Journal de la Société Française de Statistique, 2003, p 7-21.

TRAITEMENT STATISTIQUE DES QUESTIONS OUVERTES; QUELQUES PISTES DE RECHERCHE

Ludovic Lebart

CNRS-ENST, 46 rue Barrault, 75013, Paris

lebart@enst.fr

Résumé

La présence de questions ouvertes dans les questionnaires d'enquêtes correspond à des préoccupations, des objectifs et des contraintes spécifiques. Les méthodes de statistiques exploratoires (analyses en axes principaux et les méthodes de validation qui leur sont associées, cartes auto-organisées de Kohonen), peuvent apporter des contributions importantes venant en complément des approches plus manuelles ou traditionnelles. On insiste sur le rôle important de la meta-information. Les exemples d'application concernent des réponses à une question ouverte dans une enquête internationale.

Mots Clef *: Lexicométrie, analyse de données textuelles, validation, bootstrap.*

Abstract

The presence of open-ended questions in a survey questionnaire corresponds to specific concerns and constraints. Recent trends of research show that exploratory techniques (principal axes methods and unsupervised clustering as well) are widely used and have great potential in both Text Mining and processing of responses to open questions. We will focus our paper on the assessments of visualizations, and the use of meta-data. The examples of application concern open-ended questions in an international survey.

Keywords *: Open-ended questions, text mining, visualization techniques, textual data analysis.*

1. Introduction

Il est intéressant, dans un certain nombre de situations d'enquête, de laisser ouvertes certaines questions, dont les réponses se présenteront donc sous forme de textes de longueurs variables. Le traitement de ce type d'information est complexe, et actuellement bien imparfait. Les outils de calcul et les méthodes statistiques descriptives multidimensionnelles peuvent cependant apporter une certaine aide à l'analyse de ces *réponses libres*.

On désigne précisément sous le nom *d'analyse des données textuelles* l'analyse statistique descriptive multidimensionnelle de textes (principalement analyses en axes principaux et diverses techniques de classification appliquées à des profils lexicaux de texte). C'est à propos d'une application linguistique que J.-P. Benzécri, dans les années soixante, a proposé la méthode d'analyse des correspondances, en insistant sur la propriété d'équivalence

distributionnelle de la distance du Chi-deux (travaux relatés dans : Benzécri, 1981) dans l'esprit de la linguistique distributionnelle prônée par Z. Harris lors de ses premiers travaux. La disponibilité de grands corpus de textes sur support digital a d'ailleurs redonné vie, sous le nom *de linguistique de corpus*, à ces approches trop précoces pour être mises en œuvre avec les moyens de calcul de l'époque (cf., par exemple, Habert *et al.*, 1998). Les premiers domaines d'applications ouverts par cette méthodologie furent d'abord des analyses de tableaux de présence-absence ou de tables de contingence construits manuellement à partir de textes (textes littéraires, politiques, religieux, etc.). Puis les programmes de calcul se sont enrichis de modules élémentaires de gestion. Ils ont travaillé à partir du texte brut, en extrayant automatiquement des unités statistiques, la plupart du temps des formes graphiques (séquences de caractères non-séparateurs). Ceci a ouvert la voie à de nouvelles applications, le traitement des questions ouvertes, notamment, et les analyses de textes documentaires. A partir d'une ensemble de textes, et d'un seuil de fréquence pour les formes graphiques, on obtient une visualisation des proximités entre textes (vis-à-vis de leurs profils lexicaux) et entre formes graphiques (vis-à-vis de leur répartition dans les textes). L'enrichissement des unités statistiques par les segments répétés, leurs regroupements par catégorisation morphologique, l'utilisation des formes caractéristiques ou spécificités, l'adjonction des réponses modales ou des phrases ou unités de contexte caractéristiques ont perfectionné ces approches, et mis à la disposition de beaucoup d'utilisateurs des méthodes et des logiciels utiles, mais qui ont beaucoup de progrès à faire. Dans certains domaines d'application précis (comme le traitement automatique des réponses aux questions ouvertes, qui nous intéresse ici), l'efficacité de la méthode, *comme complément des approches traditionnelles*, est reconnue.

2. Quand utiliser des questions ouvertes?

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

Pour diminuer ou optimiser la durée d'interview

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

Comme complément à des questions fermées

Il s'agit le plus souvent de la question classique: *"Pourquoi ?"*. Les explications concernant une réponse déjà donnée doivent nécessairement être spontanée. Une batterie d'items risquerait de proposer de nouveaux arguments qui pourraient nuire à l'authenticité de l'explication. L'utilité de la question *pourquoi ?* a été soulignée par de nombreux auteurs, et ce sont en fait les difficultés et le coût de l'exploitation qui en limitent l'usage. Elle seule permet en effet de savoir si les différentes catégories de personnes interrogées ont compris la question fermée de la même façon. Elle est particulièrement importante dans les enquêtes internationales, car elle permet de juger les éventuelles différences sémantiques des libellés selon la langue utilisée.

Pour recueillir une information qui doit, par nature, être spontanée

Les questionnaires des enquêtes de marketing abondent en questions de ce type. Citons par exemple : "Qu'avez-vous retenu de cette campagne publicitaire ?", "Que pensez-vous de cette voiture ?".

Notons que les questions ouvertes sont considérées comme peu adaptées aux problèmes de mémorisation de comportement. "Quels sont les noms des magazines que vous avez lus la semaine dernière ?" "Quelles sont les dernières émissions de télévision que vous avez aimées ?" Pour ces questions qui font l'objet d'enquêtes périodiques, il a été prouvé maintes fois que les questions fermées donnent des taux d'oubli plus faibles (Belson et Duncan, 1962). En revanche, quand la qualité de la mémorisation est en jeu (préoccupation très courante en marketing, lorsqu'il s'agit d'évaluer l'impact d'actions publicitaires), la forme ouverte est indispensable.

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

3. Traitement pragmatique des questions ouvertes

Le prétraitement appelé "post-codage" permet de fermer *a posteriori* les questions ouvertes. Cette technique courante consiste à construire une batterie d'items à partir d'un sous-échantillon de réponses, puis à codifier l'ensemble des réponses de façon à remplacer la question ouverte par une ou plusieurs questions fermées. Pour des réponses simples, stéréotypées et peu nombreuses, cette procédure fonctionne bien. Mentionnons cependant parmi les défauts de ce type de traitement :

- La médiation du chiffreur: les décisions à prendre sont parfois difficiles.
- La qualité de l'expression, le registre du vocabulaire, la tonalité générale de l'entretien sont des éléments d'analyse perdus lors d'un post-codage (doit-on coder différemment “ je ne sais pas” et “je préfère ne rien dire” ?.
- Les réponses composites, complexes, d'une grande diversité, sont très dofficile à post-coder, et c'est souvent dans ce cas que la valeur heuristique des réponses libres est la plus grande.
- Les réponses peu fréquentes, originales, peu claires en première lecture sont considérées comme du “bruit”, et affectées à des items résiduels (“autres”) qui sont donc très hétérogènes et sont difficiles à manipuler. Ces réponses relativement peu fréquentes peuvent cependant être émises par une catégorie d'individus particulière et importante dans la problématique de l’enquête, ce qu'il n'est pas possible de savoir lors d'un post-codage "a priori" de l'information...

4. Particularités statistiques des réponses aux questions ouvertes

Dans les réponses aux questions ouvertes, le statut de la fréquence des mots est ambigu. Les fréquences lexicales observées sont pour une large partie artificielles, car la même question est parfois posée à des centaines ou des milliers de personnes... la juxtaposition des réponses constitue un "texte" redondant par construction (cf. la schématisation de la figure 1). Il existe en fait deux niveaux d'individus statistiques: les personnes interrogées, qui sont les individus habituels des statisticiens d'enquêtes, et les mots utilisés, qui sont, en quelque sorte, les individus habituels des lexicométriciens. Certains tests statistiques seront significatifs au niveau des mots, mais pas au niveau des individus, et donc ne permettront pas une inférence

des résultats à la population interrogée. Les analyses de réponses regroupées seront en fait assez voisines des analyses de "vrais textes" (littéraires, politiques, historiques), alors que les analyses de réponses individuelles seront, elles, voisines des traitements effectués en recherche documentaire. L'originalité de l'approche résultera, en fait, du grand nombre de possibilités de regroupement, et donc du grand nombre de grilles de lectures possibles. Une question ouverte représente un très grand nombre de textes artificiels potentiels, et donc aussi grand nombre de points de vue possibles sur les réponses.

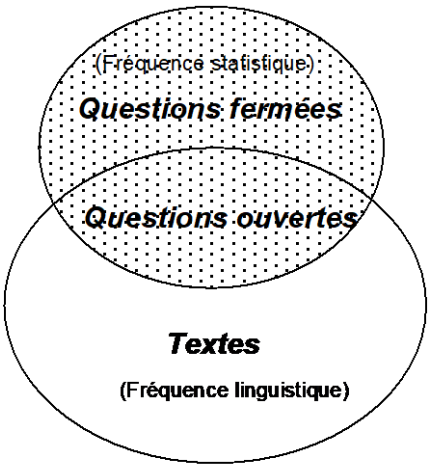

Figure 1. Questions ouvertes: le statut ambigu des fréquences

5. Les unités et les similarités

Les méthodes d'analyse descriptive des données multidimensionnelles qui constituent le coeur de la méthodologie de l'analyse des données textuelles se perfectionnent progressivement en s'efforçant d'utiliser d'une part la méta-information disponible (informations que l'on possède sur le tableau de données que l'on s'apprête à analyser, et qui ne figurent pas dans le tableau lui-même) d'autre part les techniques de validation modernes essentiellement fondées sur les méthodes de rééchantillonnage (Monte-Carlo, validation croisée, Bootstrap). On peut schématiser ces derniers développements en disant qu'ils tentent de répondre aux deux questions :

- *Comment utiliser ce que l'on sait pour essayer d'en savoir plus ?*
- *Comment valider ou généraliser les résultats obtenus ?*

Pour un statisticien, la meta-information disponible (Figure 2) dans le cas de données textuelles est quelque chose de démesuré, si on compare cette meta-information à celle des recueils de données numériques ou qualitatives habituels (fichiers d'enquête par exemple). Chaque mot utilisé, même si c'est un mot grammatical (appelé parfois mot vide en documentation, ou encore mot-outil) a droit à plusieurs lignes, ou plusieurs pages dans un dictionnaire encyclopédique.

Les règles de grammaire constituent évidemment une méta-information fondamentale ; les corpus externes également. Les mots appartiennent à des réseaux sémantiques que l'usage simultané de dictionnaires, de thésaurus et d'analyseurs morpho-syntaxiques partiellement automatisés s'efforcent de prendre en compte. Le développement des analyses exploratoires ainsi que le travail sur bases de données accentuent l'intérêt de la notion de méta-information.

Redécouvrir des structures connues est en effet utile à fin de vérification, mais ne constitue évidemment pas la fin ultime de ces analyses qui doivent apprendre à utiliser ce qui est déjà connu pour en savoir davantage.
Évaluer des similarités entre entités textuelles est un des problèmes centraux dans plusieurs disciplines comme l'analyse de données textuelles, la recherche documentaire ou l'extraction de connaissances à partir de données textuelles (fouilles de textes ou *Text Mining*).

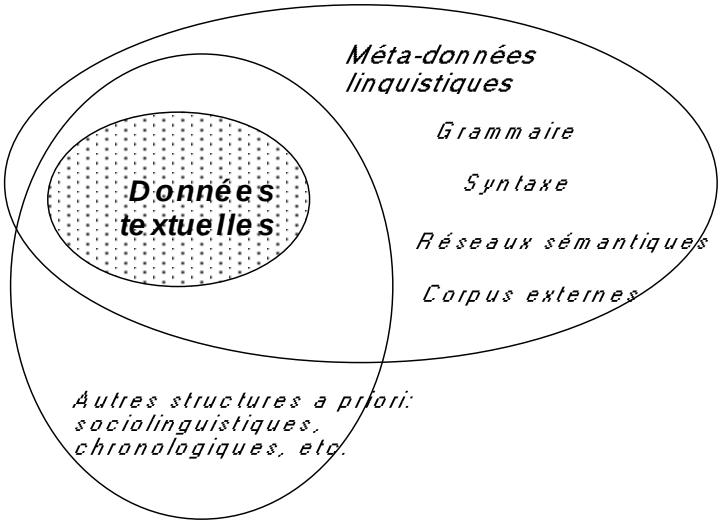

Figure 2. Donnés textuelles et meta-informations

En Analyse des Données Textuelles, la distance du chi-deux (χ^2) est un choix fréquent. En Recherche Documentaire, des similarités dérivées de mesures à base de cosinus (Salton, 1988) sont utilisées alors qu'en Text Mining on préfère souvent des mesures issues de la théorie de l'information (comme la « distance » de Kullback-Leibler à base d'entropie relative) (cf. Lebart et Rajman, 2000). Afin de produire les structures qui vont être utilisées pour représenter les textes lors du calcul des similarités, les données textuelles doivent tout d'abord être décomposées en unités lexicales plus simples. Plusieurs choix sont possibles et les différentes unités retenues auront des degrés de pertinence variables selon le domaine d'application particulier choisi.
Une approche classique pour définir les unités textuelles dans une corpus est d'utiliser les formes de surface (*formes graphiques*, ou plus simplement *mots*) pouvant être produites par des techniques simples de segmentation automatique. Cependant, ces unités élémentaires peuvent également faire l'objet de traitements additionnels permettant l'intégration de connaissances linguistiques plus sophistiquées dans les représentations. L'étiquetage morpho-syntaxique (i.e. l'affectation automatique aux mots d'étiquettes grammaticales) et/ou la lemmatisation (i.e. la réduction automatique des formes fléchies à une représentation canonique) sont des exemples de telles techniques de pré-traitement des données textuelles (cf. Labbé, 1990). Du fait que le sens des mots est fortement lié à la manière dont ils apparaissent en combinaison (par exemple, des expressions composées comme « sécurité sociale » ou «cul de basse fosse» ont des significations qui ne peuvent être simplement dérivées du sens de leurs constituants), il peut également être utile de prendre en compte des unités plus larges constituées de plusieurs mots. L'utilisation des « segments répétés » (Salem, 1984) ou des « quasi-segments » (Becue et Peiro, 1993), reposant sur la détection automatique des séquences répétitives, constituant ou non des formes ou expressions composées, est une solution possible mais des approches combinant des connaissances linguistiques et statistiques pour identifier de façon automatique les formes composées (ou termes) sont aujourd'hui également disponibles (Daille, 1994). Il est cependant à noter que

l’utilisation de techniques d’extraction plus sophistiquées pour les unités servant à la représentation des textes présuppose la disponibilité des ressources linguistiques nécessaires (ce qui n’est pas forcément le cas pour toutes les langues) et, de plus, augmente de façon sensible le nombre total d’unités à prendre en compte dans les étapes ultérieures de traitement.

6. Les analyses statistiques; les outils de base.

On a évoqué les problèmes posés par la prise en compte de la méta-information. Même si cs problèmes étaient résolus, le matériau textuel resterait encore un objet relativement complexe pour le statisticien. Il existe en effet dans tout texte une dimension séquentielle, ou syntagmatique, qui l'apparente à ce qu'on désigne en statistique sous le nom de processus, mais il s’agit d’un processus particulièrement complexe, puisque qu'il est à la fois qualitatif et multidimensionnel. L'introduction d'unités composites, comme les segments, permet une prise en compte partielle de l'ordre des mots dans les phrases. On peut aussi utiliser les informations données par un arbre syntaxique (dont l'obtention ne peut être entièrement automatique) dans les calculs de similarités entre phrases ou réponses. Notons que des procédures d'analyse statistique multidimensionnelle, comme l'analyse des correspondances (section 6.1 ci-dessous) d'une table de contingence croisant des mots (en ligne) et des parties de textes (en colonne) ne font appels qu'à des modèles statistiques rudimentaires, mais peu contraignant (l'hypothèse nulle - pratiquement toujours rejetée - est l'hypothèse d'indépendance des mots et des parties de texte). La méthode décrit par quelles associations (entre mots, entre textes) la table de contingence s'éloigne de cette hypothèse nulle. C'est une sorte de bilan global, qui ne prend pas en compte la nature séquentielle du texte, mais qui peut intégrer toutes les nouvelles variables construites après une analyse linguistique.
Les outils de base sont l'analyse des correspondances et la classification des tableaux lexicaux, les sélections de formes caractéristiques, les sélection de réponses modales.

6.1 Analyse des correspondances et classification

Les analyses des correspondances et les méthodes de classification peuvent décrire les tables de contingence croisant les réponses et les formes graphiques (Benzécri, 1981), ou des groupes de réponses (par exemple regroupement selon le niveau d'instruction des répondants) et les formes graphiques. Elles permettent de visualiser sous forme de séries de cartes planes (ou de dendrogrammes, ou de cartes auto-associatives de Kohonen) les associations entre mots (formes) et groupes ou modalités. Ainsi, une visualisation des proximités entre mots et catégories socioprofessionnelles pourra aider la lecture des réponses de chacune de ces catégories. Avec ce type de représentation, la présence de mots-outils est parfaitement justifiée: si ces mots caractérisent électivement certaines catégories, ils se positionnent dans leur voisinage, et peuvent être intéressants à interpréter; si au contraire leur répartition est aléatoire, ils seront situés dans la partie centrale du graphique, sans en encombrer la lecture. On voit quelle est l’importance fondamentale des outils de validation permettant de juger la signification statistique de la positions des points dans les visulalisations (cf. section 6.4 plus bas). De même, la présence de plusieurs flexions d'un même verbe peut aussi constituer un outil de validation. Il est d'ailleurs toujours intéressant de comparer les analyses sur textes bruts et sur textes lemmatisés, qui selon les cas, se confirment ou se complètent. Fondée sur la distance du chi-deux, l'analyse des correspondances peut aussi s'appliquer à certains tableaux binaires (comme les matrices associées à des graphes sémantiques) pour lesquels elle donne

en général des visualisations de meilleure qualité que l'analyse en composantes principales (Lebart *et al*, 1998).
La figure 3 représente une carte auto-organisée de Kohonen dans le contexte suivant : une question ouverte a été posée lors d'une enquête multinationale (cf. Hayashi *et al*.; 1992) avec le libellé suivant "*Que signifie pour vous la culture de votre pays*?". C'est le sous-échantillon relatif à la France qui est ici traité (effectif : 1009 personnes représentatives des personnes résidentes de 15 ans ou plus).

littérature grands art +55/HAUT	niveau esprit comme -30/HAUT	france	écrivains surtout histoire cinéma	peut grand	patrimoine
vivre faire arts 30-55/HA		langue	français chose		notre nos
tous musique	pense aussi	peu pays par française culture		enfants autres	même 30-55/MO
être trop gens +55/MOYE	très ont fait	monde	temps	beaucoup	mais
peinture	vie faut	sont nous bien	pour plus jeunes ils +55/BAS	pas	livres
théâtre quand moins cultivés -30/BAS	choses assez 30-55/BA	mal façon bonne		école télévision sais	sait rien lecture -30/MOYE

Figure 3. Représentation par carte de Kohonen des proximités entre mots et entre catégories.

L'ensemble des réponses à cette question représente 14742 occurrences de 2248 mots (formes graphiques) distincts. Sont traités ici les 111 mots apparaissant au moins 20 fois, qui représentent quand même 10005 occurrences gardées, auxquels on a retranché 40 mots-outils ou mots vides usuels, ce qui porte à 71 le nombre de mots gardés. Les réponses sont ici été regroupées en neuf catégories obtenues par croisement de l'âge en trois classes (moins de 30 ans, 30-55 ans, plus de 55 ans) et du niveau d'éducation en trois classes (bas, moyen, haut).

La figure 3 représente donc une carte de Kohonen suivant le même système de distance que celui de de l'analyse des correspondances de la table de contingence qui croise les occurrences des 71 mots les plus fréquents et les 9 catégories précédentes (les identificateurs des catégories sont limités à 8 caractères). On obtient ainsi une synthèse assez comparable à celle de l'analyse des correspondances qui figure plus loin (figure 4). Cette synthèse nous permet de repérer assez rapidement des cooccurrences de mots dans une même catégorie, et des proximités lexicales entre catégories. En général, la dimensionalité prise en compte par cette méthode de classification visualisée dépasse les deux dimensions d'un premier plan

factoriel. En revanche, les méthodes de validation font défaut dans l'état actuel des implémentations disponibles (les figures 3 et 4 sont issues du logiciel académique DTM).

6.2 Formes ou segments caractéristiques (ou spécificités)

Il est utile de compléter les représentations spatiales fournies par l'analyse des correspondances (qui ne sont que des approximations) par quelques paramètres d'inspiration plus probabiliste : les formes caractéristiques. Ce seront les formes "anormalement" fréquentes dans les réponses d'un groupe d'individus (Lafon, 1980). Un test élémentaire fondé sur la loi hypergéométrique permet de sélectionner les mots (formes graphiques ou lemmes) dont la fréquence dans un groupe est notablement supérieure (ou inférieure pour les mots *anti-caractéristiques*) à la fréquence moyenne dans le corpus.

6.3 Les sélections des réponses modales

Pour un groupe d'individus donné, et donc pour le regroupement de réponses correspondant, les réponses modales (ou encore phrases caractéristiques, ou documents-type, selon les domaines d'application) sont des réponses originales du corpus de base, ayant la propriété de caractériser au mieux la classe. On peut, pour chaque regroupement, calculer la distances du profil lexical d'un individu au profil lexical moyen du groupement. On peut ensuite classer les distances par ordre croissant, et donc sélectionner les réponses les plus représentatives au sens du profil lexical, qui correspondront aux plus petites distances. On obtient ainsi une sorte de résumé des réponses de chaque regroupement, formé de réponses originales (Lebart, 1982).

6.4 Un outil de validation important : les zones de confiance "bootstrap"

Les visualisations provenant des analyses en axes principaux (surtout : composantes principales et correspondances) n'ont de sens que si elles sont accompagnées de la confiance que l'on peut accorder à la position de chaque point. La technique de *bootstrap* (cf. Efron et Tibshirani, 1993) va consister, dans le contexte d'une table lexicale, à construire n réplications de l'échantillon par tirage avec remise des occurrences de formes retenues selon un schéma multinomial comportant autant de catégories que la table a de cellules, et dont les fréquences théoriques sont précisément celles de chaque cellule de la table lexicale. Les lignes et les colonnes des tables répliquées sont alors projetées comme éléments supplémentaires sur les plans factoriels de l'analyse de la vraie table lexicale (cf. Château et Lebart; 1996). Des analyses en composantes principales des réplications propres à chaque élément fournissent alors les ellipses de confiance cherchées.

La figure 4 représente de telles zones de confiances dans le premier plan factoriel de l'analyse des correspondances de la table de contingence (déjà utilisée lors de la représentation de la figure 3) qui croise les occurrences des 71 mots et les 9 catégories précédentes. Comme il n'est pas possible de représenter sur un même graphique toutes les ellipses, on a choisi quatre catégories (encadrées) en excluant les âges moyens et les niveaux d'éducation moyens. On a de même choisi cinq mots (formes graphiques) de façon à recouvrir les différentes zones du graphique *(arts, chose, choses, cinéma, télévision*).
Remarquons que les zones de confiance des catégories sont relativement petites, indiquant des positions (et des positions relatives) assez significatives pour ces catégories (la plus grande concerne les jeunes peu instruits, catégorie dont l'effectif est le plus faible (seulement 13 réponses dans l'échantillon).
On voit qu'à la mention du mot *télévision* correspond une longue ellipse de confiance, qui fait osciller la position du point entre les catégories –30/BAS (jeunes peu instruits) et +55/BAS (seniors peu instruits) mais reste du côté de ces classes de niveaux d'éducation. On note aussi le comportement inverse du mot *arts*.

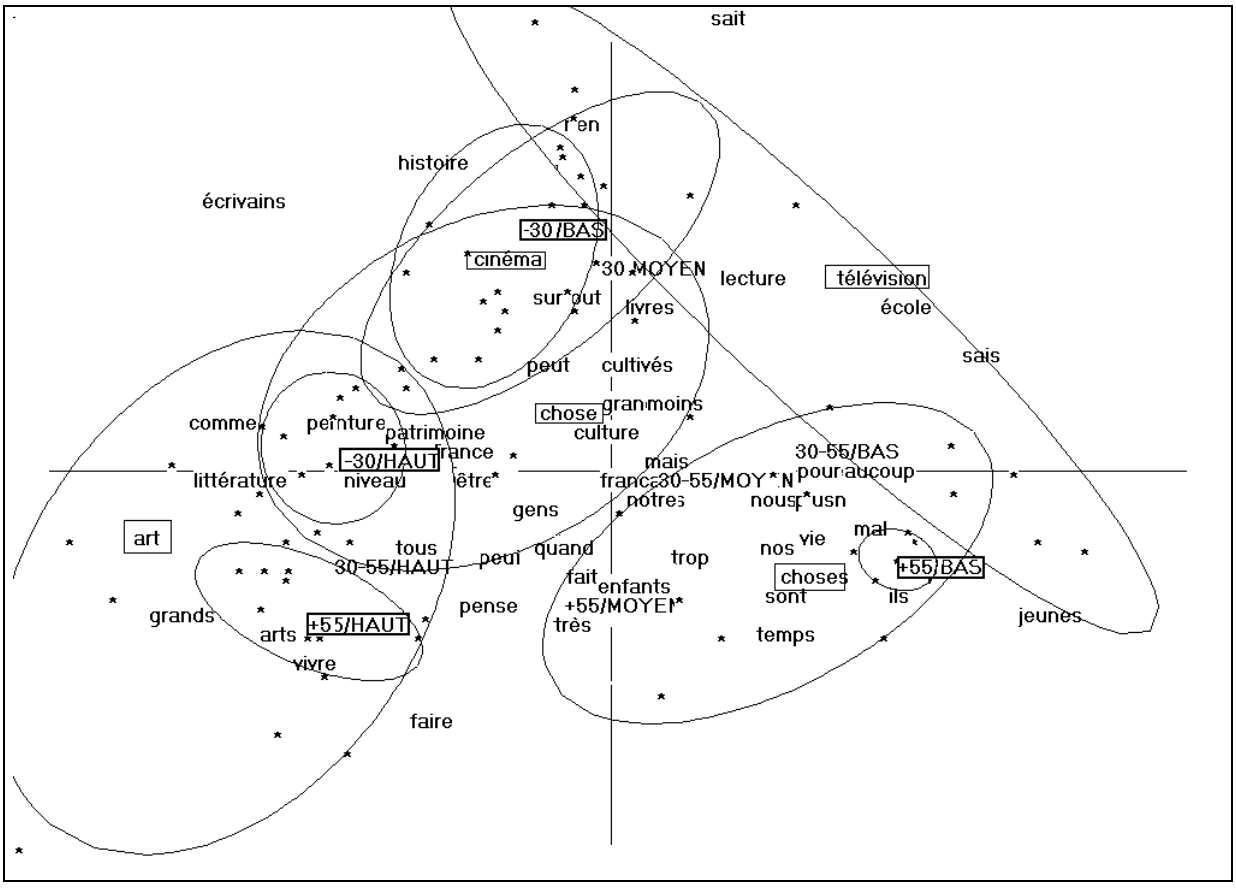

Figure 4. Zones de confiance pour 4 catégories de répondants et pour 5 mots.

Enfin, on a représenté les positions de deux formes graphiques (*chose* et *choses*) du même lemme *chose*, pour montrer qu'il peut arriver que les positions de telles formes soient significativement distinctes (les ellipses de confiance ne se recouvrent pas) (il semble que *chose* soit souvent associé à la réponse *pas grand chose*, qui exprime une opinion critique ou désabusée vis-à-vis de la culture du pays).

7. L'évolution des outils

Les analyses de type linguistique, automatisées ou non, vont permettre d'extraire du texte une partie de la meta-information mentionnée plus haut. Les logiciels correspondants, qui alimenteront les procédures d'analyse statistique structurale, font eux-mêmes de plus en plus appel à des procédures statistiques pour lever certaines ambiguïtés. Ces logiciels devront aussi faire appel à des dictionnaires de formes ou de locutions qui se sont également beaucoup développés au cours des années récentes. La prise en compte de ces nouvelles informations peut prendre la forme de nouvelles variables concernant les documents ou individus, que l'on peut alors positionner en éléments supplémentaires dans les analyses descriptives, ou, en amont, que l'on peut faire intervenir pour enrichir les indices de similarité.

7.1 Améliorer les similarités : graphes sémantiques et métrique

Les informations sémantiques peuvent aussi conférer une nouvelle structure à l'espace des formes graphiques, en munissant, par exemple, d'une métrique particulière l'espace des formes graphiques ou

des lemmes. Une analyse de similarités appliquée à une matrice de profils lexicaux peut mener à des résultats décevants. Dans le cas des réponses (souvent brèves) à une question ouverte, la matrice des profils peut en effet être extrêmement creuse: beaucoup de lignes n'auront alors aucun élément commun, et les distances entre lignes n'ont plus de sens. On doit alors envisager de mettre à profit la meta-information disponible (relations syntaxiques, réseaux sémantiques, corpus externes, thesaurus, …). Ainsi, pour rendre les dissimilarités entre profils lexicaux plus riches de sens, il peut être utile de prendre en compte des informations sémantiques sur les unités textuelles et, en particuliers, sur les similarités (sémantiques) entre ces unités. Bien que l'on ne dispose d'aucune règle universelle permettant d'établir si deux mots sont sémantiquement équivalents, il est habituellement reconnu que les co-occurrences (au sein d'une même phrase) sont des éléments pertinents pour la détermination du sens, ce qui peut déterminer l'appartenance à un certain *halo sémantique*.

7.2 Le graphe sémantique

Les nœuds d'un tel graphe sont les différentes unités textuelles (les « mots ») et ses arcs (non orientés) sont pondérés en fonction d'un indice d'intensité de co-occurrence entre les unités correspondantes. Un graphe sémantique est ainsi complètement décrit par une *matrice de poids*, M, d'ordre (p, p) où p est le nombre total d'unités distinctes. Un graphe sémantique peut être construit à partir d'une source externe d'information (un dictionnaire de synonymes ou un thésaurus par exemple) ou dérivé des associations effectivement observées dans un corpus. Le corpus utilisé peut être un corpus extérieur ou même la collection de documents sur laquelle porte l'analyse. Dans ce dernier cas, les similarités entre deux unités pourront être dérivées à partir des proximités entre leurs profils lexicaux dans la collection. Si un graphe sémantique est disponible, les techniques développées dans le cadre de l'analyse de données textuelles peuvent être adaptées par le biais du calcul d'un nouvel indice de similarité entre textes (Becue et Lebart, 1996). Ainsi, dans le cas d'un graphe sémantique externe (i.e. dérivé à partir d'informations autres que le corpus analysé lui-même), une façon simple pour prendre en compte les voisinages sémantiques est de remplacer la matrice des profils **T** par la nouvelle matrice $\mathbf{T}(\mathbf{I} + \alpha\mathbf{M})$, où **I** est la matrice identité, **M** la matrice des poids définissant le graphe sémantique et α un paramètre numérique inférieur à 1 permettant de calibrer l'importance accordée aux voisinages sémantiques. Cette approche revient à munir l'espace de représentation de dimension p d'une nouvelle *métrique* définie par $(\mathbf{I} + \alpha\mathbf{M})^2$ ce qui mène de façon immédiate à un nouvel indice de similarité qui peut être utilisé pour le calcul des similarités textuelles. Du fait de la taille des tables de données manipulées, les calculs effectifs sont souvent réalisés à l'aide des coordonnées sur les premiers axes principaux.

7.3 Les analyses en axes principaux comme filtres

De telles propriétés contribuent à souligner le rôle primordial des premiers axes principaux dans le calcul des similarités définies pour les analyses de données textuelles. Dans le domaine de la recherche documentaire, différentes techniques à base de co-occurrences ont également été utilisées, comme par exemple les vecteurs de co-occurrence moyens de l'approche « sémantique distributionnelle » proposée dans (Rajman et Rungsawang, 1995). Plusieurs de ces approches s'appuient sur des outils d'analyse factorielle très similaires à ceux proposés par Benzécri (1977) (cf. Lebart (1982) pour le traitement de matrices creuses de grandes dimensions). Par exemple, Deerwester *et al*., 1990, suggèrent indépendemment, sous le nom de « Latent Semantic Indexing », une approche dans laquelle l'hypothèse fondamentale est aussi que les relations terme/document, implicitement représentées dans la matrice des profils, sont en fait obscurcies par les phénomènes de variabilité lexicale et que la matrice des profils doit donc être traitée par le biais d'une décomposition en valeurs singulières (SVD en anglais) (cf., *e.g*., Berry, 1996) qui permet de remplacer les profils lexicaux par les coordonnées des documents dans le sous-espace engendré par les k premiers vecteurs principaux produits par la SVD. Cette nouvelle représentation a l'avantage d'encoder les relations d'association entre mots et documents d'une façon qui repose exclusivement sur les mots et non sur des informations externes : par suite de la réduction dimensionnelle, deux documents pourront alors être proches dans le sous-espace engendré même s'ils ne possèdent aucun mot commun. Il s'agit d'une sorte de lissage des textes.

BIBLIOGRAPHIE

Becue M., Lebart L., (1996). Clustering of Texts using Semantic Graphs. Application to Open-ended Questions in Surveys, in*: ,* Hayashi *et al.* (eds), *Data science, Classification and Related Methods*, Springer, Tokyo, 480-487.

Becue, M., Peiro, R.(1993). Les quasi-segments pour une classification automatique des réponses ouvertes, in *Actes des 2ndes Journées Internationales d'analyse des données textuelles, (Montpellier)*, ENST, Paris, 310-325.

Belson W.A., Duncan J.A.(1962). A Comparison of the check-list and the open response questioning system, *Applied Statistics* n°2, p 120-132.

Benzécri J.-P.& coll. (1981). *Pratique de l'analyse des données*, tome 3, Linguistique & Lexicologie, Dunod , Paris.

Benzécri J.-P., (1977). Analyse discriminante et analyse factorielle, *Les Cahiers de l'Analyse des Données*, II, 4, 369-406.

Berry M. W., Low-Rank Orthogonal Decompositions for Information Retrieval Applications, *Numerical Linear Algebra with Applications*, vol 1(1), 1996, 1-27.

Brunet E. (1981). *Le vocabulaire français de 1789 à nos jours, d'après les données du Trésor de la langue française,* Slatkine-Champion, Genève-Paris.

Château F., Lebart L., "Assessing sample variability in visualization techniques related to principal component analysis : bootstrap and alternative simulation methods". In: *COMPSTAT96, Proceedings in Computational Statistics*, A. Prats, (ed.), Physica Verlag, Heidelberg, p. 205-210, 1996

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshman R. (1990). Indexing by latent semantic analysis, *J. of the Amer. Soc. for Information Science*, 41 (6), p 391-407.

Efron B., Tibshirani R. J., *An Introduction to the Bootstrap*, Chapman and Hall, New York, 1993.

Fowler R.H., Fowler W.A.L., Wilson B.A. (1991). Integrating query, thesaurus, and documents through a common visual representation, *Proceedings of the 14th Int. ACM Conf. on Res. and Dev. in Information Retrieval*, Bookstein A. and al., , Ed, ACM Press, New York, p 142-151.

Habert B., Nazarenko A., Salem A.(1997). *Les linguistiques de corpus*. Armand Colin, Paris.

Hayashi C., Suzuki T., Sasaki M. (1992) *Data Analysis for Social Comparative Research: International Perspective*. North-Holland, Amsterdam.

Kohonen T., *Self-Organization and Associative Memory*. Springer-Verlag, Berlin, 3rd edition, 1989.

Labbé D., (1990). Normes de saisie et de dépouillement des textes politiques, *Cahier du CERAT,* Grenoble.

Lazarsfeld P.E. (1944). The controversy over detailed interviews - an offer for negotiation, *Public Opinion Quat.* n°8, p 38-60.

Lebart L. (1982). L'Analyse statistique des réponses libres dans les enquêtes socio-économiques, *Consommation,* n°1, Dunod, p 39-62.

Lebart L., Rajman M., *Computing Similarities*, in: Dale R., Moisl H., Somers H., (Editors): *Handbook of Natural Language Processing,* Marcel Dekker, New York, 2000, 477-505.

Lebart L., Salem A. (1994). *Statistique textuelle*, Dunod, Paris.

Lebart L., Salem A., Berry E., (1998). *Exploring Textual Data*, Kluwer Academic Publisher, Dordrecht.

Lebart L., Morineau A. Piron M. *Statistique Exploratoire Multidimensionnelle*, Dunod, Paris, 2000.

Muller C. (1977). *Principes et méthodes de statistique lexicale,* Hachette, Paris.

Rajman M., Rungsawang A. (1995) -*Textual Information Retrieval based on the Concept of Distributional Semantics*, in S. BOLASCO et al. (eds.) *3rd International Conference on Statistical Analysis of Textual Data (JADT'95)*, Rome, 151-162

Salem A., (1984). *La typologie des segments répétés dans un corpus, fondée sur l'analyse d'un tableau croisant mots et textes*, Les Cahiers de l'Analyse des Données, 9, n° 4, 489-500.

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