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.