

GÉRARD D'AUBIGNY

**Le traitement des questions ouvertes dans les
enquêtes et sondages : introduction**

Journal de la société française de statistique, tome 142, n° 4 (2001),
p. 1-5

http://www.numdam.org/item?id=JSFS_2001__142_4_1_0

© Société française de statistique, 2001, tous droits réservés.

L'accès aux archives de la revue « Journal de la société française de statistique » (<http://publications-sfds.math.cnrs.fr/index.php/J-SFdS>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

LE TRAITEMENT DES QUESTIONS OUVERTES DANS LES ENQUÊTES ET SONDAGES : INTRODUCTION

Gérard d'AUBIGNY *

1. Présentation générale

La présente livraison du Journal retranscrit l'essentiel des contributions à la journée d'études « Traitement des questions ouvertes dans les enquêtes et sondages » organisée par la SFDS, le 8 juin 2001 à l'université Pierre Mendès-France de Grenoble.

Il nous a en effet semblé intéressant de montrer au lecteur de cette revue, que l'analyse des données textuelles, située à la rencontre de la statistique, des sciences du langage et des sciences sociales, offre un champ très riche de recherches pluridisciplinaires ¹, en pleine effervescence et productif, qui répond en particulier à bien des questions méthodologiques posées par le traitement des questions ouvertes dans les enquêtes et sondages.

Le recours à des questions ouvertes suscite un intérêt particulier des praticiens des sondages, notamment pour les études d'image, de représentations sociales ou de notoriété. Le gain de spontanéité et de richesse d'information qu'elles procurent, par comparaison à l'utilisation classique des questions fermées, n'explique pas à lui seul l'engouement actuel pour ce mode d'interrogation. L'évolution technologique a modifié le coût et la faisabilité des méthodes de recueil. Les enquêtes assistées par ordinateur – systèmes CAPI, CATI – les enquêtes sur réseau, l'usage de scanners, etc. diversifient les choix pour la réalisation du terrain d'une enquête. Ils en changent les données humaines et économiques. Ils modifient les perspectives et influent sur la conception même du questionnaire.

À l'autre bout de la chaîne, le transfert de méthodologies mises au point en traitement automatique des langues comme en analyse des données textuelles

* LabSAD/UPMF, 1251 Avenue centrale, B.P. 47, 38040 Grenoble Cedex 9. E-mail : Gerard.d-Aubigny@upmf-grenoble.fr

1. NDLR. En tant que rédacteur en chef du Journal et partageant pleinement ce point de vue, je suis particulièrement heureux de la parution de ce volume et tiens à remercier chaleureusement tous les auteurs ainsi que Gérard d'Aubigny qui a assuré l'harmonisation de l'ensemble avec compétence et efficacité (H. Caussin).

et relayées par une large diffusion des logiciels conviviaux adaptés aux besoins des instituts (tels SPAD-T, SPHINX, ALCESTE, TROPES, etc.), a rendu le traitement de ce type d'information accessible au monde de l'enquête, peu coûteux et efficace.

Les textes regroupés dans ce numéro spécial du journal, illustrent et prolongent les échanges riches entre professionnels des instituts de sondages et annonceurs, informaticiens, linguistes, sociologues et statisticiens universitaires qui animèrent la journée. Ils se proposent d'attirer l'attention des lecteurs sur la mine de problèmes méthodologiques nés de problèmes concrets et des avancées technologiques, mais ils ne prétendent pas couvrir tout le champ des approches existantes de ces questions. En particulier, les contributions présentées lors de cette journée tiraient leur cohérence d'une culture statistique commune, centrée sur les méthodes d'analyse descriptive multidimensionnelle. Le lecteur intéressé par des modélisations plus directement inspirées d'approches probabilistes ou inférentielles dans la tradition anglo-américaine, devra donc se reporter à des ouvrages introductifs complémentaires, tels que Muller (1992) ou Manning et Schutze (1999).

2. Thèmes abordés

Le matériau collecté sous forme de réponses à des questions ouvertes fournit une donnée riche, mais complexe et fragile qui pose de multiples problèmes depuis la phase de recueil des réponses jusqu'à la phase de restitution, en passant par leur traitement statistique. Après un bref rappel de la genèse des méthodes d'analyse des données textuelles, dans la filiation des travaux de J-P. Benzécri pour les grands corpus de textes, Ludovic Lebart s'attache à montrer les particularités des données produites par ce mode de questionnement. Il montre en particulier que les deux points capitaux susceptibles de rendre informative une analyse des réponses à des questions ouvertes tiennent d'une part à la reconnaissance du caractère ambigu de la fréquence des mots, véritable unité statistique d'intérêt pour les lexicographes, et d'autre part, à la prise en compte de la méta information très nombreuse qui accompagne en général la donnée textuelle. En particulier, elle permet bien souvent de compenser en partie les effets du caractère bref, voire lapidaire de certaines réponses, afin de construire des dissimilarités entre profils lexicaux plus riches de sens. Cet exposé introductif présente de plus l'état de l'art du traitement des réponses et met l'accent sur quelques-unes des orientations actuelles prometteuses de la recherche. Par exemple, il montre le rôle très utile car bien adapté des cartes auto organisées de Kohonen pour synthétiser de l'information lexicale.

Les deux exposés suivants reprennent les aspects liés à la collecte des données. Tout d'abord, du point de vue des instituts de sondages, qui ont en charge le terrain de l'enquête, toutes les étapes du processus de collecte sont importantes par ce qu'elles imposent de contraintes au recueil de l'information. La spontanéité des réponses, qui fait l'intérêt principal des questions ouvertes, peut en être menacée. C'est particulièrement vrai quand la médiation d'un enquêteur est nécessaire : son savoir-faire, sa neutralité et sa rigueur dans le

LE TRAITEMENT DES QUESTIONS OUVERTES

respect des consignes, en résumé sa culture du terrain, influe fortement sur la qualité des réponses. En homme de terrain responsable d'études, Xavier Marc expose les problèmes rencontrés en Institut de Sondage, depuis le recrutement et la gestion des enquêteurs jusqu'à la transcription des réponses libres.

L'exposé de Geneviève Lallich-Boivin vient en complément de cette réflexion sur le codage des réponses, puisqu'il a pour objectif de proposer des solutions aux différents problèmes qui surgissent dès lors que l'on souhaite traiter automatiquement des séquences langagières. Le message principal de cette informaticienne, spécialiste du traitement automatique des langues, consiste à pointer les différences fondamentales qui distinguent les transcriptions écrites de réponses formulées sous forme orale de celles qui sont directement produites sous forme écrite. Il ne peut y avoir de bijection entre des réponses orales et leur transcription écrite du fait du statut distinct des polysémies orales et écrites.

Les deux exposés suivants concentrent leur attention sur les textes écrits de réponses à des questions ouvertes. Nous avons vu, avec l'exposé de Ludovic Lebart, que l'analyse lexicale d'un corpus de réponses ouvertes regroupe celles-ci pour constituer un pseudo texte qui est alors soumis à une analyse lexicale standard. L'originalité de cette démarche tient en fait au très grand nombre de possibilités de regroupement de réponses brèves, souvent stéréotypées mais peu homogènes, et donc au très grand nombre de grilles de lecture possibles.

Pour se prêter à un traitement statistique, le texte des réponses – perçu ici comme une succession de formes graphiques – doit être codé en sorte que la variété considérable des verbatim et des mots qui le constituent puisse se réduire à un univers d'événements réalisés avec des fréquences assez grandes pour être statistiquement exploitables. Cela suppose un travail long et rigoureux de **prétraitement**, déterminant pour la suite, car il conduit à choisir l'unité d'information – les individus d'une analyse des données – qui sera finalement objet de calculs statistiques (formes graphiques, lemmes, phonèmes, racines, N-grams, thèmes). Ce choix doit nécessairement tenir compte de la **méta information** – *id est* l'information disponible sur les données – relative aux conditions de production des textes. L'enjeu de cette opération est à l'évidence de réaliser une réduction de données avec une « perte minimale de sens ». Réaliser le prétraitement nécessaire à l'exploitation statistique de l'ensemble des formes graphiques constituant un texte, sans altérer le sens dont ce texte n'est que la trace, est en soi un challenge suffisamment ambitieux pour susciter nombre de recherches et d'écoles. Postuler de plus qu'une analyse statistique de la réduction de ces formes à un univers raisonnablement accessible peut, sinon retrouver ce sens, du moins dégager les éléments du puzzle essentiels à la compréhension du phénomène étudié, peut paraître une posture quelque peu positiviste (et n'est peut-être pas sans rappeler aux statisticiens les débats sur l'articulation entre causalité et liaison statistique).

Quoi qu'il en soit, nombreuses sont les stratégies mobilisables à cette fin et la lecture des deux contributions met en évidence les éléments d'un débat. Tout d'abord, Dominique Labbé voit la solution dans une normalisation

des formes graphiques et une lemmatisation fondée sur une nomenclature exhaustive des mots français et l'apprentissage de règles de syntaxe réalisées par un automate. Les possibilités offertes par cette approche sont illustrées par l'exploitation d'une question issue d'une enquête sur « Les femmes face au changement familial » (INED/CNAF 1985). L'ambition de cet auteur apparaît alors comme celle de saisir le sens d'un texte au travers d'une codification aussi complète que possible pour consigner toutes les finesses d'une langue dans un dictionnaire encyclopédique construit d'étude en étude, de sorte que seule l'exception (rare!) doit faire l'objet d'un traitement spécifique. Si l'exigence d'une solution aussi biunivoque que possible entre le texte et le codage des formes graphiques qui le compose est si fortement présente dans le texte de Dominique Labbé, c'est, semble-t-il, que le sens lui paraît profondément attaché au lexique. L'auteur conclut que ce n'est qu'au prix d'un appel à tous les outils de la lexicographie et de la lexicologie modernes que les réponses ouvertes « *permettront de reconstituer les principaux champs lexicaux à partir desquels il sera possible de donner aux réponses ouvertes la même puissance explicative qu'aux questions fermées* ».

À cette approche, Max Reinert oppose une vision dynamique de la production de sens par l'exploitation statistique des formes graphiques, qui considère le texte comme « *la trace d'une activité temporelle* », que cet auteur replace dans la tradition de l'analyse du discours. Dans cette perspective, le maintien constant de la relation entre une forme graphique et le contexte de sa production devient l'exigence prioritaire et le sens est moins à chercher dans l'analyse des formes graphiques que dans celle des unités contextuelles élémentaires (UCE) qui constituent de fait l'unité d'information traitée dans la méthode que cet auteur préconise et qui est reprise par les diverses contributions sous l'intitulé de méthode « ALCESTE ». Dans cette approche, le codage des formes peut être moins exigeant que chez Dominique Labbé : à titre d'illustration de l'opposition entre ces deux approches, le lecteur verra dans l'exemple donné par Max Reinert que la troncature, explicitement condamnée par Dominique Labbé, est largement utilisée dans le logiciel ALCESTE.

Les deux contributions suivantes concernent plus directement les aspects statistiques du dépouillement de questions ouvertes. Ainsi, le texte rédigé par Philippe Caillot et Michèle Moine étudie précisément l'effet de la médiation des enquêteurs sur la transcription des réponses à une question ouverte posée lors d'une enquête téléphonique. Leur étude s'appuie sur l'exploitation statistique des 400 réponses à une enquête, dont ils ont supervisé toutes les étapes, pour montrer l'existence d'un effet enquêteur sur la forme retranscrite des réponses et sur le nombre de thèmes abordés par les répondants. Ils s'interrogent de plus sur la confusion éventuelle de cet effet avec les facteurs de différenciation d'intérêt pour cette enquête.

Puisque l'apport espéré des questions ouvertes est un enrichissement de l'information collectée sous forme contrainte à l'aide de questions fermées, il peut paraître dommageable de devoir fermer *a posteriori* les réponses à une question ouverte pour en croiser les termes à une (ou des) question(s) fermée(s). L'article de Mónica Bécue-Bertaut et Jérôme Pagès, présente

une exploration des questions méthodologiques soulevées par l'exploitation simultanée de questions ouvertes et de questions fermées. Cette situation, désormais courante dans les enquêtes, nécessite quelque prudence du fait de la sensibilité des méthodes d'analyse directe des questions ouvertes à l'influence de petits groupes de répondants. Ces auteurs présentent sur un exemple une extension de la méthode d'analyse factorielle multiple pour tableaux de contingence, dont l'intérêt est de respecter l'équilibre entre les différents ensembles de données.

Ce numéro spécial se clôt, comme il se doit, sur deux articles de tonalité plus interprétative, puisqu'ils replacent les résultats d'un traitement statistique des données textuelles dans le cadre d'une analyse sociologique des réponses à une question ouverte.

En sociologues, à propos d'une étude d'opinion des agents EDF sur l'arrêt des centrales nucléaires pour Mathieu Brugidou et d'une étude sur la perception de la Santé pour Pierre Le Quéau, ces deux auteurs mènent une réflexion méthodologique sur le cheminement de leur analyse et l'émergence du sens. Ces deux contributions montrent que le rôle de la méthodologie de traitement de données utilisée est d'avoir organisé à *un moment de leur réflexion* le corpus des informations recueillies, de l'avoir distancié, et préparé ainsi le «saut épistémologique» qui va du constat clinique des résultats d'enquête à la compréhension du phénomène social pour laquelle la linguistique, la sémiologie, la sociologie sont appelées à fournir des clefs explicatives. Les phases de codage de l'information et de traitement statistique jouent un rôle pivot dans ce processus d'extraction de connaissances. Leur fonction, rappelée avec force dans chacun des articles, est d'objectiver et de contrôler la validité d'une démarche qui, par essence, implique le chercheur et ses intuitions, tant la restitution qu'il fait du sens des réponses ouvertes est elle-même une production de sens. Ce rôle d'objectivation impose au statisticien d'avoir la capacité de donner un sens universel aux options méthodologiques qu'il adopte par logiciel interposé et de le transmettre aux chercheurs qui utilisent ses méthodes ou ces logiciels...

Bibliographie

- MANNING C. D. & SCHUTZE H. (1999), *Foundations of statistical natural language processing*, The MIT Press, Cambridge, Mas.
- MULLER C. (1992), *Principes et méthodes de statistique lexicale*, Collection Unichamp, Honoré Champion éditeur, Paris.