L. Lebart • A. Salem

STATISTIQUE TEXTUELLE

Préface de Christian Baudelot

DUNOD

L. Lebart
Directeur de Recherche au CNRS,
École nationale supérieure
des Télécommunications

A. Salem
Ingénieur à l'École
normale supérieure de
Fontenay-Saint-Cloud

STATISTIQUE TEXTUELLE

Préface de Christian Baudelot
Professeur à l'École normale supérieure

DUNOD

En couverture :

À El Greco
par Sylvia ELHARAR-LEMBERG
acrylique sur papier marouflé sur toile,
1992, 130 x 162 cm

ISBN 2 10 002239 3

STATISTIQUE TEXTUELLE

Association pour la Statistique et ses Utilisations (ASU)

La qualité de l'information dans les enquêtes

L'évolution rapide des outils de communication et de traitement de l'information a profondément modifié le contexte technique et institutionnel des enquêtes par sondage.

Les domaines d'application s'étendent et se diversifient, le nombre des disciplines concernées augmente, ce qui multiplie les concepts, les terminologies et rend ainsi les mises au point plus difficiles.

Cet ouvrage fait le point sur la qualité de l'information dans ses rapports avec la conception du plan de sondage et du questionnaire, le mode d'interrogation, la compréhension et l'accueil de l'enquête, les contrôles de cohérence et enfin, l'éthique professionnelle et la réglementation.

Il s'adresse non seulement aux professionnels (réalisateurs d'enquêtes, statisticiens) mais surtout aux utilisateurs chargés d'études, chercheurs, enseignants et étudiants en sciences économiques et politiques, en gestion et marketing, en sociologie, en sciences de l'information et de la communication.

D. Grangé - L. Lebart

Traitements statistiques des enquêtes

Le traitement statistique des enquêtes se situe à un carrefour interdisciplinaire : les données issues de la socio-économie, la démographie, l'épidémiologie, du marketing, des sciences politiques y convergent pour alimenter des procédures et des logiciels. Les résultats sont ensuite livrés à l'interprétation de spécialistes qui manient des concepts et parlent des langages différents ...

Et pourtant, les outils utilisés sont, eux, étonnamment similaires, bien que rarement présentés simultanément dans ce contexte spécifique : **l'enquête**.

C'est pourquoi les auteurs de ce recueil expliquent à la fois les mécanismes de ces procédures et logiciels, et le métier de ceux qui les mettent en oeuvre, avec ses règles et aussi ses tours de main, ses subtilités.

Cet ouvrage s'adresse à ceux qui, pour leurs recherches, leurs études, leurs enseignements, sont confrontés aux enquêtes par sondage. Il est donc destiné à un vaste public et peut être lu à plusieurs niveaux : il intéressera aussi bien l'utilisateur peu soucieux du détail des formules, et le technicien qui veut comprendre «comment ça marche»...

Préface

Et le Verbe s'est fait Nombre...

Il y a dans l'activité qui consiste à traiter les mots comme des nombres - opération de base de la statistique textuelle - un a priori qui ne manquera pas d'apparaître à certains comme outrageusement réducteur voire même sacrilège. Surtout si l'on en croit Victor Hugo : *Car le mot, c'est le Verbe, et le Verbe c'est Dieu...*

Il suffit de lire ce livre et surtout d'en appliquer les principes à ses propres enquêtes pour se convaincre du contraire. Avec ses graphes d'analyse factorielle, J.P. Benzécri a rendu les individus à la statistique : longtemps ignorés à force d'être confondus dans de vastes agrégats ou pulvérisés dans des formules inférentielles qui s'intéressent d'abord aux relations entre des grandeurs abstraites (revenu et consommation, salaire et diplôme...), les individus effectuent leur rentrée sur la scène statistique sous la forme de points dans un nuage. Les positions respectives qu'ils occupent au sein de ce nuage démontrent d'abord qu'ils diffèrent tous les uns des autres. Les distances et les proximités qu'ils entretiennent avec les modalités des variables considérées permettent ensuite de comprendre en quoi chacun diffère de l'autre : par ses goûts, ses opinions politiques, son âge, son sexe, la marque de sa voiture, la profession de son père... mais la statistique est encore une histoire sans parole.

L'une des contributions majeure de la statistique textuelle est précisément d'animer tous ces graphes en donnant la parole à chacun de ces individus. Grâce à Lebart et Salem, les fameux points-individus ne sont plus muets, ils parlent. Vole alors en éclats la traditionnelle mais artificielle distinction entre le quantitatif et le qualitatif. Les méthodes ici présentées permettent de mettre en relation les propriétés sociales ou personnelles des individus telles que les saisit l'enquête statistique avec les textes par lesquels ces mêmes individus répondent aux questions qu'on leur pose sans en réduire le moins du monde l'information. Les nuances les plus subtiles de l'expression sont conservées : le singulier et le pluriel, la majuscule et la minuscule, l'usage du "je", du "on", du "nous". La formule le dit bien : *s'exprimer* c'est d'abord se livrer soi-même au-dehors. Chaque forme lexicale tire alors son sens d'un triple registre : celui que lui donne celui qui la prononce, celui que lui confère la place qu'elle occupe dans l'espace dessiné par toutes les autres formes lexicales énoncées par le même individu, celui, enfin, qu'elle tient de la place qu'elle occupe dans l'espace dessiné par toutes les autres formes énoncées par tous les autres locuteurs. Le sens jaillit des différences de profil.

Cet ouvrage a le mérite de déborder largement le cadre de l'analyse de contenu ou du traitement statistique des questions ouvertes dans les enquêtes. Il fait le point sur l'état de développement d'un chantier particulièrement foisonnant depuis dix ans. Il expose les dernières découvertes. Elles sont nombreuses et riches d'application dans les domaines les plus divers : stylométrie, recherche documentaire, modèles prévisionnels. Comment attribuer un texte à un auteur ou à une période ? Combien d'auteurs ont contribué à la rédaction du livre de la Bible attribué au prophète Isaïe ? Peut-on comparer des comportements exprimés dans des textes écrits dans des langues différentes sans les traduire ni les coder ?

C'est souvent aux confins des disciplines instituées que l'invention scientifique est la plus féconde. Lorsque deux statisticiens tout particulièrement sensibilisés aux problèmes que l'on rencontre dans les sciences humaines se réunissent autour d'un ordinateur pour élaborer les principes et les outils d'une statistique textuelle, ils occupent le coeur d'un carrefour scientifique vers lequel convergent tout naturellement des linguistes, d'autres statisticiens bien sûr mais aussi les spécialistes d'analyse du discours, d'analyse de contenu, d'analyse des textes littéraires, de recherche documentaire et d'intelligence artificielle. A ce noyau dur de producteurs de théories et d'outils est venu petit à petit s'agréger un univers polyglotte d'utilisateurs aux formations diverses : sociologues, littéraires, stylomètres, historiens, géographes, politologues, médecins, éthologues, psychologues, publicitaires, etc.

On peut savoir gré à l'ouverture d'esprit des deux auteurs (et de leurs associés !), à leur générosité intellectuelle et humaine pour avoir su accueillir autour de leur disque dur un nombre croissant de producteurs et d'utilisateurs dont ils ont souvent stimulé l'inventivité. Il suffit pour s'en convaincre de feuilleter les actes des deux journées internationales qu'ils ont suscitées, avec d'autres, à Barcelone en 1990 et à Montpellier en 1993. Ou de goûter, chez soi, le charme inattendu de nouveaux logiciels.

Au-delà de la collection de principes et d'outils statistiques présentés dans les pages qui suivent, n'oublions pas que la nature même de la matière travaillée - le texte - confère à l'entreprise des dimensions à la fois culturelles, internationales et universelles car comme le disait si bien Victor Hugo ...

Christian Baudelot

AVANT-PROPOS

Cet ouvrage s'adresse à ceux qui, pour leurs recherches, leurs travaux d'études, leur enseignement, doivent décrire, comparer, classer, analyser des ensembles de textes. Il peut s'agir de textes littéraires, scientifiques (bibliométrie, scientométrie, recherche documentaire), économiques, sociologiques (réponses aux questions ouvertes dans des enquêtes socio-économiques, entretiens divers en marketing, psychologie appliquée, pédagogie, médecine), de textes historiques, politiques...

On a tenté de faire le point sur les développements de la *statistique textuelle*, domaine de recherche vivant dont les contours exacts sont difficiles à établir tant est large l'éventail des disciplines concernées, et aussi celui des applications possibles. Les chapitres qui suivent voudraient, tout en présentant l'acquis de ce champ disciplinaire, témoigner de cette richesse d'approches, de méthodes et de domaines.

L'ouvrage reprend, en intégrant des développements récents, certains exemples du manuel *Analyse statistique des données textuelles* publié par les mêmes auteurs en 1988. Le champ des applications précédemment limité aux traitements de *questions ouvertes* a été considérablement élargi de même que l'éventail des méthodes proposées. L'ensemble, profondément remanié, inclut de nouveaux chapitres qui traitent des structures a priori et de l'analyse discriminante textuelle, thèmes qui dépassent largement l'optique essentiellement descriptive de l'ouvrage antérieur.

Plusieurs lectures devraient être possibles selon la formation du lecteur, et selon notamment ses connaissances en mathématique et statistique. Une lecture technique, complète, pour une personne ayant dans ces matières une formation équivalente à une maîtrise de sciences économiques, aux écoles d'ingénieurs ou de commerce. Une lecture pratique, d'utilisateur, pour les personnes spécialisées dans les divers domaines d'application potentiels.

Les démonstrations strictement mathématiques ne figurent pas dans le texte. On renvoie à chaque fois le lecteur curieux d'en connaître les détails

à des publications ou ouvrages plus spécialisés lorsque ceux-ci sont facilement accessibles. En revanche, la part belle est faite à la définition des concepts, à la mise en oeuvre des procédures, aux règles de lecture et d'interprétation des résultats. Le glossaire en fin d'ouvrage aidera le lecteur à préciser le contenu des notions ou des conventions de notation les plus importantes.

L'ensemble doit beaucoup à des collaborations et des cadres de travail divers : au sein du département Economie et Management, de l'Ecole Nationale Supérieure des Télécommunications (Télécom Paris) et de l'URA820 du Centre National de la Recherche Scientifique (Traitement et Communication de l'Information) de cette même Ecole ; au sein du Laboratoire "Lexicométrie et textes politiques", URL 3 de l'Institut national de la langue française (INaLF) et de l'Ecole Normale Supérieure de Fontenay-Saint-Cloud.

Nous remercions également les autres chercheurs ou professeurs auprès desquels nous avons puisé collaboration et soutien, ou simplement eu d'intéressants débats ou discussions. Citons, sans être exhaustif, C. Baudelot (ENS, Paris), M. Bécue, (UPC., Barcelone), L. Benzoni (Télécom Paris), E. Brunet (INaLF, Nice), S. Bolasco (Univ. de Salerne), L. Haeusler (Cisia, Paris), G. Hébrail (EDF, Clamart), D. Labbé (CERAT, Grenoble), A. Lelu (Univ. Paris VIII), M. Reinert (Univ. Toulouse Le Mirail).

Nous sommes heureux d'adresser ici nos remerciements à Gisèle Maïus, des éditions Dunod, pour l'accueil qu'elle a réservé à cet ouvrage.

L. L., A. S.

Sommaire

Introduction 7

Chapitre 1 : Domaines et problèmes 11

1.1 Approches du texte 11
1.1.1 Le courant linguistique 12
1.1.2 Analyse de contenu 13
1.1.3 Intelligence artificielle 14

1.2 Les rencontres de la statistique et du texte 15
1.2.1 Les premiers travaux 16
1.2.2 Les banques de données textuelles 17
1.2.3 La recherche documentaire 18

1.3 Approche statistique du texte 18
1.3.1 La chaîne de traitement 19
1.3.2 Connaissances internes et externes 20
1.3.3 Une méta-information exceptionnelle 21

1.4 Des textes particuliers : les questions ouvertes 23
1.4.1 Les questions ouvertes : un outil de recherche 24
1.4.2 Questions ouvertes et questions fermées 25
1.4.3 Quand utiliser les questions ouvertes ? 27
1.4.4 Traitement pratique des réponses libres 28
1.4.5 Les regroupements de réponses 30

Chapitre 2 : Les unités de la statistique textuelle 33

2.1 Le choix des unités de décompte 33
2.1.1 Le texte en machine 35
2.1.2 Les dépouillements en formes graphiques 35
2.1.3 Les dépouillements lemmatisés 36
2.1.4 Les dépouillements à visée "sémantique" 38
2.1.5 Très brève comparaison avec d'autres langues 40

2.2 Segmentation et numérisation d'un texte 42
2.2.1 Numérisation sur le corpus *Enfants* 44
2.2.2 Le corpus P 45

2.3 L'étude quantitative du vocabulaire 46
2.3.1 Fréquences, gamme des fréquences 46
2.3.2 La loi de Zipf. 47
2.3.3 Mesures de la richesse du vocabulaire 49

Introduction

Les méthodes de *statistique textuelle* rassemblées dans le présent ouvrage sont nées de la rencontre entre plusieurs disciplines : l'étude des textes, la linguistique, l'analyse du discours, la statistique, l'informatique, le traitement des enquêtes, pour ne citer que les principales. Notre démarche s'appuie à la fois sur les travaux d'un courant aux dénominations changeantes (*statistique lexicale*, *statistique linguistique*, *linguistique quantitative*, etc.) qui associe depuis une cinquantaine d'années la méthode statistique à l'étude des textes, et sur l'un des courants de la statistique moderne, la *statistique multidimensionnelle*.

L'outil informatique est aujourd'hui utilisé par un nombre croissant d'usagers pour des tâches qui impliquent la saisie et le traitement de grands ensembles de textes. Cette diffusion renforce à son tour la demande d'outils de gestion et d'analyse des textes qui émane des praticiens et des chercheurs de nombreuses disciplines. Confrontés à des textes nombreux recueillis dans des enquêtes socio-économiques, des entretiens, des investigations littéraires, des archives historiques ou des bases documentaires, ces derniers attendent en effet une aide en matière de classement, de description, de comparaisons...

Nous tenterons précisément de montrer comment les possibilités actuelles de calcul et de gestion peuvent aider à décrire, assimiler et enfin à critiquer l'information de type textuel.

Le choix d'une stratégie de recherche ne peut être opéré qu'en fonction d'objectifs bien définis. Quel type de texte analyse-t-on ? Pour tenter de répondre à quelles questions ? Désire-t-on étudier le vocabulaire d'un texte en vue d'en faire un commentaire stylistique ? Cherche-t-on à repérer des *contenus* à travers les réponses à un questionnaire ? S'agit-il de mettre en évidence les motivations pour l'achat d'un produit à partir d'opinions exprimées dans des entrevues ? Ou de classer des documents afin de mieux les retrouver ultérieurement ?

Bien entendu, aucune méthode d'analyse figée une fois pour toutes ne saurait répondre entièrement à des objectifs aussi diversifiés. Il nous est apparu cependant qu'un même ensemble de méthodes apportait dans un grand nombre d'analyses de caractère textuel un éclairage irremplaçable pour avancer vers la solution des problèmes évoqués.

L'ouvrage que nous avions publié chez le même éditeur en 1988 sous le titre *Analyse statistique des données textuelles* concernait essentiellement l'analyse exploratoire des réponses aux questions ouvertes dans les enquêtes. Le contenu en a été élargi tant au niveau de la méthodologie qu'en ce qui concerne les domaines d'application.

Dans ce nouvel exposé, il ne s'agit plus uniquement de décrire et d'explorer, mais aussi de mettre à l'épreuve les hypothèses, de prouver la réalité de traits structuraux, de procéder à des prévisions. Quant au champ d'application des méthodes présentées, il dépasse dorénavant le cadre des traitements des réponses à des questions ouvertes et concerne des corpus de textes beaucoup plus généraux. Enfin, on a tenté de prendre en compte les travaux qui ont été réalisés depuis la parution du premier ouvrage.

L'accès à de nouveaux champs d'application, même lorsqu'il s'agit de méthodes éprouvées, peut demander une préparation des matériaux statistiques, un effort de clarification conceptuelle, une économie dans l'agencement des algorithmes, une sélection et une présentation spécifique des résultats. Ceci est tout particulièrement vrai pour ce qui concerne le domaine des études textuelles. Dans ce domaine en effet, la notion de *donnée* qui est à la base des comptages statistiques doit faire l'objet d'une réflexion spécifique.

D'une part il est nécessaire de découper des unités dans la chaîne textuelle pour réaliser des comptages utilisables par les analyses statistiques ultérieures. De l'autre, la chaîne textuelle ne peut être réduite à une succession d'unités n'ayant aucun lien les unes avec les autres car beaucoup des *effets de sens* du texte résultent justement de la disposition relative des formes, de leurs juxtapositions ou de leurs cooccurrences éventuelles.

* *

Le premier chapitre, *Domaines et problèmes*, évoque à la fois : les domaines disciplinaires concernés (linguistique, statistique, informatique), les problèmes et les approches. Il précise dans chaque cas la nature du *matériau de base* que constituent les textes rassemblés en corpus.

Le second chapitre, *Les unités de la statistique textuelle*, est consacré à l'étude des unités statistiques que les programmes lexicométriques devront découper ou reconnaître (formes, segments répétés). Il aborde les aspects fondamentaux de l'approche quantitative des textes, les propriétés de ces

unités ; il précise leurs pertinences respectives en fonction des champs d'application.

Les troisième et quatrième chapitres, *L'analyse des correspondances des tableaux lexicaux*, et *La classification automatique des formes et des textes*, présentent les techniques de base de l'*analyse statistique exploratoire* des données multidimensionnelles à partir d'exemples que l'on a souhaité les plus simples possibles.

Le cinquième chapitre : *Typologies, visualisations*, applique les outils présentés aux chapitres trois et quatre à la description des associations entre formes et entre catégories. Il fournit des exemples d'application *en vraie grandeur* commentés du point de vue de la méthode statistique. Il détaille les règles de lecture et d'interprétation des résultats obtenus, fait le point sur leur portée méthodologique.

Pour compléter ces représentations synthétiques, le sixième chapitre, *Eléments caractéristiques, réponses ou textes modaux*, présente les calculs dits de *spécificité* ou de *formes caractéristiques* qui permettent de repérer, pour chacune des parties d'un corpus, celles des unités qui se signalent par leurs fréquences atypiques. La sélection automatique des *réponses modales* ou des textes modaux permet de replacer les formes dans leur contexte, et de caractériser, lorsque cela est possible, des parties de texte, en général volumineuses, par des portions plus petites (phrases, paragraphes, documents, réponses dans le cas d'enquêtes). On résume ainsi, dans le cas des réponses libres, l'ensemble des réponses d'une catégorie de répondants par quelques réponses effectivement attestées dans le corpus, choisies en raison de leur caractère représentatif.

Le septième chapitre, *Partitions longitudinales, contiguïté*, traite le problème des informations *a priori* qui concernent les parties d'un corpus. Dans de nombreuses applications, en effet, l'analyste possède, avant toute démarche de type quantitatif, des informations qui lui permettent de rapprocher entre elles certaines des parties, ou encore de dégager un ordre privilégié parmi ces dernières (*séries textuelles chronologiques*). On étudie dans ce chapitre, en présentant une méthode et de nombreux exemples d'application, les relations de dépendance que l'on peut observer entre ces structures et les profils lexicaux des parties.

Enfin le huitième chapitre, consacré à l'*Analyse discriminante textuelle*, étudie, au sens statistique du terme, le *pouvoir de discrimination* des textes. Comment affecter un texte à un auteur (ou à une période) ? Peut-on prévoir l'appartenance d'un individu à une catégorie à partir de sa réponse à une question ouverte ? Comment classer (ici : affecter à des classes préexistantes) un document dans une base de données textuelles ? On tente

dans ce chapitre, qui contient des exemples d'application variés, de montrer quels sont les apports de la statistique textuelle à la stylométrie, à la recherche documentaire, ainsi qu'à certains modèles prévisionnels.

Le cheminement méthodologique auquel nous invitons le lecteur verra ses étapes illustrées par des corpus de textes provenant de sphères de recherche très différentes. Les résultats présentés à ces occasions concernent des textes littéraires, des corpus de réponses libres dans des enquêtes françaises et internationales, des discours politiques.

L'ensemble des exemples devrait permettre au lecteur d'apprécier la variété des applications réalisées et potentielles, la complémentarité des divers traitements, tout en progressant dans l'assimilation et la maîtrise des méthodes, et surtout dans sa capacité à évaluer et critiquer les résultats.

Chapitre 1

Domaines et problèmes

L'étude des textes à l'aide de la méthode statistique constitue le centre d'une sphère d'intérêts que l'on désigne par *statistique textuelle*. Au fil des années le contexte général de ces recherches, les objectifs qu'elles se sont fixés, les principes méthodologiques qu'elles ont adoptés ont subi des évolutions importantes.

Ce chapitre retrace brièvement les circonstances particulières de la rencontre entre la linguistique et la statistique ; deux disciplines profondément éloignées dans leurs principes et leur histoire, ayant chacune subi plusieurs mutations importantes, toutes deux profondément marquées, pour des raisons de proximité et d'affinité évidentes, par l'avènement de l'informatique. Il souligne certains aspects des deux disciplines précitées susceptibles d'aider à mieux comprendre leurs relations et leur synergie.

Après de brefs rappels sur les préoccupations propres aux linguistes et aux statisticiens ainsi que sur les aventures que les métiers correspondants ont pu avoir en commun, seront évoqués les deux domaines d'applications qui ont été privilégiés dans cet ouvrage : l'étude des textes (littéraires, politiques, historiques...), et le dépouillement de corpus particuliers que constituent les réponses aux questions ouvertes dans les enquêtes socio-économiques.

1.1 Approches du texte

Commençons par situer brièvement la *statistique textuelle* parmi les principales disciplines en rapport avec le texte (*linguistique, analyse du discours, analyse de contenu, recherche documentaire, intelligence artificielle*). Comme on le verra dans le bref exposé qui suit, le texte constitue un passage obligé dans ces disciplines très différentes qui ont des buts, des méthodes et des perspectives de recherches nécessairement

distincts. Nombre de disciplines et domaines de recherches (théories des langages, grammaires formelles, linguistique computationnelle, etc.) allient à des degrés divers linguistique, mathématique et informatique sans utiliser cependant les modèles et les outils de la statistique. Ils seront évoqués sans faire l'objet de présentation particulière.

1.1.1 Le courant linguistique

La linguistique, "science pilote des sciences humaines", s'est précisément constituée en rupture avec toute une série de pratiques antérieures dans le domaine de l'étude de la langue. La notion de système y joue un rôle central qui interdit pratiquement de considérer des "faits" isolés. La linguistique structurale envisage, en effet, la description des unités linguistiques dans le cadre de systèmes assignant des valeurs différentielles à chacune des unités qui le constituent. On retrouve ce "point de vue" dans l'extrait ci-dessous, emprunté à Ferdinand de Saussure[1].

> "Ailleurs il y a des choses, des objets donnés, que l'on est libre de considérer ensuite à différents points de vue. Ici il y a d'abord des points de vue, justes ou faux, mais uniquement des points de vue, à l'aide desquels on crée secondairement les choses. Ces créations se trouvent correspondre à des réalités quand le point de départ est juste ou n'y pas correspondre dans le cas contraire; mais dans les deux cas aucune chose, aucun objet, n'est donné un seul instant en soi. Non pas même quand il s'agit du fait le plus matériel, le plus évidemment défini en soi en apparence, comme serait une suite de sons vocaux."

Au siècle dernier on étudie le langage le plus souvent à travers des textes. La *philologie* permet d'interpréter, de commenter les textes en restituant le vrai sens des mots qui les composent. Comme le note M. Pêcheux (1969) :

> On se demande simultanément: " De quoi parle ce texte ?", "Quelles sont les principales idées contenues dans ce texte ?", et en même temps "Ce texte est-il conforme aux normes de la langue dans laquelle il est présenté ? " ou bien "Quelles sont les normes propres à ce texte ?"
> /.../ En d'autres termes, la science classique du langage prétendait être à la fois *science de l'expression* et *science des moyens de cette expression* /.../

Après le "Cours de Linguistique Générale" de Ferdinand de Saussure (1915) la linguistique ne considère plus le texte comme l'objet de son étude. Ce qui fonctionne pour un linguiste structuraliste c'est la langue : ensemble de systèmes autorisant des combinaisons et des substitutions réglées sur des éléments définis.

---

[1] Notes de 1910 parues dans les *Cahiers Ferdinand de Saussure*, n°12 (1954), p 57-58. Cité par Benveniste (1966).