Mardi 15 mars 2005 : 16h15-16h40 *(en parallèle avec la session 11)*

La « mise en variable » des textes : Un sujet de controverses.

Gaël DE PERETTI
Insee -DG, division Conditions de vie des ménages - gael.de-peretti@insee.fr -
18 bld A. Pinard, Timbre F340, 75675 Paris Cedex - tél : 00.33.(0)1.41.17.68.55 fax : 00.33.(0)1.41.17.63.17

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. C'est peut-être l'une des raisons qui justifie le succès et l'engouement suscités par l'étude des réponses aux questions ouvertes mais sûrement pas la seule. Tout d'abord, ces questions permettent de recueillir une information riche et spontanée. Ensuite, de récentes évolutions technologiques ont modifié le coût et la faisabilité des recueils des réponses. Enfin, le traitement automatisé des réponses s'est simplifié, tout en étant moins coûteux (en temps de travail entre autres) et plus efficace (en termes de résultats). Reste que l'analyse textuelle est le sujet de nombreuses controverses. Nous focaliserons notre attention sur deux de ces controverses.

La première oppose les partisans des questions ouvertes à ses détracteurs. Elle pose le problème de ce que l'on mesure lorsque l'on étudie les réponses. Ses détracteurs considèrent que les questions ouvertes favorisent les réponses conjoncturelles et que ces dernières sont conditionnées au seul niveau d'éducation de l'interviewé. Les partisans, dont nous faisons partie, pensent que c'est le degré d'implication qui conditionne le fait de répondre ou non. Ainsi, plus la personne se sent concernée par le sujet abordé, plus sa réponse sera riche en information. Quant à l'influence de l'actualité, non seulement elle s'exerce pareillement sur les questions fermées, mais de plus, elle s'exercera prioritairement sur les sujets d'intérêt des enquêtés. Aussi, l'intégration d'une information pertinente sur le sujet abordé semble logique. Toutefois, il ne faut pas occulter un autre créateur de biais : l'enquêteur.

La deuxième controverse a lieu au sein même des partisans. Elle pose le problème du traitement statistique des questions ouvertes. Avant le développement d'outils informatisés performants, l'analyse textuelle s'est longtemps réduite au codage ou post-codage. Cette technique n'est pas sans défaut. Tout d'abord, le codage peut introduire du biais du fait de la distance qui existe entre ce que voulait dire l'enquêté et l'interprétation qu'en fait le codeur : c'est la médiation du chiffreur. Il est difficile de la nier même s'il existe des procédures pour limiter ce biais. Ensuite, il est délicat de coder des réponse complexes. Parallèlement, faut-il supprimer les réponses rares en les considérant comme du bruit ou les conserver car elles peuvent être spécifiques à une certaine catégorie de population ? Enfin, le codage cause la perte de toute la méta-information contenue dans les réponses libres (longueur des phrases, vocabulaire employé, verbes modaux, etc.). Aujourd'hui, il est possible de travailler directement à partir du corpus textuel mais, de nouveau, il y a sujet à controverse sur les traitements préalables à l'analyse textuelle. Le champ couvert par cette dernière étant particulièrement vaste, nous nous intéresserons plus particulièrement à la lexicométrie et à deux de ses outils : normalisation et lemmatisation. Nous les décrirons à l'aide d'un exemple d'application sur la question ouverte finale de l'enquête auprès des usagers des services d'hébergement et de distribution de repas chauds (dite « sans-domicile 2001 ») : « Souhaitez-vous ajouter des informations que ce questionnaire n'a pas permis de recueillir ». Cet exemple illustrera les difficultés de la « mise en variable » de textes, non seulement du fait du problème du langage (homographie et polysémie entre autres) mais aussi du fait de la nécessité de fixer des conventions qu'il faudra prendre en compte lors de l'interprétation.