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).

> En séparant la langue de la parole on sépare du même coup : 1) ce qui est social de ce qui est individuel, 2) ce qui est essentiel de ce qui est accessoire et plus ou moins accidentel[1].

On distingue à l'intérieur de la linguistique plusieurs domaines qui étudient des faits de langue de natures différentes :

- la *phonétique* étudie les sons du langage, alors que la *phonologie* étudie les phonèmes c'est-à-dire les sons en tant qu'unités distinctives.
- la *lexicologie* étudie les mots, dans leur origine, leur histoire et dans les relations qu'ils ont entre eux.
- la *morphologie* traite des mots pris indépendamment de leurs rapports dans la phrase. Elle étudie les morphèmes ou éléments variables dans les mots ; morphèmes grammaticaux (désinences ou flexions) et morphèmes lexicaux.
- la *syntaxe* étudie les relations entre les mots dans la phrase (ordre des mots, accord).
- la *sémantique* étudie la signification, le contenu du message.
- la *pragmatique* étudie les rapports entre l'énoncé et la situation de communication.

Il faut noter que, dans la pratique, les règles énoncées dans chacun de ces domaines sont en relation non seulement entre elles mais avec les règles des autres domaines.

Ainsi, l'étude du lexique ne peut être faite sans référence au sens des mots (sémantique). Celle du sens des mots passe par l'analyse de leur fonction dans la phrase (syntaxe)[2] et parfois par l'étude du contexte de l'énonciation (pragmatique).

1.1.2 Analyse de contenu

Sur le terrain de l'étude de la signification des textes délaissé par la linguistique se sont développées plusieurs méthodes d'approche des textes. Née aux États Unis au début de ce siècle l'*analyse de contenu* sert d'abord à des études portant sur les organes de presse. Selon la formule de B. Berenson et P.F. Lazarsfeld elle se présente comme :

---

[1] Cf. Saussure (1915).

[2] On regroupe parfois sous le nom de morpho-syntaxe l'ensemble des phénomènes qui relèvent de la morphologie ou de la syntaxe.

> "/.../ une technique de recherche pour la description objective systématique et quantitative du contenu manifeste de la communication."[1]

L'analyse de contenu se propose, sans s'attarder sur le matériau textuel proprement dit, d'accéder directement aux significations de différents segments qui composent le texte. Elle opère pour cela en deux temps. Tout d'abord le codeur commence par définir un ensemble de classes d'équivalence, de thèmes, dont il repérera ensuite les occurrences au fil du texte ainsi analysé. Dans un second temps, il pratique des comptages pour chacun des thèmes prévus dans la grille de départ. Les unités recensées par l'analyse de contenu peuvent être des thèmes, des mots ou même des éléments de syntaxe ou de sémantique. L'unité de décompte pour les mesures quantitatives varie elle aussi : mot, surface couverte par l'article, etc.

Comme on le voit l'analyse de contenu ainsi définie comporte une dimension statistique. Cependant sa réussite suppose que le système des catégories définies a priori est à la fois cohérent et pertinent, ce qui est difficile à assurer dans la pratique.

1.1.3 Intelligence artificielle

Les programmes de recherche en *compréhension de la parole et de l'écrit* ont des objectifs beaucoup plus ambitieux. Ce domaine d'investigation très étendu se situe au coeur des activités désignées par l'expression *industries de la langue*, avec, parmi les réalisations dont les enjeux économiques sont évidents, la traduction automatique, les correcteurs d'orthographes, la reconnaissance de la parole et de l'écriture manuscrite, la commande vocale d'automates, plus généralement les interfaces machine-utilisateurs en langage naturel, etc.[2]

Si les domaines d'étude se recoupent parfois (la recherche documentaire ou bibliométrie, par exemple, utilise aussi bien des outils statistiques que des procédures relevant de l'intelligence artificielle), les préoccupations sont cependant très distinctes, et les approches complémentaires.

Dans le domaine de l'Intelligence Artificielle on étudie par exemple le texte en vue d'applications "en temps réel" ou "décisionnelles" (on parle alors plutôt de "langage naturel") afin de rendre possible ou d'améliorer le dialogue homme-machine. Pour les informaticiens qui travaillent dans ce

---

[1] B. Berelson et P.F. Lazarsfeld, The Analysis of communications content, University of Chicago and Columbia University, Chicago and New York, 1948. En français, on pourra consulter le manuel de synthèse de L.Bardin (1989).

[2] L'ouvrage de Carré et al. (1991) dresse un panorama des résultats et des problèmes dans ces domaines de recherche en pleine effervescence.

domaine, le but est d'obtenir une représentation du sens des phrases que l'on présente au système informatique[1]. Ceci les amène parfois, comme le fait ici Jacques Pitrat (1985), à donner, dans le cadre qui est le leur, de nouvelles définitions de l'activité de compréhension.

> Je dis qu'un programme a "compris" un texte s'il a construit une représentation du sens de ce texte indépendante de toute langue naturelle.

Pour arriver à cette représentation canonique et rapprocher des phrases dont la signification peut être jugée identique, les chercheurs qui travaillent en intelligence artificielle sont amenés à créer des concepts qui ne recouvrent pas strictement les mots de la langue. On donnera au chapitre 2 (paragraphe 2.1.4) un bref exemple de ces tentatives de *représentation sémantique canonique*, largement indépendante de la forme du message.

Cette notion de compréhension qui occupe une position centrale dans la plupart des recherches impliquant intelligence artificielle et textes est, pourrait-on dire, presque étrangère à la statistique textuelle, à l'exception, dans certains cas, de phases de prétraitement qui seront évoquées au chapitre 2.

En revanche, l'indépendance des méthodes de la statistique textuelle vis-à-vis de la phase dite de "compréhension" des textes est illustrée au cours des chapitres qui suivent par la présentation, à partir des mêmes logiciels, d'applications dans des langues aussi diverses que le français, l'anglais, le japonais, l'hébreu ancien. Et l'on se réfère dans le cours du texte à d'autres applications en espagnol, arabe, grec, sans évidemment que cette liste des langues utilisées ou utilisables soit limitative. Cette indépendance n'est bien entendu que transitoire : les résultats acquis à partir de comptages et de traitements statistiques automatisés constituent seulement des pièces supplémentaires à verser au dossier du traitement global de l'information de base.

1.2 Les rencontres de la statistique et du texte

Les succès remportés par les applications de la méthode statistique dans de nombreux domaines des sciences de la nature (physique, biologie, etc.) mais

---

[1] On lira avec profit la synthèse publiée par D. Coulon et D. Kayser (1986), qui fait le point sur les méthodes utilisées en *intelligence artificielle* du point de vue des informaticiens. Cf. aussi la partie "compréhension des langues naturelles" dans Bonnet (1984), Haton (1985), et le travail de synthèse récent de McKevitt et al. (1992).

aussi dans ceux des sciences humaines (psychologie, économie, etc.) et y compris dans des disciplines qui touchent à l'utilisation du langage finissent par attirer l'attention des spécialistes de l'étude du vocabulaire[1].

Initialement découvertes comme des lois empiriques devant permettre en premier lieu des améliorations dans le domaine de la transcription sténographique (J.B. Estoup, 1916), les distributions lexicales sont par la suite étudiées sous le signe de la "psycho-biologie du langage" par G.K. Zipf (1935).

1.2.1 Les premiers travaux

Les études qui appliquent alors la méthode statistique à l'analyse des textes regroupent souvent une série d'approches quantitatives portant sur l'ensemble des unités linguistiques que l'on peut répertorier dans un même texte (phonèmes, lexèmes). Ces travaux comportent souvent plusieurs volets consacrés à des langues ou à des problèmes différents, destinés à convaincre le lecteur de la pertinence et de l'universalité des applications de la méthode statistique aux études textuelles.

Dans un second temps, la *statistique lexicale* (G. U. Yule, P. Guiraud puis Ch. Muller) entreprend de résoudre une série de problèmes posés par les stylisticiens préoccupés d'études comparatives sur le vocabulaire des "grands auteurs" et en particulier des auteurs du théâtre classique français du 17e siècle.[2]

Ce courant commencera par se fixer des objectifs qui portent encore la marque de préoccupations formulées bien avant l'apparition des méthodes quantitatives : mesures comparatives de l'étendue du vocabulaire de différents auteurs, mesure de l'évolution du vocabulaire d'un même auteur au cours de la période pendant laquelle il a produit son oeuvre, etc.

Les travaux de cette *statistique lexicale* sont aussi des manuels de statistique "à l'usage des littéraires" ; ils vont permettre d'objectiver par des comptages les appréciations portées intuitivement par les stylisticiens bien avant l'apparition des méthodes quantitatives, et parfois même de les réfuter.

---

[1] Il semble même que l'absence d'études statistiques est perçue, dans les débuts des études de statistique textuelle du moins, comme un "retard" sur les autres disciplines qu'il s'agit de rattraper au plus vite. "Il me semble pouvoir affirmer que ce serait entraver le développement de la linguistique que de continuer à tant se désintéresser des nombres quand nous parlons des phénomènes linguistiques" affirme Marcel Cohen dans une conférence prononcée en 1948, (Cohen, 1950). Rappelons également, la formule lancée par Pierre Guiraud : "La linguistique est la science statistique type ; les statisticiens le savent bien, la plupart des linguistes l'ignorent encore", (Guiraud, 1960).

[2] On consultera par exemple Muller (1964 et 1967), Bernet (1983).

Ces travaux mettront à jour de nombreuses difficultés liées à la définition des unités de décompte. La partie linguistique de ces travaux réside essentiellement dans le choix raisonné des unités de dépouillement qui serviront de base aux analyses ultérieures.

Parallèlement, les méthodes développées dans ce même cadre seront présentées par G. Herdan, sous le nom de *linguistique statistique*, comme "la quantification de la théorie saussurienne du langage". Selon Herdan (1964) cette discipline se présente comme une branche de la linguistique structurale, avec pour principale fonction la description statistique du fonctionnement (dans des corpus de textes) des unités définies par le linguiste aux différents niveaux de l'analyse linguistique (phonologique, lexical, phrastique).

Cette simplification opérée directement au plan théorique par Herdan paraît assez hardie si on la confronte au point de vue développé plus haut par le fondateur de la linguistique structurale.

1.2.2 Les banques de données textuelles

A une époque plus récente, la statistique textuelle, délaissant les dépouillements expérimentaux réalisés manuellement sur des ensembles de textes relativement peu volumineux, s'oriente vers des comparaisons portant sur de plus vastes ensembles de textes.[1] Cette orientation, qui trouve certaines justifications au plan statistique, entraîne du même coup de profondes modifications dans la pratique des dépouillements textuels.

De plus en plus, le coût important imposé par la saisie des textes sur un support lisible par un ordinateur incite le milieu scientifique à constituer, en amont des études particulières, des "banques de textes" permettant à plusieurs équipes de chercheurs d'effectuer leurs recherches propres à partir des mêmes textes saisis selon des standards très généraux. L'équipe du *Trésor général des langues et parlers français* gère ainsi, pour les besoins d'une communauté scientifique qui s'élargit chaque jour, un ensemble de textes qui compte désormais plus de 160 millions d'occurrences pour ce qui concerne les 19e et 20e siècles.[2]

Le recours à l'ordinateur, rendu nécessaire à la fois par l'ampleur des opérations de dépouillement envisagées et par le volume des calculs

---

[1] Le travail d'E. Brunet (1981) réalisé à partir des données du Trésor de la langue française fournit un exemple de telles études comparatives.

[2] Ce laboratoire a été dirigé successivement par P. Imbs, B. Quemada et R. Martin. Le logiciel STELLA (Dendien, 1986) permet d'interroger à distance la base de données textuelles FRANTEXT.

statistiques couramment effectués, implique en retour que l'on définisse de manière toujours plus précise l'ensemble des règles qui présideront au dépouillement automatique des textes stockés en machine.

A la différence des dépouillements réalisés sur des textes enregistrés en vue d'un type d'étude particulier, le dépouillement des textes provenant de banques de textes doit se faire suivant des méthodes très générales, totalement automatisables et facilement explicitables.

1.2.3 La recherche documentaire

La recherche documentaire constitue un domaine de recherche qui possède sa propre spécificité, au carrefour de l'informatique (essentiellement : les bases de données), de l'intelligence artificielle, de la linguistique quantitative, de la statistique. Les méthodes statistiques peuvent intervenir au moment de la constitution et de l'organisation de la base de documents (aides à l'analyse syntaxique et à la classification)[1] ; dans les phases d'exploration et de *navigation* à l'intérieur de la base[2]; enfin dans la phase de recherche de documents à partir de descriptifs en langage naturel ou à partir de mots-clés. Cette dernière phase sera abordée de façon plus détaillée au chapitre 8 dévolu aux méthodes d'analyse discriminante textuelle. La *bibliométrie*, la *veille technologique*[3] sont deux domaines disciplinaires qui englobent comme outil de base la recherche documentaire.

1.3 Approche statistique du texte

Pour aborder la complexité de la relation statistique-textes, commençons par prendre le point de vue du statisticien à partir de quelques notions générales. Pour un statisticien, le texte doit être appréhendé dans le domaine du discret, du qualitatif, du comptage et non de la mesure.[4]

Les méthodes d'analyse des variables qualitatives, dans l'ensemble plus complexes et difficiles à mettre en oeuvre, se sont pleinement développées à

---

[1] Outre les ouvrages fondamentaux de Van Rijsbergen (1980), Salton et MacGill (1988) , on consultera par exemple Blosseville et al. (1992), Lewis et al. (1990, 1992), Wilks and al. (1991).

[2] Cf. par exemple Cutting et al. (1992), Lelu et Rozenblatt (1986), Lelu (1991).

[3] On pourra consulter sur ce sujet l'ouvrage collectif édité par Desval et Dou (1992) et les travaux de Callon et al. (1991), Warnesson et al. (1993), Michelet (1988).

[4] Dans la plupart des disciplines qui ont fait naître et alimenté la statistique au début du siècle (biométrie, agronomie, etc.), on mesure en effet des grandeurs, on calcule des moyennes, des variances, des coefficients de corrélation à partir de variables numériques.

partir des années soixante, stimulées par les nouvelles possibilités de calcul. L'ensemble de ces circonstances explique, selon nous, le caractère récent des applications de la statistique qualitative multidimensionnelle au domaine textuel.[1]

1.3.1 La chaîne de traitement

Dans la plupart des applications statistiques, on retrouve de façon assez générale, l'enchaînement de quatre phases que l'on peut symboliser par le schéma :

problème - données - traitement - interprétation

Chacune des étapes de cette séquence pose d'ailleurs des problèmes différents selon le contexte, les préoccupations et les domaines d'application.

Le *problème* qui a motivé l'étude peut donner lieu à la formalisation a priori d'un modèle statistique ou probabiliste, ou au contraire être formulé en termes très généraux, ne débouchant que sur une description ou une exploration sommaire de l'univers concerné.

Les *données* peuvent être expérimentales (leur recueil peut même être conditionné par le modèle, par exemple dans le cas des plans d'expérience en agronomie) ou provenir de l'observation (elles peuvent dans certains cas préexister à la formulation des problèmes, comme dans le cas d'analyses secondaires d'enquêtes par exemple).

La phase que l'on désigne par *traitement* est constituée dans le cas inférentiel le plus classique, par la mise à l'épreuve d'hypothèses ou de modèles. Dans le cas descriptif ou exploratoire, cette phase est surtout une mise en forme des données destinée à faire apparaître les traits structuraux les plus importants.

Enfin la phase d'*interprétation* peut se réduire à la prise en compte des conclusions d'un test d'hypothèse, dans le cas le plus classique ; elle comportera souvent une évaluation critique des hypothèses et de l'éventuel modèle de départ. Dans le cas descriptif ou exploratoire, la phase d'interprétation comprendra inévitablement une réflexion sur la validité et la signification des structures observées, avec parfois une critique corrélative des données (pertinence, recueil, codification) et une remise en cause ou un

[1] Le cours de linguistique mathématique de J.P. Benzécri à la Faculté des Sciences de Rennes (1964) et la thèse de B. Escofier-Cordier (1965) sur l'analyse des correspondances sont des travaux de pionniers dans ce domaine.

affinement des hypothèses générales. Ces critiques et réflexions pourront, dans une phase ultérieure, donner lieu à de nouveaux traitements, de nouvelles observations, et éventuellement suggérer de nouveaux modèles.

Bien entendu, cette grille de description ne fait que baliser les grandes étapes de l'activité du statisticien. La pratique est riche de situations intermédiaires. C'est d'ailleurs dans ces situations qui compliquent la tâche de l'utilisateur que le travail devient particulièrement intéressant pour le chercheur.

1.3.2 Connaissances internes et externes, méta-information

L'automatisation de la chaîne *problème-données-traitement-interprétation*, notamment en vue du recours à une étape de type statistique dans des systèmes expert, a conduit récemment certains chercheurs à introduire la notion de *meta-data* que l'on peut traduire dans ce contexte par méta-données ou plutôt par *méta-information*. Il s'agit en bref des nombreuses informations que l'on possède sur le tableau de données que l'on s'apprête à analyser, et qui ne figurent pas dans le tableau lui-même.

Ainsi, des inégalités, des relations entre les valeurs de certaines variables, ou encore des bornes (supérieures ou inférieures) pour ces valeurs sont les formes les plus élémentaires de méta-information. Dans le dépouillement d'enquête, cette méta-information, dont la formalisation est relativement aisée, est d'ailleurs utilisée en routine pour contrôler et "nettoyer" les fichiers, ou procéder à des tests de cohérence.

Mais il y a aussi toute la connaissance externe que l'on peut avoir sur un objet d'étude, en rapport avec le recueil de données : telle mesure est suspecte ou peu fiable, on observe un effet de batterie ou d'acquiescement global sur un groupe de question, ou encore : une théorie classique laisse prévoir que..., etc.[1]

Le développement des analyses exploratoires ainsi que le travail sur bases de données accentuent l'intérêt de la notion de méta-information. Redécouvrir des structures connues est en effet utile à fin de vérification, mais ne constitue pas la fin ultime de ces analyses qui doivent apprendre à utiliser ce qui est déjà connu pour en savoir davantage.

---

[1] Des tentatives de formalisation de ces méta-informations ont été réalisées par Diday (1992), dans le cadre de ses travaux sur l'analyse des données symboliques (par opposition à analyse des données numérique). On consultera aussi Hand (1992), et également, dans le cadre de la recherche documentaire Froeschl (1992).

1.3.3 Une méta-information exceptionnelle

Ce trop bref tableau doit servir à situer l'information de type textuel dans le contexte usuel de l'activité statistique : la méta-information, dans le cas des données textuelles, est particulièrement abondante.

Chaque mot utilisé, même si c'est un mot grammatical[1] (appelé parfois mot vide en documentation, ou encore mot-outil) a droit à plusieurs lignes, ou plusieurs pages dans un dictionnaire encyclopédique. Les règles de grammaire constituent évidemment une méta-information fondamentale.

Les mots appartiennent à des réseaux sémantiques que les dictionnaires et les analyseurs morpho-syntaxiques partiellement automatisés s'efforcent de prendre en compte.

Le problème principal concerne la pertinence de ces différents niveaux de méta-information vis-à-vis du problème que l'on étudie.

Supposons, pour prendre un exemple hors du domaine textuel, que l'on étudie les histogrammes des longueurs d'ondes correspondant aux couleurs d'un tableau de Rembrandt (pour chacun des pixels d'une reproduction). Il va de soi que l'on utilise une fraction dérisoire de l'information contenue dans l'image d'origine. Il est cependant possible que la forme de l'histogramme (ou d'une fonction plus élaborée des mêmes mesures et données de base) permette de distinguer un Rembrandt d'un Rubens ou d'un Van Dyck.

L'ensemble des méta-informations disponibles n'est pas indispensable et n'a donc pas besoin d'être utilisé, si le seul but que l'on se fixe est de discriminer les tableaux.

De la même façon, les études stylométriques évoquées au chapitre 7 (homogénéité du *Livre d'Isaïe*) et au chapitre 8 (poèmes de Shakespeare) n'utilisent qu'une part infime de l'information contenue dans les textes concernés, et une part encore moindre des savoirs que nous possédons sur ces même textes.

A l'opposé, dans certains cas de recherche documentaire, par exemple, on peut travailler sur des mots-clés qui jouent le rôle de variables qualitatives classiques de présence-absence, et ainsi construire des tableaux tout-à-fait analogues à ceux que l'on peut rencontrer dans d'autres applications statistiques. Le document à classer ou à retrouver n'est alors plus un texte, mais un sac de mots, sans ordre ni syntaxe. Nous reviendrons plusieurs fois,

---

[1] Signalons que si de nombreux auteurs se servent de cette notion, la plupart du temps pour réduire leur champ d'investigation à des formes au contenu sémantique plus lourd, on ne peut espérer dresser une fois pour toutes une liste de ces formes qui donne satisfaction pour l'ensemble des études textuelles.

au cours des chapitres qui suivent, sur cette diversité de points de vue et d'outils.[1]

Modélisations de la gamme des fréquences

Même si l'on ne tient pas compte de la méta-information qui vient d'être mentionnée, le matériau textuel constitue un objet relativement complexe pour le statisticien. Il existe en effet dans tout texte une dimension séquentielle, ou syntagmatique, qui l'apparente à ce qu'on désigne en statistique sous le nom de processus[2], encore que l'énonciation d'un texte constitue un type de processus particulièrement complexe. Par ailleurs les mots du texte entretiennent entre eux (et aussi avec les autres mots de la langue) des rapports qui peuvent être éclairés par des comptages réalisés sur la totalité du texte.

Dès les débuts des études statistiques appliquées à des textes, plusieurs modèles de distribution théorique du vocabulaire ont été proposés. Les plus anciennes tentatives remontent à Zipf dès 1932 (cf. chapitre 2, paragraphe 2.3.2) et Yule (1944). La distribution de Waring-Herdan (cf. Herdan, 1964, Muller, 1977) est probablement le modèle le plus cité dans la littérature linguistique.

On donnera un exemple plus récent de modèle de gamme des fréquences du vocabulaire au chapitre 8, (paragraphe 8.3) à propos des études stylométriques. Citons encore, à titre d'exemple, dans le cas de la recherche documentaire, le modèle N-poissonien (mélange de lois de Poisson) proposé par Margulis (1992) pour la distribution des mots dans les documents.

Compte tenu de la complexité et de la nature même du matériau de base, ces modèles ne fournissent que des analogies formelles, et n'ont nullement l'ambition de participer à l'explication des causes du phénomène de création du texte.

Fort heureusement, les méthodes exploratoires, qui fournissent les outils les plus utilisés de la statistique textuelle, et qui occupent une position centrale dans cet ouvrage, ne s'appuient pas directement sur ces modèles.

---

[1] Notons qu'en revanche, la sélection des mots-clés peut demander un prétraitement faisant largement appel aux méta-informations, et qui pourra utiliser des techniques relevant de l'intelligence artificielle.

[2] On verra au chapitre 2 comment la prise en compte des unités statistiques que sont les segments répétés permet de prendre en compte de façon opératoire certains aspects séquentiels du texte.

Elles utilisent cependant des informations tirées du texte dont il faut comprendre la nature et l'organisation fréquentielle.

1.4 Des textes particuliers : les réponses aux questions ouvertes

Les réponses aux questions ouvertes, appelées encore réponses libres, sont des éléments d'information très spécifiques, qui peuvent déconcerter à la fois les statisticiens et les spécialistes des études textuelles. Les premiers peuvent être découragés par le caractère imprécis et multiforme de ces réponses, les seconds par leur caractère artificiel, et leur forte redondance globale.

Le statut de la répétition, et plus généralement celui de la fréquence avec laquelle les formes sont employées y est en effet très particulier. Les fréquences lexicales observées sont pour une large partie artificielles, car la même question est posée à des centaines ou des milliers de personnes... la juxtaposition des réponses constitue un texte redondant par construction, où les stéréotypes ne sont pas rares.

Mais les questions ouvertes constituent un prolongement indispensable des questionnaires lorsque les enquêtes vont au-delà d'une simple recherche de suffrages, lorsqu'il s'agit d'explorer et d'approfondir un sujet complexe ou mal connu.

On évoquera ici les avantages et les inconvénients de ce type d'information, mais c'est surtout le problème de leur traitement statistique qui sera abordé dans les lignes qui vont suivre.

Mille réponses à la question :*"Regardez-vous la télévision tous les jours ?*" constitueront un texte où les formes *oui* et *non* seront majoritaires, et où les fréquences relatives de ces formes auront une interprétation simple, très familière en tous cas aux spécialistes des enquêtes par sondage.

Les réponses à une question subsidiaire *"pourquoi ?"* posée après la question précédente auront, elles, un statut intermédiaire. Correspondant à mille stimuli identiques, elles pourront être stéréotypées, mais aussi comporter des contenus ou des formulations originales ou inattendues. Compte tenu des différences de formes à attendre, les simples comptages sont notoirement insuffisants. En revanche, des regroupements de réponses par catégories (âge, sexe, profession, par exemple) permettront de confronter les profils lexicaux moyens de ces catégories.

Que faire de mille ou deux mille réponses libres (saisies sous leur forme littérale, sans transformation ni codage au moment du recueil) fournies par des individus enquêtés à la question : *"A votre avis, quelles sont les raisons qui font hésiter une femme un couple à avoir des enfants ?"* ou encore à la question : "*Quelles sont vos inquiétudes en ce qui concerne les cinq prochaines années ?*"

L'approche la plus courante consiste à se ramener à une situation familière en "fermant" a posteriori la question ouverte: c'est le *post-codage*, technique fruste mais partiellement irremplaçable dont on examinera brièvement plus bas les caractéristiques essentielles ; cette technique contribue malheureusement à maintenir la dangereuse illusion d'une similitude entre questions fermées au moment de l'interview et questions fermées au moment du chiffrement, deux types de questions dont les réponses ne sont en aucune façon comparables. Ces problèmes méthodologiques généraux concernant l'*ouverture* des questions seront abordés au paragraphe 1.4.1 ci-dessous.

Une fois écartée l'idée d'une intervention manuelle (et hautement subjective) avant même la saisie de l'information, il reste à définir les diverses unités statistiques qui nous permettront de codifier et de traiter l'information textuelle. Ce point sera développé dans un cadre plus général au chapitre 2, dévolu aux unités de la statistique textuelle.

1.4.1 Les questions ouvertes : un outil de recherche

Dans la multitude des sondages réalisés dans le domaine du marketing, que celui-ci soit commercial ou électoral, lors de l'établissement des statistiques officielles, les questions ouvertes sont assez rarement utilisées.

La raison de la relative rareté de cet emploi est simple: l'exploitation des réponses recueillies est à la fois difficile et coûteuse.

Pour bien évaluer les avantages de l'information que peuvent apporter les réponses libres, on comparera tout d'abord les deux types de questionnements ouverts et fermés, et on montrera qu'ils produisent des informations de natures différentes, et même difficilement comparables. Puis on examinera les cas dans lesquels l'ouverture est inévitable, soit pour des motifs techniques, soit parce qu'elle tient à la nature même de l'information recherchée.

On dira ensuite quelques mots des techniques de post-codage, qui constituent le traitement le plus usuel des réponses libres.

Enfin, on évoquera les procédures visant à former des agrégats de réponses libres, qui permettent une lecture méthodique de l'information de base.

1.4.2 Questions ouvertes et fermées

Évoquons tout d'abord quelques-uns des nombreux travaux qui concernent la comparaison des questions ouvertes et fermées du point de vue de leur pertinence et de leur efficacité : il s'agit essentiellement de résultats empiriques, qui mettent généralement en évidence d'importantes variations lesquelles touchent la plupart du temps le contenu même du questionnement.

On sait que, dans le questionnaire d'une enquête d'attitude ou d'opinion, le *libellé* d'une question joue un rôle fondamental. En fait, il est très difficile de trouver deux libellés distincts, pour deux questions fermées dont les contenus sont similaires, qui donneront les mêmes résultats en termes de pourcentages des différentes réponses possibles. Certains auteurs refusent mêmes d'interpréter ces pourcentages de réponses comme des suffrages, et ne s'autorisent interpréter leurs variations que par catégories ou dans le temps.

La sensibilité des pourcentages de réponses vis-à-vis des libellés est bien sûr particulièrement forte dans le cas de questions d'attitudes ou d'opinions.

Ainsi, les travaux de Rugg, 1941, ont montré que la réponse *"yes"* à la question : *"Do you think the United States should forbid public speeches against democracy ?"* obtient 21 points (sur 100) de moins que la réponse *"no"* à la question *"Do you think the United States should allow public speeches against democracy ?"*.

Cette absence de symétrie entre les deux formulations, vérifiées sur d'autres thèmes, est d'autant plus forte que le niveau d'instruction de la personne qui répond est faible. Elle rend plus difficiles les études des phénomènes d'acquiescement systématique (cf. par exemple Tabard, 1975). L'équivalence pragmatique de deux questions qui appelleraient respectivement des réponses *oui* et *non* paraît impossible à atteindre.

Ce problème de libellés distincts se pose a fortiori dans le cas de deux questions dont l'une est ouverte et l'autre fermée. L'exemple donné par Schuman et al. (1981) est à cet égard fort démonstratif.

Interrogés à propos du problème le plus important auquel doivent faire face les U.S.A., 16% des Américains mentionnent *crime and violence* (réponses libres regroupées), alors que le même item proposé dans une question fermée donne lieu à 35% des réponses.

Autrement dit, le pourcentage de personnes ayant choisi l'item de réponse fermée fait plus que doubler.

L'explication donnée par les auteurs est la suivante: l'insécurité est souvent considérée comme un phénomène local et non national, ce qui fait que l'item *crime and violence* est assez peu spontanément cité.

Le fait de fermer la question et de faire figurer l'item dans la liste proposée indique que cette réponse est pertinente ou même *possible*, d'où un pourcentage de réponses plus élevé. En fermant la question, on a en fait modifié son libellé et sa signification.

Un exemple du même type peut être trouvé dans l'enquête sur les conditions de vie et aspirations des Français (cf. Lebart, 1987), à propos d'une question conduisant les personnes interrogées à préciser quelles sont, selon elles :*"Les catégories pour lesquelles la collectivité dépense le plus"*.

La question était posée sous forme ouverte en 1983, 1984 et 1987, et sous forme fermée en 1985 et 1986, la question fermée étant construite à partir des principaux items cités les années précédentes. L'item de réponse *les immigrés* a obtenu un pourcentage de 4% en 1983, 5% en 1984 (question ouverte), des pourcentages de 28% en 1985 et 30% en 1986 (question fermée), puis de 8% en 1987 (question à nouveau ouverte). Tous ces pourcentages portent sur des échantillons de 2 000 personnes interrogées chaque année.

Ici encore, le fait de spécifier des items a très probablement modifié l'éventail des réponses considérées comme "permises". Il y a bien eu une croissance du pourcentages de choix de cet item entre 1983 et 1987 (4, 5, puis 8%), mais les deux années pour lesquelles la question a été fermée donnent des pourcentages (28 et 30%) dont l'ordre de grandeur n'est pas comparable avec celui des questions ouvertes correspondantes.

Une liste étendue d'items, proposée dans une question fermée, permet de déformer à l'envi les suffrages, comme le montre par exemple Juan (1986) à propos d'une enquête française par correspondance sur les économies d'énergie: à la question : *"Pensez-vous que l'état s'oriente réellement vers un changement de politique en ce qui concerne les économies d'énergie ?"*, 23% répondent *oui* si la question est laissée ouverte, alors que ce pourcentage atteint 66% dans le cas où elle est fermée de la façon suivante : quatre items correspondent à la réponse positive (oui, très sérieusement – oui, mais prudemment – oui, mais de façon ponctuelle – oui, mais de façon incohérente...), alors qu'un seul item laconique (non) correspond à la réponse négative.

Enfin, toujours à la charge des questions fermées, il faut citer ce que l'on peut appeler l'effet d'intimidation de certains items sur certaines personnes qui n'oseront pas ne pas répondre ou prendre parti.

Dans l'enquête précitée sur les conditions de vie et aspirations des Français, une question a été posée trois années consécutives de 1978 à 1980, dont le libellé était : *"Connaissez-vous ou avez-vous entendu parler de l'amendement Bourrier concernant la sécurité sociale ?"...* pour chaque année d'enquête, environ 4% des personnes interrogées prétendent connaître cet amendement qui n'a en fait jamais existé. Ces personnes ont d'ailleurs un niveau d'instruction supérieur à la moyenne.

Mais les listes d'items peuvent jouer un rôle positif s'il est fait appel à la mémoire : c'est ce qu'établit l'expérience réalisée par Belson et al. (1962) à propos de titres de journaux lus au cours des jours précédents. A l'aide d'un véritable plan d'expérience, comportant notamment un appariement de populations comparables, il a pu être établi que le nombre de citations de quotidiens lus la veille, donné spontanément, représente 71% du nombre donné avec l'aide d'une "check list".

Ce pourcentage de citations spontanées décroît même jusqu'à 25% s'il s'agit d'hebdomadaires et de mensuels, ce qui disqualifie quelque peu le questionnement libre lorsqu'un travail de mémorisation sur longues périodes est en jeu. D'où l'usage répandu des techniques qui, par exemple, présentent des listes de titres de journaux en respectant la typographie exacte et la couleur de ces titres (cf. par exemple : Boeswillwald, 1992).

De façon tout à fait inverse, l'absence d'item de réponse peut aussi jouer un rôle positif. Elle peut en effet créer un climat de confiance et de communication, et donner de meilleurs résultats lorsque l'on aborde certains thèmes : c'est ce que l'on peut retenir des travaux de Sudman et al. (1974) à propos de questions ayant trait aux "menaces", et de Bradburn et al. (1979) à propos de questions concernant la boisson et la sexualité.

1.4.3 Quand utiliser des questions ouvertes ?

Des sociologues comme Lazarsfeld (1944) préconisent l'usage des questions ouvertes principalement dans une phase préparatoire; leur finalité principale est alors la mise au point d'une batterie d'items de réponses pour une question fermée.

Cette utilisation est toujours recommandée, mais exceptionnellement réalisée en raison de son coût : obtenir une liste d'items incluant ceux qui sont peu fréquents nécessite en effet une enquête pilote assez lourde, ce que l'enveloppe financière ou l'urgence des résultats ne permettent que rarement. On a vu plus haut que la "fermeture" d'une question peut réserver bien des surprises. Dans le système d'enquêtes précité auquel seront empruntés les différents exemples qui vont suivre, les questions ouvertes devaient leur présence à des nécessités techniques.

Il existe en effet au moins trois situations-types pour lesquelles l'utilisation d'un questionnement ouvert s'impose :

- *Pour diminuer le temps d'interview*

  Bien que, comme nous l'avons vu, les informations présentes dans les réponses libres et dans les réponses guidées soient de nature différente, les premières sont plus économiques que les secondes en temps d'interview et génèrent moins de fatigue et de tension. Dans le cas de volumineuses listes d'items (que l'on pense à des questions du type *"Quelles sont vos loisirs pendant les fins de semaine?"*, *"Quels sont les avantages de votre logement actuel?"*), la durée du questionnement peut être fortement réduite.

- *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 ?*, *Que pensez-vous de cette voiture ?* (cf. Lebart et al.1991).

- *Pour expliciter et comprendre la réponse à une question fermée*

  C'est la question complémentaire classique: *"Pourquoi?"*. En effet, les explications concernant une réponse déjà donnée par la personne interrogée doivent nécessairement être fournies de façon spontanée. Une batterie d'items risquerait de proposer de nouveaux arguments qui ne pourraient qu'entacher l'authenticité ou la sincérité de l'explication demandée.

On donnera plus bas un exemple de l'intérêt de ces informations complémentaires, intérêt mis en avant par de nombreux sociologues spécialistes des enquêtes par sondage comme Schuman (1966), qui cite les enquêtes internationales pour lesquelles les problèmes de comparabilité et de compréhension de libellés se posent de façon aiguë.

Il importe en effet dans ce cas de savoir si, d'un pays à l'autre, les personnes interrogées ont bien compris la même chose. Il est à vrai dire légitime de se poser la même question en ce qui concerne les variations que l'on peut observer dans les réponses d'une région à une autre, d'une génération à la suivante, d'un milieu socio-culturel à un autre.

1.4.4 Traitement pratique des réponses libres : le post-codage

Cette technique de traitement des réponses libres est classique et encore irremplaçable. Elle consiste à construire une batterie d'items à partir d'un échantillon de réponses (en général 100 à 200), puis à codifier l'ensemble

des réponses de façon à remplacer la question ouverte par une ou plusieurs questions fermées.

Cette procédure n'a qu'un avantage, mais il est de taille : les résultats sont facilement exploitables.

Pour des réponses simples, typées et peu nombreuses (autrement dit des réponses à une question qui aurait pu être fermée...) cette procédure n'a que peu d'inconvénients. Mentionnons cependant quelques défauts de ce type de traitement :

- *Médiation du chiffreur*

  A la médiation de l'enquêteur s'ajoute celle du chiffreur, qui doit arbitrer, interpréter, et qui introduit nécessairement une "équation personnelle" : il doit en effet prendre des décisions parfois difficiles et contestables par le spécialiste.

- *La destruction de la forme*

  L'information est mutilée dans sa forme, et souvent appauvrie dans son contenu: la qualité de l'expression, le registre du vocabulaire, la tonalité générale de l'interview sont autant de matériaux perdus lors d'un post-codage.

- *L'appauvrissement du contenu*

  Lorsque le questionnement permet des réponses composites, complexes, floues, d'une grande diversité, l'information est littéralement laminée par le post-codage; et c'est pourtant dans ce cas que la valeur heuristique des réponses libres est la plus grande.

  Prenons à titre d'exemple une question "pourquoi ?" posée à la suite d'une question assez générale sur les espoirs dans les années à venir : *"Que nous soyons en bonne santé et que nos enfants soient heureux, c'est surtout ce qui nous importe ; le reste, on n'y croit pas beaucoup"*.

  On peut, s'ils sont assez fréquents, retenir les items *santé personnelle* et *bonheur des enfants*, mais comment codifier l'idée d'exclusion des autres items qui est pourtant fondamentale et très caractéristique du contenu même de la réponse en termes de style de vie?

  On pourrait multiplier les exemples de ce type : les réponses doivent plus fréquemment qu'on ne le croit conserver leur forme, leur texture, leur tonalité, pour être vraiment comprises par le chercheur. Les réponses complexes, même formées de juxtapositions d'éléments faciles à codifier, seront de toute façon fort difficiles à exploiter.

- *Les réponses rares sont éliminées a priori*

  Les réponses rares, 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.

  Cette critique de l'opération de post-codage concerne son caractère préliminaire... on commence à post-coder, on étudie ensuite, ce qui fait que de multiples décisions d'affectation, de regroupement, sont prises (souvent par des non-spécialistes) sans analyse de l'ensemble du matériau recueilli, dans sa complexité et sa richesse.

  Ainsi, en réponse à une question concernant *"les espoirs pour les années à venir"*, un item peu fréquent tel que *diminution des injustices* ne sera pas retenu a priori, alors qu'il est intéressant de savoir qu'il est cité avec une fréquence significativement élevée (c'est-à-dire avec une fréquence qu'on ne peut imputer au hasard ou à des fluctuations d'échantillonnage) par les agriculteurs, qui sont cependant très minoritaires dans l'échantillon.

Le type de traitement proposé pour remédier[1] aux faiblesses du post-codage ne portera pas sur une réduction a priori de l'information de base, mais au contraire sur une valorisation de cette information, en utilisant toutes les autres données disponibles sur les répondants (celles-ci sont considérables dans le cas d'une enquête par sondage) et toutes les possibilités de gestion, de tri et de calcul de l'outil informatique.

1.4.5 Les regroupements de réponses

Les réponses libres peuvent être saisies sous leur forme originale sur un support informatique, tout en étant appariées avec les caractéristiques de base et les réponses aux questions fermées des personnes interrogées. Elles peuvent alors subir sans être altérées des opérations de gestion aussi utiles qu'élémentaires : des classements ou des regroupements.

On peut par exemple regrouper les réponses par catégories socio-professionnelles, et donc lire successivement les réponses des agriculteurs exploitants, des ouvriers, des cadres supérieurs, etc. Il existe en effet des catégories ou des combinaisons de catégories pertinentes vis-à-vis de chaque question ouverte: en regroupant les réponses correspondant à chacune de ces

[1] Il s'agit bien de remédier aux faiblesses du post-codage, et non de le remplacer définitivement, car, actuellement, aucune méthode purement statistique ne permet de classifier de façon exhaustive les contenus de réponses libres.

catégories, on obtient des "discours artificiels" d'autant plus typés que ces catégories ont été bien choisies.

La lecture et l'interprétation sont ainsi grandement facilitées dans la mesure où l'on verra apparaître, pour chaque catégorie, des répétitions, des leitmotivs, une plus grande concentration de certains thèmes.
Cependant, cette réorganisation de l'information brute peut être faite de nombreuses façons. Il reste donc à savoir comment regrouper de manière pertinente les réponses, puis comment faciliter ou aider la lecture des regroupements ainsi réalisés.

Comment regrouper les réponses ?

L'importance du mode de regroupement est évidemment considérable : il existe plusieurs stratégies possibles pour trouver une ou plusieurs partitions pertinentes. Ces stratégies sont d'ailleurs complémentaires, et gagnent à être utilisées simultanément.

On peut tout d'abord, prenant en compte les connaissances antérieures sur le thème étudié, utiliser, avec ou sans croisement, les critères jugés les plus discriminants. S'il s'agit par exemple de questions sur l'évolution de la famille, et si l'on soupçonne un effet d'âge ou de génération doublé d'un effet socioculturel, on pourra dans un premier temps utiliser une variable composite croisant l'âge du répondant et son niveau d'éducation.

On peut aussi chercher une partition qui soit la plus universelle possible, compte tenu de la taille de l'échantillon : c'est le principe des situations-types (ou *noyaux factuels*) dont il sera question aux chapitres 4 et 6. Les principales caractéristiques jugées pertinentes (par exemple: âge, catégorie socioprofessionnelle, sexe, niveau d'instruction, région), sont synthétisées par une technique de classification automatique, en une partition unique : cela revient à remplacer un ou plusieurs milliers d'individus par une trentaine ou une cinquantaine de groupes les plus homogènes possibles à l'égard des critères précités.

On peut au contraire faire une typologie directe (sans regroupement préalable) des réponses à partir de leurs profils lexicaux (cela n'a de sens que si les réponses ne se réduisent pas à deux ou trois formes), puis sélectionner les catégories les plus liées à cette typologie, pour procéder ensuite à des regroupements utilisant ces catégories. Ces différentes stratégies seront détaillées et discutées au chapitre 5.

Cette opération élémentaire d'agrégation des réponses facilite grandement la lecture du texte original. Cependant, la lecture de centaines ou de milliers de réponses pour chaque partition des répondants reste une tâche coûteuse, s'il

s'agit d'exploitations courantes et non de recherche approfondie. Il importe donc d'être aidé dans la comparaison des textes obtenus par regroupement. Le lecteur souhaite en effet certainement connaître les mots caractéristiques de telle ou telle catégorie; il souhaite aussi savoir quels groupes s'expriment de façon similaire...

Il va falloir pour cela préparer et disséquer la matière textuelle de façon à définir de nouvelles unités susceptibles d'être reconnues et traitées par des programmes de calcul.

Statut des analyses de réponses aux questions ouvertes

Les analyses de *réponses regroupées* seront en fait assez voisines des analyses de "vrais textes" (littéraires, politiques, historiques), alors que les analyses de *réponses individuelles* seront, elles, voisines des traitements effectués en recherche documentaire. L'originalité de l'approche résultera, en fait, du grand nombre de possibilités de regroupement, et donc du grand nombre de grilles de lectures possibles. Une question ouverte représente un très grand nombre de textes artificiels potentiels, et donc aussi grand nombre de points de vue possibles sur les réponses.