JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

LUDOVIC LEBART

Sur les analyses statistiques de textes

Journal de la société statistique de Paris, tome 135, n° 1 (1994), p. 17-36

<http://www.numdam.org/item?id=JSFS_1994__135_1_17_0>

II

COMMUNICATION

SUR LES ANALYSES STATISTIQUES DE TEXTES

Par Ludovic LEBART
CNRS, École Nationale Supérieure des Télécommunications, Paris

Lors des analyses statistiques de textes, deux grandes séries de préoccupations se font jour :

- D'une part, les applications à des textes littéraires (attributions d'auteurs, datation, par exemple), qui cherchent à s'affranchir du contenu pour saisir des caractéristiques de **forme** (souvent : de style) à partir des distributions statistiques de vocabulaire, d'indices ou de ratios, ou encore à partir de corpus partiels de mots-outils. Il s'agit de saisir les « invariants » d'un auteur ou d'une époque, dissimulés ou peu apparents, à des fins historiques, littéraires, dans le cadre d'études que l'on désigne sous le nom de *stylométrie* (cf. par exemple Holmes, 1985, pour une revue de ces travaux).
- D'autre part, les applications réalisées en recherche documentaire (*Information Retrieval* : cf. Salton, 1988), en codification automatique, dans le traitement des réponses à des questions ouvertes, qui s'intéressent principalement au **contenu,** au sens, à la substance des textes. Peu importe la façon dont une requête est rédigée, pourvu que l'on puisse atteindre dans la base de données les documents qui satisfont l'attente du requérant. Cependant, lors du traitement statistique de réponses à des questions ouvertes, ou lors des analyses d'entretiens, le socio-linguiste peut être *aussi* intéressé par la forme, par les connotations véhiculées par exemple par certains synonymes, certaines tournures. Les caractéristiques de formes peuvent en effet nuancer et infléchir le fond, ou présenter un intérêt sociologique (cf. par exemple : Achard, 1993).

Les méthodes d'analyses de réponses libres dans les enquêtes relèvent de cette seconde famille de méthodes.

1. La stylométrie

Le statisticien C. R. Rao, dans son ouvrage de réflexion générale *Statistics and Truth* (1989) se réfère, comme modèle d'étude stylométrique, aux travaux de Thisted

et Efron (1987) à propos de l'attribution à Shakespeare d'un poème découvert en 1985. Il cite également l'imposant travail de Mosteller et Wallace (1964) sur l'attribution des *Federalist Papers*. Parmi ces 77 textes politiques, publiés anonymement à New York à la fin du dix-huitième siècle, 12 textes n'ont pu être attribués à un auteur parmi deux possibles. Le traitement statistique permet de désigner l'auteur le plus probable de ces 12 textes litigieux. On pourra aussi consulter sur un thème voisin un travail plus récent de Holmes (1992) sur l'homogénéité des *Mormon Scriptures*. La seconde version du travail de Mosteller et Wallace (1984) contient également un panorama général des tentatives d'attribution d'auteur.

La plupart de ces méthodes utilisent des indices synthétiques construits à partir des longueurs des mots, des longueurs des phrases, des fréquences de mots-outils, de la richesse du vocabulaire, des distributions de fréquence des mots.

L'utilisation systématique de l'analyse des correspondances et des techniques de classification automatique (cf. Benzécri *et al.*, 1981) a considérablement enrichi ces approches.

1.1. Les unités et indices de la stylométrie

Les unités de bases seront dérivées de la forme graphique (cf. plus bas section 4) mais il convient de noter que de nombreux travaux stylométriques ont pu utiliser, pour comparer des textes, des auteurs ou des genres, des comptages qui fractionnent encore cette unité.

Smith (1983) a montré les limites de la distribution du nombre de lettres par mot, alors que la distribution du nombre de syllabes par mots, proposée de façon systématique par Fuchs (1952), a été jugée utilisable, au moins pour les attributions d'auteurs anglophones, par Brainerd (1974).

Même les distributions de fréquences globales des graphèmes (lettres et ponctuation) peuvent avoir un pouvoir discriminant important entre textes (Brunet, 1981 ; Abi Farah, 1988 [textes en langue arabe] ; Salem, 1993). Il s'agit surtout d'un exercice méthodologique, car on lit à travers les graphèmes la spécificité des vocabulaires. Brunet a cependant montré que l'effet discriminant des graphèmes subsistait même après élimination des formes les plus fréquentes, ce qui rend encore plus délicate l'interprétation de tels effets.

1.2. « Mots-outils », parties du discours

La notion de *mot-outil* (appellés encore mots-vides en documentation, mots grammaticaux) ne se prête à aucune formalisation satisfaisante. De nombreux auteurs utilisent cependant cette notion en s'appuyant sur l'intuition commune que l'on peut dresser, dans chaque langue, une liste de formes qui ont en commun la propriété d'être moins marquées au plan sémantique.

Le courant qui s'occupe des problèmes d'attribution d'auteur a longtemps privilégié l'étude de ce type d'unité, posant que leur emploi, moins maîtrisé lors de la rédaction du texte, pouvait constituer une marque d'auteur privilégiée.

C'est le sens des travaux de pionnier de Ellegard (1962), qui compare les proportions de « mots-outils » dans le corpus des *Junius Letters* (pamphlets publiés à la fin du dix-huitième siècle comportant environ 150 000 occurrences), avec celles calculées dans un autre corpus de la même époque. Cette démarche constitue également une phase importante des travaux de Mosteller et Wallace (1964, *op. cit.*) ainsi que de ceux de Holmes (1992).

Benzécri a réalisé des typologies de textes en grec ancien (1991a), latins (1991b), et espagnols (1992b) à partir d'ensembles de mots-outils, mettant en évidence à la fois les problèmes que pose la sélection de ces unités statistiques, et le pouvoir discriminant des profils de mots-outils lorsque ceux-ci interviennent comme éléments actifs d'une analyse des correspondances.

Des progrès importants ont été réalisés dans le domaine de l'analyse syntaxique automatisée des textes, comme en témoigne, par exemple, l'amélioration constante des correcteurs orthographiques que l'on trouve désormais sur la plupart des machines de traitement de texte. Des analyseurs syntaxiques permettent de calculer la proportion de noms, de verbes, d'adjectifs, etc. Ces proportions ont également été utilisées en stylométrie. Ainsi, Somers (1966) affirme que la proportion de substantifs dénote instruction et aisance dans l'expression, Brainerd (1974) étudie le pouvoir discriminant de la proportion d'articles. Précisons qu'une telle catégorisation est évidemment nécessaire pour identifier d'éventuels mots-outils. Notons que si l'isolement de mots-outils demande une désambiguïsation du texte – cas de la forme *pas*, par exemple – il existe aussi des locutions contenant des mots pleins qui sont des substituts de mots-outils, et qu'une lemmatisation préliminaire pourrait masquer (cf. par exemple Benzécri, 1992a).

1.3. La richesse du vocabulaire

Pour conclure sur ce bref survol des unités statistiques de la stylométrie, il faut mentionner les indices de richesse de vocabulaire : si V désigne le nombre de formes différentes, et T le nombre total d'occurrences, pour chaque texte, les premiers indices proposés furent les quotients du nombre de mots distincts V par le nombre total T d'occurrences :

$$R = \frac{V}{T} \qquad \textit{(type-token ratio)}$$

ou encore :

$$R' = \frac{\text{Log } V}{\text{Log } T}$$

Ce quotient a été proposé en particulier par McKinnon et Webster (1971) pour discriminer des textes écrits sous différents pseudonymes par le philosophe danois Sören Kierkegaard. Appliqué à un ensemble de seize échantillons qui varient environ du simple au double – de 6 548 occurrences à 15 432 – un tel ratio permet à ces auteurs de conclure à la possibilité de caractériser les différentes catégories de textes à partir de ce seul critère.

Dans les corpus de textes, les variations de longueurs entre les parties peuvent de plus être considérables, et les deux quotients précédents ont manifestement l'inconvénient de trop dépendre de la longueur des textes (pour une langue donnée, V est borné supérieurement, alors que T n'a pas de limite a priori).

L'indice D de Simpson (1949) est le quotient :

$$D = \frac{\sum_r r(r-1)\, V_r}{T(T-1)}$$

où V_r est le nombre de formes distinctes apparaissant exactement r fois dans le texte. C'est donc le quotient du nombre de paires d'occurrences d'une même forme par le nombre total de paires d'occurrences. Le numérateur n'est plus borné, et D dépend beaucoup moins de T que R ou R'.

D est simplement la probabilité que deux occurrences prises au hasard dans le corpus correspondent à une même forme.

Cet indice bien connu des statisticiens est très proche dans son principe de la caractéristique K introduite, très antérieurement, par Yule (1944) dans le domaine des études stylistiques et qui peut être définie comme :

$$K = 10^4 \frac{D(T-1)}{T}$$

2. Discrimination globale : recherche documentaire, codification

De nos jours, la documentation automatique s'est pratiquement érigée en discipline autonome, avec ses revues, congrès et logiciels, sa terminologie et ses concepts (cf. Salton and McGill, 1983 ; Salton, 1988).

Ce sont les dimensions et le contexte des problèmes qui ont induit cette spécificité. En général, on a affaire à de très grands tableaux clairsemés issus de comptages de vocabulaire et de mots-clés spécialisés, selon les domaines, au sein d'ensembles qui comportent plusieurs centaines de milliers de documents souvent courts et parfois stéréotypés. La finalité de cette opération a souvent un caractère très pragmatique. Il s'agit de faire fonctionner un outil, avec des taux d'échecs, des systèmes de coûts et de contraintes préalablement définis.

Dans ce cadre particulier, il est possible de faire appel à plusieurs sources extérieures d'information pour résoudre les problèmes de classement : des analyseurs syntaxiques, premiers pas vers une *compréhension* de la requête, des dictionnaires ou des réseaux sémantiques pour lemmatiser et désambiguïser les requêtes (cf. par exemple, en matière de classification de documents, les travaux de Blosseville *et al.*, 1992 ; Hebrail *et al.*, 1990), éventuellement à des corpus artificiels faisant appel à des experts (cf. les travaux anciens de Palermo et Jenkins,1964 ; cf. également Bouroche et Curvalle, 1974).

Mais beaucoup de techniques utilisées, et parmi les plus efficaces, ont recours à des outils multidimensionnels très proches de ceux préconisés par Benzécri (1977, 1981), ou dans le cadre des grands tableaux clairsemés par Lebart (1982a). Deerwester *et al.* (1990) proposent ainsi sous le nom de *Latent Semantic Analysis* une méthode très semblable à la discrimination d'après les premiers facteurs d'une analyse des correspondances (ces auteurs utilisent en fait une décomposition aux valeurs singulières, technique qui est à la base de l'analyse des correspondances et de l'analyse en composantes principales).

D'autres auteurs insistent sur la complémentarité entre modèles et méthodes descriptives dans la recherche documentaire (cf. Fuhr *et al.*, 1991), et sur l'intérêt de visualisations les plus synthétiques possible (Fowler *et al.*, 1991).

De plus en plus, les « chercheurs documentalistes » reconnaissent l'importance des phases de description et d'exploration.

3. Questions ouvertes dans les enquêtes

Il peut être 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 évidemment complexe. Les outils de calcul et les méthodes statistiques descriptives multidimensionnelles peuvent apporter une certaine aide à l'analyse de ces *réponses libres*.

On rappellera auparavant quelques-uns des problèmes posés par la rédaction des libellés des questions dans les questionnaires d'enquêtes.

On sait que le libellé d'une question joue un rôle fondamental : il est très difficile de trouver deux libellés distincts, pour deux questions fermées dont les contenus sont similaires, donnant les mêmes résultats en termes de pourcentages.

A ces précautions sur la rédaction des libellés s'ajoutent d'autres considérations :

- l'ordre des questions, qui induit une sensibilisation particulière du répondant,
- la longueur des libellés qui fait jouer, selon les cas, la mémoire auditive ou les capacités de lecture de la personne interrogée, et donc induit des biais en fonction de certaines caractéristiques de base comme l'âge, le niveau d'instruction (cf. la contribution de J.-P. Grémy dans ASU, 1992).

Le problème de la dépendance des résultats vis-à-vis des libellés se pose a fortiori dans le cas de deux questions dont l'une est ouverte et l'autre fermée. Un exemple classique concerne les réponses à la question « Quel est le problème le plus important auquel doivent faire face les USA ? » (Schuman *et al.*, 1981). L'item « violences » obtient 16 % lorsque la question est ouverte, et 32 % lorsqu'il fait partie des items de la question fermée correspondante. Cet item de réponse étant considéré comme « un problème local » plutôt que « national » n'est pas toujours considéré comme une réponse permise lorsque la question est ouverte. En somme, les libellés complets de deux questions, l'une ouverte et l'autre fermée, ne peuvent être identiques, ce qui rend extrêmement difficiles les comparaisons entre les deux types de questionnement.

3.1. Quand utiliser des questions ouvertes ?

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

Pour diminuer le temps 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 spontanées. 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 éventuels différences sémantiques des libellés selon la langue utilisée.

Pour recueillir une information qui doit être spontanée

Les questionnaires des enquêtes de marketing abondent en questions de ce type. Citons par exemple : « Qu'avez-vous retenu de ce spot publicitaire ? » (voir l'exemple qui suit), « 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 principalement 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 pré-enquête pilote assez lourde.

3.2. Traitement pragmatique des questions ouvertes

Le prétraitement appelé « post-codage » permet de fermer a posteriori les questions ouvertes. Cette technique 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 n'a que peu d'inconvénients. Mentionnons cependant parmi les défauts de ce type de traitement :

- A la médiation de l'enquêteur s'ajoute celle du chiffreur, qui doit prendre de nombreuses décisions difficiles et parfois contestables par le spécialiste.
- 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.
- Les réponses composites, complexes, d'une grande diversité, sont littéralement laminées par le post-codage 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 affectées à des items « résiduels » qui sont donc très hétérogènes et perdent de ce fait toute valeur opératoire.

Ces réponses relativement peu fréquentes peuvent cependant être émises par une catégorie d'individus très particulière, et présenter un grand intérêt au niveau de l'interprétation des résultats, ce qu'il n'est pas possible de savoir lors d'un traitement « a priori » de l'information...

4. Les unités statistiques découpées dans les textes

4.1. Les formes graphiques

L'unité de base sera la *forme graphique* définie comme une suite de caractères non délimiteurs (en général des lettres) entourée par des caractères délimiteurs (blanc, points, virgules...). Un même mot pourra en général donner lieu à plusieurs formes graphiques, selon son cas ou son genre dans le texte. Une même forme graphique peut renvoyer à plusieurs mots (en français, *avions* renvoie à un nom, mais aussi au verbe *avoir*). Cela n'est pas toujours un inconvénient grave, car les formes graphiques ne seront pas traitées isolément.

Les traitements statistiques concerneront en effet les *profils de fréquences de formes graphiques*, c'est-à-dire les vecteurs dont les composantes sont les fréquences de chacune des formes utilisées par un individu ou un groupe d'individus. Ces profils contiennent une information extrêmement riche. Plus précisément encore, les techniques mettront en évidence les *différences* entre *profils de formes graphiques*.

Si l'interprétation dans l'absolu d'un profil peut être délicate, l'interprétation des différences entre profils est plus aisée : sans spéculer sur la signification des profils, on peut très bien observer que, par exemple, les cadres et les employés ont des profils proches, éloignés de celui des ouvriers.

4.2. Les segments répétés

La notion de forme graphique peut être généralisée en procédant à des comptages portant sur des unités plus larges, composées de plusieurs formes : les segments répétés. On observe en effet dans les réponses les apparitions récurrentes d'unités comme *je ne sais pas, sécurité d'emploi, justice sociale,* dotées parfois d'un sens qui leur est propre et que l'on ne peut pas toujours déduire à partir du sens des formes qui entrent dans leur composition (Salem, 1987). Il est alors possible de reprendre les traitements avec les segments pour compléter les formes graphiques. Les résultats sont considérablement enrichis par l'introduction du contexte des formes, qui lève la plupart des ambiguïtés de sens.

Pour sélectionner formes et segments, des seuils de fréquence vont intervenir. Ils permettront d'effectuer différents filtres sur l'information de base.

4.3. Les unités lemmatisées

Un autre type de traitement préliminaire du texte consiste à procéder à une *lemmatisation.* Cette opération, très difficile à réaliser de façon entièrement automatique, consiste à remplacer les formes par l'entrée du dictionnaire correspondant (infinitif pour les verbes, masculin singulier pour les adjectifs, formes non élidées à la place des formes élidées, etc.), et parfois à supprimer certains mots-outils (articles, conjonctions, etc., cf. par exemple Reinert, 1986).

En documentation automatique, cela permet de travailler avec un nombre restreint de mots-clés dont les occurrences sont fréquentes.

En traitement de questions ouvertes, cette opération n'est pas toujours souhaitable *a priori* car elle détruit les locutions et modifie assez profondément la forme des réponses, qui peuvent intéresser le socio-linguiste.

En revanche, elle peut intervenir comme complément, car elle fournit un point de vue différent sur les textes. Dans le cas d'entretiens non directifs peu nombreux, la lemmatisation permet de travailler avec des seuils de fréquences plus élevés que ceux nécessités par l'analyse des formes graphiques.

Tableau 1. Formes apparaissant au moins 9 fois (100 réponses)
(Question ouverte « What is the main idea in this commercial? »)

Numéro	Forme	Fréquence
1	I	14
2	a	59
3	about	15
4	all	21
5	and	42
6	are	25
7	been	12
8	carbohydrate	14
9	carbohydrates	33
10	cereal	34
11	complex	25
12	crunchy	9
13	eaten	10
14	eating	19
15	energy	33
16	for	57
17	give	9
18	gives	11
19	good	52
20	grape	25
21	has	30
22	have	27
23	healthy	23
24	how	9
25	in	27
26	is	37
27	it	133
28	it's	28
29	long	14
30	morning	9
31	nothing	25
32	nutritional	9
33	nutritious	12
34	nuts	25
35	of	25
36	people	28
37	showed	11
38	taste	11
39	that	80
40	that's	13
41	he	82
42	they	50
43	to	32
44	was	19
45	with	11
46	years	11
47	you	81

5. La numérisation du texte

Cette phase de traitement préliminaire consiste à affecter à chaque nouvelle forme graphique un numéro d'ordre qui sera associé à toutes les occurrences de cette même forme. Ces numéros seront stockés dans un dictionnaire de formes, ou vocabulaire, propre à chaque exploitation. Ce dernier permettra, à l'issue des calculs ou lors des impressions, de reconstituer le graphisme des formes mises en évidence par les calculs statistiques.

Les exemples qui suivent sont empruntés au domaine du marketing : il s'agit d'étudier les réactions d'un public nord-américain après avoir visionné un « spot » publicitaire, pour un produit à base de céréale utilisé pour le petit déjeuner.

La question ouverte est : *Quelle idée principale avez-vous retenu de ce spot publicitaire ?*

Le tableau 1 représente les 47 formes apparaissant au moins 9 fois dans un échantillon de 100 réponses à cette question.

On observe comme prévu des formes se rapportant à un même mot (give, gives ; have, has ; eaten, eating), des mots-outils (of, to, and). Comme cela a été dit plus haut, la lemmatisation et l'apurement ne s'imposent pas dans une approche différentielle portant sur des échantillons importants.

Si les mots-outils sont répartis de façon aléatoire dans les diverses catégories d'individus, ils ne sont pas gênants. S'ils ne le sont pas, ils sont au contraire intéressants. De façon analogue, si deux formes graphiques se rapportant à un même mot ont des comportements identiques, elles peuvent aussi bien être remplacées par ce mot. Si elles ont des comportements différents, c'est qu'elles renvoient à des contextes d'utilisation du mot différents, ce qui mérite d'être relevé.

Le tableau 2 décrit ainsi, toujours pour les 100 réponses qui nous servent d'exemple illustratif, les différents segments observables, classés selon l'ordre alphabétique de la première forme graphique qui les compose, et sélectionnés en fonction de seuils de fréquences.

Les segments de longueur 2 (très nombreux, et pauvres du point de vue de leur apport sémantique) doivent apparaître au moins 10 fois, alors que ceux de longueur supérieure ou égale à 3 doivent apparaître au moins 4 fois pour figurer dans l'inventaire.

On voit qu'il s'agit d'éléments d'information auxiliaires, largement interdépendants, mais permettant d'identifier les contextes des formes les plus fréquentes. Une sélection s'impose : il est relativement aisé de choisir dans cette liste (établie à partir de seuils sévères, pour limiter le volume des éditions) les segments porteurs d'une information sémantique spécifique.

Tableau 2. Inventaire partiel de segments répétés – seuils minimum de fréquence de répétition = 4, segments de longueur 2 = 10, segments de longueur 3 = 4

Segment	Fréquence	Longueur	Libellé du segment
			a
1	8	3	a long time
			are
2	6	4	are good for you
			carbohydrates
3	5	3	carbohydrates in it
			complex
4	15	2	complex carbohydrates
			for
5	37	2	for you
			give
6	7	3	give you energy
			gives
7	11	2	gives you
8	9	3	gives you energy
			good
9	24	2	good for
10	22	3	good for you
			grape
11	25	2	grape nuts
			have
12	6	3	have been eating
			healthy
13	6	3	healthy for you
			is
14	9	4	is good for you
			it
15	26	2	it has
16	19	2	it is
17	14	2	it was
18	8	3	it gives you
19	8	3	it has a
20	6	3	it has complex
21	5	3	it is good
22	6	4	it gives you energy
			people
23	14	2	people have
24	8	3	people have eaten
25	5	4	people have been eating
			that
26	27	2	that it
27	6	3	that grape nuts
28	10	3	that it has
29	6	3	that it was
30	6	3	that people have
			the
31	14	2	the cereal
			they
32	13	2	they are
			you

6. Les tableaux lexicaux

Les réponses libres peuvent être numérisées de façon complètement transparente pour l'utilisateur. Le résultat de cette numérisation peut prendre deux formes différentes, matérialisées par deux matrices R et T. La matrice R a k lignes, k désignant le nombre de réponses, et un nombre de colonnes égal à la longueur de la plus longue réponse (nombre d'occurrences de formes dans cette réponse).

Pour une réponse ou un individu « i », la ligne « i » de R (tableau de pointeurs) contient les adresses (relatives à un dictionnaire ou *vocabulaire*) des formes graphiques qui composent la réponse, en respectant l'ordre et les éventuelles répétitions de ces formes. R permet donc de restituer intégralement les réponses originales.

R n'est pas rectangulaire, car chacune de ses lignes a une longueur variable. Les nombres entiers qui composent R ne peuvent dépasser v, longueur du vocabulaire (nombre de formes graphiques distinctes). La matrice T a le même nombre k de lignes que R, mais possède autant de colonnes qu'il y a de formes graphiques utilisées par l'ensemble des individus, c'est-à-dire v (v = vocabulaire) colonnes. A l'intersection de la ligne i et de la colonne j de T figure le nombre de fois où la forme j a été utilisée par l'individu i dans sa réponse. Il s'agit donc d'une table de contingence « Individus-Formes ».T peut être aisément construite à partir de R, mais la réciproque n'est pas vraie : l'information relative à l'ordre des formes dans chaque réponse est perdue dans T.

En fait, R est beaucoup plus compacte que T : ainsi, une réponse contenant 20 occurrences (pour un lexique de 1 000 formes) correspond à une ligne de longueur 20 de R et à une ligne de longueur 1 000 de T (cette dernière ligne comprenant au moins 980 zéros...). Les calculs statistiques et algorithmiques qui mettront en jeu T sont en réalité programmés à l'aide de R, moins encombrante en mémoire.

Dans la plupart des applications, les réponses isolées sont trop pauvres pour faire l'objet d'un traitement statistique direct : il est nécessaire de travailler sur des regroupements de réponses.

On désignera par Z le tableau disjonctif complet à k lignes et p colonnes décrivant les réponses de k individus à une question fermée comportant p modalités de réponses possibles.

$C = T' Z$ est un tableau à v lignes et p colonnes dont le terme général c_{ij} n'est autre que le nombre de fois où la forme « i » a été utilisée dans une réponse libre par l'ensemble des individus ayant choisi la réponse « j » à une question fermée.

Il est donc aisé, pour toute question fermée dont les réponses sont codées dans un tableau Z_q, de calculer le tableau lexical agrégé C_q par la formule :

$$C_q = T' Z_q$$

et donc de comparer les profils lexicaux de différentes catégories de population.

Ces comparaisons de profils lexicaux n'ont de sens, d'un point de vue statistique, que si les formes apparaissent avec une certaine fréquence : les hapax (formes n'ap-

paraissant qu'une fois), ou même les formes rares seront écartés de la phase de comparaisons de fréquences. Ceci a pour effet de réduire la taille du vocabulaire v. Pour une question ouverte posée à 1 000 personnes, une sélection des formes apparaissant au moins 8 fois peut, dans bien des cas, diviser par 10 la valeur de v (de 1 500, pour fixer les idées, à 150).

Trois outils vont permettre d'aider la lecture des tableaux lexicaux agrégés : l'analyse des correspondances, les listes de formes caractéristiques, les listes de réponses modales.

6.1. Analyse des correspondances des tableaux lexicaux

Les analyses des correspondances peuvent décrire les tableaux C_q qui sont des tables de contingence (dont les « individus » sont des occurrences de formes, et non plus des individus interrogés...). Elles permettent de visualiser les associations entre mots (formes) et groupes ou modalités. Ainsi, une visualisation des proximités entre mots et catégories socio-professionnelles 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 s'abîmeront dans la partie centrale du graphique, sans en encombrer la lecture. De même, la présence de plusieurs flexions d'un même verbe constitue un outil de validation.

6.2. Les listes des formes ou segments caractéristiques (ou spécificités)

Il est tentant de compléter les représentations spatiales fournies par l'analyse des correspondances par quelques paramètres d'inspiration plus probabiliste : les spécificités ou formes caractéristiques. Ce seront les formes « anormalement » fréquentes dans les réponses d'un groupe d'individus.

Toujours pour la question posées à l'issue de l'enquête commerciale précitée, le tableau 3 donne les formes les plus caractéristiques de chacune des 4 classes de réponses à une question fermée stratégique : les intentions d'achat après avoir visionné le spot publicitaire.

Tableau 3. Segments caractéristiques par catégories

Libellé du SEGMENT		Pourcentage		FREQ.		Valeur-test	PROBA
		interne	*global*	*interne*	*globale*		
TEXTE 1 : Ne vont probablement pas acheter le produit							
1	24 - people have eaten	9.38	1.97	3.	8.	2.077	.019
2	15 - it has	15.63	6.39	5.	26.	1.712	.043
3	3 - carbohydrates in it	6.25	1.23	2.	5.	1.629	.052
4	20 - it has complex	6.25	1.47	2.	6.	1.449	.074
TEXTE 2 : Hésitent							
1	8 - gives you energy	4.65	2.21	4.	9.	1.288	.099
2	20 - it has complex	3.49	1.47	3.	6.	1.217	.112
3	22 - it gives you energy	3.49	1.47	3.	6.	1.217	.112
4	7 - gives you	4.65	2.70	4.	11.	.899	.184
TEXTE 3 : Vont probablement acheter le produit							
1	9 - good for	8.50	5.90	13.	24.	1.497	.067
2	10 - good for you	7.84	5.41	12.	22.	1.449	.074
3	5 - for you	11.76	9.09	18.	37.	1.273	.102
4	27 - that grape nuts	2.61	1.47	4.	6.	1.054	.146
TEXTE 4 : Vont sûrement acheter le produit							
1	16 - it is	8.82	4.67	12.	19.	2.490	.006
2	32 - they are	5.88	3.19	8.	13.	1.840	.033
3	21 - it is good	2.94	1.23	4.	5.	1.699	.045
4	26 - that it	9.56	6.63	13.	27.	1.452	.073

A ces formes caractéristiques sont attachées des « valeurs-tests » (avant-dernière colonne du tableau 3) qui mesurent l'écart existant entre la fréquence relative d'une forme dans une classe (pourcentage interne, première colonne numérique du tableau 3) avec sa fréquence relative globale (seconde colonne) calculée sur l'ensemble des réponses ou individus.

Cet écart est normé de façon à pouvoir être considéré comme une réalisation de variable normale centrée réduite, dans l'hypothèse de répartition aléatoire de la forme étudiée dans les classes. Dans cette hypothèse, la valeur-test *VT* a 95 chances sur 100 d'être comprise entre – 1.96 et + 1.96. (ainsi pour *VT* = 1.96, la dernière colonne PROBA vaut 0.025 (*VT* et PROBA donnent des informations équivalentes, mais *VT* est plus maniable pour des probabilités très faibles, cas fréquent en analyse de textes). Ce calcul reposant sur une approximation normale de la loi hypergéométrique n'est utilisé que lorsque les effectifs concernés ne sont pas trop faibles.

6.3. Les sélections des réponses modales

Pour une classe donnée, et donc pour le regroupement de réponses correspondant, les réponses modales (ou encore phrases caractéristiques, ou documents types, selon les domaines d'application) sont des réponses originales du corpus de base, ayant la propriété de caractériser au mieux la classe.

Critère de sélection 1 : mots caractéristiques

Un premier mode de calcul de réponses modales consiste à associer à chaque réponse la valeur-test moyenne des formes caractéristiques qu'elle contient : si cette moyenne est grande, cela signifie que la réponse ne contient que des formes très caractéristiques du groupement. Les réponses de plus grandes moyennes seront donc les plus caractéristiques de la classe ou du groupement de réponse concerné.

Quand un mot très caractéristique apparaît seul dans une réponse, cette réponse est évidemment bien classée. La présence d'autres mots peut bien entendu faire baisser la moyenne des valeurs-tests, d'où cette tendance à sélectionner des réponses courtes.

Critère de sélection 2 : Distances du Chi-2 entre profils

Le principe de ces sélections est schématiquement le suivant : une réponse est une ligne de T, donc un vecteur à v composantes. Si cette réponse est formée de 25 formes différentes, seulement 25 de ces composantes seront différentes de zéro.

Un groupement de réponses (les réponses des ouvriers, par exemple) est un ensemble de vecteurs-lignes, et le profil lexical moyen de ce groupement est obtenu en calculant la moyenne des vecteurs-lignes de cet ensemble.

Si ce regroupement se fait selon les modalités d'une question fermée dont les réponses sont codées dans un tableau Z, on a vu que le tableau lexical agrégé C se calcule par la formule :

$$C = T' Z$$

Il est donc possible de calculer des distances entre des réponses et les regroupements de ces réponses. Réponses (lignes de T) et regroupements de réponses (colonnes de C, ou lignes de C', transposée de C) sont tous représentés par des vecteurs d'un même espace.

Ces distances expriment l'écart entre le profil d'une réponse et le profil moyen de la classe à laquelle cette réponse appartient. La distance choisie entre ces profils de fréquences sera la distance du Chi-2, en raison de ses propriétés distributionnelles.

Tableau 4. Exemples de réponses modales pour 2 catégories

TEXTE 1 : Ne vont probablement pas acheter le produit		
– – –	1	to tell you about how long people have eaten them. the complex carbohydrate that are in this cereal. the people who eat this cereal and the product. that's all.
– –	2	it's supposed to be healthy it has good carbohydrates in it.
– – – –	3	that it has complex carbohydrate, to keep you going all – morning, that people have eaten it a long time, the years people have eaten this cereal and some didn't know about the complex carbohydrate.
TEXTE 3 : Vont probablement acheter le produit		
– –	1	it's nutritious for you. nothing else.
– –	2	that, is good for you that's all it said to me

La distance entre un point-ligne i de T et un point-colonne m de C est alors donnée par la formule :

$$d^2(i, m) = \sum_j \left(\frac{t_{..}}{t_{.j}}\right)\left(\frac{t_{ij}}{t_{i.}} - \frac{c_{jm}}{c_{.m}}\right)^2$$

avec les notations usuelles :

$t_{..}$ désigne la somme globale des éléments de T, c'est-à-dire le nombre total d'occurrences ;

$t_{.j}$ désigne la somme des éléments de la colonne j de T (nombre d'occurrences de la forme j) ;

$t_{i.}$ la somme des éléments de la ligne i de T (longueur de la réponse i) ;

$c_{.m}$ la somme des éléments de la colonne m de C (nombre total d'occurrences de la classe ou du groupement m).

On peut, pour chaque regroupement, classer ces distances par ordre croissant, et donc sélectionner les réponses les plus représentatives au sens du profil lexical, qui correspondront aux plus petites distances.

Dans le cas du corpus qui nous sert d'exemple, le tableau 4 représente, pour deux catégories du tableau 3, quelques réponses (effectivement présentes dans le recueil

de base) les plus caractéristiques de chaque catégorie. Sans aucun travail de codification ni d'interprétation, on voit clairement que les deux catégories de personnes n'ont pas retenu la même chose du spot publicitaire, et que la mention des « carbohydrates » dans le spot n'a peut-être pas l'effet escompté.

7. Stratégie de traitement

On a vu qu'il était souvent nécessaire de regrouper les réponses pour pouvoir procéder à des analyses de type statistique. Les profils lexicaux d'agrégats de réponses ont plus de régularité et de signification que ceux des réponses isolées. Ce regroupement a priori peut être réalisé à partir des variables disponibles, retenues en fonction de certaines hypothèses. Mais ceci suppose une bonne connaissance préalable du phénomène étudié, situation qui n'est en général pas réalisée dans les études dites *exploratoires.*

7.1. Regroupement par noyaux factuels

La technique dite « des noyaux factuels » va permettre de donner des éléments de réponse à ce problème.

Etant donnée une liste de descripteurs ou de variables caractérisant les individus, le problème est de regrouper les individus en groupes les plus homogènes possible vis-à-vis de ces caractéristiques... sans en privilégier certaines a priori.

C'est précisément le type d'opération que permet de réaliser un algorithme de classification, appliqué aux lignes du tableau disjonctif Z décrivant les individus à partir d'une sélection de leurs caractéristiques.

La partition obtenue est une sorte de « partition moyenne » qui résume les principales combinaisons de situations observables dans l'échantillon, et qui permet donc de procéder à des regroupements de réponses les moins arbitraires possible.

7.2. Analyses directes sans regroupement

Si les réponses ne sont pas regroupées, mais paraissent suffisamment riches pour être traitées isolément, une analyse directe du tableau lexical T croisant formes graphiques et réponses peut être opérée.

Une telle analyse produit une typologie des réponses, en général assez grossière, et produit de façon duale une typologie de mots ou de formes graphiques.

Il est donc possible d'illustrer ces typologies par les caractéristiques des individus interrogés qui auront le statut de variables supplémentaires ou illustratives. Ce traitement direct des réponses pourra conduire à la réalisation d'un post-codage partiellement automatisé.

Notons que la proximité entre deux formes graphiques, c'est-à-dire entre deux colonnes du tableau T sera d'autant plus grande que les formes apparaîtront dans une même réponse (et non plus seulement dans un même texte), ce qui permettra de mieux représenter les voisinages syntagmatiques. L'analyse directe rendra mieux compte des contextes que les analyses de tableaux agrégés. Le traitement d'un tableau aussi grand et « clairsemé » impliquera en général la mise en oeuvre d'algorithmes de calcul particuliers, utilisant le tableau réduit R au lieu du tableau T, et évitant le calcul et le stockage d'une matrice à diagonaliser d'ordre v (cf. par exemple, Lebart, 1982a).

CONCLUSIONS

Cette approche est essentiellement différentielle, comparative. Distincte de l'analyse de contenu classique, elle est avant tout une confrontation de l'ouvert et du fermé. Elle ne vise en effet qu'à décrire les contrastes entre plusieurs textes, que ces textes soient des réponses originales ou des regroupements de réponses réalisés à partir des questions fermées de l'enquête.

Pour une question ouverte et pour une partition de la population, on obtient donc, de façon intégralement automatisable :

- Une visualisation des proximités entre formes et catégories, par analyse des correspondances du tableau lexical agrégé, éventuellement complétée par une visualisation similaire des proximités entre segments et catégories ;
- Les formes (et/ou segments) caractéristiques de chaque catégorie ;
- Les réponses modales de chaque catégorie.

Ces résultats, obtenus sans codification ni intervention manuelle, fournissent des compléments et donnent des éléments critiques nouveaux pour juger à la fois la cohérence et la pertinence du questionnement, la compréhension des réponses, ainsi que le niveau d'implication ou de participation des répondants. Ils peuvent donc participer à l'amélioration de la qualité de l'information, et fournissent des éléments originaux au dossier des analyses de contenu.

RÉFÉRENCES

ABI FARAH A. (1988) "Reconnaissance de l'auteur d'un texte d'après les caractères utilisés", *Cahiers de l'analyse de données*, n° 1, 95-96.

ACHARD P. (1993) *La sociologie du langage*, Que sais-je ? PUF, Paris.

ASU (1992) *La qualité de l'information dans les enquêtes*, Dunod, Paris.

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

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

BENZÉCRI J.-P. (1991a) "Typologies de textes grecs d'après les occurrences des formes des mots-outils", *Cahiers de l'analyse de données*, XVI, n° 1, 61-86.

BENZÉCRI J.-P. (1991b) "Typologies de textes latins d'après les occurrences des formes des mots-outils", *Cahiers de l'analyse de données*, XVI, n° 4, 439-465.

BENZÉCRI J.-P. (1992a) "Note de lecture : sur l'analyse des données dans une enquête internationale", *Cahiers de l'analyse de données*, vol XVII, n° 3, Dunod, Paris, 353-358.

BENZÉCRI J.-P. et F. (1992b) "Typologie de textes espagnols de la littérature du Siècle d'Or d'après les occurrences des formes des mots-outils", *Cahiers de l'analyse de données*, vol XVII, n° 4, Dunod, Paris, 425-464.

BLOSSEVILLE M.J., HÉBRAIL G., MONTEIL M.G., PÉNOT N. (1992) "Automatic document classification : natural language processing, statistical analysis and expert system techniques used together", *Proceeding of the ACM-SIGIR*, Copenhagen.

BOUROCHE J.-M., CURVALLE B. (1974) "La recherche documentaire par voisinage", R.A.I.R.O., V-1, 65-96.

BRAINERD B. (1974) *Weighing Evidence in Language and Literature. A Statistical Approach*, University of Toronto Press.

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.

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), 391-407.

ELLEGARD A. (1962) "A statistical method for determining authorship : the Junius Letters, 1769-1772", *Gothenburg Studies in English*, n° 13, University of Gothenburg.

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.

FUHR N., PFEIFER U. (1991) "Combining model-oriented and description-oriented approaches for probabilistics indexing", *Proceedings of the 14th Int. ACM Conf. on Res. and Dev. in Information Retrieval*, Bookstein A. and al., Ed, ACM Press, New York, p 46-56.

GUIRAUD P. (1960) *Problèmes et méthodes de la statistique linguistique*, P.U.F., Paris.

HEBRAIL G., SUCHARD M. (1990) "Classifying documents : a discriminant analysis and an expert system work together", *COMPSTAT 90*, (Momirovic K. and Midner, eds), Physica Verlag, p 63-68.

HERDAN G. (1964) *Quantitative Linguistics*, Londres, Butterworths.

HOLMES D.I. (1985) "The analysis of literary style - A Review", *J.R. Statist. Soc.*, 148, Part 4, 328-341.

HOLMES D.I. (1992) "A Stylometric analysis of mormon scripture and related texts", *J.R. Statist. Soc.*, 155, Part 1, 91-120.

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

LEBART L. (1982a) "Exploratory analysis of large sparse matrices, with application to textual data", *COMPSTAT*, Physica Verlag, 67-76.

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

LEBART L., SALEM A. (1994) *Statistique textuelle*, Dunod, Paris.

MCKINNON A., WEBSTER R. (1971) "A method of « author » identification", *The Comput. in Liter. and Linguist. Res.*, R.A. Wisbey, ed., Cambridge Univ. Press.

MOSTELLER F., WALLACE D. (1964) *Inference and disputed Authorship : The Federalists.* Addison-Wesley, Reading, Mass.

MOSTELLER F., WALLACE D.L. (1984) *Applied Bayesian and Classical Inference, the Case of the Federalist Papers*, Springer Verlag, New York.

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

RAO C.R. (1989) *Statistics and Truth*, International Cooperative Publishing House, Fairland, USA.

REINERT M. (1986) "Un Logiciel d'analyse lexicale". *Cahiers de l'analyse de données*, 4, Dunod, 471-484.

SALEM A. (1987) *Pratique des segments répétés, Essai de satistique textuelle*, Klincksieck, Paris.

SALEM A. (1993) *Méthodes de la statistique textuelle*, Thèse d'État, Université Sorbonne Nouvelle (Paris 3).

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

SALTON G., MCGILL M.J. (1983) *Introduction to Modern Information Retrieval*, International Student Edition.

SCHUMAN H., PRESSER F. (1981) *Question and Answers in Attitude Surveys*, Academic Press, New York.

SIMPSON E.H. (1949) "Measurement of diversity", *Nature*, 163, 688.

SMITH M. W. A. (1983) "Recent experience and new developments of methods for the determination of authorship", *Ass. for Lit. and Linguist. Comput. Bull.*, 11, 73-82.

SOMERS H. H. (1966) "Statistical methods in literary analysis", *The Computer and Literary Style*, (J. Leed, Eds), Kent State University Press, Kent, Ohio.

THISTED R., EFRON B. (1987) "Did Shakespeare write a newly discovered poem ?" *Biometrika*, 74, 445-455.

YULE G.U. (1944) *The Statistical Study of Literary Vocabulary*, Cambridge University Press, Reprinted in 1968 by Archon Books, Hamden, Connecticut.