ANALYSE STATISTIQUE DES RÉPONSES AUX QUESTIONS OUVERTES

Ludovic Lebart

1 - Deux grandes familles de problèmes dans l'étude statistique des textes

Dans les analyses statistiques de textes, on peut distinguer deux grandes séries de préoccupations fort différentes :

- les applications à des corpus littéraires (attributions d'auteurs, datation, par exemple) qui cherchent à s'affranchir du contenu pour saisir des caractéristiques de **forme** (souvent : de style) à partir des distributions statistiques de vocabulaire, d'indices ou de ratios, ou encore à partir de corpus partiels de mots outil. Il s'agit de saisir les "invariants" d'un auteur ou d'une époque, dissimulés ou peu apparents, à des fins historiques, littéraires, dans le cadre d'études que l'on désigne sous le nom de stylométrie (cf. par exemple Yule, 1944, et : Holmes, 1985, pour une revue de ces travaux) ;

- les applications réalisées en recherche documentaire (information retrieval : cf. Salton, 1988), en codification automatique, dans le traitement des réponses à des questions ouvertes, qui s'intéressent principalement au **contenu**, au sens, à la substance des textes. Peu importe la façon dont une requête est rédigée, pourvu que l'on puisse atteindre dans la base de données les documents qui satisfont l'attente du requérant. Toutefois, il faut signaler que lors du traitement statistique de réponses à des questions ouvertes ou d'entretiens, le socio-linguiste peut être *aussi* intéressé par la forme, par les connotations véhiculées par exemple par certains synonymes, certaines tournures. Il s'agit de caractéristiques de formes qui peuvent en fait nuancer et infléchir le fond.

Les méthodes d'analyses de réponses libres dans les enquêtes relèvent de cette seconde famille de méthodes.

2 - Questions ouvertes et questions fermées

Il peut être intéressant, dans un certain nombre de situations d'enquête, de laisser ouvertes certaines questions, dont les réponses se présenteront donc sous forme de

textes de longueurs variables. Le traitement de ce type d'information est évidemment complexe. Les outils de calcul et les méthodes statistiques descriptives multidimensionnelles peuvent apporter une certaine aide à l'analyse de ces *réponses libres.*

On rappellera auparavant quelques uns des problèmes posés par la rédaction des libellés des questions dans les questionnaires d'enquêtes.

2.1 - Le libellé des questions

On sait que le libellé d'une question joue un rôle fondamental : il est très difficile de trouver deux libellés distincts, pour deux questions fermées dont les contenus sont similaires, donnant les mêmes résultats en termes de pourcentages.

La sensibilité des pourcentages de réponses vis-à-vis des libellés est bien sûr particulièrement forte dans le cas de questions d'attitudes ou d'opinions.

Ainsi, les travaux de Rugg, 1941, ont montré que la réponse *yes* à la question *Do you think the United States should forbid public speeches against democracy?* obtient 21 points (sur 100) de moins que la réponse no à la question *Do you think the United States should allow public speeches against democracy?*

Cette absence de symétrie entre les deux formulations, vérifiées sur d'autres thèmes, est d'autant plus forte que le niveau d'instruction de la personne qui répond est faible. Elle rend plus difficiles les études des phénomènes d'acquiescement systématique (cf par exemple Tabard, 1975). L'équivalence sémantique de deux questions qui appelleraient respectivement des réponses oui et non paraît impossible à atteindre.

À ces remarques sur la rédaction des libellés s'ajoutent d'autres considérations :

- l'ordre des questions, qui induit une sensibilisation particulière du répondant ;

- la longueur des libellés qui fait jouer, selon les cas, la mémoire auditive ou les capacités de lecture de la personne interrogée, et donc induit des biais en fonction de certaines caractéristiques de base comme l'âge, le niveau d'instruction (cf. la contribution de J.-P. Grémy dans ASU, 1992).

Le problème de la dépendance des résultats vis-à-vis des libellés se pose *a fortiori* dans le cas de deux questions dont l'une est ouverte et l'autre fermée. Un exemple classique concerne les réponses à la question "Quel est le problème le plus important auquel doivent faire face les USA ?", (Schuman et al., 1981). L'item "violences" obtient 16 % lorsque la question est ouverte, et 32 % lorsqu'il fait partie des items de la question fermée correspondante. Cet item de réponse étant considéré comme "un problème

local" plutôt que "national" n'est pas toujours considéré comme une réponse permise lorsque la question est ouverte. En somme, les libellés complets de deux questions, l'une ouverte et l'autre fermée, ne peuvent être identiques, ce qui rend extrêmement difficiles les comparaisons entre les deux types de questionnement.

2.2 - Quand utiliser des questions ouvertes ?

Dans au moins trois situations courantes, l'utilisation d'un questionnement ouvert s'impose :

Pour diminuer le temps d'interview
Bien que les réponses libres et les réponses guidées fournissent des informations de natures différentes, les premières sont plus économiques que les secondes en temps d'interview et génèrent moins de fatigue. Une simple question ouverte (par exemple : "Quelles sont vos activités de loisir habituelles") peut remplacer de très longues listes d'items.

Comme complément à des questions fermées
Il s'agit le plus souvent de la question classique : "*Pourquoi ?*". Les explications concernant une réponse déjà donnée doivent nécessairement être spontanées. Une batterie d'items risquerait de proposer de nouveaux arguments qui pourraient nuire à l'authenticité de l'explication.

L'utilité de la question *pourquoi ?* a été soulignée par de nombreux auteurs, et ce sont en fait les difficultés et le coût de l'exploitation qui en limitent l'usage. Elle seule permet en effet de savoir si les différentes catégories de personnes interrogées ont compris la question fermée de la même façon.

Elle est particulièrement importante dans les enquêtes internationales, car elle permet de juger les éventuelles différences sémantiques des libellés selon la langue utilisée.

Prenons un exemple : à la question "Selon vous, la famille est-elle le seul endroit où l'on se sent bien et détendu", 93 % des personnes habitant en milieu rural, âgées et peu instruites répondent : *oui*, alors que ce n'est le cas que de 12 % des jeunes (moins de 25 ans) instruits (au moins le baccalauréat) de l'agglomération parisienne (cf. Lebart, 1986).

Cette dispersion considérable des pourcentages serait inférieure si l'on prenait en compte séparément les trois facteurs sous-jacents (âge, type d'agglomération et niveau d'instruction), ou si on ne les prenait en compte que deux par deux. À partir de ce fait statistique, plusieurs questions se posent. Des questions préliminaires telles que : l'assertion proposée a-t-elle le même sens d'une génération à une autre, d'un âge à un autre, d'une région à une autre, pour une personne ayant fait des études et pour une personne faiblement scolarisée ? Le mot *famille*, en particulier, a-t-il la même signification pour toutes ces catégories ?

Ici encore, une question ouverte complémentaire du type "Pourquoi ?" est bienvenue, et les discours des différentes catégories précitées (discours obtenus en juxtaposant les réponses) sont intéressants à comparer. Interpréter des différences de pourcentages est une activité de base dans le traitement des données d'enquêtes : pourquoi ne pas inclure dans les matériaux à interpréter le point de vue des répondants eux-mêmes ?

Pour recueillir une information qui doit être spontanée
Les questionnaires des enquêtes de marketing abondent en questions de ce type. Citons par exemple : "Qu'avez-vous retenu de ce spot publicitaire ?", "Que pensez-vous de cette voiture ?".

Notons que les questions ouvertes sont considérées comme peu adaptées aux problèmes de mémorisation de comportement. "Quels sont les noms des magazines que vous avez lus la semaine dernière ?" "Quelles sont les dernières émissions de télévision que vous avez aimées ?" Pour ces questions qui font l'objet d'enquêtes périodiques, il a été prouvé maintes fois que les questions fermées donnent des taux d'oubli plus faibles (Belson et Duncan, 1962).

En revanche, quand la qualité de la mémorisation est en jeu (préoccupation très courante en marketing, lorsqu'il s'agit d'évaluer l'impact d'actions publicitaires), la forme ouverte est indispensable.

Lazarsfeld, 1944, préconise l'usage des questions ouvertes principalement dans une phase préparatoire ; leur finalité est alors la mise au point d'une batterie d'items de réponses pour une question fermée. Cette utilisation est toujours recommandée, mais assez rarement réalisée en raison de son coût : obtenir une liste d'items incluant ceux qui sont peu fréquents nécessite en effet une pré-enquête pilote assez lourde.

2.3 - Traitement pragmatique des questions ouvertes

Les procédures dites de "post-codage" permettent de fermer *a posteriori* les questions ouvertes. Ces techniques consistent à construire une batterie d'items à partir d'un sous-échantillon de réponses, puis à codifier l'ensemble des réponses de façon à remplacer la question ouverte par une ou plusieurs questions fermées. Pour des réponses simples, stéréotypées et peu nombreuses, cette procédure n'a que peu d'inconvénients. Mentionnons cependant parmi les défauts de ce type de traitement :

L'équation personnelle du chiffreur
À la médiation de l'enquêteur s'ajoute celle du chiffreur, qui doit prendre de nombreuses décisions difficiles et parfois contestables par le spécialiste.

La destruction de la forme
La qualité de l'expression, le registre du vocabulaire, la tonalité générale de l'entretien sont des éléments d'analyse perdus lors d'un post-codage.

L'appauvrissement du contenu
Les réponses composites, complexes, d'une grande diversité, sont littéralement laminées par le post-codage et c'est souvent dans ce cas que la valeur heuristique des réponses libres est la plus grande.

Prenons l'exemple de la question "Qu'est-ce qui vous inquiète en ce qui concerne l'avenir" et la réponse relativement simple : "J'ai peur de tomber malade et d'être seule, le reste ne me fait pas peur". Les deux thèmes maladie et solitude sont certainement ici en interaction, mais devront probablement être codés séparément comme premier et second items. La seconde partie de la réponse est impossible à coder, puisqu'elle se définit par rapport au "complémentaire" des deux premières, alors qu'elle contient une information importante.

Les réponses rares sont éliminées a priori
Les réponses peu fréquentes, originales, peu claires en première lecture sont affectées à des items "résiduels" qui sont donc très hétérogènes et perdent de ce fait toute valeur opératoire.

Ces réponses relativement peu fréquentes peuvent cependant être émises par une catégorie d'individus très particulière, et présenter un grand intérêt au niveau de l'interprétation des résultats, ce qu'il n'est pas possible de savoir lors d'un traitement "*a priori*" de l'information...

Ainsi, à la question mentionnée ci-dessus sur les inquiétudes concernant l'avenir, le thème *justice* n'est cité que six fois sur 1 000. Mais le fait qu'il soit cité cinq fois sur six par des agriculteurs suggère de coder l'information correspondante, malgré sa fréquence très faible.

3 - Les unités statistiques découpées dans les textes

3.1 - Les formes graphiques

L'unité de base sera la forme graphique définie comme une suite de caractères non-délimiteurs (en général des lettres) entourée par des caractères délimiteurs (blanc, points, virgules...). Un même mot pourra en général donner lieu à plusieurs formes graphiques, selon son cas ou son genre dans le texte. Une même forme graphique peut renvoyer à plusieurs mots (en français, LIT renvoie à un nom, mais aussi au verbe lire). Cela n'est pas toujours un inconvénient grave, car les formes graphiques ne seront pas traitées isolément.

Les traitements statistiques concerneront en effet les *profils de fréquences de formes graphiques*, c'est-à-dire les vecteurs dont les composantes sont les fréquences de chacune des formes utilisées par un individu ou un groupe d'individus. Ces profils contiennent une information extrêmement riche. Plus précisément encore, les techniques mettront en évidence les *différences entre profils de formes graphiques* (Lebart et Salem, 1988).

Si l'interprétation dans l'absolu d'un profil peut être délicate (i.e. : pourquoi telle catégorie socio-professionnelle utilise-t-elle tels mots avec telles fréquences?), l'interprétation des différences entre profils est plus aisée : sans spéculer sur la signification des profils, on peut très bien observer que, par exemple, les cadres et les employés ont des profils proches, éloignés de celui des ouvriers. En simplifiant à l'extrême, on peut résumer cette approche différentielle par la formule : il n'est pas utile de comprendre ce que deux catégories ont exprimé dans leurs réponses à une question ouverte pour savoir qu'elles n'ont pas dit la même chose.

3.2 - Les segments répétés

La notion de forme graphique peut être généralisée en procédant à des comptages portant sur des unités plus larges, composées de plusieurs formes : les segments répétés. On observe en effet dans les réponses les apparitions récurrentes d'unités comme *je ne sais pas*, *sécurité d'emploi*, *justice sociale*, dotées parfois d'un sens qui leur est propre et que l'on ne peut pas toujours déduire à partir du sens des formes qui entrent dans leur composition (Salem, 1987). Il est alors possible de reprendre les traitements avec les segments pour compléter les formes graphiques. Les résultats sont considérablement enrichis par l'introduction du contexte des formes, qui lève la plupart des ambiguïtés de sens.

Pour sélectionner formes et segments, des seuils de fréquence vont intervenir. Ils permettront d'effectuer différents filtres sur l'information de base.

3.3 - Les unités lemmatisées

Un autre type de traitement préliminaire du texte consiste à procéder à une "lemmatisation". Cette opération, très difficile à réaliser de façon entièrement automatique, consiste à remplacer les formes par l'entrée du dictionnaire correspondant (infinitif pour les verbes, masculin singulier pour les adjectifs, formes non élidées à la place des formes élidées, etc.), et parfois à supprimer certains mots-outils (articles, conjonctions, etc, cf. par exemple Reinert, 1986). En documentation automatique, cela permet de travailler avec un nombre restreint de mots-clés dont les occurrences sont fréquentes.

En traitement de questions ouvertes, cette opération n'est pas toujours souhaitable *a priori* car elle détruit les locutions et modifie assez profondément la forme des réponses, qui peuvent intéresser le socio-linguiste. En revanche, elle peut intervenir comme complément, car elle fournit un point de vue différent sur les textes. Dans le cas d'entretiens non directifs peu nombreux, la lemmatisation permet de travailler avec des seuils de fréquences plus élevés que ceux nécessités par l'analyse des formes graphiques.

Des formes graphiques différentes d'un même mot peuvent être liées à un contexte et à un contenu particulier, et certains mots-outils peuvent caractériser de façon élective des attitudes ou opinions. Ainsi, en réponse à une question sur la nature du mariage, les réponses traditionnalistes contiennent souvent quand : "quand on se marie c'est pour la vie", et se présentent fréquemment comme des discours à la première personne : "je suis contre le divorce", "nous sommes croyants", contrairement aux réponses plus modernistes, qui contiennent souvent parce que, et un ton plus impersonnel : "parce que le mariage est un contrat comme un autre".

Tableau 1 : Vocabulaire des formes apparaissant au moins 20 fois
[Question : Que pensez-vous de cette enquête ? cf. § 4]

Num.	Mots (formes)	Fréq.
1	A	378
2	AI	35
3	AIME	23
4	ASSEZ	93
5	AU	45
6	AUCUNE	108
7	AUX	38
8	BEAUCOUP	40
9	BIEN	83
10	C	182
11	CA	78
12	CE	100
13	CELA	35
14	CERTAINES	70
15	CHÔMAGE	48
16	COMME	24
17	COMPLET	39
18	D	126
19	DANS	74
20	DE	557
21	DES	242
22	DIFFICILE	25
23	DIRE	22
24	DONNER	33
25	DU	88
26	EN	87
27	ENFANTS	26
28	ENQUÊTE	36
29	EST	322
30	ET	186
31	ÊTRE	34
32	FAIRE	30
33	FAIT	48
34	FAUDRAIT	23
35	FAUT	26
36	GENS	38
37	IL	179
38	ILS	22
39	INTÉRESSANT	43
40	J	72
41	JE	130
42	L	170
43	LA	234
44	LE	210
45	LES	326
46	LEUR	23
47	LONG	73
48	MAIS	59

Num.	Mots (formes)	Fréq.
49	MAL	35
50	MANQUE	33
51	ME	33
52	N	100
53	NE	172
54	NON	572
55	ON	190
56	OU	60
57	OUI	36
58	PAR	66
59	PARLE	28
60	PARLER	35
61	PAS	356
62	PEU	53
63	PEUT	43
64	PLUS	95
65	POUR	141
66	PROBLÈME	26
67	PROBLÈMES	28
68	QU	59
69	QUE	164
70	QUESTION	37
71	QUESTIONNAIRE	119
72	QUESTIONS	215
73	QUI	93
74	REMARQUES	25
75	RÉPONDRE	71
76	RÉPONSES	84
77	RIEN	81
78	RIGIDE	22
79	S	25
80	SANS	21
81	SE	29
82	SI	29
83	SONT	97
84	SUR	117
85	TEMPS	22
86	TOUJOURS	24
87	TOUT	49
88	TRAVAIL	36
89	TRÈS	69
90	TROP	178
91	UN	117
92	UNE	63
93	VA	22
94	VIE	44
95	VOUS	21
96	Y	75

4 . La numérisation du texte

Cette phase de traitement préliminaire consiste à affecter à chaque nouvelle forme graphique un numéro d'ordre qui sera associé à toutes les occurrences de cette même forme. Ces numéros seront stockés dans un dictionnaire de formes, ou vocabulaire, propre à chaque exploitation. Ce dernier permettra, à l'issue des calculs ou lors des impressions, de reconstituer le graphisme des formes mises en évidence par les calculs statistiques.

Les exemples qui suivent sont empruntés à une question ouverte de nature méthodologique posée en 1984 (Enquête Conditions de Vie et Aspirations des Français, cf. Lebart, 1986) à l'issue de l'interview, dont le libellé était :

"Vous venez d'être interrogés longuement sur vos conditions de vie, y-a-t-il des sujets importants que vous auriez aimé voir aborder ? Avez-vous des remarques à formuler ?".

Le *tableau 1* représente les 96 formes apparaissant au moins 20 fois dans un échantillon de 2000 réponses à cette question.

On observe comme prévu des formes se rapportant à un même mot (problème, problèmes), des mots-outils (dans, des, par, que, qui...). Comme cela a été dit plus haut, la lemmatisation et l'apurement ne s'imposent pas dans une approche différentielle portant sur des échantillons importants.

Si les mots-outils sont répartis de façon aléatoire dans les diverses catégories d'individus, ils ne sont pas gênants. S'ils ne le sont pas, ils sont au contraire intéressants. De façon analogue, si deux formes graphiques se rapportant à un même mot ont des comportements identiques, elles peuvent aussi bien être remplacées par ce mot. Si elles ont des comportements différents, c'est qu'elles renvoient à des contextes d'utilisation du mot différents, ce qui mérite d'être relevé.

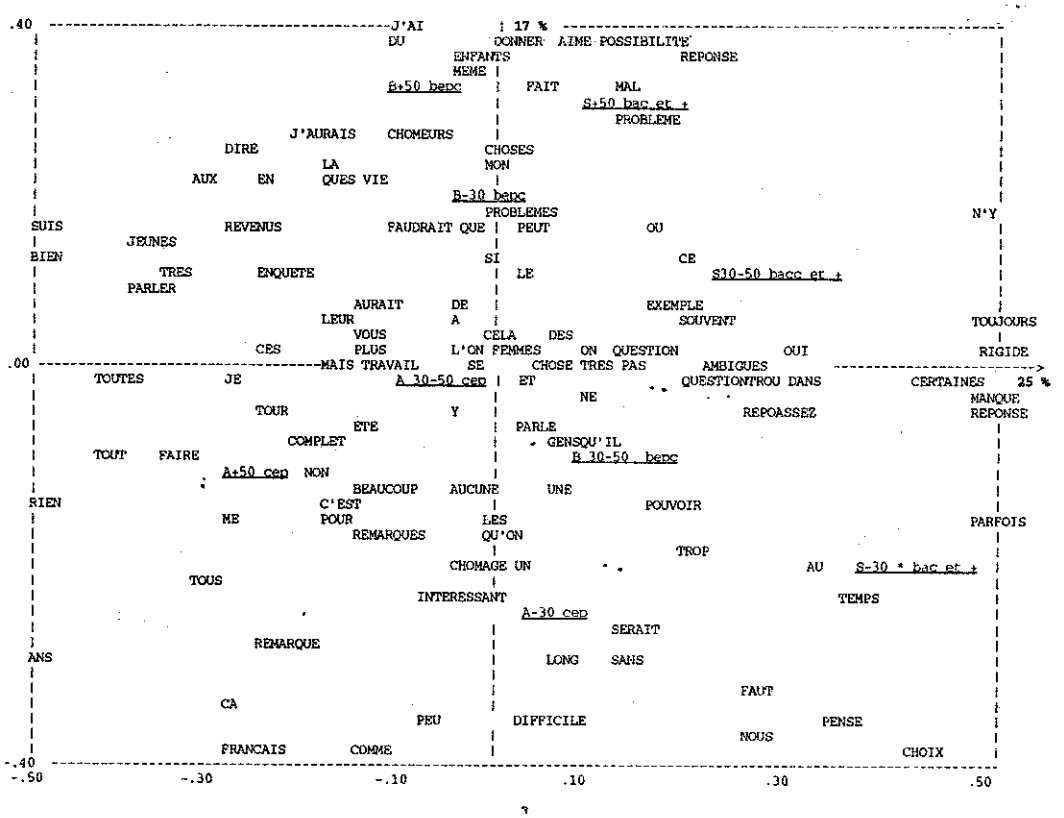
Le *tableau 2* décrit ainsi, toujours pour les 2000 réponses qui nous servent d'exemple illustratif, les différents segments observables, classés selon l'ordre alphabétique de la première forme graphique qui les compose, et sélectionnés en fonction de seuils de fréquences : les segments de longueur 2 (très nombreux, et pauvres du point de vue de leur apport sémantique) doivent apparaître au moins 50 fois, alors que ceux de longueur supérieure ou égale à 3 doivent apparaître au moins 6 fois pour figurer dans l'inventaire.

On voit qu'il s'agit d'éléments d'information auxiliaires, largement interdépendants, mais permettant d'identifier les contextes des formes les plus fréquentes. Une sélection s'impose : il est relativement aisé de choisir dans cette liste (établie à partir de seuils sévères, pour limiter le volume des éditions) les segments porteurs d'une information sémantique spécifique.

Tableau 2 : Inventaire partiel de segments répétés. Seuils minimum de fréquence de répétition : seuil général = 6,segments de longueur 2 = 50,segments de longueur 3 = 6

Seg	Freq	Long	Texte du segment
			A
1	14	3	A CERTAINES QUESTIONS
2	12	3	A DES QUESTIONS
3	7	3	A PAS DE
			BEAUCOUP
4	7	3	BEAUCOUP DE CHOSES
			C'
5	175	2	C' EST
6	11	3	C' EST ASSEZ
7	17	3	C' EST BIEN
8	12	3	C' EST INTÉRESSANT
9	8	3	C' EST LA
10	8	3	C' EST PAS
11	18	3	C' EST TRÈS
12	15	3	C' EST TROP
13	14	3	C' EST UN
14	7	4	C' EST TROP LONG
15	7	4	C' EST UN QUESTIONNAIRE
			CE
16	11	3	CE N'EST
17	8	3	CE QUESTIONNAIRE EST
18	10	4	CE N'EST PAS
19	7	4	CE QUE L'ON
			CERTAINES
20	60	2	CERTAINES QUESTIONS
21	11	3	CERTAINES QUESTIONS SONT
			DE
22	59	2	DE LA
23	7	3	DE LA VIE
24	7	3	DE QUESTIONS SUR
25	11	3	DE REPONDRE A
			DIFFICILE
26	12	3	DIFFICILE DE RÉPONDRE
			EST
27	7	3	EST ASSEZ COMPLET
28	10	3	EST DIFFICILE DE
29	7	3	EST TRÈS BIEN
30	7	3	EST TRÈS COMPLET
31	9	3	EST TROP LONG
			ET
32	7	3	ET C'EST
			FAIT
33	11	3	FAIT LE TOUR
			IL
34	13	3	IL N'Y
35	40	3	IL Y A
36	11	4	IL N'Y A
37	11	4	IL Y A DES
38	9	4	IL Y A TROP
39	8	5	IL N'Y A PAS
40	9	5	IL Y A DES QUESTIONS
			J
41	10	3	J'AURAIS AIMÉ
			LA
42	9	3	LA POSSIBILITÉ DE
			LE
43	22	3	LE QUESTIONNAIRE EST
44	7	3	LE TOUR DE
			LES
45	22	3	LES QUESTIONS SONT
46	8	4	LES QUESTIONS SONT TROP
			MAIS
47	7	3	MAIS C'EST
			N
48	11	3	N' A PAS
49	28	3	N' EST PAS
50	12	3	N' Y A
			NE
51	18	3	NE SONT PAS
52	8	4	NE SONT PAS ASSEZ
			NON
53	19	3	NON C'EST
			ON
54	9	3	ON A FAIT
55	22	3	ON NE PEUT
56	8	4	ON A FAIT LE
57	19	4	ON NE PEUT PAS
58	7	5	ON A FAIT LE TOUR
59	7	5	ON NE PEUT PAS REPONDRE
			OUI
60	11	3	OUI OU NON
			PAR
61	10	3	PAR OUI OU
			PAS
62	56	2	PAS ASSEZ
63	16	3	PAS ASSEZ DE
64	23	3	PAS DE REMARQUES
			QUE
65	8	3	QUE C'EST
66	21	3	QUE L'ON
			QUESTIONNAIRE
67	7	3	QUESTIONNAIRE EST BIEN
68	7	3	QUESTIONNAIRE TROP RIGIDE
69	7	4	QUESTIONNAIRE N'EST PAS
			QUESTIONS
70	10	3	QUESTIONS SONT TROP
			REPONDRE
71	9	4	RÉPONDRE PAR OUI OU
72	8	6	RÉPONDRE PAR OUI OU NON
			SONT
73	9	3	SONT PAS ASSEZ
			UN
74	9	3	UN PEU LONG
			Y
75	55	2	Y A
76	11	3	Y A PAS
77	10	3	Y A TROP

Figure 1 : Visualisation par analyse des correspondances de la table de contingence croisant les formes avec la variable nominale à 9 modalités : Âge. Diplôme(Question ouverte posée à l'issue d'une enquête, cf. § 4)



5 - Les tableaux lexicaux

Les réponses libres peuvent être numérisées de façon complètement transparente pour l'utilisateur. Le résultat de cette numérisation peut prendre deux formes différentes, matérialisées par deux matrices R et T. La matrice R a k lignes, k désignant le nombre de réponses, et un nombre de colonnes égal à la longueur de la plus longue réponse (nombre d'occurrences de formes dans cette réponse).

Pour une réponse ou un individu "i", la ligne "i" de R (tableau de pointeurs) contient les adresses (relatives à un dictionnaire ou *vocabulaire*) des formes graphiques qui

composent la réponse, en respectant l'ordre et les éventuelles répétitions de ces formes. R permet donc de restituer intégralement les réponses originales.

R n'est pas rectangulaire, car chacune de ses lignes a une longueur variable. Les nombres entiers qui composent R ne peuvent dépasser v, longueur du vocabulaire (nombre de formes graphiques distinctes).

La matrice T a le même nombre k de lignes que R, mais possède autant de colonnes qu'il y a de formes graphiques utilisées par l'ensemble des individus, c'est-à-dire v (v = vocabulaire) colonnes. A l'intersection de la ligne i et de la colonne j de T figure le nombre de fois où la forme j a été utilisée par l'individu i dans sa réponse. Il s'agit donc d'une table de contingence "Individus-Formes". T peut être aisément construite à partir de R, mais la réciproque n'est pas vraie : l'information relative à l'ordre des formes dans chaque réponse est perdue dans T.

En fait, R est beaucoup plus compacte que T : ainsi, une réponse contenant 20 occurrences (pour un lexique de 1 000 formes) correspond à une ligne de longueur 20 de R et à une ligne de longueur 1 000 de T (cette dernière ligne comprenant au moins 980 zéros...). Les calculs statistiques et algorithmiques qui mettront en jeu T sont en réalité programmés à l'aide de R, moins encombrante en mémoire.

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

On désignera par Z le tableau disjonctif complet à k lignes et p colonnes décrivant les réponses de k individus à une question fermée comportant p modalités de réponses possibles.

C = T' Z est un tableau à v lignes et p colonnes dont le terme général c_{ij} n'est autre que le nombre de fois où la forme "i" a été utilisée dans une réponse libre par l'ensemble des individus ayant choisi la réponse "j" à une question fermée.

Il est donc aisé, pour toute question fermée dont les réponses sont codées dans un tableau Z_q, de calculer le tableau lexical agrégé C_q par la formule :

$$C_q = T'Z_q$$

Ces comparaisons de profils lexicaux n'ont de sens, d'un point de vue statistique, que si les formes apparaissent avec une certaine fréquence : les hapax (formes n'apparaissant qu'une fois), ou même les formes rares seront écartés de la phase de comparaisons de fréquences. Ceci a pour effet de réduire la taille du vocabulaire v. Pour une question ouverte posée à 1 000 personnes, une sélection des formes apparaissant au moins 8 fois peut, dans bien des cas, diviser par 10 la valeur de v (de 1500, pour fixer les idées, à 150).

Pour la question posée à l'issue d'une enquête qui nous sert d'exemple, il y a, au départ, pour 2 000 réponses 12 866 occurrences, avec 2 035 formes distinctes ; on a vu (*tableau 1*) qu'il n'y a pour cette question que 96 formes apparaissant au moins 20 fois.

Dans la table de contingence C_q dont l'analyse produit la figure 1, (qui correspond à une sélection des formes apparaissant au moins 14 fois) on retient 128 formes, qui totalisent 8805 occurrences.

Trois outils vont permettre d'aider la lecture des tableaux lexicaux agrégés : l'analyse des correspondances, les listes de formes caractéristiques, les listes de réponses modales.

5.1 - Analyse des correspondances des tableaux lexicaux

Les analyses des correspondances peuvent décrire les tableaux C_q qui sont des tables de contingence (dont les "individus" sont des occurrences de formes, et non plus des individus interrogés...). Elles permettent de visualiser les associations entre mots (formes) et groupes ou modalités. Ainsi, une visualisation des proximités entre mots et catégories socio-professionnelles pourra aider la lecture des réponses de chacune de ces catégories.

Avec ce type de représentation, la présence de mots-outils est parfaitement justifiée : si ces mots caractérisent électivement certaines catégories, comme le mot *quand* évoqué plus haut, ils se positionnent dans leur voisinage, et peuvent être intéressants à interpréter ; si au contraire leur répartition est aléatoire, ils s'abîmeront dans la partie centrale du graphique, sans en encombrer la lecture.

De même, la présence de plusieurs flexions d'un même verbe constitue un outil de validation. Lors d'une représentation des réponses à une question sur le mariage (Lebart, 1982b), les formes *doivent, doit, devraient*, s'opposent à *peuvent, pouvoir, peut*, ce qui valide l'opposition entre les deux notions sous-jacentes.

La figure 1 représente les positions des 128 formes apparaissant au moins 14 fois, (dont les plus fréquentes sont représentées sur le tableau 1) dans le plan factoriel principal de l'analyse des correspondances de la matrice C, qui est, dans cet exemple de dimension réduite, une matrice d'ordre (128, 9). Rappelons que l'individu statistique est "l'occurrence d'une forme" (et non pas l'individu enquêté).

Pour les 9 classes de la variable Age-Diplôme, cette représentation donne une idée rapide des principales oppositions et associations entre profils de vocabulaires. Il s'agit, rappelons-le, d'une aide à la lecture des réponses regroupées suivant ces catégories. Pour les réponses à cette question, le niveau de diplôme est plus discriminant que l'âge (les trois catégories de niveau "Bacc et +" sont sur la droite du premier axe), bien que

Tableau 3 : Spécificités (formes caractéristiques) pour 4 catégories de la variable à 9 modalités : Âge - Diplôme

Libellé de la forme graphique		Pourcentage		Fréquence		V.test	Proba
		Interne	Global	Interne	Globale		
Texte numéro 1 : Moins de 30 ans, sans diplôme ou CEP							
1	INTÉRESSANT	1,43	0,49	7,0	43,0	2,373	0,009
2	NON	9,20	6,50	45,0	572,0	2,306	0,011
3	C'EST	3,27	1,99	16,0	175,0	1,829	0,034
4	CHÔMAGE	1,23	0,55	6,0	48,0	1,661	0,048
5	TROP	3,07	2,02	15,0	178,0	1,478	0,070
6	LE	3,48	2,39	17,0	210,0	1,436	0,076
7	LES	4,91	3,70	24,0	326,0	1,309	0,095
8	CHOIX	0,61	0,23	3,0	20,0	1,303	0,096
9	TOUT	1,02	0,56	5,0	49,0	1,107	0,134
10	AUCUNE	1,84	1,23	9,0	108,0	1,057	0,145
Texte numéro 3 : Moins de 30 ans, Bacc. ou Université							
1	MANQUE	1,07	0,37	10,0	33,0	2,932	0,002
2	TROP	3,41	2,02	32,0	178,0	2,875	0,002
3	CERTAINES	1,71	0,80	16,0	70,0	2,827	0,002
4	CHOIX	0,75	0,23	7,0	20,0	2,708	0,003
5	RÉPONSES	1,81	0,95	17,0	84,0	2,478	0,007
6	ASSEZ	1,92	1,06	18,0	93,0	2,386	0,009
7	QUESTIONS	3,62	2,44	34,0	215,0	2,266	0,012
8	RÉPONDRE	1,39	0,81	13,0	71,0	1,808	0,035
9	LONG	1,39	0,83	13,0	73,0	1,716	0,043
10	PAS	5,12	4,04	48,0	356,0	1,646	0,050
Texte numéro 7 : Plus de 50 ans, sans diplôme ou CEP							
1	NON	8,39	6,50	161,0	572,0	3,663	0,000
2	ÇA	1,56	0,89	30,0	78,0	3,245	0,001
3	RIEN	1,56	0,92	30,0	81,0	3,034	0,001
4	BIEN	1,46	0,94	28,0	83,0	2,414	0,008
5	POUR	2,24	1,60	43,0	141,0	2,348	0,009
6	JE	2,03	1,48	39,0	130,0	2,118	0,017
7	TRÈS	1,20	0,78	23,0	69,0	2,106	0,018
8	ANS	0,36	0,16	7,0	14,0	2,087	0,018
9	COMME	0,52	0,27	10,0	24,0	2,005	0,022
10	REMARQUES	0,52	0,28	10,0	25,0	1,876	0,030
TEXTE numéro 9 : Plus de 50 ans, Bacc. ou Université							
1	OUI	2,18	1,06	15,0	93,0	2,552	0,005
2	J'AI	0,73	0,24	5,0	21,0	2,046	0,020
3	EN	1,60	0,89	11,0	78,0	1,761	0,039
4	QUESTIONS	3,49	2,44	24,0	215,0	1,668	0,048
5	MAL	0,87	0,40	6,0	35,0	1,629	0,052
6	EST	2,03	1,29	14,0	114,0	1,554	0,060
7	PAR	1,31	0,75	9,0	66,0	1,477	0,070
8	AUX	0,87	0,43	6,0	38,0	1,460	0,072
9	QUESTIONNAIRE	2.03	1.35	14	119	1,406	0,080
10	IL	2.62	1.85	18	163	1,373	0,085

cet axe oppose principalement (opposition confirmée par les calculs de contributions non publiées ici), les deux catégories extrêmes S-30 (Jeunes instruits) et A+50 (plus de cinquante ans, aucun diplôme ou CEP).

5.2 - Les listes des formes caractéristiques (ou spécificités)

Il est tentant de compléter les représentations spatiales fournies par l'analyse des correspondances par quelques paramètres d'inspiration plus probabiliste : les spécificités ou formes caractéristiques. Ce seront les formes "anormalement" fréquentes dans les réponses d'un groupe d'individus.

À ces formes caractéristiques sont attachées des "valeurs-tests" qui 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.

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 -1.96 et + 1.96. Mais ce calcul reposant sur une approximation normale de la loi hypergéométrique n'est utilisé que lorsque les effectifs concernés ne sont pas trop faibles.

Toujours pour la question posées à l'issue de l'enquête précitée, le tableau 3 donne les 10 formes les plus caractéristiques de chacune des 4 classes extrêmes de la partition en 9 postes Age-Diplôme (classes d'âges extrêmes croisées avec les niveaux de diplôme extrêmes).

Les seuils de probabilités qui figurent à côté des valeurs-tests sont des seuils unilatéraux à droite. On vérifiera le caractère complémentaire de l'information du tableau 3 par rapport aux compromis géométriques de la figure 1, établie à partir de la même table de contingence.

5.3 - Les sélections des réponses modales

Pour une classe donnée, et donc pour le regroupement de réponses correspondant, les réponses modales (ou encore phrases caractéristiques, ou documents-types, selon les domaines d'application) sont des réponses originales du corpus de base, ayant la propriété de caractériser au mieux la classe.

Tableau 4 : Réponses modales pour 4 catégories de la variable à 9 modalités : Âge - Diplôme (10 réponses les plus caractéristiques) (Critère de sélection 1 : mots caractéristiques)

Critère	Réponse ou individu
	Texte numéro 1 : Moins de 30 ans, sans diplôme ou CEP
2.306 - 1	NON
2.306 - 2	NON
2.306 - 3	NON
2.306 - 4	NON
2.306 - 5	NON
2.306 - 6	NON
2.306 - 7	NON
2.306 - 8	NON
2.306 - 9	NON
2.306 - 10	NON
	Texte numéro 3 : Moins de 30 ans, Bacc. ou Université
2.296 - 1	TROP LONG
2.296 - 2	TROP LONG
1.811 - 3	QUESTIONNAIRE TROP RIGIDE
1.784 - 4	RÉPONSES TROP ABSOLUES
1.506 - 5	QUESTIONNAIRE TROP LONG,PAS ASSEZ DE NUANCE POUR CERTAINES RÉPONSES
1.466 - 6	QUESTIONNAIRE PAS ASSEZ NUANCÉ POUR CERTAINES QUESTIONS
1.445 - 7	QUESTIONS ASSEZ ÉVASIVES, PAS ASSEZ PERSONNALISÉES, TROP GÉNÉRAL
1.345 - 8	MANQUE DE POSSIBILITÉS DE RÉPONSES, RÉPONSES TROP TRANCHÉES
1.321 - 9	CERTAINES QUESTIONS SONT SANS ALTERNATIVE
1.259 - 10	PAS ASSEZ DE QUESTIONS OUVERTES
	Texte numéro 7 : Plus de 50 ans, sans diplôme ou CEP
3.663 - 1	NON
3.663 - 2	NON
3.663 - 3	NON
3.663 - 4	NON
3.663 - 5	NON
3.663 - 6	NON
3.663 - 7	NON
3.663 - 8	NON
3.663 - 9	NON
3.663 -10	NON
	Texte numéro 9 : Plus de 50 ans, Bacc. ou Université
0.632 - 1	IL Y A PLEIN DE QUESTIONS QUI SONT MAL POSÉES
0.600 - 2	NON, J'AI L'IMPRESSION QU'IL EST COMPLET
0.576 - 3	IL Y A PAS MAL DE QUESTIONS AMBIGUËS, QUESTIONS À PIÈGE
0.573 - 4	QUESTIONNAIRE CONDITIONNÉ A UNE CONCEPTION ACTUELLE QUI EST FAUSSE PAR RAPPORT A LA RÉALITÉ
0.556 - 5	DES QUESTIONS INDISCRÈTES
0.488 - 6	IL EST TRÈS COMPLET, CHER MONSIEUR
0.459 - 7	QUESTIONNAIRE QUI EFFECTIVEMENT MANQUE DE NUANCES,ON IGNORE LE BUT POURSUIVI PAR L'ENQUÊTE,DEUX TYPES DE QUESTIONS : QUESTIONS PERSONNELLES,QUI SEMBLENT INTÉRESSANTES ET QUI PEUVENT FINIR PAR DONNER UNE IMAGE DE LA SOCIÉTÉ ACTUELLE, QUESTIONS QUI NOUS DEMANDENT DE JUGER SUR L'ENSEMBLE, SUR LE GÉNÉRAL SONT DES QUESTIONS DIFFICILES A RÉPONDRE
0.418 - 8	LA TRANSFORMATION DE LA SOCIÉTÉ PASSE PAR LA TRANSFORMATION DES MENTALITÉS
0.417 - 9	DES QUESTIONS D'ARGENT INDISCRÈTES
0.377-10	CE QUI MANQUE DANS CE QUESTIONNAIRE C'EST L'INTÉRÊT PORTÉ AUX PAYS EN VOIE DE DÉVELOPPEMENT

Critère de sélection 1 : mots caractéristiques

Un premier mode de calcul de réponses modales consiste à associer à chaque réponse la valeur-test moyenne des formes caractéristiques qu'elle contient : si cette moyenne est grande, cela signifie que la réponse ne contient que des formes très caractéristiques du groupement. Les réponses de plus grandes moyennes seront donc les plus caractéristiques de la classe ou du groupement de réponse concerné.

Dans le cas du corpus qui nous sert d'exemple, le tableau 4 représente, pour les mêmes quatre catégories que le tableau 3, les 10 réponses (effectivement présentes dans le recueil de base) les plus caractéristiques de chaque catégorie.

Avec ce critère, les réponses seront courtes et caricaturales.

Quand un mot très caractéristique apparaît seul dans une réponse, cette réponse est évidemment bien classée. La présence d'autres mots peut bien entendu faire baisser la moyenne des valeurs-tests, d'où cette tendance à sélectionner des réponses courtes.

Critère de sélection 2 : Distances du Chi-2 entre profils

Le principe de ces sélections est schématiquement le suivant : une réponse est une ligne de T, donc un vecteur à v composantes. Si cette réponse est formée de 25 formes différentes, seulement 25 de ces composantes seront différentes de zéro.

Un groupement de réponses (les réponses des ouvriers, par exemple) 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 se fait selon les modalités d'une question fermée dont les réponses sont codées dans un tableau Z, on a vu que le tableau lexical agrégé C se calcule par la formule :

$$C = T'Z$$

Il est donc possible de 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.

Tableau 5 : Réponses modales pour 4 catégories de la variable à 9 modalités : Age - Diplôme (10 réponses les plus caractéristiques) (Critère de sélection 2 : Distances entre profil lexical de chaque réponse et profil moyen de la catégorie)

Critère de classification	Réponse ou individu caractéristique
	TEXTE NUMÉRO 1 : Moins de 30 ans, sans diplôme ou CEP
0.920 - 1	NON C'EST INTÉRESSANT DE RÉPONDRE À DES QUESTIONS
0.931 - 2	LES RÉPONSES SONT UN PEU TROP LIMITÉES, TROP DIRECTIVES, PAS ASSEZ DE CHOIX
0.936 - 3	NON
0.936 - 4	NON
0.936 - 5	NON
0.936 - 6	NON
0.936 - 7	NON
0.936 - 8	NON
0.936 - 9	NON
0.936-10	NON
	TEXTE NUMÉRO 3 : Moins de 30 ans, Bacc ou Université
0.848 - 1	QUESTIONNAIRE TROP LONG, PAS ASSEZ DE NUANCE POUR CERTAINES RÉPONSES
0.855 - 2	CE QUESTIONNAIRE PASSE SUR CERTAINES QUESTIONS SANS LES APPROFONDIR ET OBLIGE À RÉPONDRE DE FACON TROP DIRECTIVE, IL N'EST PAS TOUJOURS ASSEZ NUANCÉ
0.858 - 3	COMME VOUS LE DITES, QUESTIONNAIRE UN PEU TROP RIGIDE, UN MANQUE DE CHOIX POSSIBLES POUR CERTAINES QUESTIONS, QUESTIONS PARFOIS TATILLONNES ET DIFFICILE D'EN VOIR L'INTÉRÊT ET L'EXPLOITATION. EX : LES VOYAGES SUR LES PROBLÈMES DU TEMPS LIBRE, ON N'INSISTE PAS ASSEZ SUR LE CONTENU DU TEMPS LIBRE
0.869 - 4	LES QUESTIONS SONT TROP CATÉGORIQUES, ELLES NE PERMETTENT PAS DE NUANCER LES RÉPONSES
0.881 - 5	POUR CERTAINES QUESTIONS, L'ÉVENTAIL DES REPONSES N'ETAIT PAS ASSEZ LARGE
0.891 - 6	CERTAINES QUESTIONS N'ONT PAS UN ASSEZ LARGE ÉVENTAIL DE CHOIX, ON EST OBLIGÉ DE RÉPONDRE SANS ÊTRE VRAIMENT D'ACCORD
0.892 - 7	J'AI EU DES DIFFICULTÉS A RÉPONDRE À CERTAINES QUESTIONS QUI ME SEMBLAIENT NE PAS LAISSER UN GRAND CHOIX DE RÉPONSES, S'ADRESSE AU FRANCAIS MOYEN TYPE, LES QUESTIONS CONCERNANT LES RÉFORMES SOCIALES NE SONT PAS COMPARATIVES AVEC D'AUTRES PAYS, AVEC CE QUI A ÉTÉ FAIT ET CE QUI RESTE À FAIRE
0.893 - 8	LE QUESTIONNAIRE EST BEAUCOUP TROP LONG ET UN PEU ENNUYEUX ON NOUS OBLIGE À RÉPONDRE À CERTAINES QUESTIONS QUI NE NOUS INTÉRESSENT PAS
0.898 - 9	QUESTIONNAIRE PAS ASSEZ NUANCÉ POUR CERTAINES QUESTIONS
0.904 - 10	OUI CE QUESTIONNAIRE N'EST PAS BIEN FAIT, IL NE PERMET PAS TOUJOURS DE S'EXPRIMER ET PARFOIS IL OBLIGE À RÉPONDRE DANS UN SENS OÙ ON NE LE VOUDRAIT PAS, PARFOIS IL MANQUE DE NUANCES

Tableau 5 (suite)

TEXTE NUMÉRO 7 : Plus de 50 ans, sans diplôme ou CEP	
0.912 - 1	NON JE N'AI RIEN A DIRE DE PLUS, C'EST BIEN
0.917 - 2	LA RETRAITE À 60 ANS POUR LES AGRICULTEURS, UNE PRIORITÉ DE MAINTIEN DE TRAVAIL POUR LES HANDICAPÉS EN CAS DE LICENCIEMENT, DONNER UN MINIMUM DE REVENUS DÈS 50% DE HANDICAP ET S'OCCUPER UN PEU PLUS DES HANDICAPÉS LÉGERS
0.919 - 3	JE SUIS CONTENTE DE CET ENTRETIEN, LES QUESTIONS JE LES AI TROUVÉES À MA PORTÉE, C'EST SYMPATHIQUE,ÇA DEVRAIT SE FAIRE PLUS SOUVENT, EN PLUS C'EST UN CONTACT AGRÉABLE
0.923 - 4	NON C'EST BIEN , ÇA PARLE UN PEU DE TOUT, IL Y A DES QUESTIONS UN PEU INDISCRÈTES SUR L'ARGENT ET LA SANTÉ
0.934 - 6	AUCUNE, ÇA NE SERVIRAIT À RIEN
0.934 - 7	NON, RIEN À FORMULER
0.942 - 8	NON, ÇA NE VIENT PAS A L'IDÉE COMME ÇA
0.943 - 9	IL EST BON , CE QUESTIONNAIRE, IL EST BIEN CONSTRUIT, IL Y EN A POUR TOUS LES GOÛTS, LE DÉROULEMENT S'EST TRÈS TRÈS BIEN PASSÉ
0.943-10	NON, C'EST TRÈS BIEN
TEXTE NUMÉRO 9 : Plus de 50 ans, Bacc. ou Université	
0.831 - 1	CE QUESTIONNAIRE EST BIEN FAIT, J'AURAIS AIMÉ DONNER MON POINT DE VUE SUR LA PEINE DE MORT SURTOUT AVEC CE QUI SE PASSE EN CE MOMENT CERTAINES DE MES RÉPONSES SERAIENT PLUS EXACTES SI J'AVAIS EU LA POSSIBILITÉ DE RÉPONDRE AVEC DAVANTAGE DE NUANCES, J'AI ÉTÉ TRÈS INTÉRESSÉE MAIS JE SUIS SÛRE QUE CE SOIR J'Y REPENSERAI EN RÉFLECHISSANT PLUS LONGUEMENT ET QUE CERTAINES DE MES RÉPONSES SERONT DIFFÉRENTES APRÈS PLUS DE RÉFLEXION
0.879 - 2	JE TROUVE QUE CERTAINES QUESTIONS SONT INCOMPLÈTEMENT POSÉES :TROUVEZ VOUS QUE L'ÉLECTRICITÉ EST CHÈRE? OUI, POURQUOI? , CELA VEUT IL DIRE POURQUOI EST ELLE CHÈRE? OU BIEN POURQUOI LA TROUVEZ-VOUS CHÈRE? HEUREUSEMENT QUE LA JEUNE PERSONNE A FAIT PREUVE DE PATIENCE AVEC MOI, CAR J'AIME BIEN VOIR LES CHOSES CLAIREMENT, D'AUTRE PART LE QUESTIONNAIRE EST PAR MOMENT ORIENTÉ POLITIQUEMENT PARLANT.
0.888 - 3	QUESTIONNAIRE QUI EFFECTIVEMENT MANQUE DE NUANCES,ON IGNORE LE BUT POURSUIVI PAR L'ENQUÊTE, DEUX TYPES DE QUESTIONS : QUESTIONS PERSONNELLES, QUI SEMBLENT INTERÉSSANTES ET QUI PEUVENT FINIR PAR DONNER UNE IMAGE DE LA SOCIÉTÉ ACTUELLE, QUESTIONS QUI NOUS DEMANDENT DE JUGER SUR L'ENSEMBLE, SUR LE GÉNÉRAL SONT DES QUESTIONS DIFFICILES À RÉPONDRE
0.899 - 4	C'EST PAS SI MAL QUE CA,J'AI ÉTÉ TRES INTÉRESSÉ PAR LA PERSONNE QUI A POSE LE QUESTIONNAIRE, J'AI FAIT RAJOUTER MON OPINION PERSONNELLE POUR QUELQUES SUJETS N'AYANT PAS D'ALTERNATIVE, AUTRE QUE OUI OU NON PAS D'INDISCRÉTION C'EST BIEN TRÈS BIEN MÊME CONTINUEZ
0.902 - 5	IL EST PAS MAL, IL EST CONCRÊT À LA PORTÉE DE TOUT LE MONDE, ON N'A PAS PARLÉ DE LA MOBILITÉ DE L'EMPLOI, DE LA FORMATION PROFESSIONNELLE ET DU RECLASSEMENT DES GENS, LES PROBLÈMES DES FEMMES NE SONT PAS SUFFISAMMENT ABORDÉS
0.912 - 6	IL Y A PLEIN DE QUESTIONS QUI SONT MAL POSÉES, IL Y A TROP DE QUESTIONS, TROP GÉNÉRALES

Tableau 5 : (suite et fin)

0.918 - 7	LES QUESTIONS NE SONT PAS TOUJOURS SUFFISEMENT PRÉCISES, LES QUESTIONS SE RECOUPENT POUR MOI IL Y A DES QUESTIONS QUI SE SUPERPOSENT IL Y A PEUT ÊTRE CERTAINES QUESTIONS QUI PEUVENT ÊTRE SUPPRIMÉES ON A L'IMPRESSION DE RÉPONDRE PLUSIEURS FOIS À CERTAINES QUESTIONS
0.924 - 8	J'AURAIS AIMÉ SAVOIR PLUS PRÉCISEMENT À QUI SONT DESTINÉES CES INFORMATIONS, QUANT À L'EMPRISE DE L'ÉTAT PAR RAPPORT À L'INDIVIDU J'ESPÈRE QUE CE TYPE D'ENQUÊTE N'ABOUTIRA PAS À FAIRE EN SORTE QUE JE SOIS FICHÉ
0.925 - 9	QUESTIONNAIRE CONDITIONNÉ A UNE CONCEPTION ACTUELLE QUI EST FAUSSE PAR RAPPORT A LA RÉALITÉ
0.925 - 10	LE QUESTIONNAIRE N'EST PAS TRÈS BIEN REDIGÉ, CERTAINES QUESTIONS SONT AMBIGUËS AINSI D'AILLEURS QUE L'OBJET DE L'ENQUÊTE

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

$$d^2(i, m) = \Sigma_j (t_{..}/t_{.j})\,(t_{ij}/t_{i.} - c_{jm}/c_{.m})^2$$

avec les notations usuelles :

$t_{..}$ désigne la somme globale des éléments de T, c'est-à-dire le nombre total d'occurrences ;

$t_{.j}$ désigne la somme des éléments de la colonne j de T (nombre d'occurrences de la forme j) ;

$t_{i.}$ la somme des éléments de la ligne i de T (longueur de la réponse i) ;

$c_{.m}$ 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.

Le tableau 5 nous montre, toujours pour les quatre catégories extrêmes, les 10 réponses les plus caractéristiques selon ce critère.

On voit qu'il s'agit de réponses beaucoup plus riches et nuancées que celles du *tableau 4*, mais moins caricaturales. En fait, les deux critères sont assez complémentaires : résumé dense dans un cas, portrait plus impressionniste dans l'autre.

6 - Stratégie de traitement

On a vu qu'il était souvent nécessaire de regrouper les réponses pour pouvoir procéder à des analyses de type statistique. Les profils lexicaux d'agrégats de réponses ont plus de régularité et de signification que ceux des réponses isolées. Ce regroupement a priori peut être réalisé à partir des variables disponibles, retenues en fonction de certaines hypothèses. Mais ceci suppose une bonne connaissance préalable du phénomène étudié, situation qui n'est en général pas réalisée dans les études dites exploratoires.

6.1 - Regroupement par noyaux factuels

La technique dite des "noyaux factuels" va permettre de donner des éléments de réponse à ce problème.

Etant donnée une liste de descripteurs ou de variables caractérisant les individus, le problème est de regrouper les individus en groupes les plus homogènes possibles vis-à-vis de ces caractéristiques... sans en privilégier certaines *a priori*.

C'est précisément le type d'opération que permet de réaliser un algorithme de classification, appliqué aux lignes du tableau disjonctif Z décrivant les individus à partir d'une sélection de leurs caractéristiques.

La partition obtenue est une sorte de "partition moyenne" qui résume les principales combinaisons de situations observables dans l'échantillon, et qui permet donc de procéder à des regroupements de réponses les moins arbitraires possibles.

6.2 - Analyses directes sans regroupement

Si les réponses ne sont pas regroupées, mais paraissent suffisamment riches pour être traitées isolément, une analyse directe du tableau lexical T croisant formes graphiques et réponses peut être opérée.

Une telle analyse produit une typologie des réponses, en général assez grossière, et produit de façon duale une typologie de mots ou de formes graphiques.

Il est donc possible d'illustrer ces typologies par les caractéristiques des individus interrogés qui auront le statut de variables supplémentaires ou illustratives. Ce traitement direct des réponses pourra conduire à la réalisation d'un post-codage partiellement automatisé.

Notons que la proximité entre deux formes graphiques, c'est-à-dire entre deux colonnes du tableau T sera d'autant plus grande que les formes apparaîtront dans une même réponse (et non plus seulement dans un même texte), ce qui permettra de mieux représenter les voisinages syntagmatiques. L'analyse directe rendra mieux compte des contextes que les analyses de tableaux agrégés.

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

Notons que l'on peut également projeter en éléments supplémentaires les caractéristiques des k personnes interrogées (colonnes de la matrice Z) sur ces graphiques d'analyse directe et comprendre ainsi "qui a répondu quoi".

6.3 - La classification directe des formes

Cette technique permet de représenter la façon dont les formes graphiques se regroupent dans les réponses, et donc de compléter les plans factoriels comme ceux de la figure 1, qui ne présentent que deux dimensions à la fois.

C'est une façon très systématique de décrire les principales associations de mots et donc les principaux thèmes abordés.

6.4 - Juxtapositions de tables de contingences

Lorsqu'il n'existe pas de critère de regroupement a priori, on peut également analyser, non pas une table de contingence, mais une juxtaposition de tables de contingences. Cette juxtaposition $C = T' Z$ s'obtient toujours à partir de la matrice T, mais Z est maintenant le tableau disjonctif complet :

$$Z = (Z_1 , Z_2, . . Z_q . . .)$$

décrivant les réponses aux variables nominales à juxtaposer.

Dans les lignes de C figurent toujours les unités statistiques de base (formes graphiques, segments, ou lemmes), en colonne figurent, juxtaposées, les partitions correspondant à différentes variables.

Il ne s'agit pas d'une partition de synthèse (comme les noyaux factuels) car il y a simplement juxtaposition et non croisement. Les distances entre formes graphiques sont donc des distances moyennes, pour lesquelles chacune des partitions a la même importance. Il faut donc que ces partitions ne soient pas trop hétérogènes, pour que l'interprétation des proximités entre formes reste possible. Cette stratégie d'analyse proposée par Benzécri, a été implémentée dans le logiciel SPADT (1988). Son intérêt dans le cas ou les partitions sont constituées par de nombreuses questions fermées a été souligné par Cibois (1990).

7 - Commentaires sur l'exemple : enquête sur une enquête

On reprend la question posée à l'issue d'une interview sur les appréciations des personnes interrogées sur l'enquête elle-même et son questionnaire.

Les listages des formes et des réponses caractéristiques nous donnent un panorama des différentes réponses possibles sans que l'information de base soit pré-interprétée. Ce panorama montre que l'enquête n'est pas perçue de la même façon par les différentes catégories de personnes interrogées. En particulier, il y a un lien profond entre l'intérêt manifesté envers l'enquête, la qualité des réponses, et le contenu même de ces réponses. Ce lien est fâcheux pour ce que l'on souhaiterait être un instrument d'observation. Que dirait-t-on d'un thermomètre qui fonctionnerait plus ou moins bien selon la température qu'il est censé mesurer ?

Illustrons ce propos par les réponses modales et les formes caractéristiques de quelques catégories de répondants :

Les personnes âgées (formes caractéristiques : *aucune, retraite, impôts, jeunes*) s'expriment peu, prennent leurs distances par rapport à l'enquête, ou mentionnent des problèmes personnels. Parmi les réponses modales, citons :

"*C'est plutôt des jeunes qu'il faut interroger, à mon âge, on n'attend plus grand chose, on a fait sa vie", ou encore : "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 une retraite élevée, pas sur les petites retraites*".

Les réponses des plus jeunes (formes caractéristiques : *questionnaires, réponses, questions, indiscret, long, société*) sont fort différentes : elles sont en général assez critiques. Citons les deux premières réponses modales :

"*Ce questionnaire est vraiment trop long, je suis gêné 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".

Autre ton et autres préoccupations chez les ouvriers (avec enfants), (formes caractéristiques : *salaire, enfants, pourquoi, allocations*). Citons les réponses modales :

"*Pour deux enfants, avec la mère au foyer, les allocations familiales sont trop faibles par rapport au troisième, vu le coût de la vie, j'espère que l'enquête fera changer quelque chose.*",

"*Il y a trop d'injustice au niveau de la répartition des allocations familiales, pour un enfant, on devrait avoir la même chose*".

Les femmes au foyer se sentent négligées par le questionnaire (formes caractéristiques : *famille, foyer, femme*). Réponse modale :

"*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.*"

Cette lacune ou cette orientation du questionnaire est aussi dénoncée par les femmes actives aux revenus modestes :

"*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*".

Cette analyse, simplement résumée ici, met en évidence une mosaïque d'attitudes par rapport à l'opération statistique elle-même : réserve, récrimination, doléances, intérêt, critique distante, hostilité, agacement... qui illustrent la complexité des "fonctions de prélèvement d'information" que constituent les interviews d'une telle enquête.

8. - Conclusions

Cette approche différentielle, distincte de l'analyse de contenu classique, est avant tout une confrontation de l'ouvert et du fermé. Elle ne vise en effet qu'à décrire les contrastes entre plusieurs textes, que ces textes soient des réponses originales ou des regroupements de réponses réalisés à partir des questions fermées de l'enquête.

Pour une question ouverte et pour une partition de la population (par exemple une partition en noyaux factuels, résumant les principales catégories d'individus), on obtient donc, sans traitement préalable ni médiation :

-Une visualisation des proximités entre formes et catégories, par analyse des correspondances du tableau lexical agrégé, éventuellement complétée par une visualisation similaire des proximités entre segments et catégories ;

-Les formes (et/ou segments) caractéristiques de chaque catégorie ;

-Les réponses modales de chaque catégorie.

Ces résultats sont obtenus sans codification ni intervention manuelle.

Ils fournissent des compléments et donnent des éléments critiques nouveaux pour juger à la fois la cohérence et la pertinence du questionnement, la compréhension des réponses, ainsi que le niveau d'implication ou de participation des répondants.

Ils peuvent donc participer à l'amélioration de la qualité de l'information.

BIBLIOGRAPHIE

ACHARD P. (1993) - La sociologie du langage. *Que-sais-je? PUF*, Paris.

ASU (1992) [LEBART L., ed.], La qualité de l'information dans les enquêtes, *Dunod*, Paris.

BARDIN L. (1989) - L'analyse de contenu, *PUF*, Paris.

BELSON W.A. DUNCAN J.A.(1962) - A Comparison of the Check-list and the Open Response Questioning System. *Applied Statistics n° 2, p 120-132.*

BENZÉCRI J-P.& coll. (1981a) - Pratique de l'Analyse des Données, *tome 3, Linguistique & Lexicologie, Dunod* , Paris.

CIBOIS P. (1992) - Éclairer le vocabulaire des questions ouvertes par les questions fermées : le tableau lexical des questions. *Bull. de Method. Sociol., 26, p 24-54.*

HOLMES D.I. (1985). The Analysis of Literary Style - *A Review J.R.Statist.Soc., 148, Part 4, 328-341.*

LAFON P., Salem A. (1983) - "L'Inventaire des Segments Répétés d'un Texte", *Mots N° 6 , p. 161-177.*

LAZARSFELD P.E. (1944) - The Controversy over Detailed Interviews - An Offer for Negotiation. *Public Opinion Quat. n°8, p 38-60.*

LEBART L. (1982a) Exploratory Analysis of Large Sparse Matrices, with Application to Textual Data. *COMPSTAT, Physica Verlag, p 67-76.*

LEBART L. (1982b) L'Analyse Statistique des Réponses Libres dans les Enquêtes Socio-économiques. *Consommation, n°1, 39-62, Dunod.*

LEBART L. (1987) - Conditions de Vie et Aspirations des Français, Évolution et Structure des Opinions de 1978 à 1986. *Futuribles , sept 1987, p 25-56.*

LEBART L., Salem A.(1988) Analyse Statistique des Données Textuelles, *Dunod*, Paris.

REINERT M. (1986) - Un Logiciel d'Analyse Lexicale . *Les Cahiers de l'analyse des données, 4, p 471-484, Dunod*, Paris.

RUGG D. (1941) - "Experiments in Wording Questions" *Public Opinion Quat. 5, p 91-92.*

SALEM A. (1986) - "Segments Répétés et Analyse Statistique des Données Textuelles, Étude Quantitative à propos du Père Duchesne de Hébert", *Histoire & Mesure, Vol. I-n° 2, Paris, Ed. du CNRS.*

SALTON G. (1988) Automatic Text Processing : the Transformation, Analysis and Retrieval of Information by Computer, *Addison-Wesley.*

SCHUMAN H., PRESSER F. (1981) - Question and Answers in Attitude Surveys. 370p, *Academic Press*, New York.

TABARD N. (1975) - "Refus et approbations systématiques dans les enquêtes par sondage", *Consommation, n°4, Dunod.*

YULE G.U.(1944) The Statistical Study of Literary Vocabulary, *Cambridge University Press, Reprinted in 1968 by Archon Books*, Hamden, Connecticut.