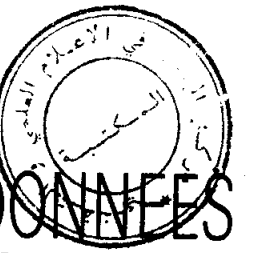
LUDOVIC LEBART
Directeur de Recherche
au C.N.R.S.

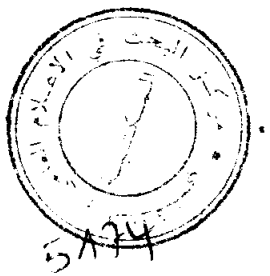
ANDRE SALEM
Ingénieur à
l'École Normale Supérieure de
Fontenay-St. Cloud



ANALYSE STATISTIQUE DES DONNEES TEXTUELLES

Questions ouvertes et lexicométrie

Préface de Christian BAUDELOT
Professeur de Sociologie à l'Université de Nantes



ISBN : 2-04-018779-0

PREFACE

Sous les dehors innocents et apparemment austères d'un nouvel outil de statistique descriptive, modestement appelé, selon ses auteurs, à faciliter le traitement des questions ouvertes dans les enquêtes d'opinion, Ludovic Lebart et André Salem fournissent des moyens rigoureux et entièrement nouveaux pour analyser le langage, dans sa structure interne aussi bien que dans les relations qu'il entretient avec les contextes de sa production.

Nombreux sont les linguistes qui ont montré que le langage n'était pas seulement un moyen de communication entre les hommes mais qu'il dotait ces derniers de la possibilité d'organiser, en l'exprimant, leur expérience sensible: "nous disséquons la nature, écrivait B.L.Whorf, selon les lignes tracées à l'avance par nos langues maternelles". Or, si les langues, en tant que systèmes de concepts et de catégories, varient selon les cultures et les pays, les façons d'user de la langue de son propre pays varient aussi fortement selon les milieux sociaux, les degrés d'instruction, le sexe, l'âge, la région, bien sûr, mais aussi et tout simplement selon les individus, les écrivains ou les poètes, les époques... Un ouvrier ne parle pas comme un avocat mais Stendhal n'écrit pas non plus comme Balzac. Ces différences ne sauraient se réduire aux stricts aspects formels que se bornent apparemment à enregistrer les comptages lexicographiques: elles renvoient bel et bien à des variations de perception et d'appréhension de la réalité, à des écarts de significations, à des distinctions de contenus investis dans des expériences sociales, individuelles ou historiques particulières. *Manque d'argent* et *problèmes financiers*, *chômage* et *manque de travail* ont beau être invoqués par les uns et les autres comme des raisons d'hésiter à avoir des enfants, on aurait tort de les considérer comme des façons différentes d'exprimer les mêmes choses: ce n'est pas, ici et là, la même réalité ni le même rapport à la réalité qui sont désignés.

Telle est bien la nature profonde des différences que permettent de mettre à jour et de rapporter aux propriétés sociales et individuelles de leurs auteurs, les outils statistiques présentés dans ce volume. Système de signes, de concepts et de représentations de la réalité qui ne saurait exister en dehors des mots qui l'expriment, le langage est soumis ici à une analyse formelle qui est nécessairement aussi une analyse de contenu. Mais une analyse de contenu où les garanties de l'objectivité sont assurées le plus longtemps possible par un retardement de la phase d'interprétation : l'information textuelle est en effet maintenue à l'état brut, sans sélection ni codage tout au long de l'analyse et chaque forme lexicale tire 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 profils et il n'est, selon Lévi-Strauss, de sociologie que de la différence. La finesse et la sophistication de l'instrument permettent, d'ailleurs en respectant dans sa totalité l'ensemble des propos tenus par chaque individu et en décomposant chaque texte en ses unités minimales, de mettre en évidence des nuances, invisibles à l'oeil nu!

C'est donc un grand outil scientifique que les auteurs de ce livre ont élaboré, avec toutes celles et tous ceux qui ont contribué à le mettre au point. Il est par là, au même titre qu'un microscope électronique ou un ordinateur, nécessairement interdisciplinaire et international : pour peu que ses utilisateurs potentiels en découvrent l'existence et sachent se l'approprier en le pliant, au prix de quelques efforts d'imagination, à la satisfaction des besoins qui sont les leurs. Il est déjà encourageant que des disciplines aussi étrangères les unes aux autres que le marketing, la médecine, la sociologie et l'analyse littéraire des textes classiques comptent aujourd'hui parmi ses utilisateurs. Quel plaisir de voir traités côte à côte dans deux librairies voisines d'un ordinateur central des fichiers contenant l'un, les réponses à une enquête cherchant à cibler le public d'une nouvelle lessive, l'autre, *Le Rivage des Syrtes* et l'ensemble des pièces de Sophocle.

Christian BAUDELOT

AVANT-PROPOS

Cet ouvrage s'adresse à ceux qui, pour leurs travaux d'études, leur recherche, leur enseignement, doivent traiter des informations de type textuel: réponses aux questions ouvertes dans les enquêtes socio-économiques, entretiens divers (marketing, psychologie appliquée, pédagogie, médecine), mais aussi études historiques ou littéraires...

Si le domaine d'application retenu ici est celui des réponses aux questions ouvertes dans les enquêtes, c'est surtout à cause de l'expérience pratique accumulée au cours des années récentes par les auteurs.
L'interdisciplinarité évidente du sujet traité implique en effet une certaine prudence, sous peine de dispersion, ou même d'écartèlement. Les utilisateurs dans d'autres domaines devraient imaginer sans trop de difficulté quels résultats et quels services ils peuvent attendre des traitements proposés dans le champ d'application qui leur est familier.

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, le lecteur intéressé étant renvoyé à des 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 doit en fait aider simultanément les deux types de lecteurs à se familiariser avec certaines notions ou simplement à se rafraîchir la mémoire...

Certains des travaux présentés dans ce volume reprennent des publications spécialisées de l'un ou l'autre des auteurs. L'ensemble doit beaucoup à des collaborations diverses dans le cadre de l'équipe "Conditions de vie et Aspirations des Français", du CREDOC (Centre de Recherche pour l'Etude et l'Observation des Conditions de Vie) et du Laboratoire "Lexicométrie et textes politiques", URL 3 de l'Institut National de la Langue Française (INALF) - Ecole Normale Supérieure de Fontenay-Saint-Cloud.

Qu'il nous soit permis de remercier, au sein de l'équipe précitée du CREDOC, Françoise Boscher, Eric Brian, Frédéric Chateau, Catherine Duflos, Ghislaine Dronault, Brigitte Ezvan, Michel Grignon, Françoise Gros, Laurence Haeusler, Yvette Houzel, Lucette Laurent, Philippe Pleuvret, qui collaborèrent au fil des années à la création et au maintien d'une source statistique qui a servi de banc d'essai aux techniques proposées dans cet ouvrage; ainsi que: Pierre Fiala, Annie Geffroy, Jacques Guilhaumou, Benoît Habert, Pierre Lafon, Josette Lefèvre, et l'ensemble des membres du Laboratoire de St. Cloud

Nos remerciements vont également à Jean-Paul Benzécri, Georges Th. Guilbaud, Edmond Lisle, Edmond Malinvaud, Maurice Tournier, qui dirigèrent nos recherches dans le cadre du CNRS ou de l'Université.

Monique Bécue, de l'Université Polytechnique de Barcelone, par ses contributions en matière de logiciels, notre collègue et ami Alain Morineau, du CEPREMAP (Centre de Prospective et de Mathématique Appliquées à la Planification) et la Direction du CISIA (Centre International de Statistique et d'Informatique Appliquées) par leur collaboration et leur soutien en matière de développement, nous ont apporté une aide précieuse. Ils ont bien voulu relire et critiquer le manuscrit avec Catherine Duflos, Laurence Haeusler, Jean-Pierre Nakache que nous sommes heureux de remercier ici, sans oublier Gisèle Maïus, des éditions Dunod, pour l'accueil qu'elle a réservé à cet ouvrage.

L. L., A. S.
Paris, mai 1988

Sommaire

INTRODUCTION

Ce travail répond à la demande de praticiens confrontés à des textes nombreux recueillis dans des enquêtes socio-économiques, des entretiens, ou dans des investigations littéraires, historiques . . .
Nous essaierons de montrer comment les possibilités actuelles de calcul et de gestion, largement ouvertes par la diffusion de l'informatique, peuvent aider à décrire, assimiler et enfin à critiquer l'information de type textuel.

Parmi les possibilités d'investigation évoquées ci-dessus, les techniques d'analyse des données, c'est-à-dire d'analyse statistique exploratoire multidimensionnelle, figurent en bonne place.

Cependant, même lorsqu'il s'agit de méthodes éprouvées (analyse des correspondances, classification hiérarchique), l'accès à de nouveaux champs d'application 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" résultent justement de la disposition relative des formes, de leurs juxtapositions.

L' "interface" entre techniques statistiques et matériaux textuels repose sur un certain nombre d'hypothèses de travail qui seront développées tout au long des chapitres qui suivent.