



Bulletin de méthodologie sociologique

Bulletin of sociological methodology

88 | 2005
October

La "mise en variables" des textes : mythe ou réalité ?

Gaël de Peretti



Édition électronique

URL : <http://journals.openedition.org/bms/773>

ISSN : 2070-2779

Éditeur

Association internationale de méthodologie sociologique

Édition imprimée

Date de publication : 1 octobre 2005

Pagination : 5-30

ISSN : 0759-1063

Référence électronique

Gaël de Peretti, « La "mise en variables" des textes : mythe ou réalité ? », *Bulletin de méthodologie sociologique* [En ligne], 88 | 2005, mis en ligne le 08 juillet 2008, consulté le 04 mai 2019. URL : <http://journals.openedition.org/bms/773>

Ce document a été généré automatiquement le 4 mai 2019.

© BMS

La "mise en variables" des textes : mythe ou réalité ?

Gaël de Peretti

After language, our greatest invention is numbers.
Numbers make measures and maps and so enable
us to figure out where we are, what we have and
how much it's worth (Wright, 1997: 52).

- 1 L'analyse textuelle a connu depuis vingt ans un succès croissant du fait du développement d'outils *ad hoc* et des sources à traiter. Cependant, le syntagme "analyse textuelle" regroupe une multitude d'approches dont le seul point commun est de travailler sur du texte. Ainsi, au sein des pratiques d'analyse textuelle informatisée, Jacques Jenny (1997) ne dénombre pas moins de sept types d'approche : lexicométrie, réseaux de mots associés, analyse propositionnelle et prédicative, ingénierie textuelle, traitement d'enquêtes, système expert. Ici, nous allons nous intéresser à une approche lexicométrique souvent contestée par les linguistes du fait de la réduction d'un discours à un simple vocabulaire. Conscient de ces limites, nous souhaitons montrer qu'elle peut prendre sens selon le corpus textuel étudié, même si elle ne doit être finalement qu'une étape intermédiaire dans une recherche plus aboutie. Notre étude s'appuie sur l'analyse d'une question ouverte dans l'enquête dite "sans-domicile 2001" sur les usagers des services d'hébergement et de distribution de repas chauds, analyse faite à l'aide du module de statistique textuelle présent dans le logiciel Spad. Notre posture vis-à-vis de cet outil est celle préconisée par Didier Demazière (2005) : "[ne pas les réduire] à des instruments d'objectivation et d'administration de preuve, mais une posture qui les considère comme une ressource mobilisable parmi d'autres". L'objectif est de décrire l'ensemble de la démarche de "mise en variable" d'un texte; c'est-à-dire, de la transformation d'un texte en objets sur lesquels nous puissions opérer des traitements statistiques.

Une question particulière dans une enquête particulière

- 2 Le questionnaire de cette enquête se termine par la question suivante : "Souhaitez-vous ajouter des informations que ce questionnaire n'a pas permis de recueillir ?". Le fait de conclure une enquête individuelle par une question ouverte est une pratique qui se développe en particulier sur les sujets sensibles.¹ Par exemple, lors d'une enquête complémentaire sur les nouveaux arrivants au RMI, le questionnaire se terminait par une question ouverte permettant de recueillir un jugement final et global sur le dispositif : "D'après votre expérience, quels sont les aspects positifs et négatifs du RMI". (Aldeghi, 1998). De même, la dernière page de l'auto-questionnaire de l'enquête ESCAPAD (Enquête Santé et Consommation au cours de l'Appel de Préparation A la Défense), essentiellement consacrée à la consommation de produits psychoactifs, réservait un espace d'expression libre avec la question ouverte suivante : "Si vous avez des remarques à faire sur le questionnaire ou le sujet, vous pouvez le faire ci-dessous. Si vous n'avez pas souhaité répondre à certaines questions, pouvez-vous expliquer pourquoi ?" (Beck *et al.*, 2000).
- 3 La question étudiée s'inspire fortement d'une question ouverte présente dans l'enquête sur le devenir des allocataires du RMI : "Souhaitez-vous ajouter, en quelques mots, des informations qui vous paraissent importantes et que notre entretien n'a pas permis de recueillir, sur votre situation, ou vos perspectives par rapport au RMI ou à la sortie du RMI". Ce type d'enquête demande une forte implication des enquêtés et peut même être assez dure moralement. C'est évidemment le cas de l'enquête SD 2001 puisqu'elle impose à l'enquêté de se remémorer un certain nombre de situations pénibles et peut exacerber son sentiment d'exclusion en pointant sur un court laps de temps l'ensemble des difficultés d'insertion dont il est victime. Le fait de finir par une question ouverte est une façon de redonner la main à l'enquêté après l'épreuve du questionnaire.² Il lui permet, entre autres et s'il en a les ressources, de s'extirper d'un cadre très fermé et de préciser sa situation particulière qui peut sortir des cases retenues par le statisticien. Elle est peut-être une solution pour répondre à cette critique formulée par un interviewé : "Le questionnaire est trop réductif.³ Il est trop vague et n'aborde ni la réalité, ni les vrais problèmes. On sent qu'il est destiné seulement à un ordinateur".
- 4 Si l'emploi de questions ouvertes dans les questionnaires des enquêtes ménages de l'Insee s'est développé ces dernières années, leur exploitation statistique reste relativement rare du fait de la difficulté des traitements de ce type de données. On peut distinguer trois temps : la collecte ; la "mise en variables" des réponses ; et l'analyse des réponses. Nous nous concentrerons plutôt sur le deuxième temps, mais il semble nécessaire de décrire les deux autres car ils sont bien évidemment liés. L'ensemble de ce processus va nécessiter la manipulation de trois savoirs complémentaires et différents⁴ : la statistique, les sciences du langage et les sciences sociales. Cette difficulté supplémentaire peut aussi être perçue comme un avantage dans le sens où elle offre un vaste champ de recherches pluridisciplinaires (d'Aubigny, 2001). Ceci peut, en partie, expliquer l'engouement récent pour ce type d'analyse.
- 5 Outre l'intérêt intellectuel que suscite ces analyses, quelles sont les raisons qui peuvent expliquer cet enthousiasme ? Tout d'abord, les questions ouvertes sont particulièrement performantes⁵ lorsque l'on souhaite recueillir une information spontanée (d'Aubigny, 2001). Ensuite, des évolutions technologiques ont modifié le coût et la faisabilité du

recueil de ces informations (Collecte Assistée Par Informatique ou CAPI). Enfin, l'engouement pour ce nouveau champ d'investigations a entraîné la multiplication d'outils informatisés⁶ qui non seulement ont simplifié le traitement, mais l'ont rendu moins coûteux (en termes de temps essentiellement) et plus efficace (en termes d'informations produites).

- 6 Cependant, l'Insee reste peu ouverte à ce nouveau domaine puisque, si on met à part l'étude de Christian Baudelot et Michel Gollac (1997) sur "Bonheur et travail", réalisée certes à partir d'une partie variable de l'enquête permanente sur les conditions de vie des ménages (EPCV), mais par une équipe de sociologues de l'ENS, il n'existe pas à ce jour, à ma connaissance, de travaux de l'institut s'appuyant sur l'analyse textuelle d'une question ouverte, alors que ces questions peuvent être au coeur de la problématique étudiée. Par exemple, dans la partie variable "Transmissions familiales" de l'enquête permanente sur les conditions de vie des ménages (EPCV), les deux questions suivantes étaient posées : "Qu'est-ce que vos parents vous ont transmis ou légué de plus important ?" ; "Qu'est-ce qui est pour vous le plus important de transmettre ou de léguer à vos enfants ?".⁷ Cette dernière a été exploitée par le DEP du Ministère de la Culture, associé à l'enquête, et une troisième question ouverte⁸ a été exploitée de façon assez sommaire dans un article récent : "Les familles ouvrières face au devenir de leur enfant", toujours par un chercheur extérieur à l'Insee (Poullaouec, 2004). Enfin, Ludovic Lebart (2000) a exploité, à la demande de l'Insee, les deux questions ouvertes de l'enquête sur le devenir des allocataires du RMI, mais dans une visée pédagogique ou, comme il le dit lui-même : "[...], explorer les possibilités et avantages de ce type de recueil". Cette frilosité au sein de l'institut peut s'expliquer par la difficulté de quantifier l'information produite par ce type de questions, mais aussi par la valorisation des travaux de type explicatif par rapport aux approches descriptives.

Pourquoi quantifier ?

- 7 Du fait de cette complexité reconnue du matériau, on peut légitimement se poser la question du pourquoi de la quantification. L'explication vient de la difficulté à appréhender l'ensemble des informations contenues dans le corpus étudié. Ainsi, dans notre cas, il y a plus de 3.000 réponses. Il paraît très difficile de dégager des résultats à partir d'une simple lecture de l'ensemble de ces textes. Ce choix repose sur le postulat d'une supériorité des méthodes quantitatives sur les méthodes qualitatives en terme d'objectivité pour traiter cette masse d'information. C'est l'idée de la neutralité des techniques. D'ailleurs, dans le contexte de la méthodologie sociologique, pour certains, ce sont les nombres qui sont la référence implicite du couple "qualitatif/quantitatif" où le terme qualitatif renverrait plutôt à l'absence de quantité (il donnerait de la "chair" aux chiffres, réputés froids), soit à un mode d'analyse mineur.⁹ Cette suprématie du nombre comme instrument d'objectivation se retrouve au sein même des différentes mesures. Il existerait une hiérarchie au sein des mesures, selon la nature du phénomène ou plus simplement de la variable observée : variable nominale, ordinale ou numérique. Il y aurait une échelle de valeur des échelles¹⁰ : premier échelon, les échelles nominales ; deuxième échelon, les échelles de rangs ; troisième échelon, les échelles métriques *stricto sensu*. Le postulat sous-jacent est la suprématie des sciences de la nature sur les autres sciences (sociales, politiques ou juridiques) dans la dimension objective. Ceci se retrouve, en particulier, en économie où l'on tente de tout monétariser afin de travailler sur des

mesures "réelles". Cette suprématie du quantitatif sur le qualitatif aurait son origine dans les sociétés occidentales des années 1300 où dans, un laps de temps réduit, l'Europe aurait produit sa première horloge mécanique ou la quantification du temps, les premières cartes marines et peintures avec perspective ou la quantification de l'espace et les premières comptabilités à double entrée ou quantification des comptes financiers (Crosby, 1997).

- 8 C'est donc un souci d'objectivité qui va guider cette quantification, du moins dans les prémices de la statistique textuelle que l'on peut attribuer en France à Benzécri, le père de l'analyse des données à la française : "Nous proposons une méthode portant sur les problèmes fondamentaux qui intéressent un linguiste. Et cette méthode (...) effectuera une abstraction quantitative, en ce sens que partant de tableaux de données les plus divers, elle construira, par le calcul, des quantités qui pourraient mesurer des entités nouvelles, situées à un niveau d'abstraction supérieur à celui des faits recensés d'abord." (Benzécri, 1981: 4). Selon ce dernier, les travaux du linguiste Z. S. Harris sur l'analyse distributionnelle de la langue sont sa source d'inspiration. L'analyse distributionnelle est un ensemble de méthodes formelles pour étudier les langues "de l'extérieur" en éliminant le recours au sens. On compare le linguiste distributionnaliste à un martien qui analyserait la langue terrestre. Le critère général de cette théorie est la position du mot. Les mots ayant la même position dans un contexte identique seront associés à une même catégorie. Cette théorie se fonde essentiellement sur des aspects syntaxiques et transformationnels, mais le point commun avec l'approche développée par Benzécri est "que l'organisation interne des éléments d'un discours 'mémorise' en quelque sorte par sa forme même les processus externes qui ont conduit à sa production" (Reinert, 1999). L'autre pilier de ces analyses automatiques de discours est la fameuse *loi de Zipf* qui énonce ainsi une des caractéristiques structurelles fondamentales de tout corpus textuel : "le produit du rang (selon l'ordre de fréquence décroissante) et du nombre d'occurrences de chaque 'élément' (mot ou forme graphique) d'un texte est à peu près constant" (Lebart et Salem, 1994: 47-51).
- 9 Benzécri ne s'est pas contenté d'importer en France les méthodes d'analyse factorielle. Il était guidé par une ambition théorique et philosophique et avait, dès le départ, pour objectif d'appliquer ces techniques à l'étude de la langue : "C'est principalement en vue de l'étude des langues que nous nous sommes engagés dans l'analyse factorielle des correspondances" (Benzécri, 1981). Son objectif assez démesuré est celui de faire naître des données une structure du réel, de dégager de "la gangue des données, le pur diamant de la nature véridique". Les méthodes de quantification du texte permettraient d'obtenir une certaine neutralité du fait de l'automatisme des calculs. En fait, la réalité est beaucoup plus complexe, en particulier lorsque l'on travaille sur du texte collecté lors d'une enquête car la subjectivité va intervenir à tous les niveaux : questionnaire, collecte, techniques de transformation des textes, techniques d'analyse des données, interprétation des résultats, etc. Par subjectivité, nous entendons intervention d'un sujet -- le concepteur d'enquête, l'enquêteur, le statisticien -- qui va orienter consciemment ou inconsciemment la perception de l'objet ou du fait qu'il veut mesurer. Les raisons de controverses sont donc multiples. Nous nous focaliserons sur deux controverses : l'intérêt des questions ouvertes ; l'utilisation de techniques de transformation du corpus textuel (la normalisation et la lemmatisation).

Qu'est-ce que l'on mesure ?

- 10 A ce jour, il est clairement établi, et ce même dans certains domaines des sciences de la nature (physique quantique, par exemple), que la mise en oeuvre d'un procédé de mesure perturbe le processus en action ou le phénomène observé. D'une certaine façon, mesurer c'est agir sur le sujet, ce qui peut modifier sa réponse. Mais au delà de ce problème inhérent à toute question fermée ou ouverte, se rajoutent trois difficultés.
- 11 La première a été et est encore le sujet de débats, particulièrement aux Etats-Unis, à propos des enquêtes d'opinion. La critique la plus fréquente sur l'utilisation de questions ouvertes dans les enquêtes concerne les difficultés d'expression de certains enquêtés qui les empêcheraient de produire une réponse claire, alors qu'ils peuvent avoir une idée précise sur le sujet abordé. Ainsi, les questions ouvertes mesureraient plus leur niveau d'éducation que leur position sur le sujet (Craig, 1985). *A contrario*, Geer (1988) montre, lors d'une étude, que les personnes qui ne répondent pas à une question ouverte, le font par manque d'intérêt¹¹ sur le sujet étudié plutôt que du fait d'une incapacité à répondre à ce type de question.
- 12 L'autre critique concerne la pertinence de l'information collectée. Généralement, les enquêtés seraient "peu susceptibles de sonder leur mémoire de façon assez précise pour se rappeler correctement des informations qui ont généré leur jugement global" sur la question posée (Smith, 1989: 84). Plus précisément, "les commentaires ne révéleraient pas leur opinion fondamentale. Au contraire, les réponses révéleraient des goûts ou des aversions plus superficiels, comme ceux que l'enquêté aurait pu lire récemment dans les journaux, ou entendre à la télévision ou lors d'une conversation avec un ami" (Smith, 1989: 84). Plus généralement, les opposants aux questions ouvertes pensent qu'elles génèrent de simples stéréotypes. A l'opposé, les partisans des questions ouvertes considèrent qu'elles permettent aux enquêtés de définir elles-mêmes leur propre champ et de nommer les problèmes qui les concernent directement (Kelley, 1983). Parallèlement, Geer montre que : premièrement, les questions fermées sont tout aussi sensibles que les questions ouvertes à l'actualité ; deuxièmement, l'influence de l'actualité s'exerce essentiellement sur les sujets d'intérêt des enquêtés et de toutes façons, si l'information est pertinente, il semble logique qu'elle soit intégrée dans le discours des personnes (Geer, 1991). Toutefois, il est nécessaire de garder présent à l'esprit que l'actualité, aussi bien collective (événement ayant eu des répercussions médiatiques) qu'individuelle (événement qui vient d'affecter directement la vie de l'enquêté), influence la réponse de l'enquêté. L'objectif ici n'est pas de relancer le vieux débat sur la supériorité des questions ouvertes ou fermées qui existe depuis au moins soixante ans.¹² Mais si, jusqu'à présent, les questions fermées ont été privilégiées, c'est principalement car elles étaient plus faciles à poser, à coder et à analyser (Schuman et Presser, 1981). Aussi, les récents progrès technologiques devraient relancer l'utilisation des questions ouvertes dès lors que l'on s'accorde sur la pertinence et la richesse des informations collectées grâce à ces dernières.
- 13 Toutefois, et malgré ces progrès, le dernier problème concerne justement la phase de quantification. Nous reprenons ici une terminologie développée, entre autres, par Alain Desrosières (2004) : "Le premier moment est celui de la quantification proprement dite. Le verbe *quantifier* est ici employé dans un sens différent de celui du verbe mesurer. L'idée de mesure, inspirée des sciences de la nature, suppose implicitement que quelque chose

de bien réel, déjà existant, analogue à la hauteur du Mont-Blanc, peut être mesuré, selon une métrologie réaliste. En revanche, le verbe *quantifier* implique une traduction; c'est-à-dire, une action de transformation, résultant d'une série d'inscriptions, de codages, de calculs, et conduisant à une mise en nombre. Celle-ci contribue à exprimer et faire exister sous une forme numérique par mise en oeuvre de procédures conventionnelles, quelque chose qui était auparavant exprimé seulement *par des mots et non par des nombres.*" Et cette quantification est d'autant plus délicate que les procédures de codage et de traduction ne sont pas fixées dans le marbre ni même parfois reconnues. Le matériau à transformer est très complexe et cette transformation est explicitement liée à ses parties aval (analyse ou interprétation) et amont (collecte).

Du codage à l'analyse textuelle

- 14 Pendant longtemps, les procédures de quantification des questions ouvertes ont consisté à les "fermer"; c'est-à-dire, à les coder. Ce traitement a suscité de nombreux débats, toujours d'actualité. Tout d'abord se pose le problème de la médiation du chiffeur (ou codeur). En effet, toute codification nécessite une interprétation par le codeur de la réponse de l'enquêté.¹³ Or, des travaux ont montré le biais introduit lors du codage de questions ouvertes du fait de la distance qui existe entre ce que voulait dire l'enquêté et l'interprétation qu'en a fait le codeur (Kammeyer et Roth, 1971). Ces travaux ont consisté à comparer le codage et une description détaillée des réponses par les enquêtés. Les auteurs ont aussi insisté sur le fait que l'on ne peut postuler que les erreurs commises se compensent et qu'ainsi l'articulation générale reste la même. Ceci les a conduit à conclure sur l'extrême prudence lors de l'interprétation et ce quelles que soient les précautions prises lors du codage. Ensuite, toute codification entraîne une perte de méta-information. En effet, par essence, classer ou regrouper des individus revient à supprimer de l'information. De même, coder du texte revient à réduire un corpus textuel plus ou moins dense à un simple thème. Toute l'information lexicale contenue dans les réponses en clair, comme la longueur des phrases, le vocabulaire employé, la densité syntaxique, l'utilisation de verbes modaux, l'articulation des idées est généralement perdue.¹⁴ Ces problèmes sont encore accentués quand la réponse à coder est complexe. Faut-il se contenter de retenir le thème principal abordé dans la réponse ? Quels sont les critères qui permettent de le détecter ? Dans le cas où l'on décide de conserver plusieurs thèmes, doit-on se fixer une limite en nombre de thèmes et, si oui, quels critères appliquer pour fixer ce nombre ? Si l'on décide de conserver l'ensemble des thèmes, doit-on tenter de les hiérarchiser ou se contenter de noter leur co-occurrence ? Enfin, il faut se poser la question des réponses rares et de leur traitement. Faut-il seulement les considérer comme du bruit ou plutôt comme une information sur une certaine catégorie de la population ?
- 15 L'ensemble de ces questions a conduit des praticiens à fixer des règles de codification afin de rendre plus robuste cette opération. Ainsi, dès le début des années 1950, un certain nombre de travaux ont été consacrés au problème de codification des réponses aux questions ouvertes. En particulier, Lazarsfeld et Barton (1955) spécifiaient quatre conditions requises pour une bonne codification : la codification doit aller du général au particulier afin de permettre une analyse plus ou moins fine selon que l'on utilise les grands thèmes ou des items plus détaillés¹⁵ ; l'articulation logique des catégories doit reposer sur un principe unique de classification et assurer à cette dernière un caractère

disjonctif et exhaustif ; elle doit s'adapter à la structure de la situation analysée ; elle doit s'adapter au cadre de référence de l'enquête. Une fois ces principes de base adoptés, reste en suspens la question des techniques de codification. Une pratique assez courante consiste à fabriquer une grille de codage¹⁶ *a priori* à partir de la confrontation des travaux de deux codeurs différents (Montgomery et Crittenden, 1977). Trois cas sont possibles : les catégories identifiées par les codeurs se correspondent totalement (c'est-à-dire qu'elles regroupent exactement les mêmes réponses) ; plusieurs catégories de l'un correspondent à une seule catégorie de l'autre ; pas de correspondance possible. Cette première analyse permet de dégager les catégories qui seront retenues en s'appuyant sur les quatre conditions requises pour gérer le cas "pas de correspondance possible" et de préciser les instructions pour les futurs codeurs. Ensuite, l'ensemble des réponses sont analysées par différents codeurs qui disposent tous de la même grille de lecture (catégories et instructions). Dans le cas où une même réponse est codée différemment, la règle de la majorité est appliquée (c'est-à-dire que le code le plus souvent cité est retenu). Des règles plus pratiques de post-codage peuvent aussi être appliquées. Ainsi Xavier Marc (2001) conseille de ne pas créer de catégorie "autre réponse" concernant plus de 5 % de la population enquêtée et, parallèlement, de ne pas retenir de thème concernant moins de 3 % de la population. L'avantage de ce mode de chiffrage est que les analyses statistiques qui vont suivre s'appuieront sur des procédures déjà consolidées, puisque cette opération de codage consiste à "fermer la question". Cependant, l'ensemble des questions soulevées précédemment pose le problème de la légitimité d'effectuer un post-codage des réponses à une question ouverte puisque, idéalement, il nécessiterait le travail de plusieurs codeurs pour assurer une certaine robustesse des résultats et que, de toutes façons, il entraînerait la perte d'une quantité importante d'information.¹⁷

- 16 Ces procédures de codage ont assez peu d'inconvénients pour des réponses simples, stéréotypées ou peu nombreuses. Mais de nombreux éléments d'analyse sont perdus lors du post-codage -- qualité de l'expression, registre du vocabulaire, syntaxe, tonalité générale de l'entretien, longueur des réponses, etc. -- éléments qui sont parfois liés à l'introduction de ces questions.¹⁸ Par exemple, dans l'enquête du CREDOC sur les "Nouveaux arrivants au RMI", l'analyse de la question ouverte sur le jugements des allocataires sur le RMI n'avait pas pour but une quantification des jugements portés, déjà étudiés au travers des nombreuses questions fermées qui constituaient le questionnaire, mais plutôt de "renseigner sur la manière dont ils les formulent, étant entendu qu'elle mettra ainsi en valeur les affects qu'elle contient" (Aldeghi, 1988: 148). L'autre inconvénient majeur concerne les difficultés à traiter les réponses complexes ou composites qui sont, selon Ludovic Lebart (2001), "littéralement laminées par le post-codage" alors même que "c'est dans ce cas que la valeur heuristique des réponses libres est la plus grande". Enfin, les réponses rares ou peu fréquentes, difficiles à analyser en première lecture sont, par construction, affectées à la catégorie "autre réponse", ce qui empêche généralement toute interprétation de cette dernière. Or, ces réponses rares peuvent être produites par des catégories particulières d'individus et donc présenter un certain intérêt lors de l'interprétation des résultats.
- 17 Parallèlement, le développement d'outils (ou logiciels) a facilité le traitement de ces données complexes que sont les réponses libres. Cependant, ces outils ont généralement été développés dans le cadre d'un des nombreux courants de recherche ou école de pensée de l'analyse. Aussi, le choix d'un logiciel d'analyse de données textuelles n'est pas innocent, puisqu'il sous-entend l'adoption d'un cadre théorique d'analyse de discours. En

particulier, Brugidou *et al.* (2000) se sont interrogés sur la complexité des paramètres qui interviennent dans le choix d'un logiciel. Dans notre cas, nous allons donner les raisons qui ont guidé notre choix dans la "mise en variable" des réponses libres en insistant, en particulier, sur les contraintes qui pesaient sur notre matériau brut.

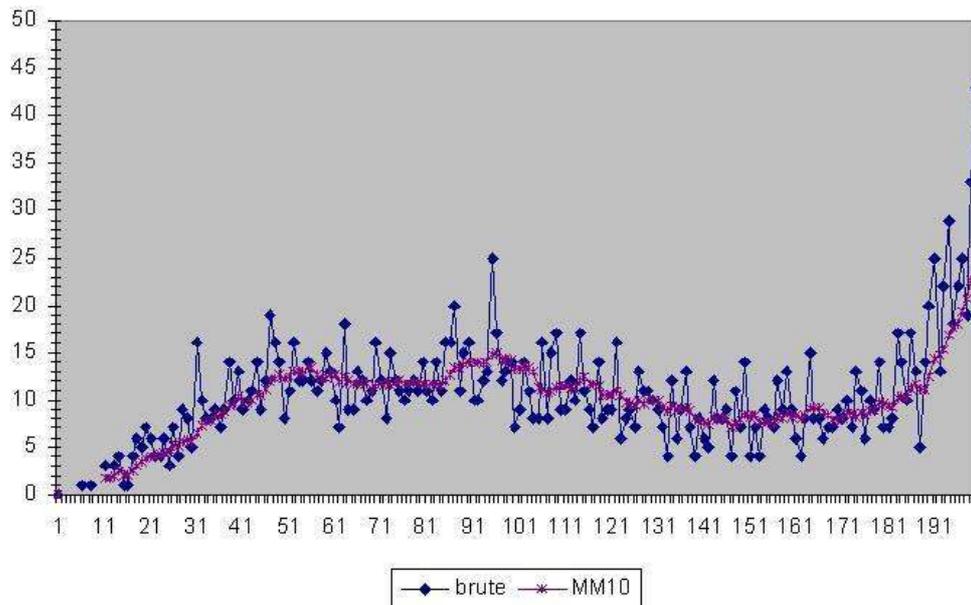
De l'oral à l'écrit

- 18 L'enquête a été collectée sur papier pour des raisons pratiques évidentes et dans des conditions parfois difficiles. En particulier, lors des entretiens dans les services de distribution de repas chauds, l'enquêteur ne disposait pas toujours d'un lieu réservé ou d'une table isolée pour réaliser son interview. Indépendamment des conditions de collecte, le protocole retenu impliquait le passage d'un discours oral à une retranscription écrite de l'enquêteur. Implicitement, cela suppose que l'on fasse l'hypothèse que l'écrit sera l'image fidèle de l'oral, or il est clair qu'il n'existe pas de bijection entre oral et écrit (Lallich-Boidin, 2001), en particulier dans un tel cadre. En effet, l'écrit est évidemment plus précis dans certains cas : marque du pluriel et du genre ; majuscule et minuscule ; sujet des verbes. La perte de ces informations est appelée la neutralisation des oppositions : opposition entre unicité et nombre, homme et femme, etc. De façon duale, l'écrit n'est pas capable de retranscrire les distinctions de l'oral (c'est ce problème auquel nous sommes confrontés).
- 19 Ainsi, comment interpréter la phrase : "plus de place(s) dans les centres". Dans le cas où cette phrase est extraite d'un corpus textuel plus riche, nous utiliserons le contexte pour décider si l'enquêté voulait souligner le manque de place dans les structures collectives ou demander l'augmentation du nombre de places dans les structures existantes, mais rien ne permet d'affirmer que l'on puisse choisir avec certitude. Dans le cas où elle est le corpus entier, nous pouvons décider d'appliquer une règle syntaxique. Si l'enquêteur a mis un "s", nous considérerons que l'enquêté réclame une augmentation du nombre de places ; si l'enquêteur n'en a pas mis, que l'enquêté constate un manque de place dans les structures d'accueil. Cette règle est simple à mettre en place, mais elle est, à la lecture de l'ensemble des réponses, un pis-aller. En effet, il existe un nombre très important de fautes d'orthographe dans l'ensemble des corpus textuels saisis qui légitime la remise en cause de cette procédure. Afin d'éviter ce genre de problème, des solutions ont été proposées : utilisation de logiciels de retranscription ; enregistrement des réponses avant saisie ; formation spécifique des enquêteurs à la saisie de tel corpus.¹⁹ Ces techniques n'étaient pas présentes lors de la collecte de l'enquête "sans-domicile 2001". De plus, aucune instruction spécifique à cette question n'était présente dans le manuel d'instructions aux enquêteurs. La saisie des réponses a donc été source de nombreuses disparités entre les différents enquêteurs.
- 20 Les réponses ont été écrites par l'enquêteur, puis ont été saisies par des opérateurs. De nombreux enquêteurs ont adopté un style télégraphique sans sujet ni conjugaison, et utilisé des abréviations afin de collecter l'ensemble du discours produit par l'enquêté. Cependant, les réponses saisies ne pouvaient dépasser 200 caractères, source d'un certain nombre de réponses tronquées. Dans notre cas, une méthode biaisée pour estimer le nombre de réponses tronquées a consisté à compter toutes les réponses contenant deux cents caractères. Elles sont au nombre de 123 sur les 2.186 réponses²⁰ abordant au moins un sujet (c'est-à-dire, différente de RAS, non, rien à ajouter, etc.), soit 5,6% des réponses. *A priori*, on peut penser qu'elle tend à surestimer le nombre de réponses tronquées,

puisque le discours saisi peut très bien s'arrêter effectivement à 200 caractères. On peut, parallèlement, compter les réponses de 200 caractères finissant par un mot tronqué ("[...] Il faut améliorer les conditions de v..."), ou une phrase incomplète ("[...] souhaite avant tout pouvoir..."). Il y a 20 phrases qui sont manifestement incomplètes et 46 phrases qui finissent par un mot tronqué. Dans ce dernier cas de figure, deux possibilités sont envisageables : soit le discours finissait par ce mot tronqué (soit par manque de place, soit parce que l'enquêteur utilisait des abréviations, ce qui est le cas dans un nombre important des réponses), soit le discours continuait. Il reste donc 57 phrases de 200 mots pour lesquelles on ne peut rien dire, sachant qu'il y a 43 phrases (respectivement 33) de 199 caractères (respectivement 198). La lecture de certains questionnaires²¹ confirme les deux cas envisagés. Ainsi, un questionnaire repéré comme tronqué finissait effectivement par le dernier mot saisi. En revanche, un autre se prolongeait longuement, la personne enquêtée donnant son opinion ou son sentiment sur de nombreux sujets : "Difficultés particulières par rapport au logement différentes à Toulouse. C'était plus facile l'insertion à Toulouse que sur Paris. Il y a l'air d'avoir plus de structures mais moins de places. On peut tourner de foyers en foyers sans d'autres issues que par le travail car il faut la caution. Quand on est dans un foyer, c'est difficile de trouver du travail (horaires, peu de calme). Différent à Toulouse où on m'avait donné un studio sans caution. Le Samu Social (115), injoignables, j'ai bataillé et on m'a dit qu'il fallait redescendre à Toulouse, trois semaines dans la rue : "Vous n'avez qu'à redescendre en stop". Le foyer Emmaüs : nickel au niveau propreté, le seul problème, il faut partir de 8h à 18h, toute la journée dans la rue, sans ressource et sans savoir quoi faire. Paris demande beaucoup de courage et de santé. On se sent perdu sans repère. Les domiciliations sont difficiles, les prix des transports aberrants sur un RMI. Si je n'ai pas de place dans un hôtel social, je redescendrais sur Toulouse".

- 21 Lorsque l'on observe la distribution des réponses en fonction de la longueur du texte saisi (voir Figure 1), on constate un nombre croissant de réponses dès lors que l'on dépasse 150 caractères avec une croissance forte à partir de 190 caractères. Ainsi, un sixième des réponses compte au moins 190 caractères (dont un tiers à 200 caractères). Une explication plausible serait le fait que la capacité de stockage d'information des enquêteurs tournait autour de 200 caractères. Cependant, elle est insuffisante. La lecture de questionnaires manuscrits nous a permis de recenser des réponses comprises entre 190 et 199 caractères qui ne semblaient ni tronquées (pas de mot coupé), ni incomplètes (elle finissait par une phrase complète ayant du sens) à la lecture du fichier, mais qui se révélaient en fait incomplètes à la lecture du questionnaire. L'opérateur de saisie avait manifestement préféré stopper sa saisie plutôt que commencer une phrase qu'il savait sans fin. Ceci pourrait être une autre explication de l'inflation des réponses de plus de 195 caractères.

Figure 1 : Fréquence des réponses en fonction du nombre de caractères en brut en moyenne mobile sur 10 caractères



Source : Enquête auprès des usagers des services d'hébergement et de distribution de repas chauds, Insee, 2001.

Note de lecture : la série brute correspond aux réponses saisies avant tout travail de correction orthographique et de normalisation, la série MM10 correspond à une moyenne mobile d'ordre 10 qui permet d'obtenir une tendance.

- 22 Au delà de ce problème de troncature, les différences liées aux prises de note (phrases complètes ou abrégées, discours indirect ou direct) nous ont conduit à préférer une approche thématique à partir d'une analyse lexicométrique du fait des fortes disparités entre les corpus textuels difficilement allouables à l'enquêté et surtout à une grande prudence quant à la généralisation des résultats du fait des possibles traductions des réponses libres des enquêtés par les enquêteurs. Mais ce type d'intervention de l'enquêteur sur le corpus textuel étudié n'est pas la seule.

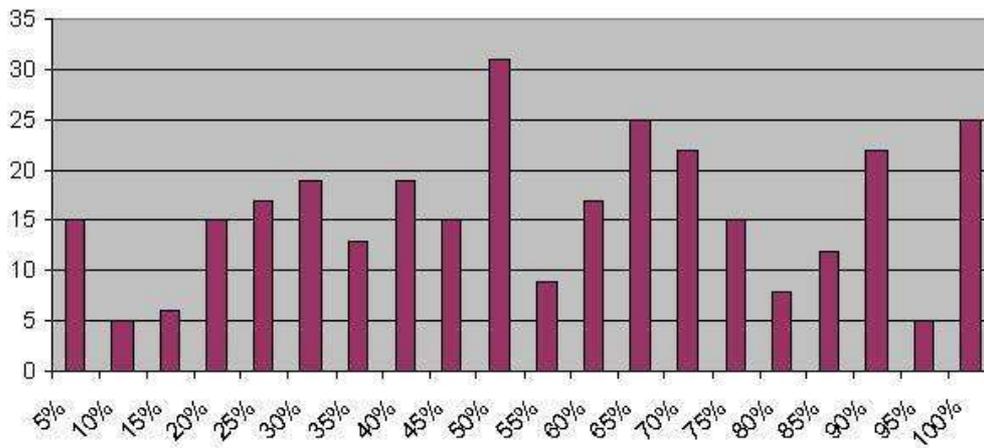
La relation enquêteur enquêté

- 23 Tout d'abord, il faut se poser deux questions (A quelle question répond l'enquêté ? Quel rôle assigne-t-il à l'enquêteur ?), deux questions qui sont fortement imbriquées. Ainsi, si l'enquêté prend l'enquêteur pour un représentant de l'Etat, il pourra en profiter pour lui faire part de sa supplique (Fassin, 2000) ou de son cahier de doléances, et interprétera la question comme une invitation à réclamer ce qui lui manque le plus (travail, domicile, papier). Au contraire, s'il prend l'agent pour un sondeur ou un agent d'une institution productrice de statistiques, indépendante de l'Etat et impuissante face à sa situation, il sera peut-être plus enclin à lui répondre de manière plus stricte sur la pertinence du questionnaire. Les réponses n'existent malheureusement pas, mais les questions doivent rester présentes à l'esprit lors de l'interprétation des résultats.
- 24 Symétriquement, il faut penser aux interférences possibles liées à l'action de l'enquêteur. Cette question ne peut être mise de côté dès lors que l'on traite d'une étude sur une

population en difficulté d'insertion sociale. En effet, lors d'une étude méthodologique sur les principes à adopter pour améliorer la qualité des enquêtes sur ce type de population (Dubéchet et Legros, 1993), les auteurs montrent l'importance de contrôler au mieux l'effet enquêteur. En fonction de son âge, de son expérience, de sa formation professionnelle, la relation qu'il va établir avec l'enquêté sera de nature différente, et peut, en raison de l'impossibilité d'une standardisation parfaite de cette relation, malgré les instructions données aux enquêteurs, conduire à un déroulement différencié des entretiens. Cet effet est d'autant plus sensible que, dans notre cas, nous nous intéressons à la question finale et qu'aucune instruction spécifique n'était donnée aux enquêteurs dans leur manuel d'instructions. Après une très longue série de questions plus ou moins personnelles, voire stigmatisantes (ou perçues comme telles), le fait de répondre à cette dernière question peut être influencé par la relation qui s'est instaurée au cours de l'entretien entre enquêté et enquêteur, ce qui dépend *a priori* fortement des caractéristiques propres de ce dernier.

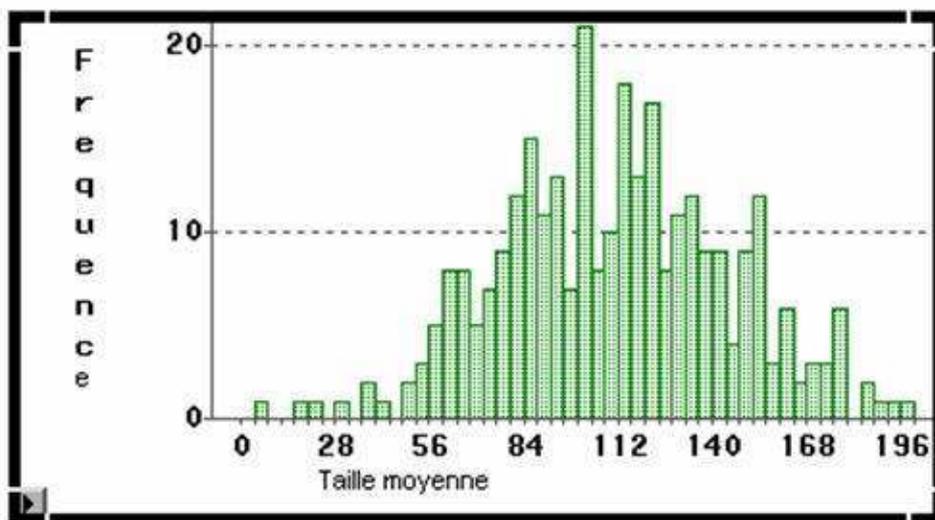
- 25 L'étude des questions ouvertes dans des enquêtes d'opinion a mis en évidence un effet enquêteur sur les réponses des enquêtés. Cependant, cet effet est généralement considéré comme mineur (Caillot et Moine, 2001). L'objectif de l'analyse textuelle, au delà de l'analyse de l'univers lexical des réponses, est de croiser cet univers lexical avec les caractéristiques des individus afin de déterminer l'influence de ces derniers sur le contenu des réponses. La conclusion de l'étude de Caillot et Moine est que, si l'on constate des effets sur la forme du corpus textuel et le nombre de thèmes abordés, cette interaction ne modifie pas significativement l'information apportée par cette question et les liaisons entre les thèmes abordés et les caractéristiques des enquêtés. Du fait de la particularité de notre enquête et de la différence entre une question d'opinion classique et notre question ouverte, il est nécessaire de vérifier si, dans notre cas, l'effet enquêteur est différent.
- 26 Une première approche de cet effet enquêteur consiste à regarder les taux de réponse à cette question par enquêteur et, plus particulièrement, la distribution de ces taux. 315 enquêteurs ont mené cette enquête et ont collecté en moyenne 13 questionnaires. Les différences entre enquêteurs sont grandes. Neuf enquêteurs ont collecté un questionnaire et un en a collecté 29. De même, les disparités sur les taux de réponse à la dernière question sont très fortes (voir Figure 2). Ainsi, quatorze enquêteurs ont un taux de réponse nul (ils ont en moyenne enquêté 5,6 personnes), et vingt-cinq enquêteurs ont un taux de réponse de 100% (ils ont en moyenne enquêté 8,7 personnes). En moyenne (non pondérée), un enquêteur a obtenu un taux de réponse de 53,3% à cette question.²²

Figure 2 : Nombre d'enquêteurs selon le taux de réponse à la question ouverte



Source : Enquête auprès des usagers des services d'hébergement et de distribution de repas chauds, Insee, 2001.

Figure 1 : Nombre d'enquêteurs en fonction de la longueur des réponses



Source : Enquête auprès des usagers des services d'hébergement et de distribution de repas chauds, Insee, 2001.

- 27 La deuxième approche revient à vérifier l'hypothèse de Caillot et Moine sur la longueur des textes. Nous avons donc analysé la distribution de la longueur des textes bruts. Ainsi, pour chaque enquêteur, nous avons calculé à partir des réponses non corrigées, la taille moyenne des réponses en nombre de caractères. La longueur des réponses varie beaucoup (de 5 caractères à 200), et un sixième des réponses compte au moins 190 caractères. Le fait d'étudier la taille par enquêteur permet de compléter cette première étude. Une nouvelle fois, nous constatons une grande disparité dans la longueur des textes.
- 28 Tout d'abord, l'effet lié à la troncature des réponses est de nouveau visible, la queue droite de distribution est épaisse. Ensuite, le nombre moyen de caractères pour un enquêteur donné est de 111 caractères (moyenne des moyennes par enquêteur) contre

117 sur l'ensemble des réponses. Cette différence est significative en terme statistique (test non paramétrique du fait des caractéristiques de notre variable), mais apporte *a priori* peu d'informations. En effet, six caractères permettent peut-être d'écrire un mot, mais pas une idée nouvelle, sauf dans le cas des réponses télégraphiques. Toutefois, cela implique que les réponses longues sont plutôt concentrées sur certains enquêteurs.

- 29 Une autre approche possible de l'effet enquêteur est d'observer l'impact d'une prise en compte de sa mesure sur la recherche des déterminants de la réponse à la question ouverte. Pour cela, nous avons envisagé plusieurs méthodes. Il faut garder à l'esprit que nous ne cherchons les déterminants de la réponse à la question ouverte que dans un volonté descriptive du phénomène. Aussi, les modèles statistiques qui vont être développés, bien que complexes, n'ont aucune prétention de "pouvoir explicatif". L'objectif est de décrire l'impact de l'effet enquêteur sur les dimensions corrélées avec le taux de réponse, à partir de différentes méthodes reposant sur des hypothèses que nous proposons de discuter. Nous avons retenu plusieurs procédures d'estimation afin de nous assurer de la robustesse toute relative des résultats.
- 30 Les méthodes retenues reposent toutes sur le même principe. Il consiste à effectuer un parallèle entre les données étudiées et les données de panel. L'avantage de ce type de données est la prise en compte d'effet individuel, ce qui dans notre cas serait un effet enquêteur. Généralement, les données de panel ont une dimension temporelle. Par exemple, pour étudier l'impact sur le sentiment d'aisance financière du revenu, de variables comme le sexe et le nombre d'enfants d'un individu i , le fait de disposer de plusieurs observations dans le temps permet de prendre en compte cet effet individuel, chaque individu ayant sa propre échelle de valeur indépendamment de son sexe, sa situation familiale, etc. Dans notre cas, pour un même enquêteur i (équivalent de notre individu dans les données de panel classique), chaque enquête q correspond à une observation (équivalent de la date t dans les données de panel classique). L'objectif sera de comparer les résultats obtenus à partir d'une régression logistique permettant de repérer les covariables (caractéristiques de l'enquêté), qui ont un impact sur la probabilité de répondre à la question ouverte, à d'autres procédures statistiques permettant *a priori* de prendre en compte l'effet enquêteur.
- 31 L'objectif ici n'est pas de décrire en détail les résultats, mais simplement de vérifier l'existence ou non d'un effet enquêteur. Toutefois, il faut garder à l'esprit que ces méthodes restent imparfaites puisqu'à la fois on ne connaît rien de l'enquêteur²³ ou du lieu précis où s'est déroulé l'entretien. Aussi, une difficulté de cette approche vient de la confusion possible entre le type de lieu d'enquête et l'enquêteur²⁴ : il est donc possible que l'on mesure plutôt un effet "lieu d'enquête" qu'un effet enquêteur, ou du moins qu'il y ait un mélange des deux effets.²⁵
- 32 Dans tous les modèles développés,²⁶ l'effet enquêteur est positif. De même, ces modèles de "panel" tendent à restreindre les dimensions ayant un impact sur le taux de réponse. Elles n'en rajoutent pas à deux exceptions près, mais ces résultats peuvent être liés à des problèmes d'effectifs du fait du nombre important de paramètres à estimer, au regard du nombre d'observations. Ces résultats sont évidemment à prendre avec précaution, mais confirment *a priori* la nécessité d'une grande prudence dans l'interprétation, et surtout la généralisation des résultats de l'analyse textuelle.

Normalisation, quasi-lemmatisation, des outils de l'analyse lexicométrique

- 33 Comme nous l'avons déjà fait remarquer, l'engouement pour l'analyse statistique de texte a entraîné le développement parallèle d'un grand nombre de logiciels ayant chacun ses spécificités propres dépendant complètement de l'approche du concepteur. Or, le choix (plus ou moins contraint) d'un logiciel a des conséquences sur le type d'analyse que l'on peut envisager. Nous avons travaillé sur le logiciel SPAD-T dont la philosophie générale consiste à repérer l'ensemble des mots (formes graphiques) utilisés dans le corpus textuel et de calculer leur occurrence. Dans ce logiciel, avant tout traitement statistique, il est nécessaire d'effectuer deux procédures (ou "méthodes", selon la terminologie propre au logiciel) qui permettent, respectivement, de repérer les mots utilisés mais aussi les groupes de mots (ou "segments répétés"). Parallèlement, il est possible d'effectuer des corrections et des regroupements de mots ou segments afin d'en réduire le nombre, tout en limitant la perte d'information et d'obtenir ce que nous appellerons notre vocabulaire d'étude. L'ensemble des traitements statistiques développés par le logiciel repose sur une analyse lexicométrique du corpus que l'on peut développer selon deux axes : le poids absolu d'une forme graphique dans l'ensemble du corpus étudié ou le poids relatif d'une forme graphique au sein de différentes catégories de population. Enfin, il est possible de créer un tableau contenant en ligne des individus avec leurs caractéristiques (âge, sexe, etc.), mais aussi la fréquence d'utilisation des mots retenus dans le vocabulaire d'étude afin de caractériser la réponse à la question ouverte des individus. Ce tableau permet de réaliser l'ensemble des traitements statistiques connus (classification, analyse factorielle, régression), chaque mot du vocabulaire étant devenu une variable.
- 34 L'approche lexicométrique peut paraître paradoxale, comme le rappelle Dominique Labbé (2001) : "Peu de mots dépassent le seuil de 1% de fréquence relative et ce ne sont probablement pas les plus intéressants puisque, selon le vieil adage classique, la quantité d'information véhiculée par un mot est inversement proportionnelle à sa fréquence d'apparition". Ceci est confirmé par la liste des vingt mots les plus fréquemment utilisés dans les réponses à notre question ouverte et qui représentent près d'un tiers des formes utilisées (32,8%) : *de, je, un, les, pas, à, pour, et, est, la, le, d', des, que, l', en, j', on, ne, il*. L'objectif premier est donc de réduire le nombre de mots (formes graphiques) que l'on prendra en compte dans les analyses statistiques, du fait des particularités de notre corpus (liées à la collecte), tout en limitant la perte d'information par rapport à notre axe d'interprétation des résultats. Cette réduction a pour but de faciliter et rendre plus robuste les calculs, qui seront menés par la suite, afin d'éviter de travailler sur des tables immenses et pleines de zéro. En revanche, il est nécessaire de bien définir les traitements que l'on va opérer, car ils auront de fait des conséquences sur les résultats produits par les calculs. Ces effets devront donc être pris en compte lors de l'interprétation des résultats. Parallèlement, les règles de traitements que l'on définit doivent être en accord avec les visées finales de l'étude. Dans notre cas, les contraintes imposées par la collecte nous ont conduit à viser une analyse de contenu de type thématique. C'est dans cette phase que l'on retrouve les tensions entre objectivité et subjectivité. Nous allons fixer un certain nombre de conventions qui vont modifier l'objet sur lequel nous travaillons. Parallèlement, ces conventions doivent apporter de la robustesse (au sens statistique) aux procédures d'analyse statistique qui seront produites par la suite. Il est nécessaire de

préciser ces conventions, mais aussi de tenter de cerner du mieux possible les conséquences qu'elles auront sur les analyses qui vont suivre.

- 35 Avant de se lancer dans une analyse textuelle des réponses, nous avons pris le parti d'effectuer des corrections sur le texte saisi. L'objectif de cette normalisation des réponses est de "débruiter" au maximum les réponses du fait des fortes disparités dans les procédures de recueil des réponses des enquêtés. Le problème majeur de la correction est qu'elle implique nécessairement une interprétation de la part du correcteur (Lallich-Boidin, 2001). Ceci conduit à appliquer un traitement standard à l'ensemble des réponses et à fixer des normes liées aux corpus et aux traitements visés.
- 36 La matière brute sur laquelle nous avons travaillé est un texte en lettres capitales, transcription intégrale (dès lors que la réponse contient moins de 200 caractères) du texte écrit par l'enquêteur. L'ensemble des textes a été retranscrit en minuscules accentuées afin d'éviter les confusions du type "a" et "à", "bornes" et "bornés". Tous les noms propres commencent par une majuscule et les mots composant le nom d'une association sont collés. Ainsi les "Restos du coeur" deviennent les "Restosducoeur". Toutes les dates sont écrites sous la forme "jour mois année" avec *jour* et *année* en chiffre. Toutes les sommes d'argent sont sous la forme xxxf avec xxx correspondant au montant évoqué et f à franc (enquête réalisée en 2001). Nous avons systématisé le recours aux acronymes ou aux sigles avec une orthographe unique en lettres capitales. Enfin, nous avons transformé l'ensemble des textes en les mettant au discours direct. Toutefois, nous avons conservé l'organisation grammaticale de la réponse. Ainsi la réponse "un travail et un logement" n'est pas modifiée, alors que la réponse "il voudrait un logement et un travail" devient "je voudrais un logement et un travail". Cette correction est un choix pratique qui n'est pas entièrement satisfaisante. Il est, en effet, impossible de savoir dans les deux cas précédents, si l'enquêteur a réellement transcrit le discours de l'enquêté. Les deux personnes ont très bien pu dire : "j'aimerais bien avoir un logement et un travail". Les deux transcriptions précédentes sont possibles, car elles conservent les thèmes évoqués. Nous avons choisi de modifier le moins possible le corpus de base. Ceci a pour conséquence que dans la suite de l'analyse, les trois réponses suivantes seront équivalentes : *un travail et un logement* ; *avoir un travail et un logement* ; *j'aimerais bien avoir un travail et un logement*. En revanche, la réponse "trouver un travail et un logement" est différente, puisqu'elle implique plus directement l'enquêté. Dans les cas précédents, il s'agit seulement d'un souhait, alors que dans le dernier exemple, l'utilisation du verbe "trouver" laisse supposer qu'il y aura une recherche de sa part. Cette distinction est nécessaire dès lors que l'on s'intéressera non pas seulement aux thèmes abordés, mais aussi à la formulation et à la constitution de "posture" (Reinert, 2001) ; c'est-à-dire, de discours types relatifs à certaines catégories de personnes. D'autres travaux auraient pu être envisagés ; en particulier, le traitement de la polysémie et de l'homographie assez fréquent dans la langue française. Dans les deux cas, l'objectif est d'ajouter des marqueurs qui permettent de repérer les différents sens d'un même mot afin d'éviter des problèmes d'interprétation des résultats. Dans le second cas, des règles syntaxiques permettent de séparer les homographes en associant à chaque forme une catégorie grammaticale. Ainsi, des chercheurs ont développé des logiciels contenant des nomenclatures de mots français en regroupant leurs différentes flexions sous un même "lemme" étiquetée de sa forme grammaticale. Par exemple, toutes les formes conjuguées d'un même verbe sont réunies sous le doublon [nom du verbe à l'infinitif, verbe]. De même, toutes les déclinaisons de l'article "le" (la, l', les) sont réunies sous le doublon [le, article]. Ensuite, l'ensemble des

règles syntaxiques de la langue française sont connues du logiciel, ce qui lui permet à la lecture du texte de séparer l'essentiel des ambiguïtés. Cela permet de distinguer automatiquement les deux sens du mot "être" dans les phrases suivantes : "[...] Faut connaître les difficultés pour un être humain de vivre dans la rue [...]" où être est un nom masculin et "[...] N'ai pas l'impression d'être aidé [...]" où être est un verbe. La polysémie est plus délicate à traiter puisque le sens va dépendre du contexte. Ainsi, le verbe "sortir" a de multiples sens dans la langue française. Le sens le plus fréquemment retrouvé dans les réponses étudiées est celui de quitter la situation de précarité actuelle, de s'en sortir : "[...] C'est difficile de s'en sortir avec des dettes [...]". Ensuite, sortir est utilisé dans les sens d'aller hors d'un lieu : "[...] Où est ce que l'on va quand on doit sortir du foyer à 8h du matin [...]". Enfin, il est utilisé dans le sens d'aller hors de chez soi pour aller se distraire : "[...] Je souhaiterais pouvoir sortir plus le soir en semaine [...]". Pour le premier cas, le logiciel SPAD, en repérant le segment répété "en sortir", permet d'éviter la confusion de sens. En revanche, il n'est possible de distinguer les deux autres cas qu'à la lecture de la réponse. C'est évidemment une limite de l'approche lexicométrique. Toutefois, comme nous allons le voir par la suite, il est possible de prendre en compte le contexte, dans certains cas, pour éviter cette confusion.

- 37 Dans un second temps, nous avons travaillé sur la lemmatisation de notre corpus ; c'est-à-dire, à donner à un mot du discours une forme canonique servant d'entrée de dictionnaire. Cette procédure correspond tout à fait au double objectif de réduction du nombre de mots et de limitation de la perte d'information. L'idée est de regrouper sous un même lemme, différents mots dont le sens est identique afin de lui donner plus de poids, mais surtout afin d'éviter de ne pas les prendre en compte du fait de la disparités des formes utilisées. En effet, nous avons fait le choix de supprimer de l'étude tous les lemmes n'apparaissant pas au moins 15 fois dans l'ensemble du corpus. Si dans certains cas, cette opération correspond à regrouper sous la même entrée (les formes conjuguées sous l'infinitif, les féminin, masculin, singulier, pluriel sous la forme la plus répandue) ; dans d'autres cas, nous opterons pour une approche quasi thématique en nous appuyant sur une approche contextuelle des réponses. En effet, il est possible de connaître, pour tous les mots de notre vocabulaire, les phrases dans lesquelles ils sont utilisés. Cette contextualisation assure une certaine robustesse à la lemmatisation. Ainsi, le lemme "conjoint" regroupe les mots ou segment suivants : concubin, concubine, compagne, compagnon, copain, copine, épouse, époux, mari, ma femme. Nous avons dû retenir seulement le segment "ma femme", car le mot "femme" recouvrait d'autres sens que celui de conjointe. La procédure retenue est simple. Nous ne voulions conserver que les mots ou segments répétés apparaissant plus de 15 fois dans l'ensemble du corpus textuel. Nous avons regroupé toutes les conjugaisons d'un même verbe sous la forme infinitive, si elle existait, sauf si une forme était largement majoritaire (fréquence trois fois supérieure aux autres formes). Dans ce cas, le lemme correspondait à cette dernière forme. Les formes au pluriel et au singulier sont regroupées sous un même lemme, sauf si leur utilisation correspondent à deux sens différents et que leurs fréquences respectives permettent de les conserver toutes les deux. Ainsi les mots "personne" et "personnes" correspondent à deux lemmes différents. Le mot au singulier correspond généralement à un usage négatif : "Après avoir vécu 20 ans dans mon pays, personne ne veut m'aider ; c'est lamentable". Le mot au singulier vise à désigner des individus avec lesquels l'enquêté est entré en contact ou une catégorie de gens : "Dans les associations, certaines personnes ne nous aident pas" ; "Que les organismes type HLM puissent accepter des personnes au RMI". Enfin, afin de définir complètement le "vocabulaire" sur lequel nous avons travaillé, nous

avons supprimé tous les mots outils. En effet, ces derniers sont très fréquemment employés et apportent peu d'information sur le contenu du texte. Ces derniers n'auraient d'intérêt que s'ils n'étaient pas aléatoirement répartis parmi les différentes catégories étudiées. Or, dans notre cas, cette répartition non aléatoire pourrait s'expliquer en (grande) partie par la disparité des techniques de recueil des réponses des enquêteurs. Nous avons donc préféré les supprimer. L'opération de lemmatisation incluant aussi des regroupements thématiques et de suppression des mots outils est parfois nommée une quasi-lemmatisation (Lebart, 2000).

Construction des données et interprétation

- 38 Les réponses complètes donnent lieu à un corpus de 47.879 occurrences (longueur total de l'ensemble des réponses en nombre de formes graphiques) pour les 2.186 réponses différentes de RAS (et toutes les variantes). Les répondants ont utilisé 4.588 mots distincts (formes graphiques), soit 9,6% des occurrences.²⁷ Parmi ces mots différents, 52% sont des hapax (forme graphique n'apparaissant qu'une fois). Les procédures de normalisation et quasi-lemmatisation ont réduit le vocabulaire étudié à 438 lemmes (ce sont soit des mots, soit des lemmes, soit des segments répétés). Ce vocabulaire représente à lui seul 57,4% des formes graphiques, sachant que les mots outils supprimés (articles, certaines prépositions ou pronoms relatifs) représentent 34,9% des formes graphiques. Les choix retenus pour la constitution du vocabulaire ont donc fortement réduit le nombre de mots (moins de 10% de l'effectif de départ), tout en conservant une grande partie du contenu thématique (au pire 75%).²⁸ L'objectif de cette réduction est de travailler sur des tableaux lexicaux plus petits, ce qui assure une meilleure robustesse des calculs. Parallèlement, dans une optique de classification, elle peut être à l'origine de la suppression de certaines classes à petits effectifs produisant un discours très particulier. Ces choix sont évidemment discutables, puisqu'ils introduisent une part de subjectivité dans un processus de quantification qui se veut, au départ, une méthode d'objectivation des résultats. Mais cette subjectivité est considérée comme une réponse possible aux problèmes posés par les données brutes et les résultats que l'on compte en tirer : elle doit être reconnue et assumée. En effet, notre hypothèse est que ces réponses sont autant de vues différentes sur le monde des usagers des services d'aide à travers le prisme de l'enquête "sans-domicile 2001" et que, parmi ces vues partielles de ce monde, certaines renvoient à une même chose, une même idée. L'approche quantitative des textes comme méthode pertinente s'appuie sur trois hypothèses que nous allons décrire en reprenant la métaphore de la ville développée par Saadi Lahlou (1995). Chaque réponse est analogue à une photographie prise par un touriste à Paris. A partir de ces centaines de clichés, on peut tenter de reconstituer les monuments de Paris. Ainsi, à partir des similarités entre les réponses, elles seront regroupées et assimilées à un objet ou une idée qu'elles seraient censées décrire. Comme la reconstitution de Paris sera d'autant plus pertinente que l'on connaît bien cette ville, l'interprétation des résultats et des classes produites sera d'autant plus pertinente que l'objet analysé est connu par l'analyste. Ceci pour rappeler que, quel que soit le degré d'objectivité des objets manipulés²⁹ et des techniques utilisées,³⁰ la phase d'interprétation fera nécessairement appel à la connaissance et aux rapports entretenus avec le sujet par l'analyste, ce qui peut être source d'une certaine subjectivité.

BIBLIOGRAPHIE

Aldeghi I. (1998), "Etude complémentaire sur les nouveaux arrivants au RMI : apports du RMI, évolution de la situation matérielle, opinions sur le dispositif", CREDOC, Collection des rapports, octobre, n. 196.

d'Aubigny G. (2001), "Introduction", *Journal de la Société Française de Statistique*, tome 142, vol. 4, pp. 1-5.

Baudelot C., Gollac M. (1997), "Faut-il travailler pour être heureux", *Insee-première*, 560, décembre.

Beck F., Legleye S., Peretti-Watel P., (2000), "Regards sur la fin de l'adolescence, consommations de produits psychoactifs dans l'enquête ESCAPAD 2000", OFDT.

Benzécri J.-P. (1981), *Pratique de l'Analyse des Données : linguistique et lexicologie*, Dunod, Paris.

Blair E., Sudman S., Bradburn N. M. et Stocking C. B. (1977), "How to ask questions about drinking and sex : Response effects in measuring consumer behavior", *Journal of Marketing Research*, 14, pp. 316-321.

Bradburn N. M. (1983), "Response Effects", in *Handbook of Survey Research*, P.H. Rossi, J.D. Wright et A. B. Anderson, eds. New York : Academic Press.

Bradburn N. M., Sudman S., et Associates (1979), *Improving interview method and questionnaire design: Response Effects to Threatening Questions in Survey Research*, San Fransisco : Jossey-Bass.

Bradburn N. M., Sudman S. (1982), *Asking Questions*, San Fransisco : Jossey-Bass.

Brugidou M., Escoffier C., Folch H., Lahlou S., Le Roux D., Morin-Andreani P., Piat G. (2000), "Les facteurs de choix et d'utilisation de logiciels d'analyse de données textuelles", *Actes des Journées Internationales d'Analyse des Données Textuelles 2000*.

Caillot P., Moine M. (2001) "Mais quelle est la réponse ?", *Journal de la Société Française de Statistique*, tome 142, vol 4, pp. 73-90.

Cicourel A. (1964), *Method and Measurement in Sociology*, New York : Free Press of Glencoe.

Craig S. C. (1985), "The decline of partisanship in the United States : A reexamination of the neutrality hypothesis", *Political Behavior*, n. 7, pp. 57-78.

Crosby A. W. (1997), *The Measure of Reality. Quantification and Western Society, 1250-1600*, Cambridge : Cambridge University Press.

Demazière D. (2005), "Des logiciels d'analyse textuelle au service de l'imagination sociologique", *Bulletin de Méthodologie Sociologique*, janvier 2005, n. 85, pp. 5-9.

Desrosières A. (2004), "Pour une politique des outils du savoir : le cas de la statistique", communication à la conférence "Politics and Knowledge: Democratizing Knowledge in Times of the Expert", Université de Bergen, 21-22 juin 2004.

Dohrenwend B. S. (1965), "Some effects of open and closed questions on respondents' answers", *Human Organization*, 24, pp. 175-184.

- Dohrenwend B. S., Richardson S. A. (1963), "Directiveness and non-directiveness in research interviewing: A reformulation of the problem", *Psychological Bulletin*, 60, pp. 475-485.
- Dubéchet P., Legros M. (1993), "La qualité des enquêtes auprès de populations en difficulté d'insertion sociale", CREDOC, *Cahier des recherches*, n. C47.
- Fassin D. (2000). "La supplique. Stratégies rhétoriques et constructions identitaires dans les demandes d'aide d'urgence", *Annales. Histoire, sciences sociales*, année 55, n. 5, septembre octobre 2000, pp. 953-981.
- Février P., d'Haultfoeuille X. (2002), "Une nouvelle modélisation de la non-réponse : l'effet enquêteur", Journées de Méthodologie Statistique, Insee, 2002.
- Geer J. G. (1988), "What do open-ended questions measure", *Public Opinion Quarterly*, vol. 52, pp. 365-371.
- Geer J. G. (1991), "Do open-ended questions measure salient issues ?", *Public Opinion Quarterly*, vol. 55 (3), pp. 360-370.
- Héran F. (1984), "L'assise statistique de la sociologie", *Economie et Statistique*, n. 168, juillet août 1984, pp 23-35.
- Jenny J. (1997), "Méthodes et pratiques formalisées d'analyse de contenu et de discours dans la recherche sociologique française contemporaine ; états des lieux et classification", *Bulletin de Méthodologie Sociologique*, n. 54, mars 1997, pp. 64-112 .
- Juan S. (1986), "L'ouvert et le fermé dans la pratique du questionnaire : analyse comparative et spécificités de l'enquête par correspondance", *Revue Française de Sociologie*, n. 27, pp 301-316.
- Kammeyer K. C. W., Roth J. A. (1971), "Coding responses to open-ended questions", *Sociological Methodology*, vol. 3, pp 60-78, American Sociological Association.
- Kelley S. (1983), *Interpreting elections*, Princeton : Princeton University Press.
- Labbé D. (2001), "Normalisation et lemmatisation d'une question ouverte : les femmes face au changement familial", *Journal de la Société Française de Statistique*, tome 142, vol. 4, pp. 37-57.
- Lahlou S. (1995), "Vers une théorie de l'interprétation en analyse statistique des données textuelles", JADT 1995, 3rd International Conference on Statistical Analysis of Textual Data.
- Lallich-Boidin G. (2001), "Données linguistiques et traitement des questions ouvertes", *Journal de la Société Française de Statistique*, tome 142, vol. n. 4, pp. 29-36.
- Lazarsfeld P. F. (1944), "The controversy over detailed Interviews -- an offer for negotiation", *Public Opinion Quarterly*, 8, pp. 38-60.
- Lazarfeld P. F., Barton A. H. (1955), "Some general principles of questionnaire classification", in Paul F. Lazarfeld et Morris Rosenberg (eds), *The Language of Social Research*, Glencoe IL : The Free Press.
- Lebart L. (2000), "Traitement statistique des questions ouvertes de l'enquête sur le devenir des personnes sorties du RMI", polycopie.
- Lebart L. (2001), "Traitement statistique des questions ouvertes : quelques pistes de recherche", *Journal de la Société Française de Statistique*, tome 142, vol. 4, pp. 7-20.
- Lebart L., Salem A. (1994), *Statistique textuelle*, Paris: Dunod.
- Marc X. (2001), "Les modalités de recueil des réponses libres", *Journal de la Société Française de Statistique*, tome 142, vol. 4, pp. 21-28.

- Montgomery A. C., Crittenden K. S. (1977), "Improving coding reliability for open-ended questions", *Public Opinion Quarterly*, vol. 41, pp. 235-243.
- Porter T. M. (1995), *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*, Princeton NJ : Princeton University Press.
- Poullaouec T. (2004), "Les familles ouvrières face au devenir de leurs enfants", *Economie et statistique*, n. 371, décembre 2004, pp 3-22.
- Reinert M. (1999), "Quelques interrogations à propos de l'"objet" d'une analyse de discours de type statistique et de la réponse "Alceste"", *Langage et Société*, n. 90, décembre 1999, pp. 57-70.
- Reinert M. (2001), "Approche statistique et problème du sens dans une enquête ouverte", *Journal de la Société Française de Statistique*, tome 142, vol. 4, pp. 59-71.
- Schuman H, Presser S. (1981), *Questions and Answers in Attitude Surveys*, New York : Academic Press.
- Sheatsley P. B. (1983), "Questionnaire construction and item writing", in P. H. Rossi, J. D. Wright et A. B. Anderson (eds.), *Handbook of Survey Research*, New York : Academic Press.
- de Singly F. (1984), "Les bons usages de la statistique dans la recherche sociologique", *Economie et statistique*, n. 168, juillet-août 1984, pp. 13-21.
- Smith E. R. A. N. (1989), *The Unchanging American Voter*, Berkeley : University of California Press.
- Wright B. D. (1997), "A History of Social Science Measurement". *Educational Measurement: Issues and Practice*, Winter 1997, pp. 33-45.

NOTES

1. Par sujet sensible, nous entendons enquête portant soit sur des pratiques (subies ou volontaires) ou des groupes pouvant être jugés comme stigmatisés. A titre d'exemple, nous pensons aux enquêtes suivantes : "Devenir des personnes sortant du RMI", "Analyse des comportements sexuels en France (ACSF)", "Enquête nationale sur les violences envers les femmes en France (ENVEFF)", "Enquête santé et consommation au cours de l'appel de préparation à la défense (ESCAPAD)", etc.
2. Le mot épreuve est doublement sensé dans notre cas : un questionnaire est parfois perçu comme un examen où l'enquêté se sent jugé par l'enquêteur, ce qui peut influencer sa réponse (de Singly, 1984) ; ce questionnaire peut être éprouvant du fait de l'accumulation de description de situations difficiles à vivre au quotidien.
3. On peut supposer que le terme "réducteur" correspond mieux à l'idée exprimée par l'enquêté.
4. Ce processus est valable pour toute étude statistique et correspond au passage de *data* (information collectée par le statisticien) à *given* (information offerte par le statisticien après apurement, redressement, imputation, etc.) pour reprendre ces anglicismes chers à Jean-Claude Deville.
5. Ceci sera discuté dans la section "Qu'est-ce que l'on mesure".
6. Jacques Jenny donne en 1997 une liste très complète des logiciels d'analyse textuelle développées et utilisés en France (Jenny, 1997).
7. A *contrario*, on peut regretter que l'enquête "Histoire de vie" (dite Identité) s'achève sur la question fermée suivante -- "Qu'est-ce qui permet le mieux de dire qui vous êtes ?" -- et une liste d'item : votre famille ; votre métier, votre situation professionnelle, vos études ; vos amis ; une passion ou une activité de loisirs ; les lieux auxquels vous êtes attachés ; vos origines

géographiques ; un problème de santé, un handicap ; votre physique, votre apparence ; vos opinions politiques ou religieuses ou vos engagements ; rien de tout cela.

8. Quelle profession ou quel genre de profession souhaiteriez-vous pour votre enfant ?

9. Cette hiérarchie est plutôt valable chez les adeptes de la sociologie quantitative. D'autres pensent, au contraire, que le processus d'enquête statistique est par essence voué à l'échec du fait du décalage entre les schèmes de perception des enquêtés et des enquêteurs (Cicourel, 1964). Enfin, Theodore M. Porter (1995) pense que l'utilisation de chiffre pour étayer son argumentation est le signe d'un moindre pouvoir institutionnel. Le chiffre serait l'arme du "faible" qui a besoin de justifier son interprétation des faits. D'autres, dont nous sommes, y voient deux approches complémentaires non hiérarchisables (voir, par exemple, Héran, 1984).

10. Cette hiérarchie implicite peut conduire à des pratiques étonnantes. Dans l'enquête PISA (Programme International pour le Suivi des Acquis des élèves) de l'OCDE, certaines questions sont formulées de plusieurs façons afin d'obtenir une mesure stricto-sensu en prenant la moyenne des réponses.

11. Les partisans de Craig pourraient répliquer que l'intérêt n'est pas indépendant du niveau d'éducation, sous l'hypothèse que plus le niveau d'éducation est élevé, plus on aurait de centres d'intérêt, mais ceci reste à discuter.

12. De nombreux articles traitent du sujet (par exemple : Blair *et al.*, 1977; Bradburn 1983; Bradburn, Sudman et Associates, 1979; Dohrenwend, 1965; Dohrenwend et Richardson, 1963 ; Juan, 1986 ; Lazarsfeld, 1944 ; Schuman et Presser, 1981 ; Sheatsley, 1983 ; Bradburn et Sudman, 1982). L'ensemble de ces recherches suggère qu'il n'existe pas de format de question supérieur à l'autre dans chaque situation et que, de ce fait, les deux formats ont leur place dans les enquêtes.

13. Dans les années 1960, une critique radicale de la sociologie quantitative (en particulier, la sociologie d'enquête par questionnaire développée par Lazarfeld) s'appuyait sur cette même idée de différence entre les schèmes de perception des acteurs (enquêtés) et des sociologues qui montre que l'erreur d'interprétation est d'autant plus grande que l'écart social et culturel est important (Cicourel, 1964). Ces critiques conduiront à une prise en compte de l'importance du questionnement, de l'interaction entre enquêteur et enquêté mais aussi des problèmes d'agrégation de réponses identiques formulées par des personnes de milieux sociaux différents.

14. Ce problème n'est pas entièrement résolu en travaillant sur l'ensemble du texte, du fait du biais possible introduit par la saisie des enquêteurs, mais aussi (voire surtout) par la quantification ou "mise en variable" des textes qui, par construction, conduit à une réduction de l'information.

15. C'est le principe de toute nomenclature construite en une arborescence de plus en plus fine.

16. Cette grille peut s'appuyer sur des réflexions reposant sur les autres questions de l'enquête ; en particulier, si la question ouverte vient compléter une batterie de questions sur un thème précis. Elle peut aussi reposer sur un sous-échantillon de réponses.

17. Ce dernier point doit être nuancé par le fait que si ce travail de réduction n'était pas fait, il n'y aurait, de fait, aucune information transmise.

18. Toutefois, rien n'empêcherait *a priori* un codeur de décrire chaque réponse selon une variable décrivant la forme (qualité de l'expression) et le fond (thème) de la réponse.

19. En particulier, il serait intéressant d'ajouter, si nécessaire, des inflexions pour rendre compte au mieux du discours produit.

20. Ceci correspond à un taux de réponse de 52%, ce qui est proche du taux de réponse (59%) à la question équivalente dans l'enquête devenir des allocataires du RMI.

21. Ces questionnaires ont été sélectionnés au hasard parmi les questionnaires en provenance d'Ile de France.

22. Ce taux est obtenu sur les réponses différentes de RAS et ses déclinaisons.

23. En effet, nous n'avons aucune connaissance des caractéristiques socio-démographiques ou de l'expérience de l'enquêteur qui aurait pu donner un autre éclairage aux résultats (Février et d'Haultfeuille, 2002).
24. Un même enquêteur enquêtait généralement des individus appartenant à la même structure ou du même type.
25. Cependant, l'introduction de la variable "type de structure" ne modifie pas les résultats.
26. Nous avons successivement testé un modèle probit à effets aléatoires ou à erreurs composées, un modèle logit à effets fixes et un modèle de Mundlak. Nous n'irons pas plus loin dans la description de ces modèles, car ils ne sont pas le sujet de notre étude.
27. A titre de comparaison, dans l'enquête devenir des allocataires du RMI, les 2.010 réponses contenaient 40.004 occurrences et 4.003 formes graphiques, soit 10% des occurrences.
28. Cette valeur est obtenue en rapportant la part de formes graphiques que constitue notre vocabulaire d'étude au vocabulaire initial privé des mots outils, ces derniers ne pouvant être vecteur de thème.
29. Ceci fait référence à l'échelle hiérarchisée des mesures en fonction de leur éloignement de la métrique *stricto sensu*.
30. De même que le statisticien préfère manipuler des nombres que des classes, il accorde généralement plus de crédit aux régressions qu'aux analyses de données à la française du fait de la possibilité de parler des effets d'une variable "toutes choses (in)égales par ailleurs".
-

RÉSUMÉS

L'analyse statistique des réponses aux questions ouvertes est le sujet de nombreuses controverses entre partisans et détracteurs, mais aussi au sein même des partisans sur le choix des méthodes d'analyse et de constructions des données. Au travers d'un exemple, nous nous focalisons sur deux controverses : l'intérêt des questions ouvertes ; l'utilisation de techniques de transformation du corpus textuel (la normalisation et la lemmatisation). Si la pertinence des questions ouvertes ne semble plus discuter, la "mise en variables" des textes reste une phase délicate de l'analyse textuelle sur laquelle un travail réflexif est nécessaire, en particulier dans la phase d'interprétation des résultats.

Turning Texts into Variables, Myth or Reality?: The statistical analysis of responses to open questions is the subject of many controversies between supporters and critics, but also among supporters concerning the choice of methods of analysis and of data construction. With an example, we focus our attention on two controversies: the interest of using open questions; the use of techniques for transforming corpora of texts (normalization and lemmatization). The pertinence of open questions is established, but turning texts into variables remains a delicate task in text analysis and requires reflexive considerations necessary particularly in the interpretation of results.

INDEX

Mots-clés : Analyse textuelle, Questions ouvertes, Quantification, Normalisation, Lemmatisation

Keywords : Text Analysis, Open Questions, Quantification, Normalization, Language, Lemmatization

AUTEUR

GAËL DE PERETTI

Chargé d'études sur la pauvreté, Institut National de la Statistique et des Etudes Economiques (Insee), gael.de-peretti@insee.fr