FAIRE PARLER LES MOTS : vers un cadre méthodologique pour l'analyse thématique des réponses aux questions ouvertes

STÉPHANE GANASSALI

Cet article vise à fournir aux chercheurs et aux praticiens un guide méthodologique pour mener leurs analyses thématiques sur les données textuelles issues de réponses aux questions ouvertes. Quatre types de méthodes sont présentées, à employer en fonction du volume du corpus et de la pré-existence d'une grille de codification (ou « code-book »). S'appuyant sur deux cas réels (dont l'un exploite 4 936 réponses), l'article propose une illustration concrète des techniques suivantes : tableau des réponses, analyse de contenu automatisée, analyse factorielle des cooccurrences et regroupement par dictionnaire thématique à partir d'une réduction lexicale. L'extraction des verbatims est évoquée enfin comme une illustration possible des résultats issus des analyses préalablement présentées.

L'un des objectifs principaux du traitement des questions ouvertes textuelles en marketing est la qualification des réponses et plus précisément leur analyse thématique. On conçoit que cette expression spontanée (7) va nous permettre de faire émerger d'une manière plus véridique les préoccupations des clients ou des consommateurs. Quelles sont-elles ? Comment peut-on les hiérarchiser ? Comment évoluent-elles avec le temps ? Sont-elles spécifiques à une catégorie de répondants ?, etc. Ces questions sont partagées par un grand nombre de chercheurs et de chargés d'étude dans des domaines très divers du marketing et des sciences sociales au sens large.

Pour le marketing en effet, de nombreuses sources de données textuelles sont disponibles. On citera notamment les entretiens semi-directifs, les bases documentaires (réclamations clients, brochures, publicités etc.), les pages de divers sites Web ou forums de consommateurs, et bien sûr, les questions ouvertes des enquêtes par questionnaire. D'importants travaux antérieurs ont bien décrit les fondements théoriques de l'analyse de données textuelles, en présentant largement ses différentes techniques et leurs apports potentiels pour la recherche en marketing (6) ou plus largement en sciences de gestion (5) ou en sciences sociales (1, 9). Bien qu'encore insuffisant, un certain nombre d'applications a été publié pour démontrer tout l'intérêt opérationnel de l'analyse textuelle dans le cadre de problématiques particulières du marketing (5, 6, 11). Elles concernent le plus fréquemment l'analyse des questions ouvertes d'une enquête de satisfaction (13) ou de comportement de consommation (3), le traitement d'entretiens semi-directifs (5, 15) ou l'étude du positionnement de firmes concurrentes à travers le contenu de leurs sites Web (4). En parallèle, depuis une dizaine d'années, les logiciels d'analyse de données textuelles se sont considérablement développés et quelque peu « démocratisés ». Ils fournissent aux chercheurs et aux chargés d'étude des outils de traitement efficaces au service de leurs stratégies de recherche (8). Toutefois, les différentes techniques d'analyse de données textuelles sont nombreuses, parfois complexes et aucun de ces articles ne propose véritablement de guide méthodologique pour l'analyse thématique. En conclusion d'un des articles francophones de référence sur le sujet, Gavard-Perret et Moscarola (6) précisent que ces méthodes « sont encore peu pratiquées et mal codifiées et c'est là leur principale limite ».

Stéphane Ganassali est Maître de conférences à l'Institut de Management de l'Université de Savoie. Membre de l'I.R.E.G.E.
Contact : sgana@univ-savoie.fr

En effet, entre la posture « épistémologique » et l'objet logiciel, en passant par les techniques, rares sont aujourd'hui les cadres méthodologiques pour éclairer et guider chercheurs et praticiens dans leur cheminement, depuis les idées vers les résultats. Ces guides sont largement connus et disponibles pour toutes les techniques quantitatives mais ils sont encore incomplets pour l'analyse des données textuelles (10, 16).
Précisons qu'afin de répondre d'une manière plus précise et plus complète à une série d'applications fréquemment rencontrées, notre contribution se limite volontairement à une catégorie de textes bien particulière : les réponses aux questions ouvertes de questionnaire. Relativement bien structurées, souvent peu volumineuses, celles-ci se prêtent plus naturellement à la proposition d'un premier cadre méthodologique.

L'ambition de cet article est alors de proposer quelques points de repère méthodologiques quant à l'analyse thématique des réponses aux questions ouvertes. Notre expérience dans le domaine du traitement des données textuelles nous a permis d'envisager la construction d'une première proposition de cadre méthodologique. Elle s'est construite progressivement à mesure de recherches et d'études réalisées sur des corpus de différentes tailles, dans des contextes différents et pour des applications diversifiées[1]. Notre objectif est ainsi d'éclairer le chercheur ou le chargé d'études sur le type de méthodes à mobiliser, en fonction de la situation à laquelle il est confronté. Plus précisément, il s'agit de décrire des procédures de traitement quelque peu standardisées, à appliquer en fonction du volume de texte disponible d'une part et de l'existence éventuelle d'un modèle d'analyse d'autre part. A partir d'exemples réels, notre article vise à présenter les différents contextes d'utilisation et à décrire les procédures les mieux adaptées.

Dans une première partie, nous identifierons tout d'abord les différentes situations qui peuvent se présenter au chargé d'études et qui dépendent principalement du volume du corpus à traiter et de la pré-existence d'une grille thématique de codification. En fonction du contexte de l'enquête, différentes méthodes seront ensuite préconisées et présentées, à l'appui de deux exemples illustratifs, dont l'un, le cas Valibest, propose l'examen de près de 5 000 réponses réelles (encadré 1). La deuxième partie sera consacrée aux méthodes applicables sur les faibles corpus : tableau des réponses et analyse de contenu automatisée. Puis la troisième section de l'article sera consacrée aux techniques appropriées aux corpus plus volumineux. En amont de cette famille de méthodes, nous rappellerons la nécessité de réaliser une réduction lexicale avant d'opter pour une recherche thématique exploratoire sans *a priori* ou/et un regroupement par dictionnaire thématique. Enfin, dans une quatrième et dernière partie, nous verrons comment les verbatims permettent finalement d'illustrer l'analyse thématique ou de produire des extraits signifiants en fonction du contexte ou du contenu de la réponse.

Quelles méthodes pour quels textes ?

En matière d'analyse de données textuelles, l'analyse thématique est un peu comme le Saint Grâal : sujet un peu mythique, beaucoup recherché, jamais vraiment éclairci… Tout chercheur ou chargé d'études rêve d'une procédure complètement automatisée qui va lui permettre en quelques secondes d'identifier les principales thématiques de son texte et de les quantifier pour des traitements ultérieurs. Malheureusement, cette procédure fantastique n'existe pas.
En revanche, depuis une dizaine d'années se sont développées des méthodes qui sont encore bien imparfaites mais qui – utilisées à bon escient – vont permettre d'obtenir des résultats de très bonne qualité. Quelle que

Encadré 1 : Présentation du cas Valibest : support illustratif principal de cet article

Présentation de la société
Valibest est une société de distribution de produits alimentaires en gros. Elle distribue sur l'ensemble du territoire national des fruits et légumes à une clientèle de professionnels composée principalement de grandes surfaces, de collectivités, de restaurateurs et de primeurs, grâce à trente agences régionales, qui servent en totalité plus de 20 000 comptes-clients. Les clients peuvent être séparés en trois grands segments : les grandes surfaces (hypermarchés, supermarchés, supérettes, etc.), la restauration hors domicile (hôtels-restaurants, brasseries, cantines, etc.) et les spécialistes (petits points de vente spécialisés en primeurs).
Le baromètre de satisfaction clientèle
La Direction Qualité Nationale a mis en place depuis fin 2002 un baromètre de satisfaction clientèle. Le principe est de mesurer deux fois dans l'année (printemps et automne) la satisfaction des clients à partir d'un court questionnaire, administré par téléphone par un centre d'appels indépendant, dont les thèmes ont été définis et choisis par un comité de pilotage interne. Celui-ci a isolé vingt thèmes (du prix à la sécurité alimentaire) qui sont évalués par chaque client interrogé. On appréhende également les critères les plus importants aux yeux des clients ainsi qu'une satisfaction globale. Par ailleurs, une question ouverte finale enregistre les remarques, commentaires ou suggestions particulières que les clients veulent bien formuler. Cet article exploite les réponses réelles enregistrées entre début 2003 et mi 2006, soient 6 345 observations. Parmi celles-ci, on peut utiliser 4 936 réponses effectives à la question ouverte finale, soit un taux de réponse de 77 %. Si l'on croise la non-réponse à cette question ouverte et la moyenne des notes de satisfaction, on voit que la majorité des non-répondants constitue un groupe de clients satisfaits et qui ne désirent transmettre à Valibest aucune remarque particulière. Les autres n'ont pas pris le temps de répondre.

Tableau 1
Synthèse des 4 principales méthodes préconisées en fonction du contexte de l'étude

	Petit volume (< 5000 mots)	Gros volume (> 5000 mots)	
Pas de grille de codification	1. Tableau des réponses	3. Analyse factorielle des cooccurrences	A partir d'une réduction lexicale
Présence d'une grille de codification	2. Analyse de contenu automatisée	4. Regroupements par dictionnaires thématiques	

soit la méthode, la production de ces résultats nécessite un peu de temps… Attention, ce temps ne se compte pas en jours mais bien en heures. Des heures qui – au premier abord – semblent un peu longues mais qui peuvent conférer à l'étude une réelle valeur ajoutée. Dans le cadre d'une enquête, l'analyse des réponses aux questions ouvertes permet de valider ou de contrer les hypothèses contenues dans notre liste de questions, elle en propose souvent d'autres, nouvelles et inattendues. Par ailleurs, les problèmes, les motivations, les suggestions sont formulées par le répondant lui-même, avec ses propres mots et expressions, dont la simple lecture est parfois très enrichissante. Mais une fois convaincus de ces principes, il reste à savoir comment procéder.

La nature des méthodes de qualification thématique des réponses à mobiliser dépend de deux facteurs de choix : le volume du texte à analyser d'une part et la présence ou non d'une grille de codification (ou *« code-book »*) d'autre part (12). Le tableau de synthèse (tableau 1) récapitule les quatre situations possibles et présente succinctement les techniques à mettre en œuvre. Les quatre sections à suivre - classées en deux parties - sont logiquement dédiées à la description et à l'illustration de chacune des méthodologies d'analyse thématique envisagées ci-dessus.

Les méthodes applicables aux faibles corpus

Seront successivement décrits dans les lignes qui suivent le tableau des réponses pour un questionnement simple et l'analyse de contenu automatisée.

Tableau des réponses pour un questionnement simple

Quand le corpus est relativement peu volumineux (guère plus de 5 000 mots), il est souvent très homogène et l'on peut obtenir facilement et rapidement l'équivalent d'un simple dépouillement à plat, par la procédure dite du « tableau des réponses ». Grâce à quelques regroupements manuels ou par racine (mots dont les n premières lettres sont identiques), on pourra constituer une nouvelle variable complète qui prendra en compte les principales réponses à la question ouverte et accéder rapidement à une analyse assez précise de la fréquence des mots ou expressions principales.

Dans l'exemple de la figure 1, nous étudions l'image des principales chaînes de télévision française. Grâce à un protocole de questionnement simple du type « Citez trois adjectifs pour qualifier la chaîne… », on obtient 500 réponses environ, assez homogènes dans leur formulation. On évite les mots grammaticaux et les formulations polysémiques. Le corpus total compte 1 290 mots. En croisant les adjectifs cités par chaînes, l'image de celles-ci nous apparaît clairement et sans surprise : Canal+ est considérée comme drôle, sportive, complète et moderne, France 5 est différente, éducative et intelligente, TF1 et M6 sont proches sur leurs qualités divertissantes, distrayantes et jeunes.

Analyse de contenu automatisée

L'analyse de contenu automatisée (5, 9) consiste à lire chaque réponse et à la coder manuellement sur la base d'un code-book. Cette procédure est adaptée pour un volume relativement limité de réponses (quelques centaines) car c'est une méthode très précise mais consommatrice de temps. Au-delà de plusieurs centaines de réponses, on préfèrera une analyse lexicale avec regroupement par dictionnaire thématique qui permettra de procéder à une sorte de codage automatique, par sélection des mots attendus dans le lexique, ou par activation d'un dictionnaire.
L'analyse de contenu automatisée peut être simple ou multiple. En configuration simple, il s'agit de recoder la réponse sur une seule dimension. Pour l'analyse de contenu multiple, on va créer plusieurs catégories de recodage afin de qualifier les réponses sur plusieurs dimensions : le thème, le sous-thème, la tonalité, les acteurs évoqués par exemple… Cette dernière méthode est fortement préconisée car le temps consacré à la lecture des réponses est ainsi optimisé, par la production d'une information plus riche. La figure 2 ci-après illustre une analyse de contenu automatisée multiple qui serait menée sur le corpus du cas Validest. A l'occasion de cette lecture manuelle, on peut également prévoir une nouvelle question ouverte pour y recopier les « perles » ou les réponses particulièrement intéressantes.

Quand l'analyse de contenu automatisée revêt une importance stratégique et pour éviter une trop grande subjectivité, il est conseillé de la faire exécuter par au moins deux personnes en parallèle. On constatera forcément quelques différences entre les deux codifications. Il s'agit alors de discuter les cas « litigieux » entre

Figure 1 : Carte des correspondances entre chaînes et adjectifs

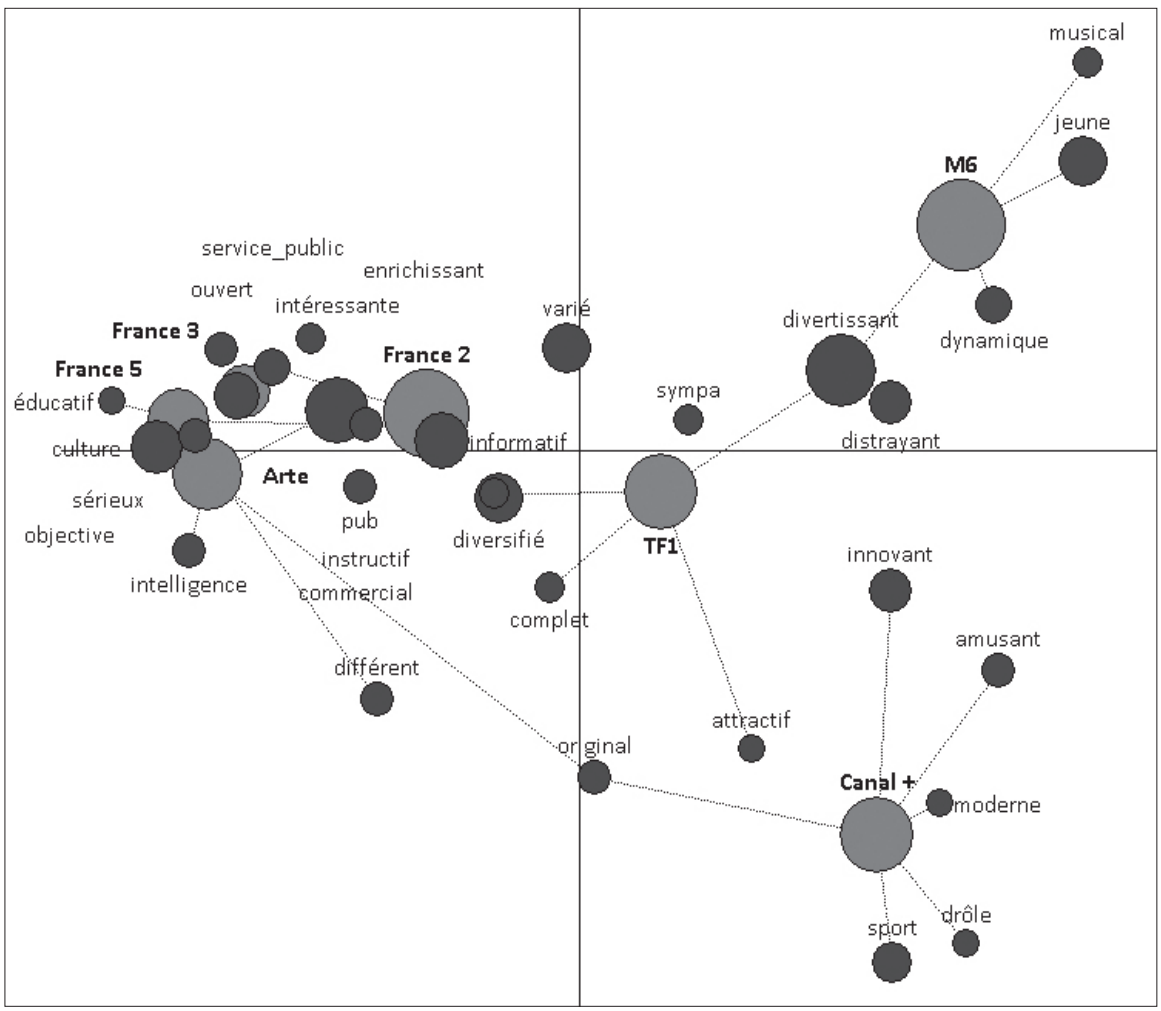
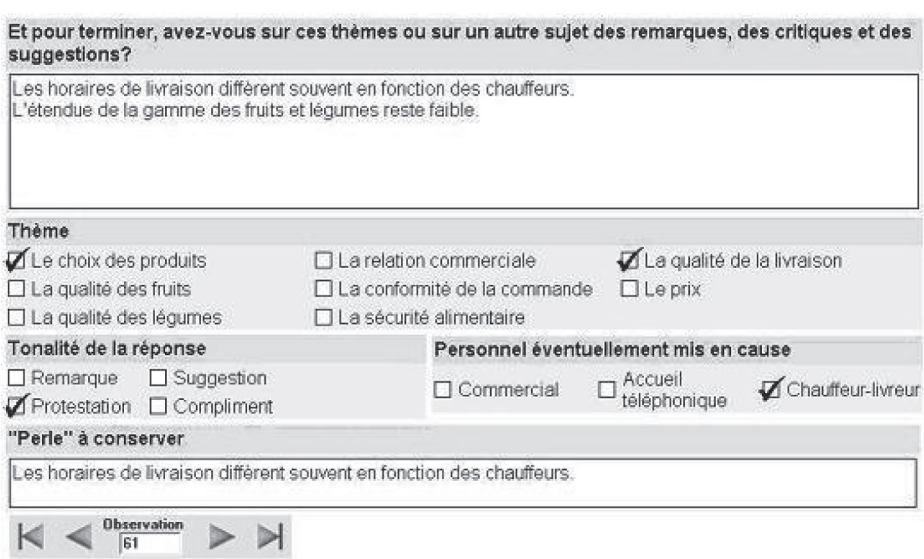


Figure 2 : Ecran d'analyse de contenu automatisée multiple



les codeurs et si possible, une tierce personne. Généralement, un accord se trouve vite sur la codification à retenir finalement pour les quelques cas qui posent problème. A l'issue de cette codification manuelle, le chargé d'études va disposer d'une ou plusieurs nouvelles variables qu'il va pouvoir analyser comme des questions classiques d'une enquête, pour établir des tris à plat et croisés, des analyses multivariées, etc.

Les techniques adaptées aux corpus volumineux

Les analyses textuelles réalisées sur un corpus volumineux nécessitent très souvent une première phase de réduction lexicale (6) que nous allons décrire précisément dans la section à suivre. Une analyse lexicale permet tout d'abord de produire la liste de toutes les formes textuelles présentes dans le texte (le lexique) afin de pouvoir appréhender celui-ci d'une manière plus organisée.

La réduction lexicale

Le lexique sera successivement présenté selon différents aspects, en fonction des recherches et des analyses envisagées. Quatre niveaux possibles de lexique peuvent être édités : du lexique brut de niveau 1 au lexique relié de niveau 4, nous allons les décrire dans les paragraphes suivants.

• Lexique brut (ou niveau 1)

Le lexique se présente en premier lieu sous une forme brute : tous les mots cités sont pris en compte, sous toutes leurs déclinaisons, les éventuelles expressions ne sont pas repérées etc. C'est une forme qui est rarement utile en marketing, sauf quand on s'intéresse au mode d'expression des locuteurs (6) dans une approche d'analyse syntaxique ou linguistique. Classiquement, on voit apparaître dans les premières places, des mots « grammaticaux » comme les articles, les pronoms, les adverbes, les prépositions et les conjonctions.
Dans le tableau 2 ci-après, dans le lexique de niveau 1, seuls quelques mots font sens sur les trente les plus cités. A la lecture rapide de ce « haut de lexique », on peut seulement s'apercevoir que l'entreprise concernée distribue des *produits* (1809) dont certains sont des *fruits* (1471)… Il faut aller plus loin dans la lecture et dans la réduction lexicale.

• Lexique réduit (niveau 2)

Réduire le lexique consiste à se concentrer sur les mots chargés de contenu et à ignorer les mots grammaticaux ou « mots-outils », à l'exception des adverbes comme « non », « plus », « trop », etc., qui associés aux mots pleins peuvent altérer la signification de la remarque ou de la suggestion. La consultation filtrée de ce lexique de niveau 2 approche le texte par les noms, les verbes et les adjectifs, qui vont plus rapidement nous permettre de prendre connaissance des principales actions, des thématiques essentielles et de la qualification de celles-ci par les locuteurs. En ne sélectionnant toujours que les trente mots les plus fréquents dans le cas Valibest, on élargit assez nettement la compréhension des problèmes. On voit d'ores et déjà apparaître des questions portant sur :

- la qualité (1 022 occurrences) des produits : les bananes sont citées 417 fois, l'adjectif « murs » est évoqué à 414 reprises,
- les commandes : 444 pour le mot au pluriel et 1 060 pour « commandé »,
- la livraison : 1177 occurrences.

Dans le cas de l'analyse qui nous intéresse, comme dans la plupart des cas, la consultation du lexique de niveau 2 (réduit) sera bien plus économique et bien plus riche d'enseignement. Demeurent quelques ambiguïtés que seules des réductions lexicales complémentaires vont permettre de lever partiellement.

• Lexique lemmatisé (niveau 3)

Toujours dans un objectif d'économie et de rapidité dans la prise de connaissance du texte, la lemmatisation consiste à ramener chaque mot à son « lemme », c'est-à-dire à sa racine grammaticale. Par ce procédé, on va simplifier le texte en ramenant le singulier et le pluriel d'un nom à leur singulier, toutes les formes d'un adjectif à leur masculin singulier et toutes les formes conjuguées d'un verbe à leur infinitif. Le lexique de niveau 3 ainsi produit va ignorer les déclinaisons d'un même mot et sera simplifié pour être recentré sur l'essence de son contenu. Par rapport aux 6 909 mots du lexique brut, le lexique lemmatisé grâce à ses regroupements automatiques, ne travaille « plus que » sur 4 246 mots. Comme on le voit dans le tableau 2, si l'on extrait les trente mots les plus cités, de nouvelles thématiques apparaissent par rapport à la lecture du lexique réduit : les horaires, le contact commercial par exemple.

• Lexique relié (niveau 4)

Finalement, afin d'appréhender plus précisément le lexique, il est souvent nécessaire de prendre en compte les expressions contenues dans le texte. Celles-ci peuvent être des groupes nominaux usuels comme « pomme de terre », ou « fruits de mer » par exemple, ou des segments de texte (dits « segments répétés »), spécifiquement apparus dans le corpus en fonction du sujet de l'étude : « 5^{e} gamme » ou « pas assez mûr » dans notre exemple. Pour ne pas faire d'erreur d'interprétation à la lecture du lexique, il est important de repérer ces expressions et de les faire apparaître séparément dans un lexique « relié » de niveau 4, afin par exemple que la « pomme » de « pomme de terre » ne soit pas comptabilisée avec la véritable « pomme »… A l'issue de cette dernière procédure, le lexique de niveau 4 va inclure des mots isolés et des expressions (comme dans le tableau 2) qui seront désormais une base fiable pour un travail ultérieur de regroupement thématique, que celui-ci soit mené manuellement ou automatiquement. Ces différentes méthodes

Tableau 2
Extraits de lexiques brut, réduit, lemmatisé et relié pour le cas Valibest

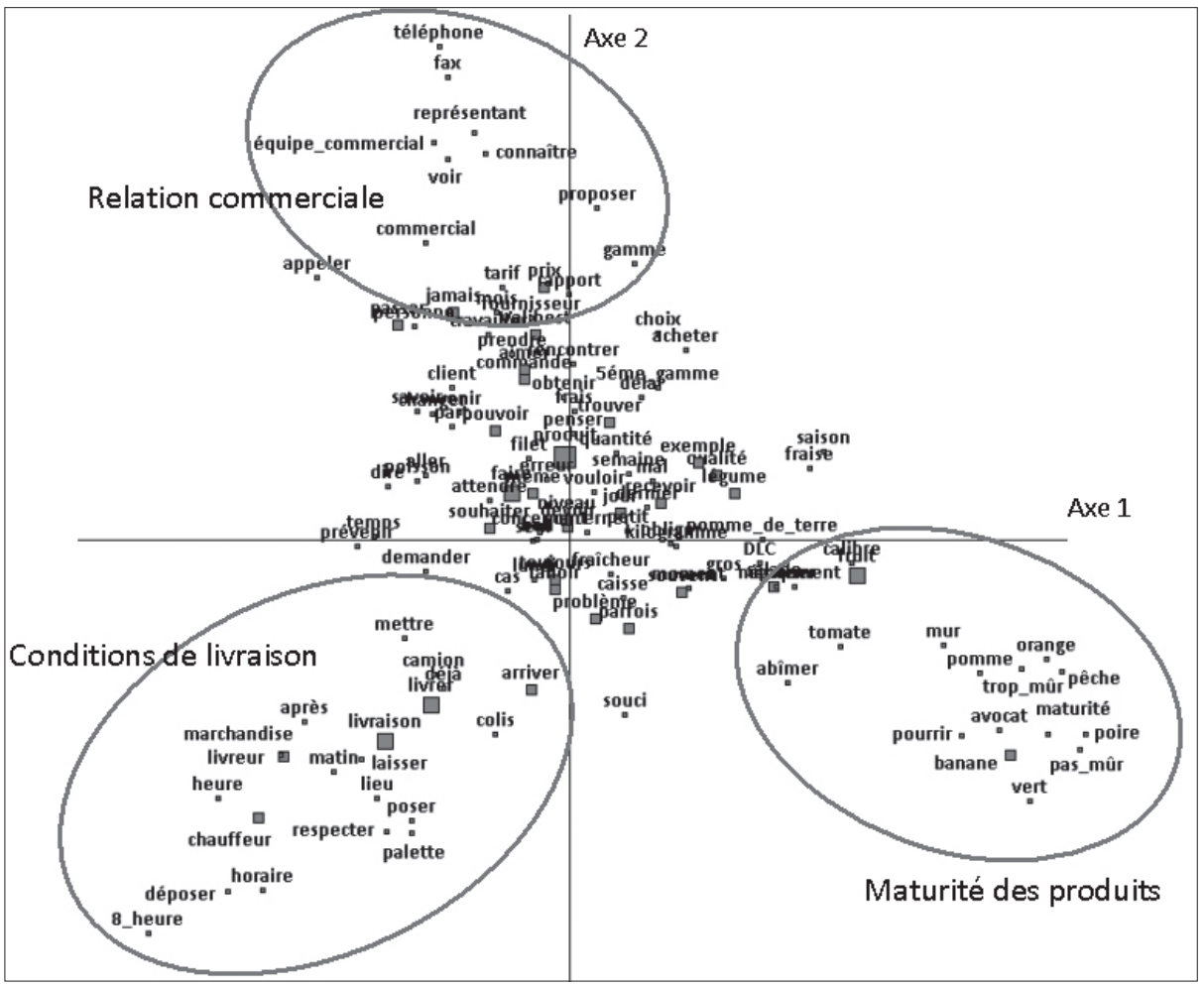
Lexique brut (niveau 1)	*Nb occ.*	Lexique réduit (niveau 2)	*Nb occ.*	Lexique lemmatisé (niveau 3)	*Nb occ.*	Lexique relié – extraits (niveau 4)	*Nb occ.*
de	*8225*	pas	*3998*	être	*5579*	(…)	
les	*6306*	produits	*1809*	pas	*3998*	8_heure	*137*
je	*5122*	fruits	*1471*	produit	*1987*	filet	*137*
la	*4744*	plus	*1470*	fruit	*1496*	frais	*137*
a	*4694*	trop	*1430*	trop	*1430*	penser	*137*
des	*4535*	livraison	*1177*	livraison	*1370*	tarif	*136*
et	*4298*	commandé	*1060*	livrer	*1332*	pomme_de_terre	*135*
pas	*3998*	Valibest	*1038*	plus	*1236*	7_heure	*134*
le	*3811*	qualité	*1022*	problème	*1079*	orange	*134*
est	*3262*	très	*828*	Valibest	*1038*	abîmer	*132*
ne	*3060*	souvent	*827*	faire	*1029*	délai	*132*
que	*3007*	livre	*717*	qualité	*1026*	déposer	*132*
en	*2618*	parfois	*717*	heure	*970*	équipe_commercial	*128*
j	*2570*	prix	*704*	commande	*924*	pas_mûr	*128*
il	*2547*	toujours	*640*	pouvoir	*878*	souci	*127*
sont	*2512*	légumes	*638*	recevoir	*836*	connaître	*125*
ai	*2240*	être	*609*	très	*828*	fraise	*125*
un	*2237*	problème	*570*	souvent	*827*	5e_gamme	*124*
nous	*2235*	problèmes	*509*	commander	*822*	acheter	*124*
n	*2155*	trouve	*507*	arriver	*780*	fraîcheur	*124*
produits	*1809*	même	*504*	parfois	*715*	caisse	*123*
l	*1762*	arrive	*501*	prix	*704*	noter	*122*
pour	*1695*	niveau	*468*	légume	*643*	9_heure	*121*
fruits	*1471*	marchandise	*463*	mûr	*641*	conditionnement	*121*
plus	*1470*	commandes	*444*	toujours	*638*	gros_problème	*121*
trop	*1430*	jamais	*433*	devoir	*590*	stock	*121*
d	*1384*	aimerais	*426*	trouver	*589*	pourri	*120*
au	*1343*	faire	*421*	commercial	*580*	rupture	*120*
sur	*1321*	bananes	*417*	même	*556*	effort	*119*
qui	*1307*	murs	*414*	jour	*544*	pas_assez_mûr	*119*

vont d'ailleurs être présentées dans la section suivante. Dans le cas d'un corpus volumineux, avant de procéder à un autre traitement (notamment thématique), il est recommandé d'enchaîner ces procédures de réduction lexicale, successivement du niveau 1 au niveau 4, en partant du lexique brut, que l'on réduira, puis que l'on lemmatisera, avant d'identifier et relier finalement les expressions récurrentes.

Recherche thématique exploratoire sans a priori

Sur un grand corpus, l'analyse thématique peut obéir à deux logiques de recherche différentes. La première est purement exploratoire. Elle consiste à « plonger dans le texte » sans *a priori* afin d'en découvrir rapidement les principaux sujets récurrents (11). L'une des méthodes

Figure 3 : Représentation des cooccurrences sur les deux premiers axes factoriels



les plus employées pour atteindre cet objectif consiste à réaliser une « analyse des cooccurrences » sur le haut du lexique de niveau 4, c'est-à-dire, selon le volume du corpus, sur les 50, 100 ou même 200 premiers mots ou premières expressions du lexique. L'analyse des cooccurrences s'opère par une analyse factorielle des correspondances.

Dans l'exemple ci-dessus et vue l'étendue du corpus, nous avons sélectionné les 200 premières formes. Notons avec intérêt que 99,7 % des réponses ont utilisé au moins une de ces formes. On voit que cette analyse très rapide sur les deux premiers axes factoriels (voir figure 3) nous permet – à partir d'une connaissance minimum du sujet – d'identifier trois premières thématiques principales :

- les problèmes de contact (insuffisant ?) avec le représentant : voir, représentant, téléphone, commercial, jamais,
- la maturité des fruits : trop_mûr, pêche, mur, pomme, avocat, poire, maturité, pourrir, banane, vert,
- les conditions de livraison : chauffeur, respecter, palette, poser, livraison, marchandise, horaire etc.

Il est bien entendu recommandé d'aller plus loin dans l'identification d'autres thématiques, en affichant par exemple les troisième et quatrième axes factoriels et ainsi de suite. On arrive alors à une analyse plus fine, pour identifier généralement une petite dizaine de thématiques principales. Dans notre exemple, à partir des axes 3 et 4, on peut repérer au moins deux nouvelles thématiques, qui sont la conformité de la commande et l'étendue de la gamme. A partir de cette analyse des cooccurrences, on pourra ensuite, réaliser une typologie des réponses afin de mieux qualifier les remarques et suggestions des clients. L'encadré 2 page suivante détaille cette procédure.

La seconde famille de méthodes d'analyse thématique est de nature plus confirmatoire, elle consiste à travailler à partir d'un modèle pré-existant. Il s'agit souvent d'une grille thématique pré-définie (ou *code-book*) qui - à partir d'une bonne connaissance du sujet - établit une liste de thèmes que l'on va aller rechercher dans le texte, soit par une analyse de contenu automatisée si le corpus est de petite taille ; soit par des regroupements lexicaux, ef-

Encadré 2 : Typologie à partir d'une analyse factorielle des cooccurrences

A partir de l'analyse des cooccurrences décrite précédemment, on procédera à la construction d'une typologie (méthode par les centres mobiles pilotée à vue dans cet exemple). En guise d'illustration, six groupes seront ici constitués, les cinq premiers correspondant aux cinq premières thématiques identifiées (commercial, livraison, maturité, conformité et choix), le sixième est un groupe « central » recueillant les réponses qui traitent d'un tout autre sujet. Une nouvelle variable fermée est créée qui enregistre l'appartenance de chaque répondant à l'un des six groupes issus de l'analyse typologique. On peut ensuite, comme présenté dans la carte à droite ci-dessous, croiser cette variable avec l'agence commerciale. Grâce à cette démarche produite en quelques minutes, le responsable national de Valibest peut déceler des premières pistes de travail propres à certaines agences commerciales, qui semblent touchées par quelque problème de qualité spécifique.

Figure 4 : Représentation de la typologie et croisement avec l'agence commerciale d'origine

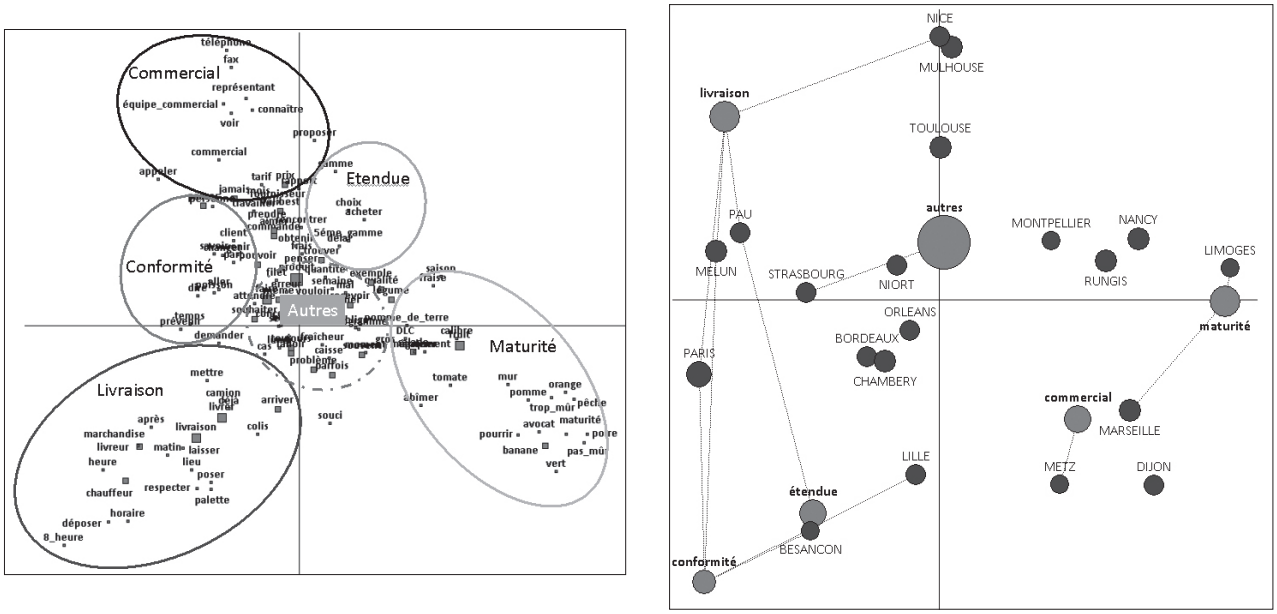


Tableau 3
Dictionnaire de regroupement thématique pour le cas Valibest

#Choix_produits=*choix=gamme=variété=étendue*
#Qualité_fruits=*fruit=banane=poire=maturité=pêche=pomme=mur=orange=fraise=kiwi=nectarine=ananas=raisin=clémentine=brugnon=prune=abricot=citron=pamplemousse=mangue=agrumes=framboise=mandarine*
#Qualité_légumes=*légume=salade=tomate=avocat=pomme_de_terre=carotte=chou=courgette=concombre=batavia=endive=céleri=poireau=oignon=laitue=aubergine=radis=poivron=betterave*
#Relation_commerciale=*représentant=commercial=téléphone=fax=appeler=équipe_commercial=interlocuteur=service_commercial=contact=réclamation=informer=joindre=réactivité=téléphoner=téléphonique=répondeur=relation_commercial=télévendeu*=vendeur*
#Conformité_commande=*commande=recevoir=commander=respecter=prévenir=rupture=conditionnement=retourner=retour=préparation=conforme=contrôler=contrôle=préparateur=conformité*
#Sécurité_alimentaire=*DLC=sécurité_alimentaire=DLC_être_trop_court=hygiène=sanitaire*
#Qualité_livraison=*livraison=livrer=chauffeur=colis=palette=horaire=heure=matin=livreur=camion=8_heure=7_heure=déposer=9_heure=11_heure=6_heure=10_heure=7h30=retard=réceptionner=fragile=8h30=tournée=6h30=écraser=10h30=livrer_trop_tôt=livrer_trop_tard=palettisation=colisage*
#Prix=*prix=tarif=promotion=prix_élever=cher=élever=plus_cher=élevé=proposition=négocier=trop_cher=promotionnel*

fectués sur le lexique relié, manuellement ou plus souvent grâce à un dictionnaire thématique.
Suite à une réflexion menée en général collectivement dans le cadre d'un groupe de travail, le chercheur ou le chargé d'étude sait quels sont les principaux thèmes à identifier dans le texte formulé par les répondants. Il définit alors un *« code-book »* ou grille thématique pour une codification des réponses ouvertes textuelles. Dans le cas Valibest, en version simplifiée, elle pourrait être la suivante : *choix des produits / qualité des fruits / qualité des légumes / relation commerciale / conformité de la commande / sécurité alimentaire / qualité de la livraison / prix.*

Regroupements par dictionnaire thématique à partir d'une réduction lexicale

Quand le corpus est très volumineux – comme c'est souvent le cas – il convient de faire appel à une méthode plus automatisée. Celle-ci s'appuie également sur une grille thématique mais ne va pas requérir une lecture systématique de toutes les réponses. Il s'agit alors de constituer un dictionnaire thématique. A chaque thème est associée une série de mots et d'expressions qui constituent autant de traces ou d'indices potentiels pour indiquer la présence du thème dans la réponse. Le thème « prix » par exemple pourra être repéré par les formes textuelles : « prix », « tarif », « promotion » ou « plus_cher ». Voici dans le tableau 3 l'exemple du dictionnaire thématique, constitué pour le cas Valibest. Notons que l'analyse exploratoire des cooccurrences (décrite dans la section précédente) peut constituer une aide précieuse à la rédaction des dictionnaires, dans la mesure où – par définition – elle permet de repérer les mots et les expressions qui se regroupent autour d'une thématique.

Ce dictionnaire thématique va être appliqué automatiquement à l'ensemble des 5 000 réponses textuelles. Il va nous permettre de qualifier très rapidement – mais un peu « grossièrement » – un grand ensemble de réponses. Certes, l'analyse de contenu automatisée aurait été plus fine mais ô combien plus coûteuse : la lecture et la codification manuelle de 5000 réponses nécessiterait plusieurs jours de travail, pour un résultat très similaire. Dans le cas présent (tableau 4), l'application de ce dictionnaire thématique constitué permet de qualifier près de 95 % des réponses, ce qui est très satisfaisant. Précisons que nous nous sommes limités dans notre exemple aux mots et expressions cités au moins à dix reprises.

A partir de cette analyse thématique par dictionnaire, on va pouvoir comparer rapidement les motifs d'insatisfaction par agence ou suivre l'évolution des problèmes dans le temps.

Les intensités lexicales pour calculer le poids relatif des thèmes évoqués

On peut aborder la partie finale de l'analyse thématique par une méthode quelque peu différente qui consiste – grâce aux intensités lexicales – à estimer la proportion de la réponse évoquant à un thème donné. Au-delà de la présence ou de l'absence du thème, telles qu'elles sont constatées par l'analyse thématique par dictionnaire, il s'agit d'aller un peu plus loin. L'intensité lexicale mesure le poids du thème dans la réponse et nous permet donc d'affiner quelque peu l'analyse. L'une des procédures consacrées consiste à calculer pour chaque réponse, l'intensité des thèmes définis dans la grille d'analyse utilisée (tableau 3). Le calcul des intensités lexicales nous indique que le thème le plus fréquent est la qualité de la livraison qui pèse près de 7 % du texte brut des réponses. Notons que la hiérarchie des thèmes est d'ailleurs totalement identique à celle fournie par l'analyse dichotomique (présence/absence) présentée dans la section précédente. Ces intensités lexicales permettent d'effectuer également des analyses bi-variées. On peut compa-

Tableau 4
Fréquence des thématiques sur la totalité des répondants

8 THEMES

Taux de réponse : **94,5%**

	Nb	
Qualité_livraison	2 316	46,9%
Conformité_commande	2 103	42,6%
Qualité_fruits	1 797	36,4%
Qualité_légumes	1 474	29,9%
Relation_commerciale	1 223	24,8%
Prix	856	17,3%
Choix_produits	513	10,4%
Sécurité_alimentaire	331	6,7%
Total	4 936	

Tableau 5
Intensités lexicales des thèmes croisées par types de clients

Intensité des thèmes x Type de client			
	GRANDES SURFACES	RESTAURATEURS	SPECIALISTES
Sécurité_alimentaire_I	0,14	0,52	0,09
Choix_produits_I	0,96	0,85	1,47
Qualité_fruits_I	3,44	6,14	1,31
Qualité_livraison_I	6,98	6,73	11,92
Conformité_commande_I	3,38	4,21	4,53
Qualité_légumes_I	2,62	2,97	3,04
Relation_commerciale_I	1,87	2,70	1,07
Prix_I	3,44	1,48	0,00

contacts/clients / Sécurité_alimentaire_I p = <0,1% ; F = 13,34 (TS)
contacts/clients / Qualité_fruits_I p = <0,1% ; F = 38,07 (TS)
contacts/clients / Relation_commerciale_I p = <0,1% ; F = 8,17 (TS)
contacts/clients / Prix_I p = <0,1% ; F = 42,39 (TS)
contacts/clients / Conformité_commande_I p = <0,1% ; F = 8,74 (TS)

Tableau 6
Synthèse des principales procédures de traitement selon la méthode retenue

Méthode d'analyse	Principales étapes
Tableau des réponses	1. Dépouillement simple des réponses puis regroupement manuel ou par racine 2. Création de nouvelles variables fermées pour traitements ultérieurs
Analyse de contenu automatisée	1. Définition du code-book pour analyse simple ou multiple 2. Lecture et codification en une (ou plusieurs) nouvelle(s) variable(s) fermée(s)
Analyse factorielle des cooccurrences	1. Réduction lexicale 2. Nouvelle variable fermée sur les n premières formes du lexique réduit, lemmatisé et relié (de niveau 4) 3. A.F.C. sur les n premiers axes pour identification des thématiques principales 4. Le cas échéant, analyse typologique 5. Création de nouvelles variables fermées ou numériques pour traitements ultérieurs
Regroupement par dictionnaire thématique à partir d'une réduction lexicale	1. Définition du dictionnaire thématique 2. Réduction lexicale (jusqu'au niveau 4) 3. Regroupement du lexique par dictionnaire thématique 4. Création de nouvelles variables fermées ou numériques (intensités lexicales) pour traitements ultérieurs

rer par exemple les intensités thématiques pour trois catégories de clients de Valibest : les grandes surfaces, les restaurateurs et les spécialistes. Dans le tableau 5, on peut identifier clairement les préoccupations dominantes de chacune des catégories : les grandes surfaces sont concernées par les prix, les restaurateurs s'intéressent plutôt à la qualité des fruits, à la sécurité alimentaire et les spécialistes enfin, se focalisent plutôt sur la qualité de la livraison.
En guise de synthèse, le tableau 6 récapitule les procédures principales de traitement à effectuer en fonction de la méthode retenue.

Les verbatims

Produire des extraits de réponses textuelles est un besoin fréquent du chercheur, du chargé d'études et de son client. Au-delà de l'aspect purement statistique des méthodes précédemment exposées, les verbatims permettent de rendre compte d'une manière plus authentique des perceptions ou des expériences de consommation, telles qu'elles ont été vécues (14). Outre les analyses thématiques ou les mesures lexicales, qui revêtent un intérêt de nature « stratégique », la restitution brute des réponses aux questions ouvertes est parfois suffisante

pour rendre compte des résultats et permettre aux décideurs une réaction opérationnelle aux commentaires enregistrés. Bien entendu, quand le volume des réponses est relativement limité (quelques centaines), il peut s'avérer fort utile de proposer les textes complets en annexe du rapport d'étude. Ceux-ci peuvent s'avérer essentiels également pour la production d'un récapitulatif des réponses individuelles. Par exemple, chaque client de Valibest ayant répondu à l'enquête de satisfaction va donner lieu à une « fiche contact » qui sera archivée et traitée par le service commercial. En fonction du contenu des remarques établies par le client, celui-ci sera rappelé et suivi, afin que ses éventuels problèmes puissent être résolus au plus vite. La restitution des réponses consiste en un ensemble de fonctions appelées « verbatims ». Les citations peuvent être choisies en lisant le texte ou sélectionnées de manière systématique selon le contexte ou le contenu. Dans les faits, comme de nombreux autres décideurs, les responsables des agences régionales Valibest se plongent en premier lieu sur la consultation de leurs verbatims plutôt que sur leurs nombreuses moyennes sur 10… Un extrait de ces textes est proposé dans l'encadré 3.

Verbatim par contexte

On veut savoir « qui dit quoi », c'est-à-dire, sélectionner les citations selon les circonstances ou toute autre information contenue dans les réponses aux questions fermées de l'enquête. Ce type de sélection peut être plus ou moins complexe : on peut n'utiliser qu'un seul critère, par exemple ce que disent les femmes, ou en combiner plusieurs en s'intéressant aux femmes de moins de 30 ans et diplômées. Dans les enquêtes de satisfaction, le contexte pertinent sera plutôt d'isoler les réponses de tous ceux ou celles qui se déclarent insatisfaits, l'identité apparaissant alors comme la signature de chaque citation…

Verbatim selon le contenu

Cette approche privilégie le contenu qui peut être repéré automatiquement par la présence dans le texte d'un ou plusieurs mots ou par un travail de codification préalable par analyse de contenu ou regroupement thématique automatisé. Pour Valibest, et pour un besoin précis du chef de produit, on voudra par exemple extraire toutes les réponses contenant le mot « orange ».
Les deux approches par le contexte et par le contenu peuvent bien sûr être combinées pour restituer le texte par fragments relatifs aux différents angles de vue ou problématiques de l'étude. Ainsi, l'extraction des verbatims peut être exhaustive ou filtrée et classée selon un ensemble de critères de sélection, selon le contexte ou le contenu, comme par exemple :

- textes d'une certaine catégorie de répondants : supermarchés d'une certaine enseigne, chaîne de restauration, clients les plus récents, les moins satisfaits etc.,
- textes visant à illustrer les différents sujets identi-

Encadré 3 : Quelques verbatims pour les thèmes de la qualité de la livraison, du prix et de la sécurité alimentaire

Thème = Qualité de la livraison :

** Les livraisons sont faites beaucoup trop tôt, car il n'y a personne d'arriver à 4h30 du matin mais ce serait bien si elles étaient décalées vers 6h, surtout pour des produits comme le poisson, la chaîne du froid ne serait pas rompue et on serait plus protéger par rapport au vol.*

** Je suis un magasin test pour les produits Valibest et je ne travaille plus depuis 3 semaines avec Valibest. L'équipe commerciale ne s'en est même pas aperçue. J'avais un gros problème avec les livraisons de produits « frais ». Il y a une semaine (et sans justifier quoique ce soit) j'ai changé pour Plus Primeurs qui m'a déjà visité trois fois, qui me livre des produits hyper frais et trois fois moins cher. J'ai donc téléphoné au groupe XYZ dont je dépends pour les prévenir.*

Thème = Prix :

** Le seul grief à faire concerne les prix, qui sont plus élevés que ceux pratiqués par mon autre fournisseur.*

** ZZZ Discount est quand même un bon client donc il faudrait faire un effort au niveau des prix et du rapport qualité/prix des produits.*

** Le seul point faible de Valibest est les prix qui restent élevés.*

** Si les prix étaient un peu moins élevés, je pourrais prendre plus de produits chez Valibest et moins à la centrale.*

Thème = Sécurité alimentaire :

** J'ai eu deux amendes concernant le DLC suite à une livraison que je n'avais pas encore contrôlée (les produits étaient déjà hors délai).*

** Je trouve que le traitement des bons de livraisons en cas de problèmes est beaucoup trop long. La date de DLC est souvent beaucoup trop courte.*

** Je trouve qu'il y a un manque de régularité dans le calibre et la maturité des fruits ; la DLC des produits 4e et 5e gammes est trop courte.*

fiés à l'issue de l'analyse thématique (encadré 3),
- réponses contenant un ou plusieurs mots important(s),
- réponses les plus longues ou les plus spécifiques (contenu), etc.

Conclusion

Certes, l'ajout de questions ouvertes dans une enquête entraîne un surcoût dans la collecte et surtout dans l'analyse des résultats. Le recours aux techniques d'analyse de données textuelles présentées dans cet article permet de le réduire, grâce à des procédés partiellement informatisés. Néanmoins, ce surcoût demeure et peut représenter – selon le nombre de questions ouvertes et leur niveau d'exploitation – 15 à 30 % du budget dédié à l'analyse des données. Dans une enquête de satisfaction, il est aujourd'hui fréquent de retrouver de une à trois questions ouvertes importantes. Il appartient alors au chargé d'études ou au chercheur, en fonction des enjeux de son enquête et des moyens dont il dispose, d'arbitrer en faveur ou non de l'exploitation approfondie des questions ouvertes.
Il est probable que par rapport à une enquête ou une recherche constituée uniquement de questions fermées, l'analyse des données textuelles exploitée à bon escient, produise un supplément de valeur. Les procédures de questionnement ouvert fournissent en effet un angle de vue qui peut compléter, renforcer ou amender une démarche quantitative traditionnelle. Dans le cas de Valibest, les clients sont globalement satisfaits. Sur les nombreuses questions échelles qui proposent pourtant une notation entre 0 et 10 sur une vingtaine de critères, la variance est assez faible puisque les notes varient globalement entre 6,8 et 8,2 pour les moyennes les plus extrêmes. L'analyse des questions ouvertes permet d'obtenir plus d'informations, elle recentre souvent le débat sur les aspects les plus importants ressentis par l'interlocuteur, qui sont recueillis en « version originale ». Et s'ils sont produits selon des procédures de traitement standardisées, elle peut fournir d'autres indicateurs tout aussi objectifs pour une mesure dynamique de la satisfaction, par exemple. Toutefois, il est primordial de préciser que les questions ouvertes n'ont pas vocation à se substituer à une démarche qualitative classique, basée sur des entretiens individuels en profondeur ou des *focus groups*, indispensables notamment pour bien préparer une phase quantitative ou pour mieux appréhender la complexité des phénomènes de consommation.

Notre contribution se limite certes au traitement des questions ouvertes dans les enquêtes par questionnaire. A ce titre, le cadre méthodologique que nous proposons sera particulièrement utile à l'utilisateur de progiciels d'enquête, qui proposent l'intégration des fonctions d'analyse des données textuelles avec d'autres techniques statistiques quantitatives. D'autres logiciels focalisés sur l'analyse des textes permettent néanmoins de mettre en œuvre une partie des méthodes décrites dans cet article. Quel que soit l'outil, il s'agit désormais de dépasser l'opposition quantitatif/qualitatif et selon une approche de fertilisation croisée (2), d'enrichir une démarche avec les atouts de l'autre. L'analyse des données textuelles est une illustration de cette nouvelle approche et il serait pertinent d'éveiller des vocations, pour

Encadré 4 : Un glossaire de l'analyse de données textuelles

Analyse de contenu automatisée : codage du texte réalisé au fur et à mesure de sa lecture, selon une grille simple ou multiple, souvent pré-établie.
Analyse des cooccurrences : présence simultanée, mais non forcément contiguë, dans un fragment de texte (séquence, phrase, paragraphe, voisinage d'une occurrence, partie du corpus, etc.) des occurrences de deux formes données.
Analyse lexicale : technique consistant à constituer la liste de toutes les formes textuelles présentes dans le texte (= le lexique) afin de pouvoir l'appréhender d'une manière plus organisée pour la consultation, l'analyse thématique ou l'extraction notamment.
Analyse thématique : identification et codage des thèmes principaux évoqués dans le texte. Peut s'opérer par lecture et codification par l'analyse de contenu, ou par un mode plus automatique grâce à un dictionnaire.
Intensité lexicale : poids d'un mot (ou d'un regroupement de mots) sur l'ensemble de la réponse d'un individu donné.
Lemmatisation : procédé qui consiste à remplacer chaque mot par son « lemme », c'est-à-dire à sa racine grammaticale (l'infinitif par exemple pour un verbe conjugué).
Réduction lexicale : méthode progressive de simplification du lexique qui permet à l'analyste de repérer les formes (mots, lemmes ou expressions) les plus utilisées. Celle-ci peut inclure une suppression des mots grammaticaux, une lemmatisation et un repérage des segments répétés.
Regroupement par racine : regroupement des mots dont les n premières lettres sont identiques.
Regroupement par dictionnaire : procédure informatisée pour regrouper les formes d'un lexique, grâce à un dictionnaire qui précise l'intitulé et les composants (mots et expressions) de chaque groupe.
Segments répétés : expressions ou segments de texte repérés à plusieurs reprises dans le corpus.
Tableau des réponses (ou des fragments) : édition brute de l'ensemble des réponses différentes à une question ouverte, avec indication de leur fréquence.
Verbatim : extrait des réponses sélectionné en fonction d'un critère de contenu (présence d'un mot-clé) ou de contexte (catégorie de locuteurs par exemple).

explorer plus fréquemment ce champ méthodologique qui reste encore peu exploité dans le domaine du marketing. Notre article est une contribution en ce sens, visant à baliser un peu mieux les différentes démarches d'analyse possibles. Il ne constitue bien entendu qu'une première étape qui nécessite des améliorations et des extensions que nous serons heureux d'avoir suscitées.

◆

Note

1. Il s'agit par exemple d'enquêtes de satisfaction pour des opérateurs de télécommunications, des distributeurs de produits alimentaires ou des industriels du secteur agroalimentaire. Nous avons également réalisé des analyses stratégiques de sites Web pour des équipementiers automobiles et des fabricants de matériel médical, ainsi que des analyses de forums ou de bases documentaires pour des banques ou des fabricants de vins de champagne…

Références

(1) Alexa M. (1997), Computer-Assisted Text Analysis Methodology in the Social Sciences, *Working Paper*, ZUMA University of Mannheim.

(2) Bolden R. et Moscarola J., (2000), Bridging the Quantitative-Qualitative Divide : the Lexical Approach to Textual data Analysis, *Social Science Computer Review*, 18, 4, 450-460.

(3) Bourgeon D. et Filser M. (1995), Les apports du modèle de recherche d'expériences à l'analyse du comportement dans le domaine culturel : une exploration conceptuelle et méthodologique, *Recherche et Applications en Marketing*, 10, 4, 5-25.

(4) Crié D., (2002), Le positionnement des sites Web par l'analyse lexicale, *Décisions Marketing*, 25, 71-82.

(5) Gauzente C. et Peyrat-Guillard D. (2007), *Analyse statistique de données textuelles en sciences de gestion*, Editions EMS.

(6) Gavard-Perret M.-L. et Moscarola J. (1998), Enoncé ou énonciation ? deux objets différents de l'analyse lexicale en marketing, *Recherche et Applications en Marketing*, 13, 2, 31-47.

(7) Geer J. G. (1988), What do open-ended questions measure?, *Public Opinion Quarterly*, 52, 365-371.

(8) Helme-Guizon A. et Gavard-Perret M.-L. (2004), L'analyse automatisée de données textuelles en marketing : comparaison de trois logiciels, *Décisions Marketing*, 36, 75-90.

(9) Jenny J., (1997), Méthodes et pratiques formalisées d'analyse de contenu et de discours dans la recherche sociologique française contemporaine, *Bulletin de Méthodologie Sociologique*, 54, 64-112.

(10) Lebart L. (1997), Statistical Processing of Responses to Open Questions in Survey Analysis, *Lexicometrica*, 0. disponible en ligne.

(11) Mathieu J.-P. (2004), L'analyse lexicale par contexte: une méthode pertinente pour la recherche exploratoire en marketing, *Décisions Marketing*, 34, 67-77.

(12) Moscarola J., Baulac Y. et Papatsiba V. (2002), Exploration sans *a priori* ou recherche orientée par un modèle : contributions et limites de l'analyse lexicale pour l'étude de corpus documentaires, *Actes des 5e J.A.D.T.*, Rennes.

(13) Mossholder K.W., Settoon R.P., Harris S.G. et Armenakis A.A. (1995) : Measuring Emotion in Open-ended Survey Responses : an Application of Textual Data Analysis, *Journal of Management*, 21, 335-355.

(14) Scherry J. F. Jr et Schouten J. W. (2002), A Role for Poetry in Consumer Research, *Journal of Consumer Research*, 29, 218-234.

(15) Sinkovics R. R., Penz E. et Ghauri P.N. (2005), Analysing Textual Data in International Marketing Research, *Qualitative Market Research: an International Journal*, 8, 1, 9-38.

(16) Van Perlo-ten Kleij, F. (2004), Text Analysis of Open-ended Survey Responses, *in Contributions to Multivariate Analysis with Applications in Marketing*, (Chapitre 7), Dissertation, Université de Groningen.

■