le vecteur lexical de description ne soit pas clairsemé, et donc que les calculs de distances aient un sens sans nécessiter de regroupement en classes. Il s'agit bien sûr d'une optique différente, pouvant conduire à des résultats intéressants, qu'il ne nous a pas été donné d'expérimenter, mais qui fait partie de notre programme de recherche actuel.

BIBLIOGRAPHIE SOMMAIRE

BENZECRI (J.P.) et Coll., *Linguistique et lexicologie,* Pratique de l'analyse des données, tome 3, Cf. pp. 414-419 : Vers l'analyse automatique des textes : le traitement des réponses libres aux questions ouvertes d'une enquête, Dunod, 1980.

LEBART (L.), Une procédure d'analyse lexicale écrite en langage FORTRAN, *Les cahiers de l'analyse des données*, Vol. VI, n° 2, Dunod, pp. 229-243, 1981.

LEBART (L.) et HOUZEL VAN EFFENTERRE (Y.), Le système d'enquête sur les aspirations des Français : une brève présentation, *Consommation,* n° 1, 1980, pp. 3-25.