Les traitements statistiques de données textuelles.

(L. Lebart, CNRS-ENST ; lebart@enst.fr)

Le matériau statistique « texte » est omniprésent, presque banal, depuis le développement d'internet et de la toile (*web*). L'étude quantitative et statistique de ces textes semble avoir fait irruption récemment, et pourtant les études statistiques de textes datent de plusieurs décennies, avec notamment en France les travaux de P. *Guiraud (Problèmes et méthodes de la statistique linguistique*, PUF, 1960), C. Muller (*Principes et méthodes de statistique lexicale*, Hachette, 1977) puis de J.P. Benzécri (Pratique de l'Analyse des Données, Tome 3 : *Linguistique et lexicologie*, Dunod, 1981).

Après la « stylométrie », consacrée à l'étude de la forme des textes, en vue d'identifier un auteur ou de dater une œuvre, sont apparues les techniques de documentation automatique (*Information retrieval* en Anglais), visant à rechercher dans une base de documents (articles scientifiques, résumés, brevets, …) le ou les éléments pertinents à partir d'une requête exprimée sous forme de textes libres. Le champ disciplinaire « Traitement du Langage Naturel » est alors apparu, et s'est développé, au départ, comme un des domaines d'application privilégié de l'intelligence artificielle. La complexité du matériau, le besoin d'assimiler d'immenses corpus de textes, la pertinence du concept d'apprentissage ont naturellement ouvert ce champ aux méthodes statistiques. La statistique multidimensionnelle, les chaînes de Markov cachées, les méthodes d'analyse discriminantes interviennent ainsi pour construire les outils de base que sont les moteurs de recherche sur le *web*, les analyseurs morphosyntactiques, les correcteurs orthographiques, ainsi que dans des champs d'application pratiques comme le traitement des réponses aux questions ouvertes dans les enquêtes socio-économiques.

Les questions ouvertes

Il est utile, 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 recueil des données

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

Pour diminuer ou optimiser la durée de l'entrevue d'enquête

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 furent vos principales activités dimanche dernier ?") peut remplacer de longues listes d'items.

Comme complément à des questions fermées

Il s'agit le plus souvent de la question: *"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.

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 cependant que les questions ouvertes sont considérées comme peu adaptées aux problèmes de mémorisation de comportement. "Quels magazines avez-vous 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. En revanche, quand la qualité de la mémorisation est en jeu, la forme ouverte reste indispensable.

Voici quatre exemples de réponses à la question « Quelle est pour vous la chose la plus importante dans la vie ? » (Question posée à des échantillons d'environ mille personnes dans sept pays en1991).

1) La santé, ne pas manquer d'argent, avoir une bonne ambiance familiale, je voudrais pouvoir aider les enfants abandonnés, leur redonner le goût à la vie, pouvoir aider les personnes âgées handicapées, secourir les gens autour de soi.

2) C'est de faire ce qu'on veut. Lire, voyager si je pouvais. Les loisirs si on pouvait.

3) La santé puisqu'il faut toujours travailler quand on est commerçant. Une bonne entente en famille. Avoir assez d'argent pour vivre.

4) La famille, ma famille, mon foyer, vivre avec la société : mon entourage les voisins, pour faire quelque chose qu'il y ait moins de malheureux, donner du travail aux jeunes surtout.

Ces exemples illustrent à la fois la complexité et la richesse des réponses.

Les unités statistiques

Les programmes travaillent à partir du texte brut, en extrayant automatiquement des unités statistiques, la plupart du temps des *formes graphiques* (séquences de caractères non-séparateurs). On utilise le vocable *forme graphique* parce que le mot « mot » lui-même est ambigu. Il désigne en effet selon les contextes *l'occurrence* d'un mot (quand on dit qu'un texte a huit cent mots, on parle bien sûr d'occurrences, et non de mots différents), le type (qui correspond à la forme graphique) et le *lemme* (*avoir* est le lemme de *avait*, et, dans certains cas seulement, de *avions*). La première réponse de l'exemple ci-dessus contient 38 occurrences, mais la forme graphique « les » apparaît trois fois, « pouvoir » apparaît deux fois. Le lemme de « bonne » est bon (le masculin singulier, selon une convention française), celui de « voudrais » est « vouloir ».

Dans le cas de l'exemple précédent, pour 1009 réponses, on obtient 14337 occurrences de 1394 formes distinctes (ou types). Il est bien connu que la distribution de fréquence des mots est très dissymétrique (loi dite de Zipf, apparentée à la distribution de Paréto). Ainsi, en ne retenant que les formes apparaissant au moins 20 fois, il reste un texte de 10 994 formes, avec seulement 97 formes distinctes (ainsi 7 % des mots distincts correspondent à 77 % du texte global). En particulier, près de la moitié des formes grahiques distinctes n'apparaissent qu'une fois (ce sont les « hapax »).

Le post-codage

Le prétraitement empirique 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 l'exemple ci-dessus, la seconde réponse, la plus simple, donnerait les items « lecture », voyage », « loisirs », sous réserve que ces items apparaissent avec une certaine fréquence dans l'échantillon de réponses. En revanche la première réponse est plus délicate à post-coder.

Les outils statistique de base

Les outils de base sont la sélection de formes caractéristiques, la sélection de réponses modales, l'analyse des correspondances et la classification des tableaux lexicaux.

Formes ou segments caractéristiques (ou spécificités)

Les formes caractéristiques sont les formes "anormalement" fréquentes dans les réponses d'un groupe d'individus (technique proposé par P. Lafon en 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. Il s'agit de test classique de comparaisons de fréquences, maisla répétition de ce test conduit à prendre des seuils de signification très sévères (phénomène de *comparaisons multiples* bien connu des statisticiens).

Dans l'exemple évoqué plus haut, la fréquence moyenne du mot *travail* dans le corpus était de 3.4 %; pour le groupe des femmes de plus de 55 ans, la fréquence n'est que de 1.2 %. Cette différence est en fait hautement significative (on peut exprimer le test de comparaison de fréquences en termes d'écart-types : dans l'hypothèse d'homogénéité des fréquences, la valeur1.2% est à 4.5 écart-types de la valeur moyenne 3.4). Comme il s'agit d'une fréquence anormalement faible, on parlera de mots anti-caractéristiques. [L'individu statistique est ici l'occurrence de mots. Les femmes de plus de 55 ans ont émis 1349 mots dans leurs réponses. La variance de la fréquence d'un mot dont la frequence "théorique" est de 0.034 est donnée par la formule classique $0.034(1 - 0.034) /1349$. On voit dans ces conditions qe la fréquence observée de 0.012 est à 4.5 écart-types de 0.034].

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, la terminologie variant selon les domaines d'application) sont des réponses originales du corpus de base, ayant la propriété de caractériser au mieux le groupe. On peut, pour chaque regroupement, calculer la distance 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 (L. Lebart et A. Salem, *Statistique Textuelle*, Dunod, 1994). Toujours dans le cas de notre exemple, "Etre heureux, avoir un bon travail, réussite professionnelle et familiale" est ainsi une réponse caractéristique des jeunes hommes; "la santé, la famille" est une réponse caractérisant les plus âgés. On utilise en pratique plusierus réponses caractéristiques par groupe.

Analyse des correspondances et classification

Le volume des données demande que l'on fasse appel à de puissants outils de description. Les méthodes d'analyses des correspondances et de classification peuvent décrire les tables de contingence croisant les réponses et les formes graphiques, 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 dans le cas des méthodes de classification, ou de *cartes auto-associatives de Kohonen*, méthode "neuronale" de visualisation) 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.

Conclusions et ouvertures

Pour des réponses simples et stéréotypées, nous l'avons vu, les procédures de post-codage

peuvent fonctionner. 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 difficile 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.

Sans qu'il soit nécessaire de procéder à un post codage, on peut, actuellement, à partir d'une ensemble de textes, et d'un seuil de fréquence pour les formes graphiques, obtenir 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*,(cf. A. Salem, *Pratique des segments répétés,* Klincksieck, 1987), 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. 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.

Parallèlement aux travaux relevant de *l'Industrie de la Langue*, que nous avons évoqués plus haut, et qui relèvent d'une *ingéniérie statistique* complexe, il existe donc des applications textuelles de la statistique qui restent à portée de main. Elles nécessitent certes des logiciels spécifiques, mais la nature familière et vivante du matériau de base compense en quelque sorte la relative complexité des traitements et les difficultés d'interprétation.

Proche des bases de données, de l'intelligence artificielle et des réseaux de neurones, de la théorie de l'apprentissage, des techniques récentes d'extraction et de gestion des connaissances, le domaine textuel illustre bien la polyvalence et la puissance de la méthodologie statistique. Même quand les méthodes prennent parfois les noms plus exotiques de *fouille de texte* ou de *text mining*, le statisticien est toujours sollicité quand il s'agit de connaître la portée réelle des faits observés et des traits structuraux obtenus, de savoir ce que l'on a le droit de dire ou le devoir de ne pas dire, c'est-à-dire finalement de donner un statut scientifique aux résultats.