L. Lebart • A. Salem

STATISTIQUE TEXTUELLE

Préface de Christian Baudelot

DUNOD

Ludovic Lebart

Directeur de Recherche au CNRS,
Ecole Nationale Supérieure
des Télécommunications

André Salem

Ingénieur à l'Ecole Normale Supérieure
de Fontenay-Saint-Cloud

Statistique textuelle

Préface de Christian Baudelot

Professeur à l'Ecole Normale Supérieure

Ouvrage publié initialement par Dunod en 1994

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).

L. L., A. S.
Paris, Janvier 1994

Sommaire

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).

> 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.

Chapitre 2

Les unités de la statistique textuelle

La nécessité de comparer des textes sur des bases quantitatives se présente aux chercheurs dans des domaines scientifiques très divers. Dans chaque cas particulier, le recours aux méthodes quantitatives est motivé par des préoccupations différentes et les objectifs poursuivis souvent très distincts (études stylométriques comparées de textes dus à différents auteurs, typologies des réponses d'individus à une même question ouverte, recherche documentaire, etc.).

L'expérience du traitement lexicométrique d'ensembles textuels réunis à partir de problématiques différentes montre, cependant, que, moyennant une adaptation minime, un même ensemble de méthodes trouve des applications pertinentes dans de nombreuses études de caractère textuel. C'est à l'exposé de ces méthodes que seront consacrés les chapitres qui suivent.

2.1 Le choix des unités de décompte.

Segmentation, identification, lemmatisation, désambiguïsation.

La méthode statistique s'appuie sur des mesures et des comptages réalisés à partir des objets que l'on veut comparer. Décompter des unités, les additionner entre elles, cela signifie, d'un certain point de vue, les considérer, au moins le temps d'une expérience, comme des occurrences identiques d'un même type ou d'une *forme* plus générale. Pour soumettre une série d'objets à des comparaisons statistiques il faut donc, dans un premier temps, définir une série de liens systématiques entre des cas particuliers et des catégories plus générales.

Dans la pratique, l'application de ces principes généraux implique que soit définie une *norme* permettant d'isoler de la chaîne textuelle les différentes unités sur lesquelles porteront les dénombrements à venir. L'opération qui

permet de découper le texte en *unités minimales* (c'est à dire en unités que l'on ne décomposera pas plus avant) s'appelle la *segmentation* du texte. A cette phase, qui permet d'émietter le texte en unités distinctes succède une phase de regroupement des unités identiques : la phase d'*identification* des unités textuelles.

Pour un même texte, les différentes normes de dépouillement ne conduisent pas au mêmes décomptes. Pour chaque domaine de recherche particulier, elles ne présentent pas toutes le même degré de pertinence, ni les mêmes avantages (et inconvénients) quant à leur mise en oeuvre pratique.

On comprendra, par exemple, qu'un chercheur qui explore un ensemble d'articles rassemblés au sein d'une base de données ayant trait au domaine de la Chimie, exige de voir regroupés en une même unité le singulier du substantif *acide* et son pluriel *acides* afin de pouvoir interroger lors d'une même requête chacun des textes sur la présence ou l'absence de l'une ou l'autre des formes dans l'ensemble des textes qu'il étudie.

Dans le domaine de l'étude des textes politiques, au contraire, les chercheurs ont constaté que le singulier et le pluriel d'un même substantif renvoient souvent à des notions différentes, parfois en opposition (cf. par exemple l'opposition dans les textes récents de *défense de la liberté* /*défense des libertés* qui renvoie à des courants politiques opposés). On préférera souvent, dans ce second cas, indexer séparément les deux types d'unités qui seront étudiées simultanément.

Au-delà de ces considérations propres à chaque domaine, une fois définie la norme de segmentation, les méthodes de la *statistique textuelle* s'appliquent sans adaptation particulière aux décomptes réalisés à partir de chacune des normes, même si le rappel précis des contours de chaque unité textuelle, reste présent dans l'esprit du chercheur lors des phases suivantes de l'analyse.

Le problème de la définition des unités les plus aptes à servir de base aux analyses quantitatives a longtemps opposé les tenants du découpage en *formes graphiques* (directement prélevables à partir du texte stocké sur support magnétique) et les tenants de la lemmatisation qui jugent nécessaire de mettre en oeuvre des procédures d'identification plus élaborées, rassemblant l'ensemble des flexions d'une même unité de langue.

L'évolution simultanée des centres d'intérêt dans la recherche lexicométrique, des méthodes de la statistique textuelle mais aussi les avancées technologiques dans le domaine de la micro-informatique, permettent aujourd'hui de mieux cerner les spécificités et la complémentarité des deux approches, et d'entrevoir l'ensemble de ces problèmes sous un jour nouveau.

2.1.1 Le texte en machine

Sur toutes les machines permettant de saisir et de stocker du texte, on dispose désormais d'un système de caractères, qui compte en général une centaine d'éléments. Parmi ces caractères, certains correspondent aux lettres de l'alphabet : majuscules, minuscules, lettres diacrisées (munies d'accents, par exemple) propres à la langue que l'on traite ; d'autres servent à coder les chiffres ; d'autres encore permettent de coder des signes comme le pourcentage, le dollar etc. ; enfin, certains caractères servent à coder les divers signes de ponctuation usuels.

Les systèmes actuels individualisent un caractère particulier le "retour-chariot" qui permet de séparer des *paragraphes* (précisément définis comme l'ensemble des caractères situés entre deux retours-chariots).

Dans la pratique, un ensemble de normes typographiques unique se met progressivement en place pour réaliser l'opération que l'on appelait précédemment : l'*encodage* des textes. Ces normes subissent et subiront de plus en plus les effets des progrès technologiques dans le domaine la saisie des textes.

Aujourd'hui, lorsqu'ils ne sont pas directement composés sur clavier d'ordinateur, les textes peuvent être mis à la disposition des chercheurs par simple *scannage* (opération comportant la reconnaissance des différents caractères) de leur support papier. De ce fait, la masse des textes mis à la disposition des chercheurs pour pratiquer des études lexicométriques est désormais sans commune mesure avec celle dont disposait la communauté scientifique il y a tout juste vingt ans.

C'est circonstances renforcent le besoin de disposer d'outils méthodologiques relativement simples permettant d'inventorier, de comparer, d'analyser des informations textuelles toujours plus volumineuses.

2.1.2 Les dépouillements en formes graphiques

Le dépouillement en formes graphiques constitue un moyen particulièrement simple de constituer des unités textuelles à partir d'un corpus de textes. Suivant les objectifs de l'étude on accordera à cette approche un statut qui pourra varier : vérification de la saisie, première approche du vocabulaire, base des comparaisons statistiques à venir.

Pour réaliser une segmentation automatique du texte en occurrences de formes graphiques, il suffit de choisir parmi l'ensemble des caractères un sous-ensemble que l'on désignera sous le nom d'ensemble des *caractères délimiteurs* (les autres caractères contenus dans la police seront de ce seul fait considérés comme caractères non-délimiteurs).

Une suite de caractères non-délimiteurs bornée à ses deux extrémités par des caractères délimiteurs est une *occurrence*. Deux suites identiques de caractères non-délimiteurs constituent deux occurrences d'une même *forme*. L'ensemble des formes d'un texte constitue son *vocabulaire*.

La segmentation ainsi définie permet de considérer le texte comme une suite d'occurrences séparées entre elles par un ou plusieurs caractères délimiteurs. Le nombre total des occurrences contenues dans un texte est sa *taille* ou sa *longueur*.

Cette approche suppose qu'à chacun des caractères du texte correspond un statut et un seul, c'est à dire que le texte à été débarrassé de certaines ambiguïtés de codage (par exemple : points de fin de phrase et points pouvant être présents à l'intérieur de sigles ou d'abréviations : S.N.C.F, etc.)[1].

2.1.3 Les dépouillements lemmatisés

Privilégiant le point de vue lexicographique, on peut, dans certaines situations, considérer qu'il est indispensable, avant tout traitement quantitatif sur un corpus de textes, de soumettre les unités graphiques issues de la segmentation automatique à une *lemmatisation*, c'est-à-dire de se donner des règles d'identification permettant de regrouper dans de mêmes unités les formes graphiques qui correspondent aux différentes flexions d'un même lemme.

Pour lemmatiser le vocabulaire d'un texte écrit en français, on ramène en général :

- les formes verbales à l'infinitif,
- les substantifs au singulier,
- les adjectifs au masculin singulier,
- les formes élidées à la forme sans élision.

Cette manière d'opérer, qui vise à permettre des décomptes sur des unités beaucoup plus soigneusement définies du point de vue de la "langue", peut paraître séduisante au premier abord. Cependant, la pratique de la lemmatisation du vocabulaire d'un corpus rencontre inévitablement des problèmes dont la solution est parfois difficile[2].

---

[1] Notons qu'un premier dépouillement en formes graphiques constitue souvent le moyen le plus sûr pour répertorier les problèmes de ce type qui peuvent subsister au sein d'un corpus de textes que l'on désire étudier. On prêtera également attention aux différences induites par l'utilisation des majuscules (début de phrase ou nom propre) aux traits d'unions, etc.

[2] Cf. par exemple Bolasco (1992) et les travaux de son laboratoire (Universita di Salerno).

En effet, s'il est relativement aisé de reconstituer, "en langue", l'infinitif d'un verbe à partir de sa forme conjuguée, le substantif singulier à partir du pluriel etc., la détermination systématique du lemme de rattachement pour chacune des formes graphiques qui composent un texte suppose, dans de nombreux cas, que soient levées préalablement certaines ambiguïtés.

Certaines de ces ambiguïtés résultent d'une homographie "fortuite" entre deux formes graphiques qui constituent des flexions de lemmes très nettement différenciés (par exemple : *avions* issu du verbe *avoir,* et *avions:* substantif masculin pluriel). Pour d'autres il s'agit de dérivations ayant acquis des acceptions différentes à partir d'une même souche étymologique (cf. les différents sens du mot *voile*, par exemple).

Dans certains cas il faut lever des ambiguïtés touchant à la fonction syntaxique de la forme, ce qui nécessite une analyse grammaticale de la phrase qui la contient.

Certaines de ces ambiguïtés d'ordre sémantique peuvent être levées par simple examen du contexte immédiat. D'autres nécessitent que l'on regarde plusieurs paragraphes, voire l'ensemble du texte[1].

Parfois enfin l'ambiguïté entre plusieurs sens d'une forme, plus ou moins entretenue par l'auteur, ne peut être levée qu'au prix d'un choix arbitraire.

Dans son *Initiation à la statistique linguistique*, Ch. Muller (1968) expose les difficultés liées à l'établissement d'une telle norme de dépouillement

> La norme devrait être acceptable à la fois pour le linguiste, pour ses auxiliaires, et pour le statisticien. Mais leurs exigences sont souvent contradictoires. L'analyse linguistique aboutit à des classements nuancés, qui comportent toujours des zones d'indétermination; la matière sur laquelle elle opère est éminemment continue, et il est rare qu'on puisse y tracer des limites nettes; elle exige la plupart du temps un examen attentif de l'entourage syntagmatique (contexte) et paradigmatique (lexique) avant de trancher. La statistique, dans toutes ses applications, ne va pas sans une certaine simplification des catégories; elle ne pourra entrer en action que quand le continu du langage à été rendu discontinu, ce qui est plus difficile au niveau lexical qu'aux autres niveaux; elle perd de son efficacité quand on multiplie les distinctions ou quand on émiette les catégories; opérant sur des ensembles très vastes elle tolère difficilement une casuistique subtile qui s'arrête à analyser les faits isolés.

[1] Dans certains analyseurs morpho-syntaxiques automatiques, l'ambiguïté peut être levée à partir d'une modélisation de certaines régularités statistiques (cf. par exemple : Bouchaffra et Rouault, 1992).

Dans la pratique, les tenants de la lemmatisation[1] s'appuient pour effectuer leurs dépouillements sur le découpage "en lemmes" opéré par un dictionnaire choisi au départ. Les décisions lexicographiques prises par ce dictionnaire serviront de références tout au long du dépouillement.

De fait, certaines études scindent en deux unités distinctes les formes contractées qui correspondent à une seule unité graphique du texte (*au -> à +le; aux -> à les*, etc.). D'autres regroupent en une même unité certaines unités graphiques qui correspondent à des locutions et dont la liste varie selon les études.

Pour des ensembles de textes peu volumineux, la pratique de la lemmatisation "manuelle" peut constituer une source de réflexions utiles portant à la fois sur le découpage lexicographique opéré "en langue" par les dictionnaires et sur les emplois particuliers attestés dans les corpus de textes considérés. Cependant, ces appréciations positives doivent être sensiblement nuancées lorsqu'on envisage des dépouillements portant sur des corpus de textes étendus.

L'examen des problèmes liés à la lemmatisation montre qu'il ne peut exister de méthode à la fois fiable et entièrement automatisable permettant de ramener à un lemme chacune des unités issues de la segmentation d'un texte en formes graphiques.

2.1.4 Les dépouillements à visée "sémantique"

On a regroupé sous cette rubrique différentes démarches utilisées dans des domaines n'entretenant parfois que peu de liens entre eux, mais pour lesquels le texte de départ subit un précodage substantiel. Le double but assigné à cette dernière opération est, à la fois d'*informer* le texte stocké en machine sur toute une série de traits sémantiques liés à chacune des unités dont il est composé, tout en utilisant, par ailleurs, des procédures définies de manière relativement formelle.

Les "frames"

La représentation par "frames", proposée par Roger Schank (1975), se fixe pour objet "la représentation du sens de chaque phrase". Cette démarche constitue un exemple extrême d'une méthodologie qui tend à insérer le

[1] On trouvera le point de vue des "lemmatiseurs" dans les différents travaux de Ch. Muller. On doit également à cet auteur une critique, relativement mesurée, du point de vue opposé, que l'on trouvera dans la préface *De la lemmatisation* qu'il a donné à l'ouvrage Lafon (1981). Le point de vue des "non-lemmatiseurs" est exposé dans les travaux du laboratoire de lexicométrie politique St.Cloud. Voir par exemple Geffroy et al. (1974) et la réponse de M. Tournier (1985a) à Ch. Muller.

discours dans des catégories sémantiques définies a priori et de manière définitive. La représentation de Schank est censée prévoir à l'avance, dans ses grandes lignes du moins, toutes les questions qui peuvent se poser à propos d'une unité textuelle dans le cadre de ce qu'il appelle *la vie de tous les jours*. L'exemple qui suit donne une première idée du fonctionnement et du cadre conceptuel d'un tel mécanisme.[1]

La phrase : *Marie pleure* reçoit dans ce type de représentation le codage esquissé sur la figure 2.1

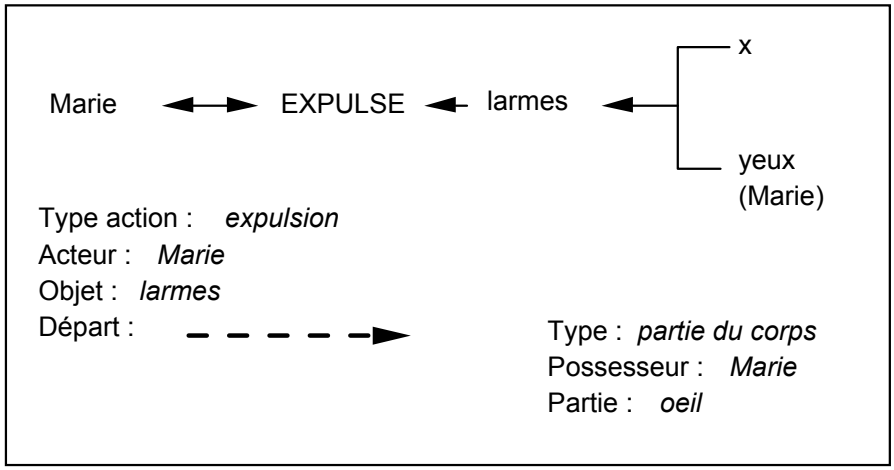


Figure 2.1

Représentation schématisée de la phrase "Marie pleure"

Diverses publications dans le domaine de l'intelligence artificielle témoignent de l'utilité de telles entreprises pour la mise en place de systèmes experts adaptés à un domaine de connaissance particulier.

Cependant les préoccupations et la pratique de l'analyse de textes dans nombreux domaines des sciences humaines nous laisse penser qu'un tel prédécoupage a priori de la réalité conduit à une mutilation considérable du matériau textuel soumis à l'analyse.

Analyse de contenu

Bien qu'elles fassent intervenir un codage beaucoup plus complexe du monde réel, les pratiques d'analyse textuelle utilisées au sein du courant que l'on a coutume d'appeler *analyse de contenu*[2] ne sont pas sans présenter des similitudes avec ce qui vient d'être évoqué plus haut. Dans ce cas encore, déjà évoqué au premier chapitre, les unités textuelles sont rassemblées, avant le recours aux comptages, dans des classes définies a priori, ou après une

[1] Cet exemple est reproduit dans l'ouvrage de Pitrat (1985).
[2] Cf. par exemple Bardin (1977), Weber (1985).

première lecture du texte, afin que les comptages portent directement sur les "contenus".

Ici encore les problèmes naissent de la grande latitude laissée en pratique à l'utilisateur dans la définition des catégories de comptages.

Familles morphologiques

M. Reinert a développé une méthodologie qui se fixe pour objet, de dresser une typologie des *unités de contexte* (séquences de textes de longueurs comparables, qui peuvent souvent coïncider avec les phrases) contenues dans un texte, fondée sur les associations réalisées par ces dernières parmi les unités appartenant à une même famille morphologique[1].

Dans une première phase entièrement automatisée, le logiciel *Alceste*, après avoir écarté de l'analyse toute une série de mots-outil contenus dans un glossaire, regroupe les formes susceptibles d'appartenir à une même famille morphologique, ignorant délibérément tout ce qui peut apparaître comme désinence finale. Lors du dépouillement de la traduction en français d'un roman de L. Durrell, l'algorithme a ainsi rassemblé dans une même classe notée *admir+* les formes graphiques :

admir+ : *admirable, admirait, admirateurs, admiration, admire, admirée, admiriez, admires*

Certaines classes construites par cet algorithme à partir des données de départ, résultent de coïncidences graphiques (dans le cas présent : *acide* et *acier*) et seront éliminées dans un second temps, lors de l'examen détaillé par le chercheur, des classes proposées par le programme.

Cette manière de procéder augmente considérablement le nombre des liens que l'on peut établir entre les différentes unités graphiques d'un même texte, circonstance particulièrement intéressante dans le cas de l'analyse de corpus de petite taille.

2.1.5 Très brève comparaison avec d'autres langues

Les problèmes qui touchent à l'utilisation de la forme graphique dans les études lexicométriques s'éclairent d'un jour particulier lorsqu'on compare le cas du français à celui d'autres langues. Les quelques exemples qui suivent n'ont d'autre prétention que de faire entrevoir au lecteur la variété des problèmes auxquels s'affronte la délimitation automatique de l'unité de décompte. Comme on le verra, ces différences entre langues concernent à la

[1] Cf. Reinert (1986, 1990). On trouvera une brève description du logiciel *Alceste* en annexe.

fois les particularités morpho-syntaxiques qui touchent en profondeur à la structure de chacune des langues concernées et la tradition orthographique qui leur est propre.

L'anglais regroupe, par exemple, par rapport au français, de nombreuses flexions d'une même forme verbale, sous une forme graphique unique : (*speak* versus : *parle, parles, parlons, parlez, parlent, etc.)*. Il en va de même pour les flexions qui correspondent, en français, au genre de l'adjectif.

En espagnol, les pronoms personnels : *nos, se*, pronoms enclitiques, s'accolent à la fin des formes verbales avec lesquelles ils fonctionnent : (*referirse, referirnos* versus *se référer, nous référer*). Ce mécanisme qui est à l'origine de la formation d'un grand nombre de formes graphiques différentes, complique par ailleurs le repérage direct des pronoms personnels en tant que tels.[1]

Les langues à déclinaisons dispersent les différents cas d'un même substantif en un plus ou moins grand nombre de graphies différentes. Ainsi, le russe éparpille-t-il en plusieurs formes graphiques les flexions d'un même substantif selon la fonction grammaticale qu'il remplit dans la phrase, faisant par ailleurs l'économie d'un grand nombre de prépositions.

ministry	les ministres
sovet ministrov	le conseil des ministres
predsedatelh soveta ministrov	le président du conseil des ministres
predsedatelå soveta ministrov	du président du conseil des ministres

Selon l'intérêt que l'on attend du repérage des catégories grammaticales d'un texte, pour l'analyse lexicométrique, on considérera cette circonstance comme particulièrement favorable ou, au contraire, comme une gêne considérable pour le repérage des différentes flexions d'une même unité textuelle.

En sus des déclinaisons, l'allemand présente, par rapport aux langues que nous venons de mentionner, la particularité, de créer des mots composés en agglomérant plusieurs substantifs, ou un radical verbal et un substantif.

Trinkwasser	eau potable,
Zusammengehörigkeitsgefühl	sentiment d'appartenance à

L'ordre d'agglomération des substantifs en allemand étant l'inverse de l'ordre des mots dans les composés français, un tri par la fin permet de repérer de nombreux composés ayant le même radical de base.
Lorsque l'on considère, par exemple, la liste des noms composés que l'on peut prélever dans un texte donné et qui contiennent la forme

[1] Cf. sur ce sujet particulier Romeu (1992).

(*Bildung=formation*), la nécessité de scinder chacune de ces unités composée en fragments plus élémentaires avant de soumettre les textes à l'analyse lexicométrique paraît beaucoup moins évidente.

Bildung	formation, forme, culture
Hochschulbildung	enseignement des grandes écoles
Weiterbildung	formation continue
Preisbildung	formation des prix
Ausbildung	culture
Wortbildung	morphologie (lexicale)
Gesichtsbildung	forme du visage
Herzensbildung	noblesse d'âme

D'autant que l'on trouvera dans ces mêmes textes des mots composés pour lesquels la forme se trouve placée au début du mot composé.

Les expériences que nous avons pu réaliser à partir de textes rédigés dans différentes langues ont montré que, la plupart du temps, les particularités morpho-syntaxiques de chacune des langues concernées ne constituaient pas un obstacle majeur à l'approche des textes par les méthodes de la statistique textuelle. Comme on le verra plus loin, les typologies réalisées à partir des décomptes textuels se révèlent peu sensibles aux variations de l'unité de décompte.

Répétons-le, la forme graphique ne constitue en aucun cas une unité *naturelle* pour le dépouillement des textes ; l'avantage des décomptes en formes graphiques réside avant tout dans la facilité incomparable qu'il y a à les automatiser.

2.2 Segmentation et numérisation d'un texte

La mise en oeuvre des traitements informatisés peut être grandement simplifiée par l'application d'une technique de base appelée : la *numérisation du texte.*

Cette technique consiste à faire abstraction, pendant l'étape des calculs, de l'orthographe des formes décomptées pour ne retenir qu'un numéro d'ordre qui sera associé à toutes les occurrences de cette forme. Ces numéros seront stockés dans un dictionnaire de formes, propre à chaque exploitation. Ce dernier permettra, à l'issue des calculs, de reconstituer le graphisme des formes du texte mises en évidence par les calculs statistiques.

Tableau 2.1
Exemples de réponses à la question *Enfants*

Ind=01 *les difficultés financières et matérielles*
Ind=02 *les problèmes matériels, une certaine angoisse vis à vis de l'avenir*
Ind=03 *la peur du futur, la souffrance, la mort, le manque d'argent*
Ind=04 *l'avenir incertain, les problèmes financiers*
Ind=05 *les difficultés financières*
Ind=06 *les raisons matérielles et l'avenir qui les attend*
Ind=07 *des problèmes financiers*
Ind=08 *l'avenir difficile qui se prepare, la peur du chomage*
Ind=09 *l'insecurite de l'avenir*
Ind=10 *le manque d'argent*
Ind=11 *la guerre éventuelle*
Ind=12 *la charge financière que ca représente, la responsabilité morale aussi*
Ind=13 *la situation économique, quand le couple ou la femme n'est pas psychologiquement prêt pour accueillir un enfant*
Ind=14 *raisons éthiques*
Ind=15 *les problèmes de chômage, les problèmes d'ordre matériel*
Ind=16 *le fait de l'insécurité face au futur, la peur des responsabilités que cela implique*
Ind=17 *l'égoïsme (un chien vaut toujours mieux qu'un enfant), le manque d'argent faussement interprété*
Ind=18 *à Paris les conditions de logement, c'est fatigant, ca amène souvent à interrompre son travail, dans la situation mondiale actuelle c'est problématique*
Ind=19 *l'incertitude de l'avenir, les difficultés sociales*
Ind=20 *les raisons pécuniaires, la crainte de l'avenir à tous les niveaux*
Ind=21 *le chômage, les menaces de guerre*
Ind=22 *le refus de l'énorme responsabilité qui consiste à mettre un enfant au monde*
Ind=23 *les difficultés financières, la peur de s'investir affectivement*
Ind=24 *les raisons financières, l'augmentation du chômage*
Ind=25 *les difficultés financières*
Ind=26 *les problèmes d'argent surtout la peur du lendemain*
Ind=27 *l'argent, le manque d'argent*
Ind=28 *la trop grande responsabilité morale*
Ind=29 *le manque d'argent des raisons de santé parfois*
Ind=30 *vouloir garder une certaine indépendance, faibles ressources, conditions de logement*

La question ouverte a pour libellé : *Quelles sont les raisons qui, selon vous, peuvent faire hésiter une femme ou un couple à avoir un enfant ?*

Elle a été posée en 1981 à 2000 personnes représentant la population des résidents métropolitains de 18 ans et plus. Pour plus d'informations sur l'enquête utilisée (enquête sur les conditions de vie et aspirations des Français), cf. Lebart et Houzel van Effenterre (1980), Babeau et al. (1984), Lebart (1982b, 1987).

La numérisation du texte s'appuie sur des critères formels tels que : l'ordre alphabétique, le nombre des caractères qui composent la forme, etc. Dans ce qui suit nous avons choisi de numéroter les formes du texte en utilisant une combinaison de ces critères.

Les formes sont d'abord triées en fonction de leur fréquence dans l'ensemble du corpus. L'ordre alphabétique départage les formes de même fréquence.

On donnera un exemple de numérisation d'un corpus de réponses à une question ouverte (le corpus ***Enfants***), puis on introduira un corpus **P**, (comme *pédagogique*) de dimensions très réduites, qui servira par la suite à introduire plusieurs notions et définitions.

2.2.1 Numérisation sur le corpus *Enfants*

Dans le corpus constitué par les 2000 réponses à la question ouverte ***Enfants*** dont un extrait figure sur le tableau 2.1, la forme *de* est la plus fréquente avec 915 occurrences.

La forme *les* vient en 5ème position avec 442 occurrences. Les formes: *problèmes* (108 occ.), *l'* (674 occ.) et *avenir* (318 occ.) ont respectivement les numéros : 31, 2 et 8. La numérisation de la séquence :

les problèmes de l'avenir, les problèmes financiers.

que nous avons vue plus haut donnera donc la séquence numérique :

5 31 1 2 8 , 5 31 39

Tableau 2.2

Numérisation de 10 réponses à la question *Enfants* présentées au tableau 2.1

```
1  *   5  42  19  12  56
2  *   5  31 230  28 198   283 138     1  2   8
3  *   3  23  24 166   3  1251   3  1074  4  22  6   13
4  *   2   8  70   5  31    39
5  *   5  42  19
6  *   5  42  19   5  17    56  12     2  8  47  5  716
7  *  11  31  39
8  *   2   8  88  47  81  1147   3    23 24   9
9  *   2  52   1   2   8
10 *   4  22   6  13
```

On trouve au tableau 2.2 la numérisation effectuée à partir des premières réponses présentées au tableau 2.1.

Ainsi, les traitements lexicométriques réalisés à partir des séquences textuelles numérisées vont se trouver simplifiés. Par ailleurs, le volume du stockage du texte en mémoire d'ordinateur se trouve considérablement réduit.

Notons ici que certains logiciels considèrent les ponctuations comme des formes à part entière, d'autres les ignorent complètement. On leur conférera ici un statut spécial de caractères délimiteurs.

2.2.2 Le corpus P

Pour de simples illustrations des définitions données dans ce chapitre, nous travaillerons sur un "texte" appelé, dans la suite de ce chapitre le "corpus" **P**, où les lettres capitales représentent des formes dont ni le graphisme exact, ni la signification n'ont d'importance pour ce qui suit. Il s'agit d'un "texte" simple, qui fournit cependant quelques exemples de redondances que nous utiliserons plus loin.

Tableau 2.3

Le "corpus" pédagogique P

```
 A  B  C  A  C ;  C  D  E  A  B   C . D   F.
(1  2  3  4  5)  (6  7  8  9  10  11)(12  13)
```

Ce "corpus" compte treize *occurrences,* numérotées de 1 à 13 ; sa taille, que l'on désignera par la lettre *T*, vaut :

$$T = 13.$$

Parmi ces occurrences, il en est qui sont des occurrences de la même forme graphique. Ainsi, les occurrences qui portent les numéros : 1, 4 et 9 sont toutes des occurrences de la forme A.

Au total on dénombre dans ce corpus six formes différentes : A, B, C, D, E, F. Ces formes constituent le vocabulaire du corpus P. Désignant par *V* l'effectif du vocabulaire, nous écrirons dans le cas présent :

$$V = 6.$$

2.3 L'étude quantitative du vocabulaire

La lexicométrie comprend des méthodes qui permettent d'opérer des réorganisations formelles de la séquence textuelle et aussi de procéder à des analyses statistiques portant sur le vocabulaire à partir d'une segmentation.

En amont de ces phases de travail, il est possible de calculer à partir du texte numérisé un certain nombre de paramètres distributionnels qui relèvent en quelque sorte d'une application et d'une adaptation de la statistique descriptive élémentaire aux particularités des recueils textuels.

2.3.1 Fréquences, gamme des fréquences

Revenons à l'exemple du corpus **P**. La forme A possède trois occurrences : elle a une *fréquence* égale à trois, les *adresses* de cette forme dans le corpus sont : 1, 4 et 9. L'ensemble des adresses d'une forme constitue sa *localisation*.

Dans ce corpus la forme C est, avec ses 4 occurrences, la forme la plus fréquente ; la *fréquence maximale* est donc égale à 4. Ce que l'on note habituellement :

$$F_{max} = 4$$

La forme E n'apparaît qu'une seule fois dans le texte ; c'est une forme de fréquence 1 ou encore un *hapax*[1]. Dans ce corpus les hapax sont au nombre de deux : les formes E et F. Désignant par V_1 l'effectif des formes de fréquence 1 on écrira :

$$V_1 = 2$$

On notera de la même manière : V_2, V_3, V_4, … les effectifs qui correspondent respectivement aux fréquences deux, trois et quatre. Dans notre corpus :

$$V_2 = 2 ; \quad V_3 = 1 ; \quad V_4 = 1 .$$

On peut également représenter la *gamme des fréquences* de la manière suivante :

Fréquence	:	1	2	3	. . .	F_{max}
Effectif	:	V_1	V_2	V_3	. . .	$V_{F_{max}}$

Pour le corpus P, la gamme des fréquences s'écrit :

[1] Du grec *hapax legomenon* , "chose dite une seule fois".

Fréquence	:	*1*	*2*	*3*	*4*
Effectif	:	*2*	*2*	*1*	*1*

Les fréquences et les effectifs correspondant à chacune de ces fréquences sont liés entre eux par des relations d'ensemble. Certaines de ces relations sont simples. La somme des effectifs correspondant à chacune des fréquences est égale au nombre des formes contenues dans le corpus, ce que l'on note :

$$\sum_{i=1}^{Fmax} V_i = V$$

D'autre part, la somme des produits (fréquence x effectif) pour toutes les fréquences comprises entre 1 et *Fmax*, bornes incluses, est égale à la longueur du corpus.

$$\sum_{i=1}^{Fmax} V_i \times i = T$$

2.3.2 La loi de Zipf

Si le corpus que l'on dépouille compte quelques milliers d'occurrences, on observe que la gamme des fréquences acquiert des caractéristiques plus délicates à décrire. Ces caractéristiques sont associées au nom de G. K. Zipf.

La première découverte[1] de Zipf fut de constater que toutes les gammes de fréquences obtenues à partir de corpus de textes présentent des caractéristiques communes.

Très grossièrement, on peut dire que si l'on numérote les éléments de la gamme des fréquences préalablement rangés dans l'ordre décroissant, on aperçoit une relation approximative entre le rang et la fréquence. On dit parfois :

"le produit (rang x fréquence) est à peu près constant"

Ainsi dans l'*Ulysses* de Joyce qui compte 260000 occurrences les décomptes opérés dès 1937 par H.L. Handley (cf. Guilbaud, 1980) donnent par exemple :

au rang	10	la fréquence	2653
au rang	100	la fréquence	265
au rang	1000	la fréquence	26
au rang	10000	la fréquence	2

[1] Mandelbrot (1961), dans un exposé reprenant les diverses tentatives de modélisation de la gamme des fréquences, souligne les apports de J.B. Estoup (op. cit.) dans ce domaine.

Le diagramme de Pareto fournit une représentation très synthétique des renseignements contenus dans la gamme des fréquences. Ce diagramme est constitué par un ensemble de points tracés dans un repère cartésien.

Sur l'axe vertical, gradué selon une échelle logarithmique, on porte la fréquence de répétition *F*, qui varie donc de 1 à *Fmax*, la fréquence maximale du corpus. Sur l'axe horizontal, gradué selon la même échelle logarithmique, on porte, pour chacune des valeurs de la fréquence *F* comprises entre 1 et *Fmax*, le nombre *N(F)* des formes répétées au moins *F* fois dans le corpus. La courbe obtenue est donc une courbe cumulée.

De nombreuses expériences faites dans le domaine lexicométrique montrent que, quel que soit le corpus de textes considéré, quelle que soit la norme de dépouillement retenue, les points ainsi tracés s'alignent approximativement le long d'une ligne droite. Le diagramme de Pareto est l'une des formes sous laquelle peut s'exprimer une loi très générale touchant à l'économie du vocabulaire dans un corpus de textes.

L'étude de la forme générale de la courbe ainsi obtenue, des "meilleurs" ajustements possibles à l'aide de courbes définies par un petit nombre de paramètres, a longtemps retenu l'attention des chercheurs. On trouve la trace de préoccupations similaires dans la plupart des ouvrages qui traitent de statistique lexicale[1]. L'enjeu principal de ces débats est la détermination d'un modèle théorique très général de la gamme des fréquences d'un corpus. L'élaboration d'un tel modèle permettrait en effet d'interpréter par la suite, en termes de stylistique, les écarts particuliers constatés empiriquement sur chaque corpus.

Malheureusement la démonstration faite par chaque auteur de l'adéquation du modèle qu'il propose s'appuie, la plupart du temps, sur des dépouillements réalisés selon des normes qui varient d'une étude à l'autre. Par ailleurs, la présentation de ces "lois théoriques", dont la formule analytique est souvent complexe, apporte peu de renseignements sur les raisons profondes qui sont à l'origine de ce phénomène.

La figure 2.2 donnera un exemple de diagrammes de Pareto permettant de comparer des textes traditionnels avec des corpus formés de juxtaposition de réponses à des questions ouvertes.

[1] Cf., par exemple, Mandelbrot (1968), Petruszewycz (1973), Muller (1977), Ménard (1983), Labbé et al. (1988).

2.3.3 Mesures de la richesse du vocabulaire.

Considérons maintenant le "texte" constitué par la réunion des 2 000 réponses à la question ***Enfants*** évoquée au début de ce chapitre (cf. tableau 2.1). Dans ce qui suit on désignera cet ensemble de réponses par le symbole Q_1. L'ensemble Q_1 compte 15 523 occurrences. Le nombre des formes *V* est égal à 1 305.

On appellera Q_2 un ensemble du même type, constitué par la réunion des réponses à une autre question ouverte de ce même questionnaire :

Etes-vous d'accord avec l'idée suivante : La famille est le seul endroit où l'on se sente bien et détendu ? Pouvez-vous dire pourquoi ?

Ce second ensemble compte 22 202 occurrences pour 1651 formes. Ces données numériques n'ont de valeur que si on les compare à d'autres données du même type. On peut alors caractériser chacune des questions de l'enquête par le nombre total des occurrences présentes dans les réponses des enquêtés ou encore par le nombre des formes différentes dans chacun de ces ensembles.

On n'entreprendra pas ici l'analyse comparée de ces comptages pour les différentes questions posées lors de l'enquête précitée. On se bornera à comparer les principales caractéristiques lexicométriques pour Q_1 et Q_2 avec d'autres textes prélevés de manière moins artificielle dans des textes rédigés par un même locuteur.

Dans le tableau 2.4 on trouve ces caractéristiques lexicométriques calculées pour six "textes" provenant d'horizons très divers.

Les deux premiers échantillons correspondent aux questions Q_1 et Q_2 que nous venons de décrire.

Les deux suivants, que l'on désignera par S_1 et S_2, sont des fragments de 20 000 occurrences prélevés dans un corpus de textes socio-politiques[1].

Les deux derniers G1 et G2 sont des prélèvements de 20 000 occurrences consécutives découpés dans un roman français contemporain[2].

---

[1] Ce corpus réuni par J. Lefèvre et M. Tournier au laboratoire de St.Cloud, comprend les résolutions adoptées dans leurs congrès respectifs par les principales centrales syndicales ouvrières françaises au cours de la période 1971-1976. L'étude de ce corpus de textes syndicaux a donné lieu à de nombreuses publications (cf., notamment, Bergounioux et al., 1982.)

[2] Il s'agit du roman de Julien Gracq : *Le Rivage des Syrtes*, Paris , Corti, 1952.

Tableau 2.4

Les principales caractéristiques lexicométriques pour deux question d'un questionnaire socio-économique (Q1, Q2), deux fragments de textes syndicaux (S1, S2) et deux fragments de texte littéraire (G1, G2).

	occ.	formes	hapax	fmax	
Q1	15 523	1 305	648	*de*	915
Q2	22 202	1 651	881	*on*	1 128
S1	20 000	3 061	1 576	*de*	1 180
S2	20 000	3 233	1 642	*de*	1 238
G1	20 000	4 605	3 003	*de*	968
G2	20 000	4 397	2 831	*de*	908

Ces quelques données ne prétendent nullement résumer à elles seules l'éventail des situations possibles. Elles aideront seulement à situer très grossièrement les caractéristiques lexicométriques des textes constitués par la concaténation de réponses à une question ouverte parmi d'autres exemples de comptages pratiqués sur des textes.

Comme on le voit, les "textes" Q1 et Q2 sont beaucoup moins riches en formes que les échantillons de textes syndicaux eux-mêmes beaucoup moins riches que les fragments de texte littéraire.

Ceci est lié, entre autres, au fait que les textes Q1 et Q2 comptent beaucoup moins d'hapax que les quatre autres fragments présentés au tableau 2.5.

Pour caractériser plus précisément la gamme des fréquences du texte Q1 par rapport au texte S1, on utilisera le diagramme de Pareto présenté plus haut.

On voit sur la figure 2.2 (et sur le tableau des fréquences maximales) que les fréquences maximales (i.e. fréquence de la forme la plus fréquente) sont proches pour les deux textes bien que Q1 compte 5 000 occurrences de moins que S1.

Les formes de faible fréquence sont plus nombreuses dans le texte S1. Cet avantage s'estompe à mesure que l'on se rapproche de la fréquence 20.

Entre la fréquence 20 et la fréquence 100, la courbe relative au texte Q1 est au-dessus de la courbe qui correspond au texte S1. Ceci traduit un excédent de formes de fréquence moyenne pour la question Q1. Entre la fréquence 100 et la fréquence maximale, la courbe correspondant au texte S1 reprend

l'avantage, ce qui traduit un excédent relatif des formes dont la fréquence est comprise dans cet intervalle par rapport à la question Q1.

Guide de lecture du Diagramme de Pareto

Sur l'axe vertical, gradué selon une échelle logarithmique, on lit la fréquence de répétition *F*.

Sur l'axe horizontal, gradué de la même manière, on trouve, pour chaque fréquence *F* comprise entre *1* et *Fmax*, le nombre des formes répétées au moins *F* fois dans le corpus.

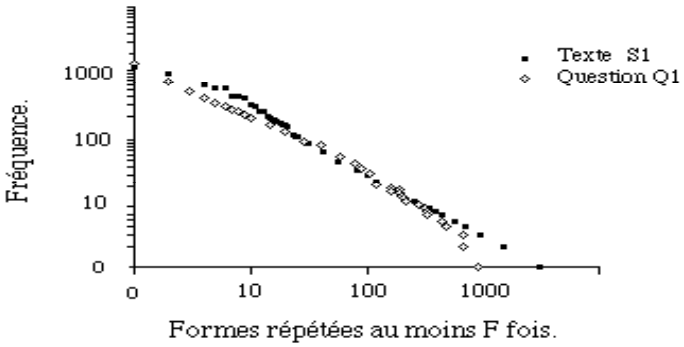


Figure 2.2

Diagramme de Pareto pour les "textes" S_1 (20 000 occ. de texte suivi) et Q_1 (15 000 occ. prises dans des réponses à une question ouverte).

On retiendra donc que le texte Q1 possède un vocabulaire plus réduit et une répétitivité plus grande dans l'intervalle des fréquences moyennes que son homologue S1.

2.4 Documents lexicométriques

Plusieurs documents permettent, à partir d'une simple réorganisation de la séquence textuelle, de porter sur le corpus des textes que l'on étudie un regard différent de celui qui résulte de la simple lecture séquentielle. Il faut noter que l'habitude de produire de tels documents est beaucoup plus ancienne que l'analyse lexicométrique assistée par ordinateur. Cependant pour automatiser entièrement la confection de tels états il est indispensable de formaliser l'ensemble des règles qui président à leur établissement.

2.4.1 Index d'un corpus

Les index, qui constituent par rapport au corpus de départ une réorganisation des formes et des occurrences, permettent de repérer immédiatement, pour chacune des formes, tous les endroits du corpus où sont situées ses occurrences.

Pour retrouver ces occurrences dans le texte de départ, on utilise généralement, de préférence au système des adresses, un système de coordonnées numériques plus étroitement liées à l'édition du texte comme : le tome, la page, la ligne, la position de l'occurrence dans la ligne, etc. Ces renseignements qui permettent de retourner plus facilement au document d'origine sont les *références* associées à chacune des occurrences.

Dans les index, les formes peuvent être classées selon des critères différents. On appelle *index alphabétique* un index dans lequel les formes sont classées selon l'ordre lexicographique (i.e. l'ordre alphabétique couramment utilisé dans les dictionnaires).

On appelle *index hiérarchique* un index dans lequel les formes sont rangées en premier lieu par ordre de fréquence décroissante. Dans ce type d'index, on départage en général les formes de même fréquence d'après l'ordre lexicographique. L'ordre que l'on vient de définir est l'ordre lexicométrique.

Tableau 2.5

L'index hiérarchique du corpus P

fréquence	*forme*	*références*			
4	C	3	5	6	11
3	A	1	4	9	
2	B	2	10		
2	D	7	12		
1	E	8			
1	F	13			

Le tableau 2.5 contient l'index hiérarchique réalisé à partir du corpus **P**. Dans cet index, c'est le numéro d'ordre dans le corpus qui sert de référence pour chacune des occurrences.

2.4.2 Contextes, concordances

Pour chacune des formes du corpus, on peut, en se servant des index, localiser l'ensemble de ses occurrences dans le texte d'origine. Ce travail

effectué, il peut être intéressant d'étudier systématiquement les contextes immédiats dont ces occurrences ont été extraites. Cependant pour une forme pourvue d'une fréquence élevée, cette opération, qui exige un va-et-vient constant entre le texte et l'index, se révèle vite fastidieuse.

Tableau 2.6

Concordance de la forme C dans le corpus P.

contexte avant	*forme-pôle*	*contexte après*	*référence*
A B	**C**	A C ;C D E A B	3
A B C A	**C**	; C D E A B C .D	5
A B C A C;	**C**	D E A B C .D F	6
C D E A B	**C**	. D F.	11

(Rappel du corpus P) **:**

A B **C** A **C** ; **C** D E A B **C** . D F.

Il est possible d'opérer des réorganisations des formes et des occurrences du texte : les occurrences d'une même forme se trouveront rassemblées en un même endroit accompagnées d'un petit fragment de contexte immédiat dont on fixera la longueur en fonction des besoins particuliers.

Ici encore, la manière de procéder a largement précédé l'apparition de l'ordinateur dans le domaine des études textuelles. Le recours à la machine permet simplement d'obtenir en quelques secondes et avec des risques d'erreur considérablement réduits des états dont la confection suffisait parfois à occuper nos ancêtres leur vie durant.[1]

On appellera *forme-pôle* la forme dont on regroupera les contextes. On trouve ci-dessus l'exemple d'une concordance qui regroupe toutes les occurrences relatives à une même forme-pôle, la forme C du corpus P.

Dans l'exemple qui suit, les lignes de la concordance sont triées d'après l'ordre alphabétique de la forme qui suit la forme-pôle (successivement A, C, D, D). Comme plus haut, la référence située à droite est le numéro d'occurrence dans le texte.
Malgré leur volume, les concordances se révèlent à l'usage des instruments très pratiques pour l'étude des textes. En effet, grâce aux réorganisations de la séquence textuelle qu'elles permettent, les concordances fournissent, sur l'emploi d'une forme donnée, une vision plus synthétique que celle qui

[1] On trouvera des éléments sur l'histoire des concordances dans Sekhraoui (1981).

résulte de la lecture séquentielle. En particulier, elles permettent d'étudier plus facilement les rapports qui peuvent exister entre les différents contextes d'une même forme.

Tableau 2.7
Concordance de la forme *enfants* dans le corpus des réponses à la question *Enfant*

```
L 304   c' est pas drole d' avoir des enfants a l' heure actuelle a
L 420   la peur de l' avenir pour les enfants a quelle place peuvent
L2415  onomique* risques d' avoir des enfants anormaux* avenir incer
L2281  s raisons medicales( risque d' enfants anormaux)* manque de c
L1760  est pas la peine de mettre des enfants au monde dans un clima
L2307  c' est bien beau de mettre des enfants au monde mais si c' es
L1879  chomage, on ne peut mettre les enfants au monde sans leur ass
L 429  affection entre le mari et les enfants* aucune garantie d' av
L 710  as* les gens n' aiment pas les enfants* aucune raison n' est
L1461  l' age, l' argent, la peur des enfants aussi, le confort a de
L1295  ou d' argent* ne pas aimer les enfants, avoir peur d' en avoi
L2165  ur, elles sont tributaires des enfants c' est astreignant* le
L 498  and on n' en a pas assez, deux enfants c' est le maximum* les
L1009   il y en a qui n' aime pas les enfants, c' est le travail, la
L 846   desirent pas, un mariage sans enfants, c' est pas un mariage
L2013  ve au chomage avec beaucoup d' enfants c' est plus difficile*
L1903   la vie et le chomage pour les enfants* c' est trop dur d' el
L1647  er une meilleure education aux enfants, ca coute cher actuell
L 285  st difficile de s' occuper des enfants* ca depend de la menta
L1720  roblemes pour faire garder les enfants, ca permet au couple d
L2170  ieres, on choisit d' avoir des enfants, choix delibere de la
L  68   financieres, ne pas aimer les enfants* chomage* difficultes
L1839  probleme financier, avenir des enfants (chomage)* le cout de
L1196  rs, peur de l' avenir pour les enfants* crise actuelle peur d
L 921  fants, difficile d' elever des enfants de nos jours car les e
L 441   l' avenir du chomage pour les enfants* des femmes ne sont pa
L 921   REP* peur du chomage pour ses enfants, difficile d' elever d
L 289  les, materielles, la garde des enfants difficile due au trava
L2072  P* la peur du chomage pour ses enfants* difficulte d' assumer
L 966  eulent plus se priver pour les enfants* economique, pas de de
L 744   couples qui n' aiment pas les enfants* egoisme* de peur d' a
L 612  mage* la femme ne veut plus d' enfants, elle veut sa liberte,
L1650  structures d' accueil pour les enfants en bas age dans le cas
L1467   actuelle n' est fait pour les enfants en bas age les infrast
L 166  nconvenients presentes par des enfants en bas age( trop absor
```

On trouve au tableau 2.7 les premières lignes d'une *concordance* réalisée pour les contextes de la forme : *enfants* dans le "texte" Q1. La forme *enfants* compte 148 occurrences parmi l'ensemble des réponses. C'est dire que la concordance complète pour cette forme compte 148 lignes. Les lignes de contexte sont ici triées par ordre alphabétique croissant en fonction de la forme qui suit la forme-pôle.

Remarquons que cette réorganisation des occurrences du texte initial fait apparaître des liens *(mettre les/des enfants au monde, enfants en bas âge)* qui auraient peut-être échappé à une simple lecture séquentielle du texte.

Remarquons également qu'un tri réalisé en fonction de l'occurrence qui précède la forme-pôle aurait eu pour effet de disperser les occurrences de ces expressions à différents endroits de la concordance et d'en rassembler d'autres qui se trouvent dispersées par le mode de tri adopté ici.

2.4.3 L'accroissement du vocabulaire

Lorsqu'on ajoute des occurrences à un texte, son vocabulaire (le nombre de formes distinctes) a tendance à augmenter. L'étude de l'accroissement du vocabulaire au fil des occurrences d'un même texte constitue l'un des domaines traditionnels de la statistique lexicale.

Pour le corpus P, cet accroissement est décrit de la manière suivante :

	A	B	C	A	C	;	C	D	E	A	B	C	.	D	F.
occurrences	1	2	3	4	5		6	7	8	9	10	11		12	13
formes	1	2	3	3	3		3	4	5	5	5	5		5	6

Pour des corpus de plus grande taille, l'accroissement du vocabulaire subit une double influence.

a) La prise en compte de nouvelles occurrences tend à augmenter le nombre total des formes du corpus (plus il y a d'occurrences, plus les formes distinctes sont nombreuses).

b) Lorsque la taille du corpus augmente, le taux des formes nouvelles apportées par chaque accroissement du nombre des occurrences a tendance à diminuer.

Il s'ensuit que le nombre des formes contenues dans un corpus n'est pas proportionnel à sa taille.

On appellera : *grandeurs lexicométriques de type T* les grandeurs qui croissent *à peu près* en proportion de la longueur du texte. Comme on le verra plus loin les fréquences des formes les plus fréquentes d'un corpus sont des grandeurs de ce type.

On appellera *grandeurs lexicométriques de type V* les variables, telles le nombre des formes, dont l'accroissement a tendance à diminuer avec la longueur du texte.

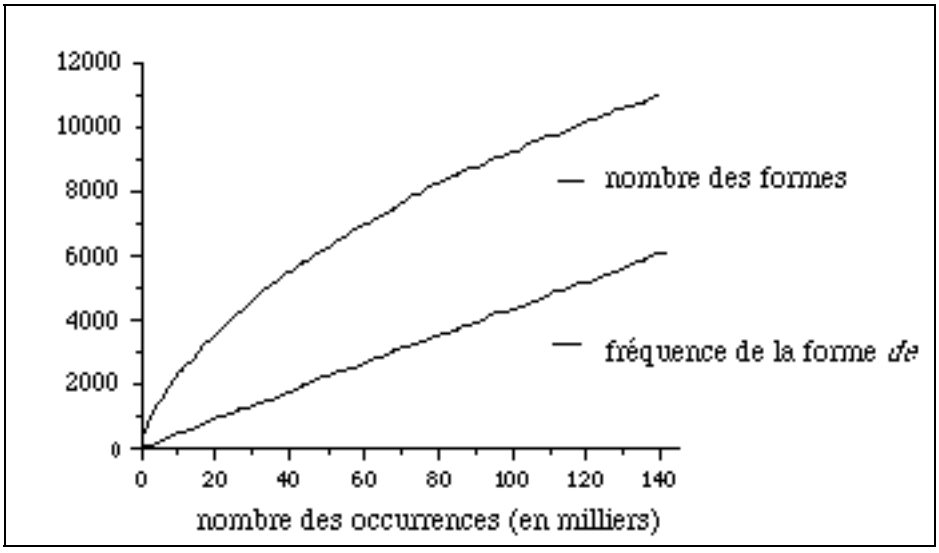


Figure 2.3

Le nombre des formes et le nombre des occurrences de la forme *de* en fonction de la longueur du corpus *Duchesne*.

La figure 2.3 fournit un exemple de tels comptages réalisés sur une grande échelle[1]. Le corpus sur lequel ont été effectués ces comptages compte 11 070 formes pour 141 182 occurrences.

Comme on le voit sur cette figure, le nombre des occurrences de la forme *de*, forme la plus fréquente du corpus, croît linéairement alors que l'accroissement du nombre des formes tend à diminuer au fur et à mesure que le corpus s'allonge.

2.4.4 Partitions du corpus

Le nombre des occurrences dans une partie du corpus est la longueur de la partie. On notera t_j la longueur de la partie numéro j.

Si chaque occurrence du corpus est affectée à une partie et une seule, la somme des longueurs de parties est égale à la longueur du corpus.

$$\sum_{i=1}^{n} t_i = T$$

[1] Ce corpus réuni par J.Guilhaumou (1986) comprend 96 livraisons d'un journal paru pendant la révolution française : *Le Père Duchesne* de Jacques Hébert.

Dans un corpus divisé en *N* parties, correspondant par exemple à *N* périodes ou à *N* locuteurs différents, on parlera des sous-fréquences d'une forme dans chacune des parties.

La suite de *N* nombres constituée par la succession des sous-fréquences prises dans l'ordre des parties est la ventilation des occurrences de cette forme, ou plus simplement la *ventilation* de cette forme, dans les parties du corpus.

La répartition d'une forme est égale au nombre des parties du corpus dans lesquelles elle est présente. Si une forme n'est attestée que dans une seule des parties du corpus elle appartient au *vocabulaire original* de cette partie. Si une forme appartient au vocabulaire de chacune des parties du corpus c'est une forme *commune*.

On notera que toutes ces qualités dépendent très étroitement de la partition opérée sur le corpus de départ. Sur un même corpus, en effet, on sera souvent conduit, en fonction d'objectifs de recherche particuliers, à opérer plusieurs partitions qui serviront de base à des analyses différentes.

Si l'on divise, par exemple, le corpus **P** en trois parties : P1, P2, P3 délimitées par les deux signes de ponctuation situés respectivement après les occurrences 5 et 11, on peut calculer les trois sous-fréquences de la forme A et constituer sa *ventilation* dans les parties du corpus.

Cette ventilation s'écrit :

A : (2 1 0)

La forme A n'est attestée que dans deux des trois parties du corpus **P**. La forme F appartient au *vocabulaire original* de la partie P3.

2.4.5 Tableaux lexicaux

Il est commode de ranger les décomptes des occurrences de chacune des *V* formes dans chacune des *N* parties du corpus dans un tableau rectangulaire qui compte *V* lignes et *N* colonnes. Dans ce tableau, les formes sont classées de haut en bas selon l'ordre lexicométrique. On appelle ce tableau : le *Tableau Lexical Entier* (TLE). Les nombres *V* et *N* sont les dimensions du TLE. Des tableaux lexicaux partiels peuvent être construits, par exemple en ne retenant que les formes dont la fréquence est au moins égale à un certain seuil.

La partition en trois parties du corpus P permet de construire un TLE de dimensions (6 x 3). La ligne *longueur des parties* qui figure sous le TLE est la marge horizontale du tableau. On obtient chacun des nombres situés sur cette

ligne en additionnant tous les nombres qui se trouvent dans la colonne située au-dessus de lui. La marge horizontale du tableau indique donc bien la longueur de chacune des parties.

Tableau 2.8

Le Tableau Lexical Entier (TLE) du corpus P muni de la partition en 3 classes et ses marges.

forme	P1	P2	P3	*fréquence des formes*
C	2	2	0	4
A	2	1	0	3
B	1	1	0	2
D	0	1	1	2
E	0	1	0	1
F	0	0	1	1
longueur des parties	5	6	2	13

Sur la droite du TLE la colonne *fréquence des formes* est la marge verticale. On obtient chacun des nombres situés sur cette colonne en additionnant les nombres du TLE situés sur la même ligne que lui.

Les nombres qui forment la marge verticale du tableau, indiquent donc la fréquence de chacune des formes dans le corpus. On trouvera au chapitre 3 (tableau 3.1) et surtout au chapitre 5 (tableau 5.3) des exemples de tableaux lexicaux construits à partir des données d'enquêtes.

2.5 Les segments répétés

Les formes graphiques ne sont évidemment pas des unités de sens : même si l'on écarte celles qui ont des fonctions purement grammaticales, elles apparaissent dans des mots composés, dans des locutions ou expressions qui infléchissent, voire modifient totalement leurs significations. En effet, on observe les récurrences d'unités comme : *sécurité sociale, niveau de vie, etc.* dotées en général d'un sens qui leur est propre et que l'on ne peut déduire à partir du sens des formes qui entrent dans leur composition.

Dans les études textuelles, il sera utile de compléter les résultats obtenus à partir des décomptes de formes graphiques par des comptages portant sur des unités plus larges, composées de plusieurs formes.

Il est alors utile de soumettre ces unités aux mêmes traitements statistiques que les formes graphiques. Mais ces décisions de regroupement de certaines occurrences de formes graphiques ne peuvent être prises a priori sans entacher de subjectivité les règles du dépouillement.

Dans ce paragraphe on introduira des objets textuels de type nouveau : *les segments répétés*[1]. On montrera ensuite au cours des chapitres suivants comment utiliser les produits de cette formalisation dans les différentes méthodes d'analyses statistiques.

2.5.1 Phrases, séquences

La distinction des caractères du texte en caractères délimiteurs et non-délimiteurs (paragraphe 2.1.2) permet de définir une série de descripteurs relatifs aux formes simples. Pour aborder la description des segments, composés de plusieurs formes et répétés dans un corpus de textes, il s'avère utile de préciser le statut de chacun des caractères délimiteurs.

Si l'on se contente, en effet, de recenser toutes les suites de formes identiques, sans tenir compte des signes de ponctuation qui peuvent venir s'intercaler entre ces dernières, on sera amené à considérer comme deux occurrences d'un même segment deux suites de formes telles par exemple : [A, B, C D] et [A B C D] ce qui peut devenir une source de confusion.

Pour éviter d'avoir à pratiquer des décomptes séparés entre les occurrences de segments contenant des signes de ponctuation et celles des segments identiques aux signes de ponctuation près, on se bornera en général à la recherche des seuls segments ne chevauchant pas de ponctuation faible ou forte.

La question de la définition des limites de la phrase, réputée difficile chez les linguistes, ne trouve pas de solution originale dans le domaine lexicométrique. Pour les procédures formelles que l'on élabore, on se contentera, faute de mieux, de donner à certains des signes de ponctuation (le point, le point d'exclamation, le point d'interrogation) le statut de séparateur fort ou séparateur de phrase.

Reprenons l'exemple du corpus **P** présenté plus haut :

```
A  B  C  A  C ;  C  D  E  A  B   C . D   F.
1  2  3  4  5    6  7  8  9  10  11  12  13
```

On voit que le corpus P est composé de deux phrases.

---

[1] Les procédures de repérage des segments répétés sont issues d'un travail réalisé à l'ENS de Fontenay-St.Cloud, cf. Lafon et Salem (1986), Salem (1984, 1986, 1987). Cf. aussi Bécue (1988).

Parmi les caractères délimiteurs nous choisirons également un sous-ensemble correspondant aux ponctuations faibles et fortes (en général : la virgule, le point-virgule, les deux points, le tiret, les guillemets et les parenthèses auxquelles on ajoutera les séparateurs de phrases introduits plus haut) que l'on appellera l'ensemble des délimiteurs de séquence. La suite des occurrences comprises entre deux délimiteurs de séquence est une séquence.

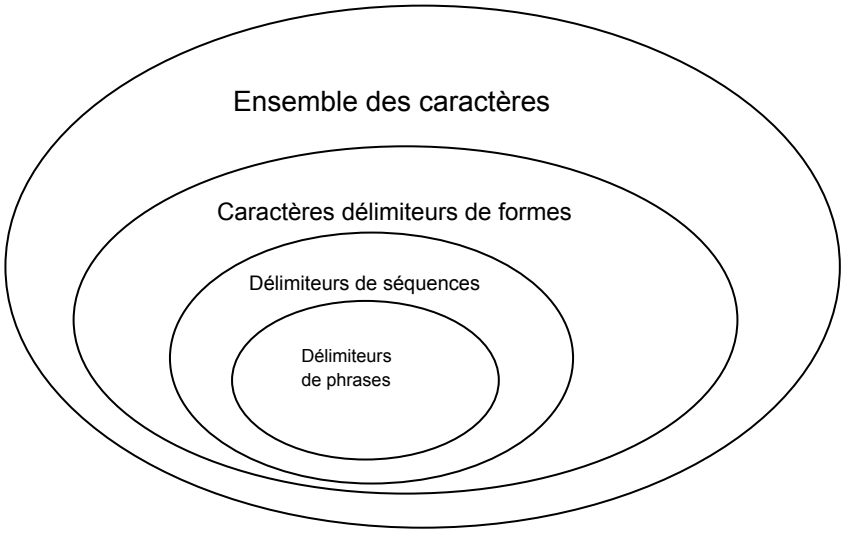


Figure 2.4

Classification des caractères.

Les délimiteurs de formes qui ne sont ni séparateurs de phrase ni délimiteurs de séquence seront, quant à eux, traités comme le caractère que l'on appelle blanc ou espace. On voit que le corpus P est composé de trois séquences. La classification des caractères opérée ci-dessus est résumée par la figure 2.4.

2.5.2 Segments, polyformes

Selon la définition donnée au paragraphe précédent, toutes les suites d'occurrences non séparées par un délimiteur de séquence sont des occurrences de segments, ou de polyformes. Les segments dont la fréquence est supérieure ou égale à 2 dans le corpus sont des segments répétés dans le corpus.

Reprenons le corpus P. Selon la définition que nous venons d'introduire, la première des séquences ABCAC du corpus P se décompose en segments de la manière suivante :

ABCAC, ABCA, BCAC, ABC, BCA, CAC, AB, BC, CA, AC.

Pour l'ensemble du corpus P, seuls les segments : ABC, AB, BC sont des segments répétés.

De la même manière, pour le "corpus" constitué par la réponse à la question Q_1 :

les problèmes financiers, les problèmes matériels

la première séquence contient les segments :

les problèmes financiers, *les problèmes*, *problèmes financiers*

Le segment de longueur 2 : *les problèmes* est le seul segment répété du corpus. On vérifie par ailleurs que toute séquence de longueur L contient obligatoirement: (L-1) segments de longueur 2, (L-2) segments de longueur 3, ... , etc.

Il est désormais difficile de distinguer les deux termes "segments" et "polyformes" qui fonctionnent presque comme des synonymes. On peut noter cependant qu'ils éclairent chacun de manière différente les objets que l'on étudie.

La dénomination de *segment* souligne bien le fait que ces unités sont obtenues en opérant des coupures dans un texte préexistant. Il faut se souvenir cependant que ces coupures sont pratiquées par un algorithme qui ne s'appuie que sur des critères formels tels que la longueur et l'absence de délimiteur de séquence intermédiaire[1].

Ainsi, dans la séquence : *le petit chat est mort*, les définitions retenues nous amènent à considérer, dans un premier temps du moins, le segment constitué par le syntagme nominal : *le petit chat* sur le même plan que le segment obtenu à partir des trois occurrences *petit chat est* qui ne constitue pas une unité *en langue*. Cependant, le repérage des segments les plus répétés dans un corpus de textes, en plus de l'éclairage quantitatif qu'il apporte sur ces problèmes, permet de mettre en évidence des unités qui constituent souvent des syntagmes autonomes.

A l'inverse, la dénomination de *polyforme* implique davantage l'idée qu'une suite de formes, dont on n'a pas encore nécessairement repéré l'existence ni a fortiori toutes les occurrences éventuelles dans le texte que l'on étudie, possède, dans ce texte sinon en langue, une unité de fonctionnement qui lui est propre.

Ainsi, dans les textes socio-politiques, il est intéressant de localiser, en plus des occurrences des formes *sécurité* et *sociale*, les occurrences de la

[1] Les procédures de traitement lexicométrique ne peuvent, dans l'état actuel des choses, tenir compte des frontières de syntagmes (i.e. groupes de mots en séquence formant une unité à l'intérieur de la phrase) que le grammairien repère par une simple lecture.

polyforme *sécurité sociale* qui fonctionne dans ces textes comme une unité qu'il est dommageable de scinder en deux formes isolées.

2.5.3 Quelques définitions et propriétés relatives aux segments

Comme dans le cas des formes, la fréquence d'un segment est le nombre des occurrences de ce segment dans l'ensemble du corpus de textes.

La terminologie employée pour décrire la localisation des formes simples s'applique sans difficulté à la localisation des segments, si l'on convient que l'adresse de chacune des occurrences d'un segment est donnée par celle de la première des occurrences de forme simple qui le composent. Ainsi, on dira sans risque de confusion que dans le corpus **P** le segment AB, de longueur 2 et de fréquence 2, possède deux occurrences dont les adresses sont respectivement : 1 et 9. La manipulation de la notion de segment exige cependant que soient définies et précisées quelques notions ou propriétés.

Eléments d'un segment, sous-segments

Le premier segment de longueur 3 du corpus **P**, le segment ABC, possède deux occurrences dans ce corpus. Les formes A, B et C sont, respectivement, les premier, deuxième et troisième éléments du segment ABC, les segments AB et BC les premier et deuxième sous-segments de longueur 2 de ce même segment.

Expansions, expansions récurrentes, expansions contraintes

Considérons le segment ABC. La forme C constitue pour le segment AB une *expansion droite* de longueur 1. On définit de la même manière l'expansion gauche d'un segment. Le segment ABC constitue pour le segment AB un *voisinage* de longueur 3.

Les deux occurrences du segment AB situées aux places 1 et 9 du texte P sont suivies de la même forme, en l'occurrence la forme C. On dira donc que dans ce corpus, la forme C est une *expansion récurrente* (droite) pour le segment AB. Le segment ABC est un *voisinage récurrent* pour le segment AB.

Dans la mesure où il n'existe aucune autre expansion du segment AB dans ce corpus, la présence d'une occurrence de AB implique obligatoirement la présence d'une occurrence du segment ABC. On dira que C est une *expansion contrainte* (de longueur 1 à droite) pour le segment AB.

Segments contraints, segments libres

De la même manière les sous-segments d'un segment donné se divisent en deux catégories : les *segments contraints* et les *segments libres*. Un sous-segment est contraint (dans un autre segment S de longueur supérieure) si toutes ses occurrences correspondent à des occurrences du segment S.

Si au contraire un segment possède plusieurs expansions distinctes, qui ne sont pas forcément récurrentes, on dira que c'est un segment libre. Pour notre second exemple, *problèmes* possède une expansion récurrente à gauche, *les problèmes,* qui est de longueur deux. Ce segment est lui-même un segment contraint dans le corpus.

Tableau des segments répétés (TSR)

Comme dans le cas des formes simples, on rangera dans ces tableaux les comptages portant sur la ventilation des occurrences d'un même segment dans les différentes parties d'un corpus. On verra plus loin comment ces tableaux sont utilisés dans différentes analyses statistiques sur les segments répétés.

On convient de classer les segments répétés d'abord par ordre de longueur décroissante (en se bornant en général aux segments de longueur inférieure ou égale à sept) puis, à l'intérieur d'une même classe de longueur, de les classer par ordre de fréquence décroissante ; enfin, si nécessaire, les segments seront départagés par l'ordre lexicographique. Les segments contraints dans des segments plus longs, qui apportent une information redondante, seront écartés de ce type de tableau.

Pour le corpus P le tableau des segments répétés peut donc s'écrire de la manière suivante :

segment \ *partie*	P1	P2	P3
ABC	1	1	0

En effet, tous les autres segments du corpus P sont écartés de ce TSR soit parce qu'ils sont des segments hapax, soit parce qu'ils sont contraints dans le segment ABC.

2.6 Les inventaires de segments répétés (ISR)

Une fois repérés par les procédures informatiques, les segments répétés doivent être répertoriés dans des "inventaires" destinés à être consultés.

Pour classer l'ensemble des segments répétés dans un texte, il est nécessaire de définir un ordre de classement qui permettra de retrouver sans trop de difficultés chacun des segments sans avoir pour cela à parcourir des listes trop volumineuses.
L'expérience montre cependant, que même pour un type de texte défini à l'avance, aucun ordre "canonique" d'édition fixé a priori ne peut prétendre satisfaire l'ensemble des utilisateurs. En effet, comme dans le cas de l'édition

d'une concordance et, dans une moindre mesure, de la confection d'index hiérarchiques de formes simples, le mode de classement des unités rassemblées dans les ISR doit satisfaire à la fois à deux types d'exigences.

Tout d'abord, pour permettre une consultation aisée de ces listes, l'ordre qui préside à leur établissement doit s'appuyer sur des caractères formels ne prêtant à aucune ambiguïté (afin de permettre un traitement en machine), faciles à décrire et à mémoriser, pas trop nombreux et, bien évidemment, productifs du point de vue du classement. Pour ces raisons, on retiendra habituellement des critères tels que : la fréquence, l'ordre alphabétique, la longueur, l'ordre d'apparition dans le texte, etc.

Mais le mode de classement des segments a par ailleurs, une autre fonction dont l'importance est loin d'être négligeable. En effet, en combinant et en hiérarchisant, par exemple, lors de l'édition d'une liste, deux ou plusieurs de ces critères formels de classement, on obtient parfois un "effet visuel" qui peut attirer l'attention du chercheur.

Cet "effet" renvoie, selon les cas, à l'existence dans le corpus étudié d'une locution usuelle ou bien encore à l'emploi massif d'une association syntagmatique, dont on n'aurait sans doute pas remarqué l'existence si les lignes de l'inventaire avaient été triées autrement.

Pour l'étude de corpus de textes comptant plusieurs centaines de milliers d'occurrences il est inutile d'imprimer systématiquement des inventaires exhaustifs dont le volume rendrait la consultation peu pratique. Pour mener à bien des recherches de ce type, il est préférable d'automatiser directement la quantification et la recherche des "effets" décrits plus haut. La notion de segment répété dans l'ensemble du corpus se révèle être un outil particulièrement adapté à la mise en évidence de ces récurrences.

Pour toutes ces raisons, il est important de bien maîtriser la combinatoire des modalités de sélection et de classement des segments répétés à partir de critères formels. On détaillera ici plusieurs modes de classement et de sélection en commençant par les plus simples.

2.6.1 Inventaire alphabétique des segments répétés

En consultant l'index alphabétique des formes simples présentes dans un corpus de textes, on est souvent conduit à se demander si certaines de ces formes n'apparaissent pas massivement dans des syntagmes qui les contiennent simultanément.

Tableau 2.9. Extrait d'un inventaire alphabétique des segments répétés réalisé pour le corpus des réponses à la question *Enfants*.

F	L		Segment
14	3		*avenir des enfants*
			318 208 148
2	2		*avenir difficile*
			318 28
4	2	—> 2	*avenir du*
	3		18 155
2	3		*avenir du couple*
			318 155 95
13	2	—> 7	*avenir économique*
			318 61
2	3		*avenir est incertain*
			318 190 45
35	2		*avenir incertain*
			318 45
2	2		*avenir leur*
			318 71
			===> *quel avenir leur . . .*
13	2	—> 7	*avenir pour*
			318 161
7	4		*avenir pour les enfants*
			318 161 442 148
3	5		*avenir qui lui est réservé*
			318 76 15 190 3*

Guide de lecture du tableau 2.9

- La première colonne indique la fréquence du segment dans le corpus.
- Dans la seconde, on lit la longueur du segment (nombre de formes qui le composent).
- On trouve éventuellement dans la troisième colonne, après une flèche orientée vers la droite ---> , la fréquence de l'expansion droite (de longueur immédiatement supérieure) la plus fréquente.
- Sous le segment lui-même, on peut lire les fréquences respectives de chacun des éléments qui le composent. Lorsqu'une forme n'apparaît que dans le contexte constitué par le segment donné, sa fréquence est suivie d'un astérisque.
- Les segments sont parfois munis de références numériques permettant de localiser leurs différentes occurrences dans le corpus.
- Lorsqu'un segment possède une expansion contrainte à gauche (cf. plus haut) une double flèche ====> précède la mention de cette expansion gauche. C'est à cette entrée de l'ISR que l'on trouvera les références du segment contraint omises ici pour cause de redondance.

Ainsi par exemple, dans les réponses aux questions ouvertes étudiées aux chapitres précédents, la présence simultanée des formes *coût* et *vie* induit

immédiatement la question de la présence ou non dans le texte étudié du syntagme *coût de la vie* et de sa localisation en tant que tel.
A côté de chacun des segments, on trouve dans l'inventaire des renseignements sur sa fréquence, sa localisation et sur certaines de ses expansions éventuelles.
Si la polyforme que l'on recherche est répétée dans le texte, l'ISR alphabétique permet de répondre à ce genre de question sans trop de difficultés.
On peut voir au tableau 2.9 un extrait de l'ISR alphabétique réalisé à partir des réponses à la question ***Enfants*** qui nous sert d'exemple de référence. Dans ce document, les segments sont triés par ordre lexicographique sur le premier terme; si deux segments ont un terme initial identique, ils sont départagés par l'ordre lexicographique de leur deuxième terme et ainsi de suite[1].

Signalons encore que dans l'ISR alphabétique les segments possédant eux-mêmes des expansions droites de même fréquence sont éliminés. En effet dans l'inventaire:

ABC	6 fois
ABCD	4 fois
ABCDE	4 fois

la mention du segment "intermédiaire" ABCD est redondante du fait de la présence à la ligne suivante du segment de même fréquence ABCDE.

2.6.2 Inventaire hiérarchique des segments répétés

L'inventaire alphabétique des segments répétés se révèle peu adapté au repérage des segments les plus longs ou encore des segments les plus fréquemment utilisés dans un corpus.

Dans l'inventaire hiérarchique, les segments sont triés en premier lieu par ordre de longueur décroissante, les segments de même longueur étant rangés par ordre de fréquence décroissante. Enfin, l'ordre lexicographique départage les segments qui seraient encore en concurrence à ce niveau.
On peut voir au tableau 2.10 un fragment de l'ISR hiérarchique du corpus des réponses à la question ***Enfants*** correspondant aux segments de longueur 5.

[1] Dans la version des programmes utilisée pour cette expérience, nous nous sommes bornés au recensement des segments répétés composés de sept formes au maximum, afin de ne pas alourdir un algorithme déjà complexe. Les séquences répétées plus longues, en général assez peu nombreuses, sont pour l'instant recensées dans une étape ultérieure. L'expérience montre qu'on peut cependant les repérer assez aisément à partir de leurs sept premières composantes.

Les informations qui concernent chacun des segments sont identiques à celles décrites à propos des inventaires alphabétiques[1].

Tableau 2.10

Extrait d'un inventaire hiérarchique des segments répétés réalisé pour le corpus des réponses à la question *Enfants*.

F	L		Segment
10	5		*le coût de la vie* 474 14 915 666 179
9	5		*le travail de la femme* 474 152 915 666 60
8	5	—> 2	*les difficultés de la vie* 442 82 915 666 179
8	5	—> 3	*peur de ne pas pouvoir* 160 915 193 325 41
7	5	—> 2	*le fait de ne pas* 474 25 915 193 325
6	5		*la cherté de la vie* 666 7 915 666 179

Inventaire hiérarchique par partie

Il est utile de disposer, pour chacune des parties du corpus, d'un inventaire hiérarchique des segments répétés ne prenant en compte que les occurrences des segments répétés dans cette seule partie.

Comme dans l'ISR hiérarchique total, les segments sont d'abord triés en fonction de leur longueur. Les segments de même longueur sont ensuite triés d'après leur fréquence dans la seule partie concernée par l'inventaire.

L'ordre alphabétique départage les segments de même longueur qui ont une fréquence égale.

Tableau 2.11

Inventaire distributionnel avant la forme : *enfants*
(Question *Enfants*)

[1] Il faut noter que l'on ne peut, dans le cas des inventaires hiérarchiques, éliminer les informations redondantes selon les critères retenus pour la confection des inventaires alphabétiques. Voisins dans les inventaires alphabétiques, les segments répétés et leurs expansions récurrentes, droites ou gauches, se trouvent dispersés dans les inventaires hiérarchiques, en raison de leurs longueurs respectives différentes.

```
                    les enfants    ---- ---- ---- ---- ----   59
                pour les enfants    ---- ---- ---- ----   22
          avenir pour les enfants    ---- ---- ----    7
         chômage pour les enfants    ---- ---- ----    4
      du chômage pour les enfants    ---- ----    2
         travail pour les enfants    ---- ---- ----    2
                aimer les enfants    ---- ---- ---- ----    7
fait de ne pas aimer les enfants    ---- ---- ----    2
                  pas les enfants    ---- ---- ---- ----    7
           aiment pas les enfants    ---- ---- ----    5
             aime pas les enfants    ---- ---- ----    2
                  que les enfants    ---- ---- ---- ----    4
               élever les enfants    ---- ---- ---- ----    3
         embêter avec les enfants    ---- ---- ---- ----    2
                    des enfants    ---- ---- ---- ---- ----   55
               avenir des enfants    ---- ---- ---- ----   14
                avoir des enfants    ---- ---- ---- ----    9
           pour avoir des enfants    ---- ---- ---- ---2
              occuper des enfants    ---- ---- ---- ----    3
            vis a vis des enfants    ---- ---- ---- ----    3
              éduquer des enfants    ---- ---- ---- ----    2
               élever des enfants    ---- ---- ---- ----    2
                faire des enfants    ---- ---- ---- ----    2
             la garde des enfants    ---- ---- ---- ----    2
           de mettre des enfants    ---- ---- ---- ----    2
                  par des enfants    ---- ---- ---- ----    2
         la situation des enfants    ---- ---- ---- ----    2
                   sans enfants    ---- ---- ---- ---- ----    6
                    ses enfants    ---- ---- ---- ---- ----    6
peur du chômage pour ses enfants    ---- ---- ---- ----    2
                    aux enfants    ---- ---- ---- ---- ----    2
                de leurs enfants    ---- ---- ---- ---- ----    2
```

2.6.3 Inventaires distributionnels des segments répétés

Les inventaires distributionnels réalisés à partir d'une forme rassemblent les expansions récurrentes de cette forme. On trouve au tableau 2.12 un inventaire distributionnel[1] réalisé pour les expansions gauches de la forme *enfants*. L'inventaire s'ouvre sur l'expansion de longueur 2 la plus fréquente *les enfants* qui compte 59 occurrences. Les lignes qui suivent sont constituées par les expansions de longueur 3 de cette même polyforme.

Comme on peut le vérifier, pour chaque polyforme de longueur donnée, les expansions de longueur immédiatement supérieure sont classées par ordre de fréquence décroissante. Les expansions de même longueur n et de même fréquence sont départagées par la fréquence de l'expansion gauche de longueur n+1.

Lorsqu'une polyforme possède une expansion de longueur 1 contrainte à gauche (dans le texte pour lequel a été réalisé l'inventaire) elle apparaît

[1] Cette forme d'inventaire est issue d'un travail en collaboration avec P. Fiala.

précédée de cette expansion. Sur la gauche du tableau on trouve les fréquences de chacune des formes qui composent les polyformes répétées.[1]

Symétriquement, on peut construire un inventaire distributionnel qui classe, selon les mêmes principes, les expansions récurrentes situées à droite d'une forme-pôle. Cet inventaire concerne en général des expansions différentes qui éclairent sous un autre jour l'utilisation de la forme dans le corpus.

2.6.4 Tableau des segments répétés

Le tableau des segments répétés (TSR) permet de juger de la ventilation de chacun des segments dans les différentes parties du corpus. L'ordre des segments dans ce tableau est le même que celui utilisé dans l'inventaire hiérarchique. C'est ce tableau qui sert de base à la plupart des études statistiques faites à partir des segments. Les segments contraints dans des segments plus longs sont écartés de ce tableau.

Tableau 2.12

Ventilation de quelques segments de longueur 4 dans les parties (Question ouverte *Enfants*)

F	L	Diplôme	A	A	A	B	B	B	S	S	S
		Age	-30	-50	+50	-30	-50	+50	-30	-50	+50
33	*4*	*la peur de l*	2	7	6	6	3	1	2	5	1
20	*4*	*je ne sais pas*	1	2	13	1	2	0	1	0	0
15	*4*	*les conditions de vie*	0	1	7	2	0	1	2	0	2
12	*4*	*coût de la vie*	3	2	3	2	0	1	0	0	1
12	*4*	*le manque de travail*	2	4	5	1	0	0	0	0	0
12	*4*	*peur de ne pas*	1	1	3	2	1	0	1	0	3
12	*4*	*travail de la femme*	1	0	3	2	2	1	0	2	1
11	*4*	*difficultés de la vie*	0	3	2	1	1	2	0	1	1
10	*4*	*le coût de la*	2	2	2	2	0	1	0	0	1
9	*4*	*de ne pas pouvoir*	1	0	2	1	1	0	1	1	2
8	*4*	*de plus en plus*	0	5	2	0	0	0	1	0	0

Le tableau 2.12 donne un extrait du TSR du corpus correspondant aux segments répétés de longueur 4. La première colonne donne un numéro d'ordre qui permet de retrouver le segment dans les analyses statistiques, on trouve ensuite le segment répété puis la fréquence du segment dans le corpus. Les dernières colonnes donnent les sous-fréquences du segment dans chacune des parties du corpus.

---

[1] On notera qu'à la différence des concordances, les inventaires distributionnels des expansions récurrentes ne fournissent pas d'information sur les expansions non-répétées. Ainsi, par exemple, l'inventaire présenté au tableau 3.12 ne concerne que 132 des 148 occurrences de la forme *enfants* dans le corpus que nous étudions.

2.7 Recherche de cooccurrences, quasi-segments

Les recherches sur les segments répétés se sont développées à partir des difficultés rencontrées dans le domaine de la recherche des cooccurrences (attirances particulières entre couples de formes au sein d'une unité de contexte donnée). Sans prétendre à l'exhaustivité, car les méthodes utilisées dans ce but varient fortement en fonction des domaines d'application, on mentionnera très brièvement, dans les paragraphes qui suivent, quelques-unes des méthodes élaborées pour répondre à cette préoccupation.

2.7.1 Recherche autour d'une forme-pôle

Pour une forme pôle donnée, plusieurs méthodes permettent de sélectionner un ensemble de formes qui ont tendance à se trouver souvent dans un voisinage de cette forme.

Pour sélectionner ces formes il faut commencer par définir une unité de contexte, ou voisinage, à l'intérieur duquel on considérera que deux formes sont cooccurrentes. Cette unité peut ressembler à la phrase[1] ou encore être constituée par un contexte de longueur fixe (x occurrences avant, et x occurrences après la forme-pôle).

Labbé (1990) propose une méthode particulièrement simple destinée à mettre en évidence ce qu'il appelle l'*univers lexical* d'une forme donnée. Pour chaque forme *forme1* du corpus, l'ensemble des phrases du corpus peut être divisé en deux sous-ensembles : P_1, sous-ensemble de celles qui contiennent *forme1* et P_0, sous-ensemble des unités desquelles *form1* est absente.

Pour chacune des autres formes du corpus, on applique ensuite le test de l'écart-réduit aux sous-fréquences dans les deux ensembles P_0 et P_1 en tenant compte de leurs longueurs respectives. Si les fréquences des formes considérées ne sont pas trop faibles cette méthode permet de sélectionner, pour chaque forme pôle donnée un ensemble de formes qui se trouvent situées de manière privilégiée dans les mêmes phrases.

La *méthode des cooccurrences n°1*, proposée par Lafon et Tournier (1970)[2] différait de la méthode exposée plus haut sur deux points principaux :

---

[1] Si l'on accepte de se satisfaire de critères relativement simples permettant de repérer les limites des phrases dans la plupart des cas, mais cette question est plus difficile qu'elle ne le semble au premier abord.

[2] Cf. Demonet et al. (1975)

a) elle distinguait les positions relatives par rapport à la forme-pôle séparant de ce fait des cooccurrences *avant* et des cooccurrences *après*.

b) elle construisait un coefficient de cooccurrence faisant intervenir à la fois la cofréquence des deux formes et leur distance moyenne mesurée en nombre d'occurrences.

La *méthode des cooccurrences n°3*, Lafon (1981) affecte un indice de probabilité au nombre des rencontres, chaque fois à l'intérieur d'une phrase (séquence de formes entre deux ponctuations fortes), de chaque couple et de chaque paire (couple non-orienté) de formes du texte. Cette méthode permet de sélectionner, à l'aide d'un seuil en probabilité des ensembles de couples et de paires de formes présentant des affinités dans le corpus de textes que l'on étudie. Tournier (1985b) a appliqué ce type de recherche de cooccurrences à la construction de *lexicogrammes* (réseaux orientés de formes cooccurrentes).

D'une manière générale, la recherche d'associations privilégiées est un enjeu important dans les applications relevant de l'industrie de la langue. Ainsi, certaines incertitudes rencontrées lors de la lecture optique de caractères peuvent être levées (au moins en probabilité) par la considération des formes voisines déjà reconnues, si l'on connaît les probabilités d'association. La désambiguïsation lors d'une analyse morpho-syntaxique peut être réalisée dans les mêmes conditions.

Les travaux de Church et Hanks (1990) se situent dans ce cadre général. Ces auteurs proposent d'utiliser comme mesure d'association entre deux formes *x* et *y* l'information mutuelle *I(x, y)*, issue de la théorie de la communication de R. Shannon :

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

où *P(x)* et *P(y)* sont les fréquences des formes *x* et *y* dans un corpus, et *P(x,y)* la fréquence des occurrences voisines de ces deux formes, x précédant y (il n'y a donc pas symétrie vis-à-vis de x et y), le voisinage étant défini par une distance comptée en nombre de formes. Ainsi, pour les textes en anglais, ces auteurs préconisent de considérer comme voisines deux formes séparées par moins de cinq formes.[1]

2.7.2 Recherches de cooccurrences multiples

[1] Cette procédure paraît efficace pour identifier certaines séquences "verbe + préposition" (*phrasal verbs*). Sur un corpus de 44 millions d'occurrence, les auteurs montrent ainsi que *set* est suivi d'abord de *up (I(x,y)* = 7.3), puis de *off* (*I(x,y)* = 6.2), puis de *out, on in,* etc.

On a rangé sous ce chapitre des méthodes qui se fixent pour objet le repérage direct des cooccurrences de deux ou plusieurs formes du texte. Ce dernier type de recherche s'est tout particulièrement développé dans le domaine de la recherche documentaire.

Dans le domaine de l'indexation automatique on utilise des coefficients qui permettent de repérer des cooccurrences de plusieurs termes à la fois Ainsi, par exemple, Chartron (1988) a proposé un coefficient dit *coefficient d'implication réciproque*.

Pour un groupe composé de k formes numérotées 1, 2, 3, ... , k on calcule le coefficient E de la manière suivante :

$$E = \frac{N_{12..k}}{N_1 N_2 \ldots N_k}$$

avec : $N_{12...k}$ = nombre de cooccurrences des k formes dans le corpus
$N_1, \ldots, N_k$ = nombre d'occurrences de chacune des k formes

D'autres chercheurs ont proposé des approches de ce même problème à partir de la notion de représentation arborescente.[1]

2.7.3 Quasi-segments

Les procédures qui sélectionnent les segments répétés dans un corpus ne permettent pas de repérer, dans l'état actuel, des répétitions légèrement altérées, par l'introduction d'un adjectif ou par une modification lexicale mineure touchant l'un des composants. Bécue (1993) a proposé un algorithme qui permet de repérer des *quasi-segments* (répétés). Cet algorithme permet, par exemple, de rassembler en une même unité *(faire, sport)* les séquences *faire du sport* et *faire un peu de sport*. Il s'agit d'une direction de recherche prometteuse, mais les quasi-segments sont encore plus nombreux que les segments, et leur recensement pose des problèmes de sélection et d'édition.

2.8 Incidence d'une lemmatisation sur les comptages

Pour clore ce chapitre, nous avons choisi de présenter une comparaison de comptages, effectués à la fois en termes de formes graphiques et en termes d'unités lemmatisées, sur un corpus relativement important du point de vue de sa taille.

[1] Cf. par exemple : Barthélémy et al. (1987) et Luong X.(1989).

Cette comparaison permettra de mesurer l'incidence effective du travail de lemmatisation sur les décomptes lexicométriques. Nous analyserons plus loin, (Chapitre 7, paragraphe 7.6) les conséquences de cette modification de l'unité de décompte sur les typologies portant sur l'ensemble des parties de ce même corpus.

2.8.1 Le corpus *Discours*

Le corpus que l'on a retenu rassemble les textes de 68 interventions radiotélévisées de F. Mitterrand survenues entre juillet 1981 et mars 1988.

A partir des textes de ces interventions retranscrits sur support magnétique, Dominique Labbé[1] a effectué un travail de lemmatisation selon des normes qu'il décrit dans Labbé (1990b). Un patient travail partiellement assisté par ordinateur, lui a permis de rattacher chacune des formes graphiques du texte d'origine à un lemme.

On trouve au tableau 2.14 un exemple de fichier-résultat obtenu à la suite d'un tel processus. Chaque ligne du fichier-résultat correspond à une occurrence graphique ou à une ponctuation du texte d'origine.

— la première colonne contient la forme graphique attestée dans le texte;
— la seconde le lemme auquel le processus de lemmatisation rattache cette forme graphique;
— la dernière colonne donne la catégorie grammaticale du lemme.

Tableau 2.13. Les principaux codes grammaticaux utilisés pour la lemmatisation du corpus *Discours*

Codes grammaticaux

1 Verbe :
— 11 *forme fléchie*

[1] L'ensemble des données textuelles que nous avons utilisées dans ce paragraphe a été rassemblé par D. Labbé, chercheur au *Centre de Recherche sur le Politique, l'Administration et le Territoire (CERAT -Grenoble)*, pour une étude sur le vocabulaire de F. Mitterrand. Cet ensemble est constitué à la fois par le texte de discours prononcés au cours du premier septennat et un fichier de lemmatisation qui indique pour chacune des occurrences du texte un lemme de rattachement et une catégorisation grammaticale de l'occurrence considérée. D. Labbé nous a rendu accessible l'ensemble des données dont il disposait afin de nous permettre de faire les quelques expériences dont nous rendons compte ici. D. Labbé a publié plusieurs ouvrages portant sur le discours politique. Deux de ces ouvrages, Labbé (1983) et Labbé (1990) portent plus particulièrement sur le discours de F. Mitterrand.

— 12 *forme au participe passé*
— 13 *forme au participe présent*
— 14 *forme à l'infinitif*

2 Substantif :
— 21 *substantif masculin*
— 22 *substantif féminin*
— 23 *substantif homographe masculin*
— 24 *substantif homographe féminin*

3 Adjectif :
5 Pronom
6 Adverbe
7 Déterminant
8 Conjonction
9 Locution

P Ponctuation
— p *ponctuation mineure (interne à la phrase)*
— P *ponctuation majeure (délimitant la phrase)*

Il est alors possible de reconstituer la séquence de formes lemmatisées qui va être soumise aux analyses statistiques.

Tableau 2.14 . Exemple de fichier de lemmatisation

Forme	*Lemme*	*Cat.Gr.*	*Forme*	*Lemme*	*Cat.Gr.*
je	je	5	le	le	7
crois	croire	11	quatorze	quatorze	7
qu'	que	82	juillet	juillet	21
on	on	5	c'	ce	5
ne	ne	6	est	être	11
peut	pouvoir	11	sans	sans	81
que	que	82	aucun	aucun	7
souhaiter	souhaiter	14	doute	doute	21
cela	cela	5	-	-	p
/.	/.	P			

Le caractère $ sera utilisé pour séparer la suite de caractères qui indique le lemme de rattachement de la partie numérique, laquelle indique une catégorie grammaticale.

A partir du texte ainsi transcodé, il va être possible, en utilisant les mêmes algorithmes que ceux que nous avons utilisé pour effectuer les décomptes de

formes graphiques, d'obtenir des comptages portant cette fois sur les lemmes ainsi codés.

Tableau 2.15

Discours - Extrait du discours de F. Mitterrand(14/07/81) avant et après l'opération de lemmatisation

je crois qu'on ne peut que souhaiter cela. le 14 juillet c'est sans aucun doute - et c'est fort important - l'occasion d'une revue, d'un défilé, d'une relation directe entre notre armée et la nation.

---

je$5 croire$11 que$82 on$5 ne$6 pouvoir$11 que$82 souhaiter$14 cela$5. le$7 quatorze$7 juillet$21 ce$5 être$11 sans$81 aucun$7 doute$21 - et$82 ce$5 être$11 fort$6 important$3 - le$7 occasion$22 de$81 un$7 revue$22, de$81 un$7 défilé$21, de$81 un$7 relation$22 direct$3 entre$81 notre$7 armée$22 et$82 le$7 nation$22

2.8.2 Comparaison des principales caractéristiques quantitatives

Le tableau 2.16 permet une première comparaison entre les décomptes effectués sur les lemmes et ceux réalisés à partir des formes graphiques du corpus.

Tableau 2.16

Principales caractéristiques lexicométriques des dépouillements en formes graphiques et en lemmes

	formes	lemmes
nombre des occurrences :	297258	307865
nombre des formes :	13590	9309
nombre des hapax :	5543	3255
fréquence maximale :	11544	29559

Comme on peut le vérifier sur ce tableau, le nombre des occurrences est plus élevé pour le corpus lemmatisé (+10 607 occurrences). Cette circonstance n'étonnera pas outre mesure car deux grands types d'opérations influent de manière inverse sur le nombre des occurrences d'un corpus dans le cas d'une lemmatisation du type de celle qui a été considérée plus haut.

— le repérage de certaines unités composées de plusieurs formes graphiques (*à l'instar, à l'envi, d'abord, d'ailleurs, etc.*) tend à réduire le nombre des occurrences du corpus lemmatisé.

— à l'inverse, l'éclatement en plusieurs unités distinctes de chacune des nombreuses occurrences des formes graphiques contractées (*au = à + le, des = de + les, etc.*.) tend pour sa part à augmenter le nombre des occurrences du fichier lemmatisé par rapport au texte initial.

L'excès important du nombre des *occurrences* constaté pour le fichier lemmatisé par rapport au texte d'origine tend simplement à prouver, comme on pouvait d'ailleurs s'y attendre, que le second des phénomènes évoqués ci-dessus est nettement plus lourd que le premier.

On peut expliquer par des raisons analogues le fait que le nombre total des *formes* du corpus lemmatisé est, quant à lui, nettement inférieur au nombre des formes graphiques différentes dans le corpus non-lemmatisé. Ici encore deux phénomènes jouent en sens contraire lors du passage aux unités lemmatisées.

— le regroupement de plusieurs formes graphiques correspondant aux différentes flexions d'un même lemme (flexions verbales, marques du genre et du nombre pour les adjectifs, etc.) réduit le nombre des formes du corpus lemmatisé.

— la désambiguïsation rendue possible par la détermination de la catégorie grammaticale pour chacune des formes (*le*-pronom vs *le*-article*, etc.*) augmente plutôt ce nombre.

Au plan quantitatif, on le voit, c'est le processus de regroupement des formes graphiques qui l'emporte nettement lors d'une lemmatisation.

Ces conclusions valent également, nous semble-t-il, pour le nombre des hapax du corpus lemmatisé nettement inférieur lui aussi à son homologue mesuré en formes graphiques.

La fréquence maximale du corpus lemmatisé correspond aux occurrences du lemme *le-article* qui regroupe celles des occurrences des formes graphiques : *le, la, les, l'* qui ne correspondent pas à des occurrences des pronoms homographes[1]. Cette fréquence se révèle plus de 2 fois supérieure à son homologue calculée à partir du corpus non-lemmatisé.

Mesure de l'accroissement du vocabulaire

La figure 2.5 rend compte de l'accroissement du vocabulaire, mesuré à l'aide des deux types d'unités. Comme on le voit, l'allure générale de la courbe d'accroissement rappelle bien dans les deux cas les courbes qui ont été

[1] On se souvient que dans la plupart des corpus de textes écrits en français que nous avons dépouillés en décomptant les formes graphiques, la forme la plus fréquente est la forme *de*. Tel est également le cas pour le corpus ***Discours***.

observées pour les corpus dépouillés précédemment. Cependant, le nombre des formes nettement inférieur dans le cas du corpus lemmatisé fait que les deux courbes paraissent assez différentes au premier abord.

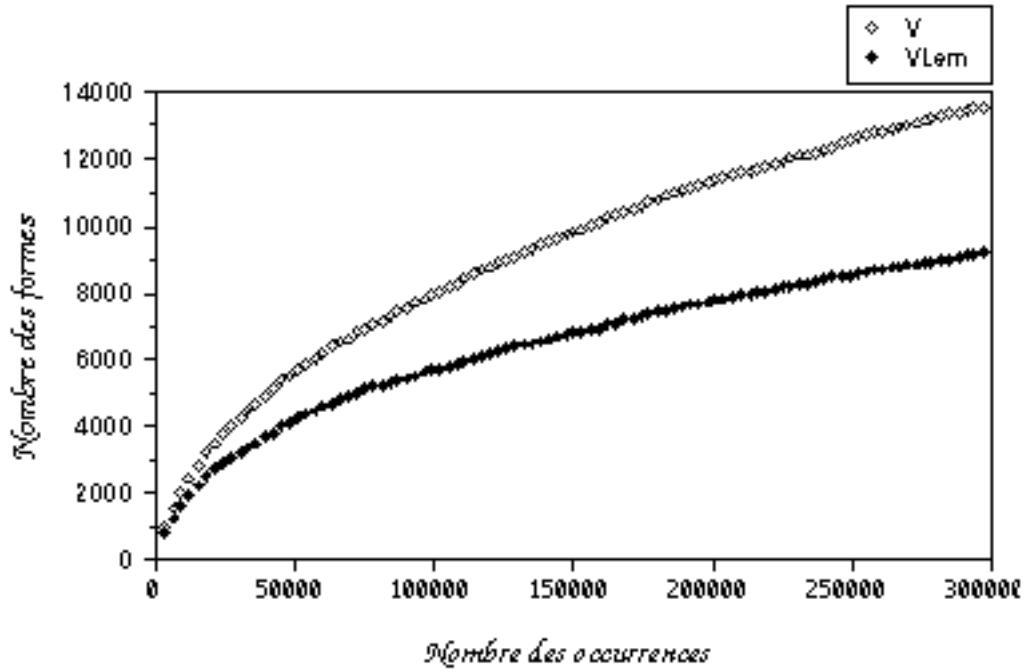


Figure 2.5

L'accroissement du vocabulaire mesuré en formes graphiques et en lemmes.

Ces circonstances incitent à limiter les comparaisons, en ce qui concerne les principales caractéristiques lexicométriques, à des comptages réalisés selon des normes de dépouillement identiques.

On verra cependant au chapitre 7, à propos des séries textuelles chronologiques (paragraphe 7.6) que les typologies réalisées sur une même partition du corpus présentent une grande stabilité que l'on retienne l'une ou l'autre des normes de dépouillement du texte.

Chapitre 3

L'analyse des correspondances des tableaux lexicaux

Le tableau croisé, ou table de contingence, est l'une des formes de structuration des données les plus courantes dans l'analyse des variables qualitatives. En confrontant deux partitions d'une même population ou d'un même échantillon, le tableau croisé permet en effet de travailler sur des *variations par catégorie*, éléments indispensables en vue d'une première interprétation des résultats. Les analyses et descriptions de tableaux croisés sont d'ailleurs à la base du traitement statistique des données d'enquêtes.

Les méthodes d'analyse des données (ou encore : analyses descriptives multidimensionnelles) sont dévolues, pour l'essentiel, à la description de tableaux : tableaux de mesures, de classement, d'incidence, de contingence, ou de présence-absence. Elles ont profondément modifié les méthodes de traitement des données d'enquêtes en mettant à la disposition des utilisateurs des panoramas globaux par thème, des outils pouvant éprouver la cohérence de l'information, des procédures de sélection des tabulations les plus pertinentes.

Les typologies (classification des individus en prenant en compte simultanément plusieurs réponses ou plusieurs caractéristiques de base), les outils de visualisation (plans factoriels) permettent de jeter un regard plus *macroscopique* sur l'information de base.

Les tableaux lexicaux auxquels nous allons appliquer ces techniques sont des tables de contingence particulières ; l'individu statistique donnant lieu à des comptages pour chaque case du tableau sera l'*occurrence* d'une unité textuelle : forme, lemme, segment répété, etc.

Les lignes du tableau correspondront par exemple aux formes graphiques dont la fréquence dans le corpus est supérieure à un seuil donné, les colonnes aux parties du texte : locuteurs, catégories de locuteurs, auteurs, document, etc. Dans le cas général, la case (i,j) contiendra donc le nombre des occurrences de la forme i dans la partie de texte j.
Le présent chapitre rappelle brièvement les principes de base et les modalités pratiques d'utilisation de ces méthodes, avant de mettre en relief leur

contribution propre au renouvellement des techniques de dépouillement d'enquêtes.

3.1 Principes de base des méthodes d'analyse des données

Le principe commun à toutes les méthodes de statistique descriptive multidimensionnelle est schématiquement le suivant :

Chacune des dimensions d'un tableau rectangulaire de données numériques permet de définir des distances (ou des proximités) entre les éléments de l'autre dimension. Ainsi, l'ensemble des colonnes (qui peuvent être des variables, des attributs) permet de définir à l'aide de formules appropriées des distances entre lignes (qui peuvent être des individus, des observations). De la même façon, l'ensemble des lignes permet de calculer des distances entre colonnes.

On obtient ainsi des tableaux de distances, auxquelles sont associées des représentations géométriques complexes décrivant les similitudes existant entre les lignes et entre les colonnes des tableaux rectangulaires à analyser.

Le problème est alors de rendre assimilable et accessible à l'intuition ces représentations, au prix d'une perte d'information de base qui doit rester la plus petite possible. Brièvement, on peut dire qu'il existe deux familles de méthodes qui permettent d'effectuer ces réductions :

- *Les méthodes factorielles*, largement fondées sur l'algèbre linéaire, produisent des représentations graphiques sur lesquelles les proximités géométriques usuelles entre points-lignes et entre points-colonnes traduisent les associations statistiques entre lignes et entre colonnes. C'est à cette famille de méthodes qu'appartient l'analyse des correspondances, présentée dans ce chapitre.

- *Les méthodes de classification* qui opèrent des regroupements en classes (ou en familles de classes hiérarchisées) des lignes ou des colonnes, sont présentées au chapitre suivant.

Ces deux familles de méthodes peuvent d'ailleurs être utilisées avec profit sur les mêmes tableaux de données, et se compléter utilement. Il va de soi que les règles d'interprétation des représentations obtenues par le biais de ces techniques de réduction n'ont pas la simplicité de celles de la statistique descriptive élémentaire.
L'interprétation des histogrammes, les diagrammes en bâtons, les graphiques de séries chronologiques ne nécessitent qu'un apprentissage rudimentaire;

alors que dans le cas de l'analyse des correspondances, par exemple, il sera nécessaire de connaître des règles de lecture des résultats plus contraignantes que ne le laisse croire le caractère souvent suggestif des représentations obtenues. Il faudra au lecteur non seulement un apprentissage, mais aussi une véritable expérience clinique (on veut dire par là une expérience variée et difficilement formalisable : l'analyse de chaque tableau de données présentant quelques aspects particuliers).

Un paradoxe pédagogique

Pour bien assimiler le fonctionnement d'une méthode, il est nécessaire de raisonner sur un exemple de dimension réduite, parfaitement maîtrisable par l'utilisateur... Mais les techniques d'analyse exploratoire multidimensionnelle ne présentent de l'intérêt que parce qu'elles permettent de traiter de très grands ensembles de données. Une image résume cette situation pédagogique paradoxale : il est difficile de démontrer l'efficacité d'un filet de pêche dans un aquarium de salon!

Nous allons néanmoins montrer comment fonctionnent ces méthodes sur des tableaux de dimensions très réduites qui illustreront les principes, mais ne permettront pas d'apprécier pleinement l'efficacité des outils utilisés. Les exemples du chapitre 5 mettront sans doute mieux en évidence leurs possibilités réelles en matière de synthèse et d'exploration.

3.2 L'analyse des correspondances

L'analyse des correspondances est une technique de description des tables de contingence (ou encore : tableaux croisés) et de certains tableaux binaires (dits tableaux de "présence-absence"). Cette description se fait essentiellement sous forme de représentation graphique des associations entre lignes et entre colonnes.

3.2.1 Bref historique

Présentée et étudiée de façon systématique comme une technique souple d'analyse exploratoire de données multidimensionnelles par Benzécri (1973) dans un ouvrage qui reprend les travaux de son laboratoire au cours des dix années précédentes, l'analyse des correspondances s'est trouvée depuis de nombreux précurseurs, et a donné lieu à des travaux dispersés et indépendants les uns des autres. On peut citer sans être exhaustif les noms de Guttman (1941) et de Hayashi (1956) qui parlent tous deux de *méthodes de*

quantification, Nishisato (1980) qui parle de *dual scaling*, Gifi (1990) qui parle d'*homogeneity analysis* pour désigner l'analyse des correspon-dances multiples.[1]

3.2.2 L'analyse des correspondances exposée à partir d'un exemple numérique simple

Le tableau 3.1 que l'on va s'efforcer de décrire avec l'aide de l'analyse des correspondances est un tableau de fréquences croisant, en ligne, 14 formes utilisées dans leurs réponses à une question ouverte (la question ***Enfants*** introduite au paragraphe 2.2 du chapitre 2) par 2 000 personnes, et en colonne, 5 niveaux de diplômes[2] déclarés par chacune de ces personnes.

Tableau 3.1

Croisement (formes-Niveau de Diplôme). Effectifs bruts

	Sans Dipl.	CEP	BEPC	Bacc	Univ	Total
Argent	51	64	32	29	17	193
Avenir	53	90	78	75	22	318
Chômage	71	111	50	40	11	283
Conjoncture	1	7	5	5	4	22
Difficile	7	11	4	3	2	27
Economique	7	13	12	11	11	54
Egoisme	21	37	14	26	9	107
Emploi	12	35	19	6	7	79
Finances	10	7	7	3	1	28
Guerre	4	7	7	6	2	26
Logement	8	22	7	10	5	52
Peur	25	45	38	38	13	159
Santé	18	27	20	19	9	93
Travail	35	61	29	14	12	151
Total	323	537	322	285	125	1592

Rappelons que l'unité statistique retenue ici n'est pas l'individu, mais l'occurrence d'une forme graphique. Les colonnes constituent bien une partition de l'ensemble des personnes interrogées, mais pas les lignes : un répondant peut utiliser dans ses réponses plusieurs formes de la liste retenue

[1] Cf. aussi l'historique de Benzécri (1982), les synthèses de Tenenhaus et Young (1985) et de Escoufier (1985). On trouvera dans Greenacre (1984, 1993), Lebart et al. (1984), Benzécri (1992c) des exposés et des exemples en langue anglaise.

[2] Sans diplôme, CEP (Certificat d'études primaires), BEPC (Brevet d'études du premier cycle ou équivalent), Bac (Baccalauréat ou équivalent), Univ. (Université, grandes écoles ou équivalent).

ici. Les lignes constituent cependant une partition de l'ensemble des occurrences de formes.

On lit ce tableau de la manière suivante : la forme *Argent*, par exemple, a été utilisée 51 fois dans leurs réponses par les personnes appartenant à la catégorie "sans diplôme".

Les totaux de chaque ligne représentent les nombres d'occurrences de chaque forme alors que les totaux de chaque colonne représentent les nombres totaux de formes (de la liste) utilisées par les diverses catégories d'enquêtés.

Le tableau 3.2 représente les profils-lignes, exprimés ici en pourcentages : ils sont obtenus en divisant chaque élément du tableau par la somme de la ligne correspondante : le profil-ligne de la forme *Peur*, par exemple, est obtenu en divisant chaque terme de la ligne par 159.

Tableau 3.2

Les profils-lignes du tableau 3.1

	Sans Dipl.	CEP	BEPC	Bacc	Univ	Total
Argent	26.4	33.2	16.6	15.0	8.8	100.0
Avenir	16.7	28.3	24.5	23.6	6.9	100.0
Chômage	25.1	39.2	17.7	14.1	3.9	100.0
Conjoncture	4.5	31.8	22.7	22.7	18.2	100.0
Difficile	25.9	40.7	14.8	11.1	7.4	100.0
Economique	13.0	24.1	22.2	20.4	20.4	100.0
Egoisme	19.6	34.6	13.1	24.3	8.4	100.0
Emploi	15.2	44.3	24.1	7.6	8.9	100.0
Finances	35.7	25.0	25.0	10.7	3.6	100.0
Guerre	15.4	26.9	26.9	23.1	7.7	100.0
Logement	15.4	42.3	13.5	19.2	9.6	100.0
Peur	15.7	28.3	23.9	23.9	8.2	100.0
Santé	19.4	29.0	21.5	20.4	9.7	100.0
Travail	23.2	40.4	19.2	9.3	7.9	100.0
Total	20.3	33.7	20.2	17.9	7.9	100.0

La comparaison de deux profils-lignes va nous renseigner sur la façon dont les formes correspondantes s'associent aux catégories[1].
Cette comparaison est plus difficile à partir du seul tableau 3.1, car les fréquences des formes sont très variables. Ainsi, il n'apparaît pas immédiatement à la lecture du tableau 3.1 que *Conjoncture* est particulièrement citée par les personnes diplômées d'université, ce que le tableau 3.2 permet de lire facilement

[1] Pour rendre ce tableau plus lisible, les nombres obtenus ont été multipliés par 100.

Lorsque deux formes seront représentées par des points voisins en analyse des correspondances, cela signifiera que les profils-lignes correspondants sont voisins.

Le tableau 3.3 représente les profils-colonnes : ceux-ci sont obtenus de façon tout à fait analogue en divisant les éléments de chaque colonne par leur somme, et en multipliant par 100 le résultat obtenu.

Tableau 3.3

Profils-colonnes du tableau 3.1

	Sans Dipl.	CEP	BEPC	Bacc	Univ	Total
Argent	15.8	11.9	9.9	10.2	13.6	12.1
Avenir	16.4	16.8	24.2	26.3	17.6	20.0
Chômage	22.0	20.7	15.5	14.0	8.8	17.8
Conjoncture	.3	1.3	1.6	1.8	3.2	1.4
Difficile	2.2	2.0	1.2	1.1	1.6	1.7
Economique	2.2	2.4	3.7	3.9	8.8	3.4
Egoisme	6.5	6.9	4.3	9.1	7.2	6.7
Emploi	3.7	6.5	5.9	2.1	5.6	5.0
Finances	3.1	1.3	2.2	1.1	.8	1.8
Guerre	1.2	1.3	2.2	2.1	1.6	1.6
Logement	2.5	4.1	2.2	3.5	4.0	3.3
Peur	7.7	8.4	11.8	13.3	10.4	10.0
Santé	5.6	5.0	6.2	6.7	7.2	5.8
Travail	10.8	11.4	9.0	4.9	9.6	9.5
Total	100.0	100.0	100.0	100.0	100.0	100.0

La comparaison de deux profils-colonnes nous renseigne sur les proximités existant entre les différentes catégories de diplômes vis-à-vis du vocabulaire employé.

En analyse des correspondances, le voisinage de deux points représentant des catégories de diplômes traduira une similitude de profils-colonnes.
L'analyse des correspondances va donc décrire simultanément les similitudes de profils-lignes et de profils-colonnes, et fournir une représentation schématique des informations contenues dans les tableaux 3.2 et 3.3.

La figure 3.1 (plan factoriel engendré par les deux premiers axes de l'analyse des correspondances du tableau 3.1) donne une représentation visuelle des associations entre lignes et colonnes de ce tableau.
Pour une personne qui en maîtrise bien les règles de lecture, c'est un procédé rapide d'assimilation de l'information. On ne peut cependant espérer de la description d'une table dont les dimensions sont aussi modestes des grandes surprises ou des révélations. L'analyse des correspondances a ici une

fonction purement descriptive : la consultation des résultats est simplement plus aisée.

Montrons, à propos de cet exemple, la simplicité des principes et des règles d'interprétation de la méthode.

On notera f_{ij} le terme général de la table des fréquences dont les éléments sont préalablement divisés par l'effectif total k (k =1 592 pour le tableau 3.1). Cette table possède n lignes et p colonnes (ici, n =14 et p =5).

Selon les notations usuelles, $f_{i.}$ désigne la somme des éléments de la ligne i et $f_{.j}$ la somme des éléments de la colonne j de cette table.

Le profil de la ligne i est l'ensemble des p valeurs :

$$\frac{f_{ij}}{f_{i.}}, \quad j = 1,...p.$$

Le profil de la colonne j est l'ensemble des n valeurs :

$$\frac{f_{ij}}{f_{.j}}, \quad i = 1,...n.$$

Les éléments des profils, multipliés par 100, ne sont autres que les pourcentages-lignes et les pourcentages-colonnes qui figurent dans les tableaux 3.2 et 3.3.

Comment lire la figure 3.1 ?

Si deux points-lignes i et i' ont des profils identiques ou voisins, ils seront confondus ou proches sur chacun des axes factoriels. De façon tout à fait analogue, si deux points-colonnes j et j' ont des profils identiques ou voisins, ils seront confondus ou proches.

L'origine des axes correspond aux profils moyens (marges de la table de fréquences).

Le profil-ligne moyen a pour composantes : $f_{.j}$, $j = 1,..p.$
Le profil-colonne moyen a pour composantes : $f_{i.}$, $i = 1,..n.$

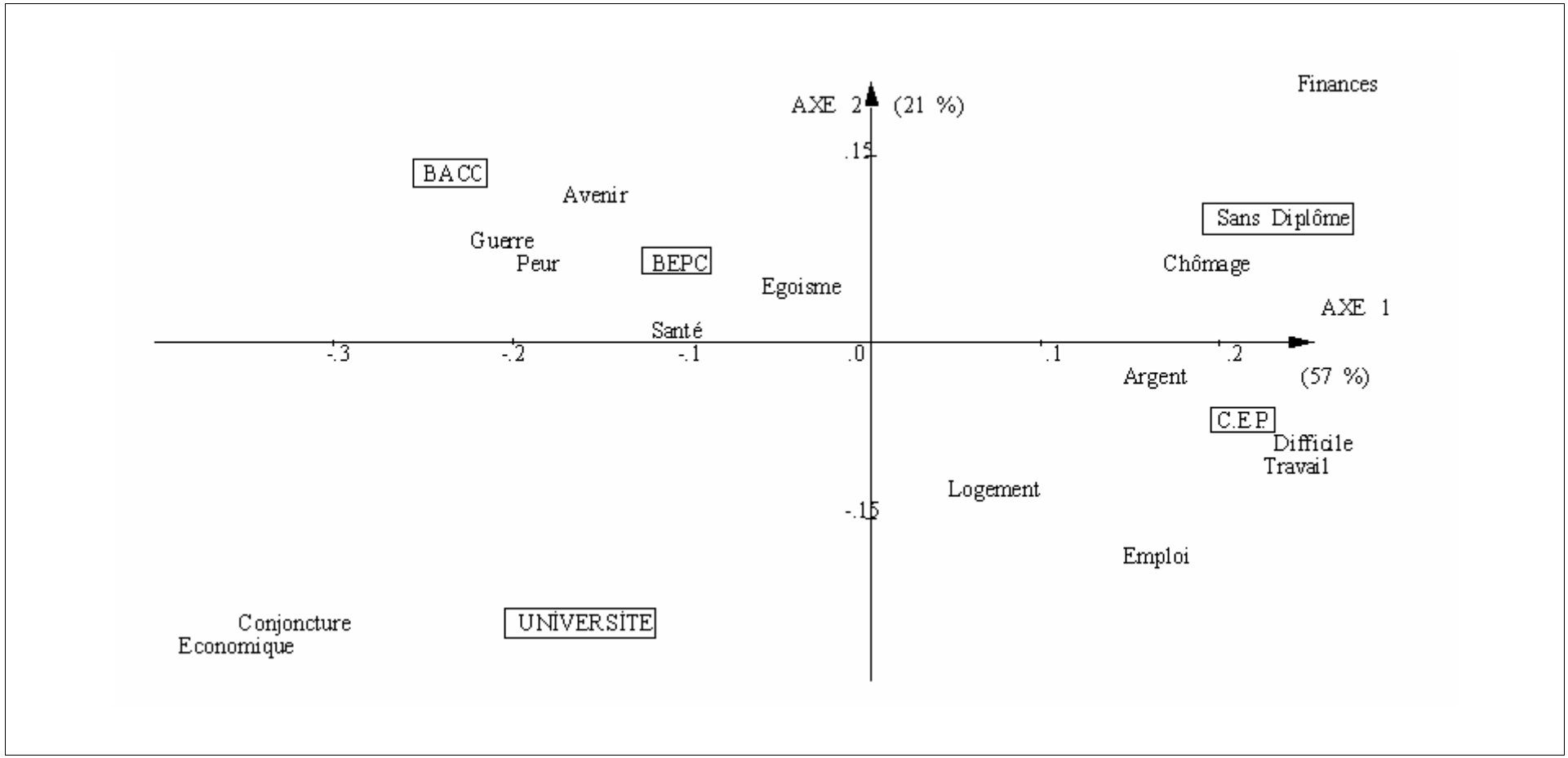


Figure 3.1

Proximités entre formes et entre diplômes (Réponses libres du corpus *Enfants*)
Analyse des correspondances du tableau 3.1

Ainsi un point-colonne comme *BEPC,* assez proche de l'origine, a un profil voisin de celui de la colonne *Total* (marge verticale) du tableau 3.3. De la même façon, un point-ligne comme *Santé* a un profil voisin de la ligne *Total* (marge horizontale) sur la dernière ligne du tableau 3.2.

Les points occupant des positions périphériques auront donc les profils les plus différents du profil moyen, et seront donc les plus typés. Tel est le cas pour des formes comme *Finances* et *Economique*, diamétralement opposées par la représentation, car correspondant respectivement aux formulations populaires et "instruites" de préoccupations sans doute voisines.

Que signifient les proximités ?

Il reste cependant à définir ce que l'on appelle "proche", c'est-à-dire à expliciter la façon dont sont calculées les distances dont des graphiques tels que la figure 3.1 fournissent des approximations planes.

Les profils qui sont des suites de *n* ou *p* nombres, selon qu'il s'agit de lignes ou de colonnes, permettent de définir des points dans des espaces à n ou p dimensions (plus précisément à *n-1* et *p-1* dimensions, car les sommes des composantes de chaque profil valent 1).

Les distances entre points seront donc définies dans ces espaces de dimensions élevées. La phase d'analyse proprement dite consistera à réduire ces dimensions de façon à permettre une représentation visuelle, tout en déformant le moins possible les distances.

La distance entre deux points-lignes i et i' est donnée par la formule:

$$d^2(i,i') = \sum_{j=1}^{p} (1/f_{.j})(f_{ij}/f_{i.} - f_{i'j}/f_{i'.})^2$$

De la même façon, la distance entre deux points-colonnes j et j' est donnée par :

$$d^2(j,j') = \sum_{i=1}^{n} (1/f_{i.})(f_{ij}/f_{.j} - f_{ij'}/f_{.j'})^2$$

Cette distance, appelée distance du chi-2, ressemble beaucoup à la distance euclidienne usuelle (somme des carrés des différences entre composantes des profils) à ceci près qu'une pondération intervient. Cette pondération est l'inverse de la fréquence correspondant à chaque terme : $(1/f_{.j})$ pour chaque terme dans la somme qui définit $d^2(i, i')$, et $(1/f_{i.})$ pour chaque terme dans la somme qui définit $d^2(j, j')$.

La distance du chi-2 possède une propriété remarquable, dite d'*équivalence distributionnelle.* Disons en bref que cette propriété assure une invariance

des distances entre lignes (resp. colonnes) lorsque l'on agrège deux colonnes (resp. lignes) ayant des profils identiques. Cette propriété d'invariance assure une certaine stabilité des résultats vis-à-vis des nomenclatures : ainsi, deux classes de diplômes ayant même profil lexical peuvent être considérées comme distinctes ou confondues sans que cela modifie les représentations obtenues.[1]

Pourquoi une représentation simultanée?

Nous savons maintenant interpréter la proximité entre deux points-lignes ou entre deux points-colonnes, ainsi que leurs positions respectives par rapport à l'origine des axes.

Mais la figure 3.1 nous montre des points-lignes et des points-colonnes simultanément, et donc des proximités qu'il est bien tentant d'interpréter : que le point-ligne *Chômage* soit au voisinage du point-colonne *Sans Diplôme* n'a peut-être rien de surprenant. Mais la proximité entre *Santé* et *BEPC* est moins évidente.

En fait, il n'est pas licite d'interpréter ces proximités croisées entre un point-ligne et un point-colonne, car les deux points ne sont pas dans le même espace au départ. En revanche, il est possible d'interpréter la position d'*un* point-ligne par rapport à l'*ensemble* des points-colonnes ou d'*un* point-colonne par rapport à l'*ensemble* des points-lignes. La principale justification de cette représentation simultanée est donnée par les *relations de transition*, liant les coordonnées d'un point dans un espace (celui des lignes par exemple) à celles de tous les points de l'autre espace (celui des colonnes pour notre exemple).

Si ϕ_i désigne la coordonnée d'un point-ligne i sur l'axe horizontal de la figure 3.1 (premier axe factoriel), et si ψ_j désigne celle d'un point-colonne j sur ce même axe, on a le système de relations remarquablement symétriques :

$$\phi_i = \beta \sum_{j=1}^{p} (f_{ij} / f_{i.}) \psi_j \qquad (1)$$

$$\psi_j = \beta \sum_{i=1}^{n} (f_{ij} / f_{.j}) \phi_i \qquad (2)$$

Le coefficient β est un coefficient positif, supérieur à 1.

Sans ce coefficient, on voit que les points-lignes seraient des barycentres des points-colonnes, et réciproquement. Une telle représentation doublement

[1] Pour une justification plus argumentée de ces pondérations, cf. par exemple Benzécri (1973), et Escofier (1978) qui présente plusieurs distances ayant cette propriété.

barycentrique n'est pas possible, car la prise de barycentre a un effet contractant. Pour que les relations (1) et (2) soient possibles simultanément, il faut donc un coefficient dilatateur β supérieur à 1.

Dans le cas des formules de transition correspondant au premier axe de l'analyse des correspondances, on trouve d'ailleurs le coefficient β le plus proche de 1 qui soit compatible avec ces deux formules.

On peut ainsi présenter directement l'analyse des correspondances comme la recherche des valeurs de Φ_i et de Ψ_j correspondant au plus petit coefficient dilatateur β.

Les formules (1) et (2) sont également valables pour les coordonnées sur l'axe vertical de la figure 3.1, avec bien sûr des valeurs différentes pour les Φ_i, Ψ_j, et pour le coefficient β.

Autre propriété des coordonnées des points-lignes (et des points-colonnes) sur les axes factoriels : elles sont centrées, c'est-à-dire vérifient les relations :

$$\sum_{i=1}^{n} f_i.\Phi_i = 0$$

$$\sum_{j=1}^{p} f_{.j}\Psi_j = 0$$

3.2.3 Validité de la représentation

Le tableau 3.4 donne les valeurs de deux séries de paramètres dont nous n'avons pas encore parlé : les *valeurs propres,* désignées par λ_α (comprises entre 0 et 1 en analyse des correspondances).

Tableau 3.4

Valeurs propres et pourcentages de variance (ou d'inertie)

	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE	
1	.0354	57.04	57.04	**********
2	.0131	21.13	78.17	**********
3	.0073	11.76	89.94	*****
4	.0062	10.06	100.00	****

Ces valeurs propres valent ici $\lambda_1 = 0.035$ pour le premier axe et $\lambda_2 = 0.013$ pour le second ; les *pourcentages de variance,* appelés encore *pourcentages d'inertie* (rapport de chacune des valeurs propres à leur somme globale)

correspondant à ces valeurs propres valent ici 57% et 21% pour ces deux mêmes axes.

Un autre paramètre, la *trace t* (somme de toutes les valeurs propres) vaut ici 0.062 : il y a 4 valeurs propres non nulles. Alors que la trace représente l'inertie totale (ou variance totale) du nuage, les valeurs propres représentent les inerties (ou variances) correspondant à chaque axe.

Propriétés de la trace t

Lors de l'analyse des correspondances d'une table de contingence *(n,p)*, le produit de la trace *t* par l'effectif total *k* n'est autre que le classique chi-2 (χ^2 de Karl Pearson, avec *(n-1)(p-1)* degrés de liberté) utilisé pour tester l'indépendance des lignes et des colonnes de la table.

Ce paramètre est donc calculé par la formule:

$$\chi^2 = k\ t$$

de façon explicite : $$\chi^2 = k \sum_{i=1}^{n} \sum_{j=1}^{p} (f_{ij} - f_{i.} f_{.j})^2 / f_{i.} f_{.j}$$

On a, pour le tableau 3.1 :

$$k\ t = 1592 \times 0.062 = 98.7$$

pour *52* degrés de libertés [*52 = (14 - 1) x (5 - 1)*].

Bien entendu, l'hypothèse d'indépendance est rejetée : l'analyse des correspondances est précisément là pour nous aider à comprendre pourquoi cette hypothèse est rejetée.

On voit sur cet exemple la complémentarité entre statistique inférentielle et analyse exploratoire de donnée : dans le cas présent, celle-ci prend le relais de celle-là dès que la complexité du problème exclut de procéder à des tests d'hypothèses, ou à une modélisation qui s'avérerait inadaptée.

Les valeurs propres

Pour un axe donné, il existe la relation suivante entre le coefficient β des relations de transition et la valeur propre:

$$\beta = 1/\sqrt{\lambda}$$

Une valeur propre proche de 1 assure donc une bonne représentation barycentrique le long de l'axe correspondant. Une telle situation permet d'interpréter les positions relatives des deux ensembles de points-lignes et de

points colonnes[1]. Mais cela n'est pas vraiment le cas pour notre exemple : si les points-diplômes étaient représentés comme de vrais barycentres des points-formes, ils seraient beaucoup moins excentrés... d'où le caractère peut-être trompeur, pour les non-initiés, de la représentation simultanée[2].

Les pourcentages de variance (ou d'inertie)

Ils mesurent l'importance relative de chaque valeur propre dans la trace. Il est assez exceptionnel que, comme c'est le cas ici, le premier plan factoriel "explique" 78% de la variance totale.

En général, ces pourcentages sont au contraire une mesure pessimiste de la part d'information représentée. De nombreux contre-exemples montrent que des pourcentages médiocres correspondent parfois à des représentations qui rendent compte de façon satisfaisante de la structure des données.

Autres aides à l'interprétation

Deux séries de paramètres permettent d'interpréter avec plus de sûreté les résultats, en complétant l'information donnée par les coordonnées des éléments sur les axes factoriels:

a) Les *contributions* (ou contributions absolues), qui décrivent la part prise par un élément (ligne ou colonne) dans la construction d'un axe factoriel.

b) Les *cosinus carrés* (ou contributions relatives) qui mesurent la qualité de la représentation de chaque élément par les axes.

---

[1] Sans pour autant permettre l'interprétation des proximités entre points lignes et points colonnes pris deux à deux.

[2] La racine carrée de la plus grande valeur propre peut également être interprétée comme le plus grand coefficient de corrélation existant entre les lignes et les colonnes de la table (corrélation canonique). Ce coefficient se calcule de la façon suivante : supposons qu'à chacune des 5 catégories de diplôme corresponde une valeur numérique (5 valeurs différentes possibles), et qu'il en soit de même pour les 14 formes utilisées (opération dite de codage ou de quantification). A chaque occurrence, on peut faire correspondre deux valeurs numériques (une pour les lignes, une autre pour les colonnes). On peut donc calculer un coefficient de corrélation entre ces deux ensembles de valeurs. Les 1592 occurrences peuvent être regroupées en 70 groupes (les 70 cases du tableau 3.1) correspondant à des couples de valeurs distinctes. La valeur maximale que peut atteindre ce coefficient de corrélation (qui est associée à un "codage optimal" des mots et des diplômes) est l. Cette recherche de corrélation maximale remonte à Hirschfeld (1935). On peut présenter l'analyse des correspondances à partir de cette approche.

Le tableau 3.5 présente les valeurs de ces différents paramètres pour l'exemple présenté plus haut.

Guide de lecture du tableau 3.5

La colonne P. REL (Poids relatifs) désigne les marges des lignes et des colonnes, qui figuraient déjà dans les tableaux 3.2 et 3.3 précédents.

La colonne DISTO (Distances à l'Origine) contient les carrés des distances à l'origine des axes, c'est à dire les distances de chaque profil au profil moyen (ou: marge). On peut vérifier par exemple sur le tableau 3.2 que les formes *finances* et *conjoncture* ont des profils très différents de la marge.

Les deux premières COORDONNEES sont celles des points, dont la figure 3.1 fournit une représentation approchée.

Les CONTRIBUTIONS (ou encore : contributions absolues), dont les sommes en colonne valent 100, traduisent l'importance des différents éléments dans la construction de chaque axe.

Les COSINUS CARRES (ou encore : contributions relatives), dont les sommes en ligne valent 1, montrent l'importance des différents axes dans l'explication de chaque élément.

3.2.4 Variables actives et illustratives

L'analyse des correspondances permet de trouver des sous-espaces de représentation des proximités entre profils. Mais elle permet aussi de positionner dans ce sous-espace des lignes ou des colonnes *supplémentaires* du tableau de données.

Une fois trouvés les axes factoriels Φ et Ψ et la valeur propre λ (et donc β), les formules de transition (1) et (2) ci-dessus peuvent être appliquées à des lignes ou des colonnes supplémentaires :

La formule (1) permet, à partir de $\Psi, \beta,$ de calculer, pour chaque axe factoriel, la coordonnée ϕ_{i+} d'une ligne supplémentaire $i+$, à partir du profil (f_{i+j}/f_{i+}) (cf. formule 1').

Tableau 3.5

Principaux paramètres de l'analyse des correspondances du tableau 3.1
Les individus actifs sont les lignes du tableau 3.1 alors que les fréquences actives en sont les colonnes

			COORDONNEES				*CONTRIBUTIONS*				*COSINUS CARRES*			
COLONNES ACTIVES	P.REL	DISTO	1	2	3	4	1	2	3	4	1	2	3	4
Sans Diplôme	20.29	.06	.21	.08	-.07	.10	25.1	10.1	14.7	29.9	.68	.10	.08	.14
CEP	33.73	.03	.14	-.06	-.02	-.08	18.3	8.1	1.5	38.4	.64	.11	.01	.24
BEPC	20.23	.04	-.11	.03	.15	.06	6.8	1.3	59.9	11.9	.31	.02	.57	.10
Bacc	17.90	.10	-.27	.12	-.08	-.06	38.0	20.1	14.4	9.6	.76	.15	.06	.03
Univ	7.85	.17	-.23	-.32	.09	.09	11.9	60.5	9.5	10.3	.31	.59	.05	.05

			COORDONNEES				*CONTRIBUTIONS*				*COSINUS CARRES*			
Formes	P.REL	DISTO	1	2	3	4	1	2	3	4	1	2	3	4
Argent	12.12	.03	.12	-.02	-.10	.08	4.5	.4	16.9	13.9	.43	.01	.33	.23
Avenir	19.97	.04	-.18	.10	.05	-.01	17.6	14.6	7.6	.1	.72	.22	.06	.00
Chômage	17.78	.05	.21	.07	.00	-.04	22.6	6.8	.0	4.0	.87	.10	.00	.03
Conjoncture	1.38	.28	.40	-.33	.02	-.07	6.3	11.5	.0	1.0	.58	.40	.00	.02
Difficile	1.70	.07	.25	.07	-.06	.00	3.0	.6	.8	.0	.88	.06	.05	.00
Economique	3.39	.26	.35	-.32	-.08	.15	12.0	26.6	3.3	13.0	.48	.40	.03	.09
Egoisme	6.72	.05	-.06	.03	-.18	-.11	.7	.3	29.5	13.6	.07	.01	.66	.26
Emploi	4.96	.11	.14	-.22	.21	-.06	2.6	17.6	30.8	2.7	.16	.41	.40	.03
Finances	1.76	.20	.24	.21	.04	.32	2.8	5.7	.5	29.0	.28	.21	.01	.51
Guerre	1.63	.06	-.22	.07	.10	.02	2.2	.7	2.1	.2	.75	.09	.15	.01
Logement	3.27	.06	.01	-.13	-.09	-.19	.0	4.1	3.5	19.4	.00	.27	.13	.60
Peur	9.99	.05	-.20	.06	.03	-.01	11.7	2.6	1.5	.1	.90	.07	.02	.00
Santé	5.84	.02	-.11	.00	.02	.05	2.1	.0	.4	2.4	.80	.00	.03	.17
Travail	9.48	.06	.21	-.11	.05	.02	12.0	8.6	3.0	.6	.75	.20	.04	.01

$$\phi_{i+} = \beta \sum_{j=1}^{p} (f_{i+j} / f_{i+.}) \psi_j \qquad (1')$$

$$\psi_{j+} = \beta \sum_{i=1}^{n} (f_{ij+} / f_{.j+}) \phi_i \qquad (2')$$

La formule (2) permet de la même façon, à partir de Φ, β, de calculer la coordonnée ψ_{j+} d'une colonne supplémentaire $j+$ (cf. formule 2') dont le profil est $(f_{ij+} / f_{.j+})$.

On peut ainsi illustrer les plans factoriels par des informations supplémentaires n'ayant pas participé à la construction des plans, ce qui va avoir des conséquences très importantes au niveau de l'interprétation des résultats.

Les éléments ou variables servant à calculer les plans factoriels sont appelés *éléments actifs* ou *variables actives*. Ils doivent former un ensemble homogène pour que les distances entre individus ou observations aient un sens, et donc que les proximités graphiques observées s'interprètent facilement.

Les éléments ou variables projetés a posteriori sur les plans factoriels sont les éléments *supplémentaires* ou *illustratifs*.

Ces éléments illustratifs (lignes ou colonnes) n'ont nul besoin de former un ensemble homogène dans la mesure où le calcul est effectué séparément pour chacun d'eux. Cette dichotomie entre variables actives et variables illustratives est fondamentale d'un point de vue méthodologique.

Exemple

Les quatre formes graphiques d'effectifs faibles dont on peut voir la ventilation sur le tableau 3.6 n'ont pas participé à l'analyse précédente.

Tableau 3.6

Quatre lignes supplémentaires (ou illustratives)

	SansDipl	CEP	BEPC	Bac	Univ	TOTAL
Confort	2	4	3	1	4	14
Mésentente	2	8	2	5	2	19
Monde	1	5	4	6	3	19
Vivre	3	3	1	3	4	14

On désire cependant savoir comment elles se situent par rapport aux autres formes déjà représentées sur le plan factoriel de la figure 3.1.

Leurs profils-lignes peuvent être positionnés dans le même espace à 5 dimensions, et peuvent donc subir la même projection sur le plan de la figure 3.1.

De façon analogue, le tableau 3.7 contient trois colonnes supplémentaires (des catégories d'âge), qui n'ont pas été incluses dans l'ensemble des colonnes actives en raison de l'hétérogénéité des thèmes : l'interprétation des proximités entre lignes, donc entre formes, aurait été plus délicate. Deux formes sont-elles proches à cause de leur répartition par diplômes ou par classes d'âge? Il n'est pas facile de trancher si les distances entre formes sont calculées à partir des deux variables simultanément[1].

Tableau 3.7

Trois colonnes supplémentaires (ou illustratives)

forme	Age-30	Age-50	Age+50
Argent	59	66	70
Avenir	115	117	86
Chômage	79	88	177
Conjoncture	9	8	5
Difficile	2	17	18
Economique	18	19	17
Egoisme	14	34	61
Emploi	21	30	28
Finances	8	12	8
Guerre	7	6	13
Logement	10	27	17
Peur	48	59	52
Santé	13	29	53
Travail	30	63	58
Total	433	575	663

La figure 3.2 ci-après nous montre que les trois premières formes supplémentaires appartiennent plutôt aux réponses des personnes dont le niveau de diplôme est élevé, alors que la quatrième, *mésentente*, est moins caractéristique.
On trouvera au tableau 3.8 un ensemble de paramètres plus techniques qui permettent d'apprécier la particularité de chacune de ces ventilations.
On ne s'étonnera pas de ne pas trouver de "contributions" dans ce tableau, puisque les éléments illustratifs ont par définition une contribution nulle : ils

[1] On peut être cependant amené à étudier la juxtaposition de plusieurs tableaux de contingence (cf. chapitre 5, paragraphe 5.4), pour obtenir une première visualisation d'un ensemble de caractéristiques (un ensemble de variables socio-démographiques par exemple, ou une batterie de questions fermées).

ne participent pas à la construction des axes, mais sont positionnés a posteriori.

Il est de bonne discipline de commencer par utiliser, en qualité de tableaux actifs, des tableaux homogènes, ne décrivant les proximités que d'un point de vue. On pourra, dans un second temps, illustrer la représentation obtenue par des informations supplémentaires.

Sur la figure 3.2, les trois classes d'âge s'ordonnent comme les diplômes le long de l'axe horizontal : aux âges croissants correspondent des niveaux de diplômes décroissants. Il s'agit d'un trait structurel de la population étudiée; les moins âgés sont les plus instruits, mais ceci complique l'interprétation en terme de causalité... Est-ce que l'effet du niveau d'instruction sur le contenu et la forme des réponses libres n'est pas un effet plus direct lié à l'âge ? Pour aller plus avant dans l'interprétation, il faudra croiser les variables Age et Diplôme, pour obtenir une partition des répondants en $3 \infty 5 = 15$ colonnes... mais on sort là de l'exemple d'école qui nous sert à illustrer le fonctionnement de la méthode.[1]

Tableau 3.8

Paramètres issus de l'analyse pour les éléments illustratifs

Colonnes illustratives

			COORDONNÉES			*COSINUS CARRES*		
Libellés des colonnes	P.REL	DISTO	*1*	*2*	*3*	*1*	*2*	*3*
moins de 30 ans	27.2	.08	-.11	.06	.10	.14	.04	.13
de 30 a 50 ans	36.1	.03	.02	-.05	.02	.01	.09	.01
plus de 50 ans	41.6	.11	.18	.05	-.10	.29	.02	.09

Lignes illustratives

			COORDONNÉES			*COSINUS CARRES*		
Formes	P.REL	DISTO	*1*	*2*	*3*	*1*	*2*	*3*
Confort	.8	.64	-.21	-.70	-.07	.07	.78	.01
Mésentente	1.1	.16	-.15	-.12	-.17	.13	.09	.18
Monde	1.1	.31	-.52	-.14	-.08	.88	.07	.02
Vivre	.8	.68	-.31	-.50	-.52	.14	.37	.40

[1] Une partition croisée Age-Diplôme en 9 postes sera utilisée dans l'exemple "grandeur réelle" du chapitre 5.

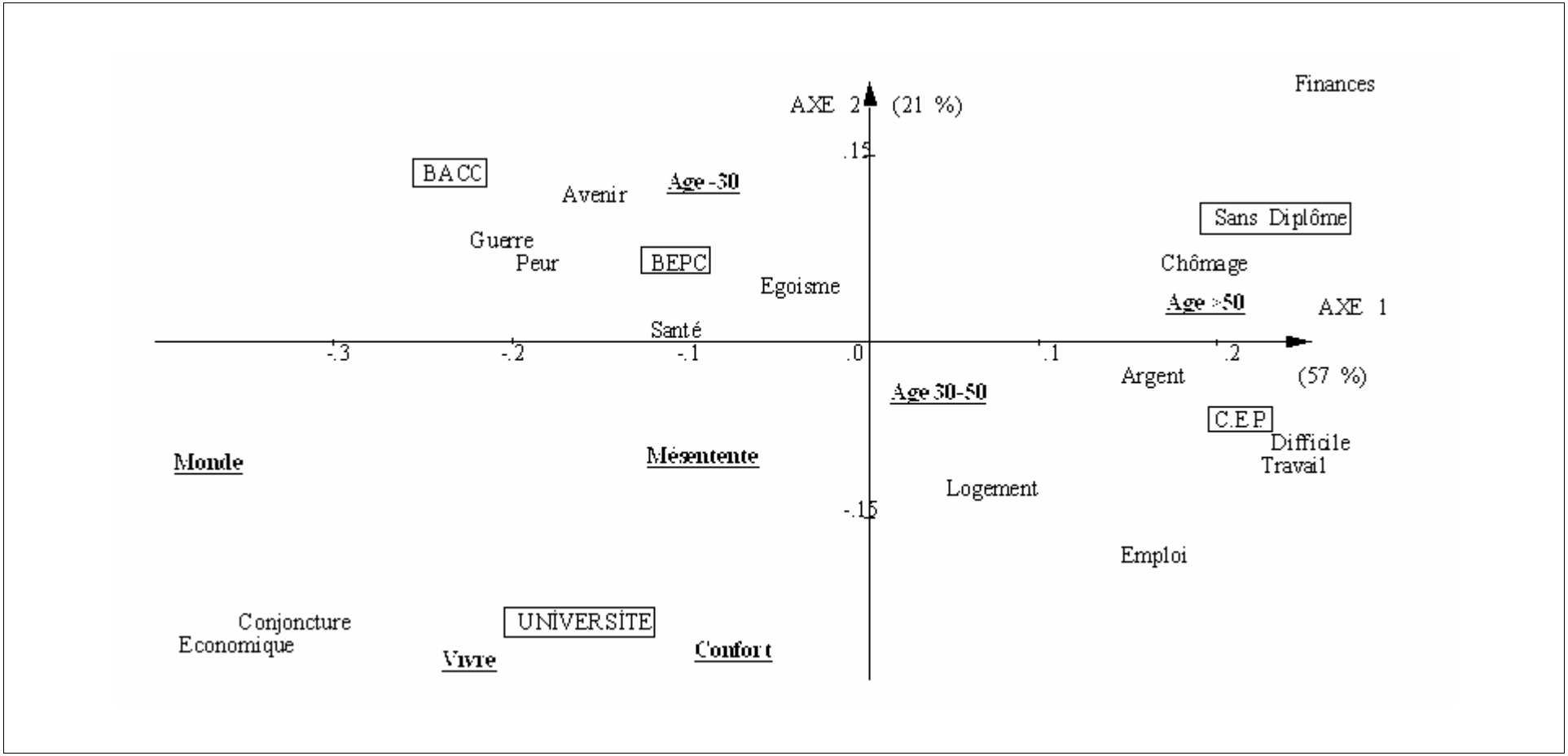


Figure 3.2

Associations entre formes et diplômes (suite)
Positionnement des éléments illustratifs (soulignés) dans le plan de la figure 3.1

3.3 Analyse des correspondances multiples

L'analyse des correspondances multiples permet de décrire de vastes tableaux binaires, dont les fichiers d'enquêtes socio-économiques constituent un exemple privilégié : les lignes de ces tableaux sont en général des individus ou observations (il peut en exister plusieurs milliers) ; les colonnes sont des modalités de variables nominales, le plus souvent des modalités de réponses à des questions. On peut faire remonter les principes de cette méthode à Guttman (1941), mais aussi à Burt (1950) ou à Hayashi (1956). Il s'agit en fait d'une simple extension du domaine d'application de l'analyse des correspondances, avec cependant des procédures de calcul et des règles d'interprétation spécifiques.

Cette extension se fonde sur la propriété suivante. On pose à k individus deux questions ayant respectivement n et p modalités de réponse, et l'on suppose que pour chaque question, les différentes modalités s'excluent mutuellement (une seule réponse possible).

Il est alors équivalent de soumettre à l'analyse des correspondances la table de contingence (n,p) croisant les deux questions, ou d'analyser le tableau binaire **Z** *à k lignes et (n+p) colonnes qui décrit les réponses.*[1]

L'analyse de **Z** est plus coûteuse, mais plus intéressante, car elle se généralise immédiatement au cas de plus de deux questions.

Le tableau 3.9 représente ainsi le cas de trois questions ayant respectivement 4, 3 et 4 modalités de réponse.

Tableau 3.9

Tableaux R (Codage réduit), Z (disjonctif complet) .

R=

2 3 4
2 1 3
3 1 2
4 2 4
1 2 3
2 2 3
3 1 1
1 1 1
4 1 2
2 2 3
3 2 2
4 1 4

Z =

0 1 0 0	0 0 1	0 0 0 1
0 1 0 0	1 0 0	0 0 1 0
0 0 1 0	1 0 0	0 1 0 0
0 0 0 1	0 1 0	0 0 0 1
1 0 0 0	0 1 0	0 0 1 0
0 1 0 0	0 1 0	0 0 1 0
0 0 1 0	1 0 0	1 0 0 0
1 0 0 0	1 0 0	1 0 0 0
0 0 0 1	1 0 0	0 1 0 0
0 1 0 0	0 1 0	0 0 1 0
0 0 1 0	0 1 0	0 1 0 0
0 0 0 1	1 0 0	0 0 0 1

[1] Chaque ligne de **Z** comporte dans ce cas deux "1" dans les colonnes correspondant aux réponses choisies, et *n+p-2* "0".

Le *tableau disjonctif complet* **Z** comporte donc ici trois blocs. Les réponses des 12 individus (lignes) peuvent alors être codifiées dans un tableau réduit **R**, (12 lignes et 3 colonnes) qui contient simplement les numéros des modalités. Les procédures de calcul n'utiliseront en fait que ce tableau peu encombrant.

Représenté sur le tableau 3.10, le troisième tableau **B** est le produit du tableau disjonctif **Z** par son transposé :

$$\mathbf{B} = \mathbf{Z}'\mathbf{Z}$$

Ce dernier tableau est symétrique (on l'appelle *tableau de Burt* ou *tableau de correspondance multiple*).

Il contient 9 blocs : les blocs diagonaux sont des matrices diagonales, dont les éléments diagonaux sont les effectifs de réponses correspondant à chaque modalité. Les trois blocs distincts parmi les blocs restant ne sont autres que les trois tables de contingence croisant les trois questions deux à deux.

Tableau 3.10

Tableaux B=Z'Z (de Burt)

B = Z'Z =

2 0 0 0	1 1 0	1 1 1 1
0 4 0 0	1 2 1	0 0 3 1
0 0 3 0	2 1 0	1 2 0 0
0 0 0 3	2 1 0	0 1 0 2
1 1 2 2	6 0 0	2 2 1 1
1 2 1 1	0 5 0	0 1 3 1
0 1 0 0	0 0 1	0 0 0 1
1 0 1 0	2 0 0	2 0 0 0
1 0 2 1	2 1 0	0 3 0 0
1 3 0 0	1 3 0	0 0 4 0
1 1 0 2	1 1 1	0 0 0 3

L'analyse des correspondances du tableau **Z** donne les mêmes axes factoriels normés que celle du tableau **B**, c'est-à-dire finalement des graphiques de proximités très similaires, aux échelles près ; mais les valeurs propres homologues sont différentes : à la valeur propre λ_α de l'analyse de **Z** correspond la valeur propre $(\lambda_\alpha)^2$ de l'analyse de **Z'Z**.

Notons que dans le cas de deux questions, le tableau **R** n'a que deux colonnes, le tableau **Z** est formé de deux blocs, $\mathbf{Z} = (\mathbf{Z}_1,\mathbf{Z}_2)$ et **B** en comporte quatre.

Les deux analyses précédentes donnent également la même représentation que l'analyse des correspondances de la petite table de contingence **C**

croisant les deux questions[1], avec, cette fois, une valeur propre égale à : $(2\lambda - 1)^2$.

3.3.1 Structure de base d'un échantillon d'enquête.

Cet exemple illustre l'utilisation de l'analyse des correspondances multiples pour le traitement des données d'enquêtes. L'exemple précédent a attiré l'attention sur le fait que derrière un effet "niveau d'éducation" pouvait se cacher un effet d'âge... (dans notre échantillon, les moins instruits sont en moyenne les plus âgés). Étant donnée la structure de la population française, ces mêmes effets, on le sait, ne sont pas indépendants du genre (sexe), du niveau de vie... d'où l'idée de décrire le réseau d'interrelations entre toutes les caractéristiques de base des enquêtés, puis de positionner les autres thèmes de l'enquête en tant qu'éléments illustratifs.

Toujours pour faciliter la présentation des résultats, l'échelle réelle de l'application sera réduite : le tableau **Z** aura 144 lignes et 24 colonnes actives représentant les modalités de réponses à 5 questions (tableau 3.11).

Un tableau complémentaire $\mathbf{Z}^+$ comprendra en outre 25 questions illustratives notant la présence ou l'absence des formes graphiques extraites des réponses à une question ouverte posée aux 144 personnes en 1990. Cette question ouverte, inspirée par un travail antérieur de C. Baudelot (1988) a pour libellé : *Que signifie pour vous réussir sa vie ?*

(l'élément z^{+}_{ij} correspondant de $\mathbf{Z}^+$ vaut 1 si l'individu i a utilisé la forme j pour la réponse libre précitée, et 0 sinon. $\mathbf{Z}^+$ aura donc 50 colonnes).

Dans une exploitation en vraie grandeur, il n'est pas rare de positionner en éléments illustratifs plusieurs centaines de modalités. L'intérêt réel de la méthode réside dans ces possibilités de tri et de filtrage systématique à grande échelle.

Le paradoxe pédagogique évoqué au paragraphe 3.1 est encore valable... il y a conflit entre la taille et la pertinence de l'exemple.

---

[1] La matrice **C** est un bloc non-diagonal de **B**. Elle s'écrit $\mathbf{C} = \mathbf{Z}_1'\mathbf{Z}_2$. On notera que dans ce cas les *lignes et les colonnes* de la table de contingence **C** ne sont *que des colonnes* pour la matrice **Z**. La représentation simultanée des lignes et des colonnes n'est donc pas un simple artifice graphique: elle trouve une justification dans cette présentation de l'analyse des correspondances.

Tableau 3.11 . Structure de base. Liste des questions actives et illustratives

I - 5 questions actives (15 Modalités)

	nombre de modalités
1 . genre (Sexe)	2
2 . niveau de Diplôme	4
3 . statut matrimonial	4
4 . avez-vous eu des enfants?	2
5 . age en 3 classes	3

II - 25 variables illustratives

(Présence ou absence de la forme dans la réponse à la question ouverte)

	nombre de modalités		nombre de modalités
argent	2	mes	2
arriver	2	mari	2
atteindre	2	métier	2
bien	2	peau	2
bon	2	professionnelle	2
bonheur	2	réussir	2
boulot	2	réussite	2
équilibre	2	santé	2
familiale	2	sentir	2
famille	2	situation	2
fonder	2	travail	2
ma	2	travailler	2
me	2		

Lecture du tableau 3.12

Le tableau 3.12 ne représente que la moitié inférieure du tableau de Burt (lequel, rappelons-le, est symétrique) relatif aux 5 questions actives. On trouve dans ce tableau tous les tableaux de contingence (il y en a 10) croisant les 5 questions actives deux à deux.

Sur la diagonale se trouvent les questions croisées avec elles-mêmes, et donc les effectifs correspondant à chaque modalité : on peut ainsi lire qu'il y a 70 hommes et 74 femmes dans l'échantillon. Sur 78 célibataires, il y a 42 hommes et 36 femmes.

Lecture du tableau 3.13

Les règles de lecture du tableau 3.13 sont presque en tout point identiques à celles du tableau 3.5 précédent. Seuls les calculs de contributions cumulées pour les modalités de chaque question sont nouveaux. Leur interprétation est immédiate.

Tableau 3.12

Croisements deux à deux des cinq questions actives.

(Tableau de Burt B)

	masc	femi	CEP*	BEPC	Bacc	Univ	celi	mari	divo	veuf	enfa	enfn	A–30	A–50	A+50
masc	70	0													
femi	0	74													
CEP*	15	12	27	0	0	0									
BEPC	12	24	0	36	0	0									
Bacc	18	25	0	0	43	0									
Univ	25	13	0	0	0	38									
celi	42	36	10	19	24	25	78	0	0	0					
mari	27	28	12	13	18	12	0	55	0	0					
divo	1	6	2	3	1	1	0	0	7	0					
veuf	0	4	3	1	0	0	0	0	0	4					
enfa	23	29	15	17	15	5	2	42	5	3	52	0			
enfn	47	45	12	19	28	33	76	13	2	1	0	92			
A–30	39	35	8	16	23	27	65	8	1	0	3	71	74	0	0
A–50	26	23	12	10	17	10	9	35	5	0	34	15	0	49	0
A+50	5	16	7	10	3	1	4	12	1	4	15	6	0	0	21
	masc	femi	CEP*	BEPC	Bacc	Univ	celi	mari	divo	veuf	enfa	enfn	A–30	A–50	A+50

Tableau 3.13

Paramètres issus de l'analyse des correspondances multiples pour les éléments actifs

MODALITES			*COORDONNEES*				*CONTRIBUTIONS*				*COSINUS CARRES*			
IDEN — LIBELLE	P.REL	DISTO	1	2	3	4	1	2	3	4	1	2	3	4
1 . Genre														
masc — homme	9.72	1.06	—.22	—.49	—.58	—.14	.9	7.3	13.2	.9	.04	.23	.32	.02
femi — femme	10.28	.95	.20	.46	.55	.14	.8	6.9	12.5	.9	.04	.23	.32	.02
			contribution cumulée =				1.7	14.3	25.7	1.8				
3 . Niveau de Diplôme														
CEP* — Sans diplôme ou CEP	3.75	4.33	.68	.31	—1.19	—.40	3.3	1.1	21.5	2.8	.11	.02	.32	.04
BEPC — BEPC	5.00	3.00	.29	.63	.79	—.40	.8	6.2	12.7	3.9	.03	.13	.21	.05
Bacc — Baccalauréat	5.97	2.35	—.07	—.37	.54	1.09	.1	2.6	7.2	33.2	.00	.06	.13	.50
Univ — Université	5.28	2.79	—.68	—.39	—.52	—.56	4.6	2.5	5.8	7.8	.16	.05	.10	.11
			contribution cumulée =				8.8	12.4	47.2	47.7				
7 . Etes vous actuellement														
celi — Célibataire	10.83	.85	—.80	.18	.02	.02	13.4	1.1	.0	.0	.76	.04	.00	.00
Mari — Marié(e)	7.64	1.62	.89	—.55	—.10	.32	11.6	7.3	.3	3.7	.49	.19	.01	.06
Divo — Divorcé(e)	.97	19.57	1.01	.14	1.82	—3.09	1.9	.1	13.2	43.9	.05	.00	.17	.49
veuf — Veuf(ve)	.56	35.00	1.65	3.82	—2.27	.53	2.9	25.5	11.6	.7	.08	.42	.15	.01
			contribution cumulée =				29.8	33.9	25.1	48.4				
8 . Avez—vous eu des enfants														
enfa — oui	7.22	1.77	1.19	—.21	.00	.01	19.5	1.0	.0	.0	.79	.03	.00	.00
enfn — non	12.78	.57	—.67	.12	.00	—.00	11.0	.6	.0	.0	.79	.03	.00	.00
			contribution cumulée =				30.6	1.6	.0	.0				
9 . Age en 3 classes														
A—30 — Moins de 30 ans	10.28	.95	—.84	.13	.04	.02	13.9	.6	.1	.0	.74	.02	.00	.00
A—50 — Entre 30 et 50 ans	6.81	1.94	.78	—.86	.10	—.16	8.1	15.8	.3	.8	.32	.38	.01	.01
A 50 — Plus de 50 ans	2.92	5.86	1.13	1.53	—.37	.29	7.1	21.5	1.6	1.2	.22	.40	.02	.01
			contribution cumulée =				29.1	37.8	2.0	2.0				

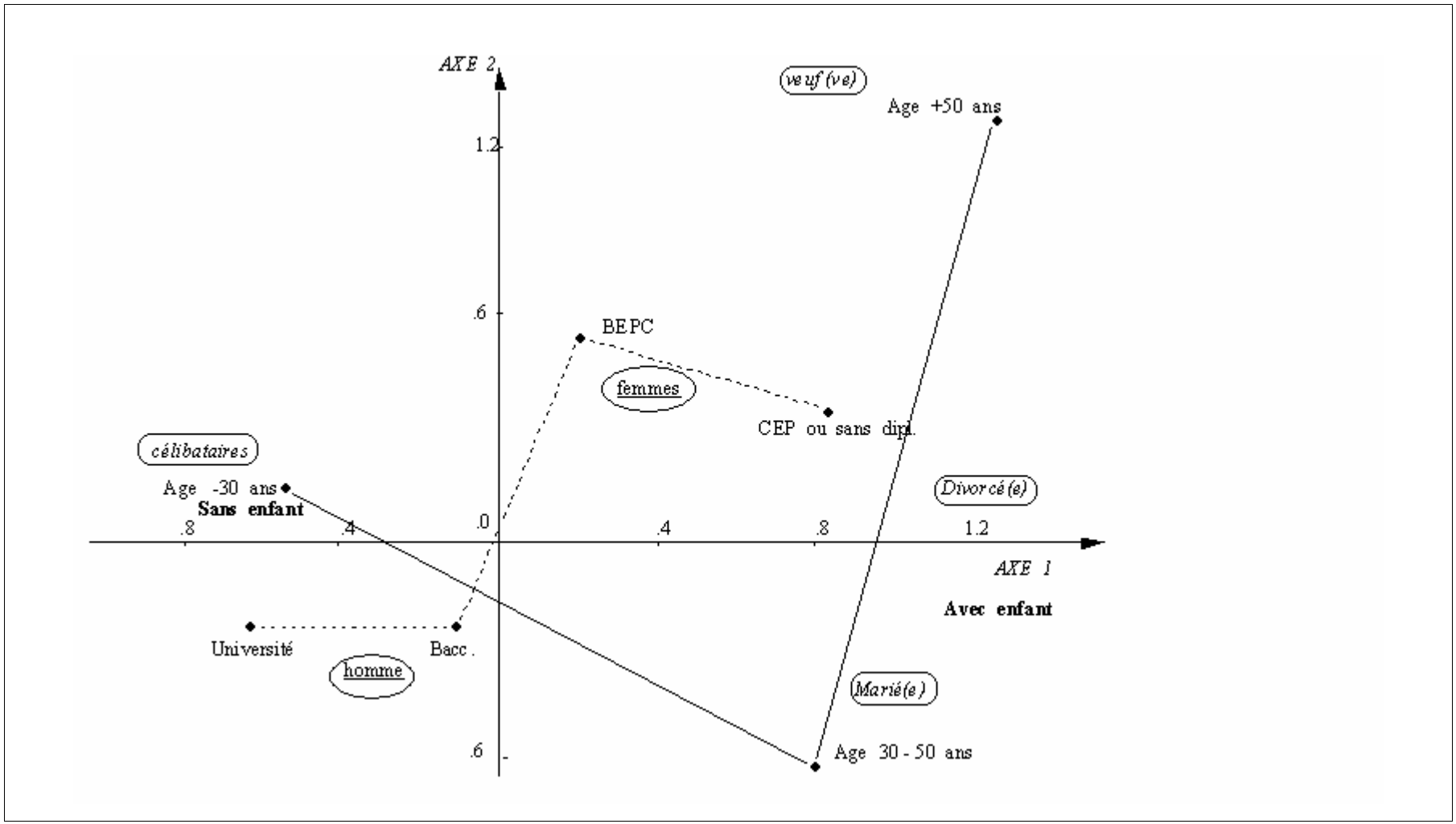


Figure 3.3

Structure de base de l'échantillon.

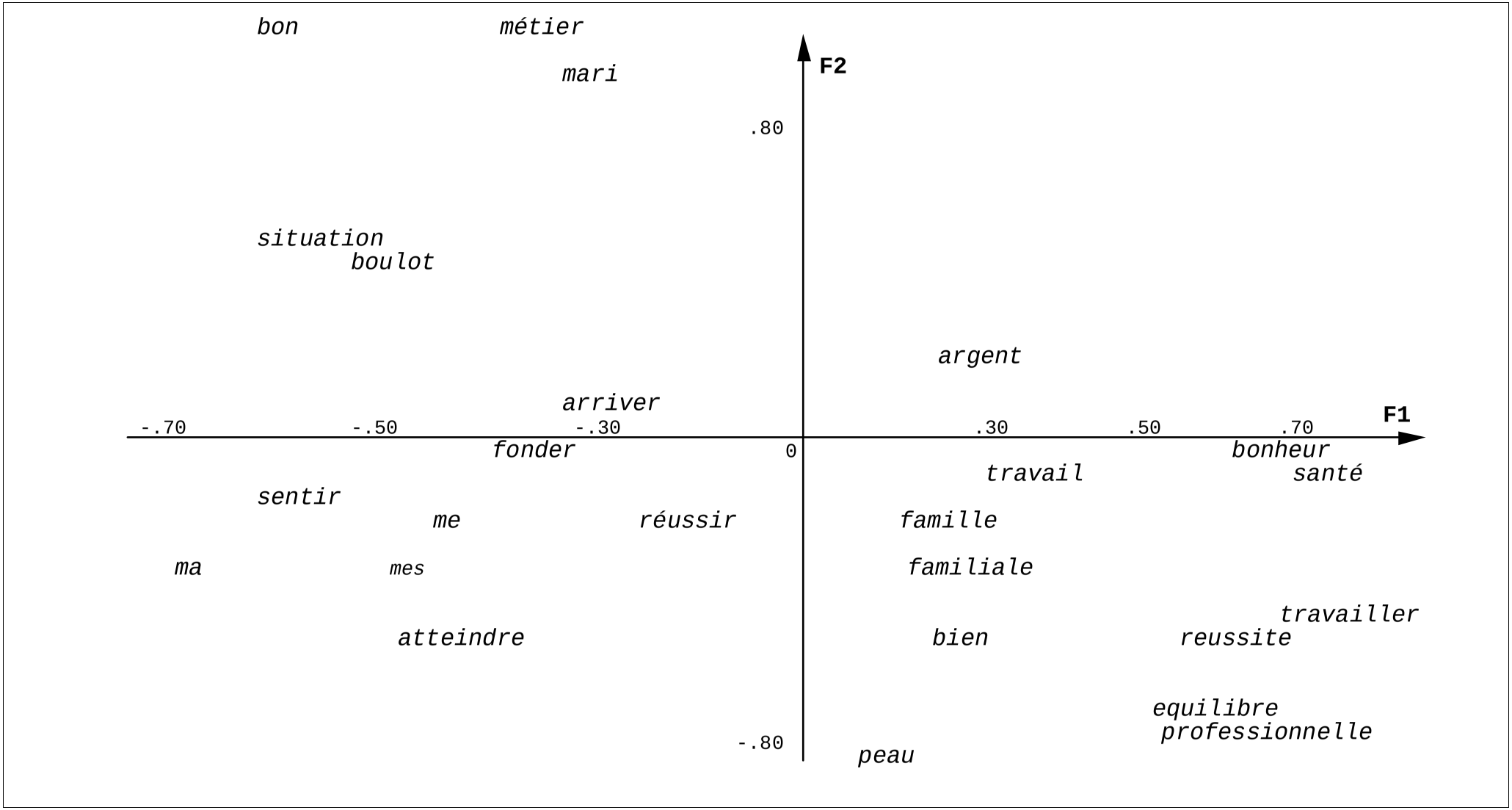


Figure 3.4

Positionnement des formes graphiques du tableau 3.15 dans le plan de la figure 3.3

Lecture de la figure 3.3

La structure du *nuage* des modalités actives est décrite par le plan factoriel de la figure 3.3, qui résume donc les 10 tableaux croisés.

Malgré le caractère sommaire des variables utilisées, les liaisons entre niveau de diplôme, âge, genre et équipement sont plus complexes que l'on ne l'imagine a priori. Un effet de cycle de vie, et aussi de génération se cache derrière des variables à deux modalités comme la variable *genre* : la catégorie "féminin" est lestée par un surnombre de personnes âgées sans diplôme. Il faut donc en tenir compte chaque fois que l'on interprète un éventuel "effet" de cette variable sur telle opinion ou attitude.

La figure 3.3 n'est cependant qu'une approximation plane d'un ensemble de proximités qui ne peuvent pas toujours être représentées dans un plan... il y a donc certaines déformations.

Même si elle peut donner lieu à quelques commentaires, la figure 3.3 n'est pas un résultat en soi : elle n'est qu'une préparation de l'information destinée à accueillir les variables illustratives. Auparavant, disons quelques mots de la validité de la représentation.

3.3.2 Validité de la représentation

Dans le cas des correspondances multiples, on évitera d'utiliser les pourcentages de variance pour juger de la pertinence des axes : ces pourcentages n'ont pas le même sens que lorsqu'il s'agit d'une table de contingence car le codage binaire introduit un *bruit* qui réduit la part d'explication attachée à chaque valeur propre.

Tableau 3.14

Valeurs propres et pourcentages correspondants

Numéro	Valeur propre	Pourcentage	Pourcentage cumulé	
1	.51	25.95	25.95	**********
2	.31	15.89	41.83	******
3	.24	12.27	54.11	*****
4	.21	10.59	64.69	****
5	.19	9.88	74.58	****
6	.16	8.03	82.61	****
7	.12	6.36	88.97	***
8	.11	5.61	94.58	**
9	.06	3.27	97.85	*
10	.04	2.15	100.00	*

Les deux premiers pourcentages (arrondis) sont respectivement de 26% et 16%, pour des valeurs propres de 0,51 et 0.31 (cf. tableau 3.14).

En pratique, la stabilité du plan factoriel s'éprouve par des techniques de simulation (perturbations aléatoires du tableau de données) ou de validation par échantillon-test (positionnement d'individus n'ayant pas participé à la construction des axes).

Ici, la taille réduite de l'échantillon (144 individus) ne permet pas de procéder à de fines inductions à partir de proximités observées.

3.3.3 Positionnement des variables illustratives.

La représentation simultanée des lignes et des colonnes liée à l' analyse des correspondances n'a pas été utilisée : seules les colonnes du tableau **Z** ont été portées sur la figure 3.3. Les 144 points-lignes n'ont pas été tracés pour des raisons d'encombrement graphique, mais surtout parce que les individus correspondants sont anonymes... seules leurs caractéristiques présentent de l'intérêt. Ce sont précisément les autres informations disponibles sur les personnes interrogées qui vont être "projetées" en éléments illustratifs.

Le fait que le tableau **Z** ne comporte que des "0" et des "1" va permettre une interprétation particulièrement simple de la disposition relative des colonnes illustratives.

La formule de transition (2) précédente nous montre en effet que la coordonnée Φ_j d'une colonne j (réponse illustrative) s'obtient en multipliant par β la simple *moyenne arithmétique* des coordonnées des individus qui ont choisi la réponse j. C'est ainsi que l'on positionne toute catégorie illustrative.

Lecture de la figure 3.4

On lit sur la figure 3.4 le positionnement des formes graphiques dont les libellés et les coordonnées figurent dans le tableau 3.15. Ces éléments de réponse sont donc projetés comme variables illustratives sur le plan factoriel de la figure 3.3 (les modalités des variables actives n'ont pas été reproduites sur la figure 3.4 afin d'en alléger la lecture).

Les liaisons existantes entre l'emploi des mots et les caractéristiques des personnes qui répondent sont visibles immédiatement dans un cadre *qui tient compte des interrelations* entre ces caractéristiques.
Les consultations classiques de tableaux croisés sont en effet hypothéquées par le fait "qu'une variable peut en cacher une autre"... elles sont de plus largement redondantes lorsque les caractéristiques successives sont liées entre elles. Le système de projection de variables supplémentaires permet donc d'économiser du temps et d'éviter des erreurs d'interprétation.

La série de variables illustratives portée sur la figure 3.4 est constituée par 25 formes graphiques utilisées par les mêmes individus dans leurs réponses à la question ouverte : "*Que signifie pour vous réussir sa vie ?*". Ces 25 formes ont été sélectionnées en raison de leurs positions excentrées, et donc caractéristiques.

Lors d'une application en vraie grandeur, des centaines de formes seront projetées et triées de cette façon.

Tableau 3.15

Coordonnées factorielles et contributions pour les formes supplémentaires

Formes	MASSES	DISTO	COORDONNEES			COSINUS CARRES		
			F1	F2	F3	F1	F2	F3
argent	.009	7.68	.23	.17	.39	.007	.004	.020
arriver	.004	21.18	-.25	.09	.38	.003	.000	.007
atteindre	.003	43.28	-.50	-.36	-.25	.006	.003	.001
bien	.013	6.67	.32	-.34	.15	.016	.017	.004
bon	.004	30.55	-.70	.84	.77	.016	.023	.020
bonheur	.003	32.85	.78	.03	.47	.019	.000	.007
boulot	.005	25.45	-.57	.36	.23	.013	.005	.002
équilibre	.004	13.49	.52	-.46	.58	.020	.016	.025
familiale	.005	14.10	.10	-.24	.22	.001	.004	.004
famille	.009	8.22	.17	-.15	.03	.004	.003	.000
fonder	.002	61.00	-.40	.01	.53	.003	.000	.005
ma	.005	17.09	-.89	-.23	-.23	.047	.003	.003
me	.002	34.56	-.47	-.18	-.50	.006	.001	.007
mes	.002	39.23	-.54	-.30	.03	.007	.002	.000
mari	.003	30.40	-.30	.70	1.13	.003	.016	.042
métier	.005	11.97	-.31	.91	.24	.008	.070	.005
peau	.006	11.09	.06	-.57	.21	.000	.029	.004
professionnelle	.006	13.13	.58	-.55	.70	.026	.023	.038
réussir	.024	2.41	-.20	-.14	.06	.016	.009	.002
réussite	.004	14.54	.52	-.36	.57	.019	.009	.022
santé	.004	22.75	1.30	-.09	-.84	.075	.000	.031
sentir	.003	22.88	-.74	-.08	-.17	.024	.000	.001
situation	.007	16.72	-.74	.42	.36	.033	.010	.008
travail	.014	5.07	.37	-.06	.30	.027	.001	.018
travailler	.004	41.29	.75	-.29	-.73	.014	.002	.013

On trouve par exemple des possessifs (*ma, mes*) dans une zone où les personnes interrogées sont jeunes et instruites. Ils sont accompagnés de verbes *(atteindre, arriver, réussir*). Seul, le verbe *travailler* caractérise les personnes plus âgées, qui s'expriment beaucoup moins à la première personne.

Toutes ces proximités suggèrent en fait des conjectures à vérifier ensuite sur les données initiales, et sur un échantillon plus important.

En résumé...

On peut résumer la démarche suivie lors de cette application de l'analyse des correspondances multiples par les deux étapes :

a) établissement d'une typologie de la population selon un point de vue caractérisé par un ensemble homogène de variables actives (ici : les caractéristiques de base des personnes interrogées).

b) identification des sous-groupes de cette population à partir de toutes les autres informations pertinentes disponibles. En fait, c'est la méthode qui permettra de sélectionner les variables supplémentaires ayant des coordonnées "significatives" sur les axes factoriels, ce qui permet d'envisager des explorations systématiques, avec de nombreux croisements de variables.

Les analyses des correspondances simples et multiples permettent donc d'obtenir des images suggestives, mais fragiles des tables de contingences et des tableaux de codages binaires.

L'utilisation conjointe des méthodes de classification présentées au chapitre suivant permettra à l'utilisateur de procéder à des inductions plus sûres et à des lectures plus confortables.

Chapitre 4

La classification automatique des formes et des textes

Les méthodes de classification automatique, constituent à côté des méthodes factorielles, la seconde grande famille de techniques d'analyse des données. Ces méthodes permettent de représenter les proximités entre les éléments d'un tableau lexical (lignes ou colonnes) par des regroupements ou classes.

Cet ensemble se divise lui-même en deux familles principales :

a) Les méthodes de *classification hiérarchique*, qui permettent d'obtenir à partir d'un ensemble d'éléments décrits par des variables (ou dont on connaît les distances deux à deux), une hiérarchie de classes partiellement emboîtées les unes dans les autres.

b) Les méthodes de *partitionnement*, ou de classification directe, qui produisent de simples découpages ou partitions de la population étudiée, sans passer par l'intermédiaire d'une hiérarchie. Ces dernières sont mieux adaptées aux très grands ensembles de données (plusieurs milliers d'individus à classer).

Les méthodes correspondant aux deux familles peuvent être combinées en une approche mixte qui sera évoquée au paragraphe 4.3.

Les résultats fournis par les méthodes de classification se révèlent, dans la pratique, des compléments indispensables aux résultats fournis par l'analyse des correspondances qui ont été décrits au chapitre précédent.

En effet, dans l'étude des tableaux lexicaux, l'analyse des correspondances permet avant tout de dégager de grands traits structuraux qui portent à la fois sur les deux ensembles mis en correspondance. Cependant, lorsque le nombre des éléments représentés est important, il devient délicat, dans la pratique, d'apprécier leurs positions réciproques au seul vu des résultats graphiques. Dans ce cas en effet, les plans factoriels deviennent peu lisibles en raison du grand nombre de ces éléments. Il en va de même, dès que le texte du corpus analysé est un peu long même si le nombre des parties est relativement

restreint. En effet, le nombre des formes croît alors inévitablement de manière importante, même si on restreint l'analyse aux formes les plus fréquentes.

Une seconde raison qui motive l'emploi conjoint des méthodes factorielles et de la classification tient à l'enrichissement dimensionnel des représentations obtenues. Les regroupements obtenus se font en effet à partir de distances calculées *dans tout l'espace*, et non seulement dans le ou les premiers plans factoriels. Ils seront donc de nature à corriger certaines des déformations inhérentes à la projection sur un espace de faible dimension.

La dernière raison de cette complémentarité est pratique : il est plus facile de faire décrire automatiquement des classes par l'ordinateur, que de faire décrire un espace continu, même si la dimension de celui-ci a été fortement réduite. Un exemple tentera de montrer l'intérêt de cette "dissection" de l'espace.

Dans ce chapitre on se limitera à la description et à l'utilisation de méthodes de classification automatique qui utilisent la même métrique que l'analyse des correspondances, la métrique du chi-2 définie au chapitre précédent, de façon à assurer une bonne compatibilité des résultats.

Après un bref rappel sur les méthodes elles-mêmes (4.1), la classification sera utilisée pour décrire tour à tour les proximités sur les formes et sur les parties d'un corpus (4.2). On montrera ensuite quel rôle elle peut jouer dans le cas de fichiers de type "enquête" (4.3).

4.1 Rappel sur la classification hiérarchique

Comme l'analyse des correspondances, la classification hiérarchique s'applique aux tableaux à double entrée tels les tableaux lexicaux décrits dans les chapitres précédents. On peut soumettre à la classification soit l'ensemble des colonnes du tableau (qui correspondent la plupart du temps aux différentes parties d'un corpus) soit celui des lignes de ce même tableau (qui correspondent en général aux formes et parfois aux segments du même corpus).

Le principe de l'agrégation hiérarchisée est très simple dans son fondement. On part d'un ensemble de n éléments (que l'on appellera ici des éléments de base ou encore *éléments terminaux*) dont chacun possède un poids, et entre lesquels on a calculé des distances (il y a alors $n(n-1)/2$ distances entre les différents couples possibles). On commence par agréger les deux éléments les plus proches. Le couple ainsi agrégé constitue alors un *nouvel élément* dont on peut recalculer à la fois le poids et les distances à chacun des éléments qu'il

reste à classer.[1] A l'issue de cette étape, le problème se trouve ramené à celui de la classification de *n-1* éléments. On agrège à nouveau les deux éléments les plus proches, et l'on réitère ce processus (*n-1* fois au total) jusqu'à épuisement de l'ensemble des éléments. L'ultime et *(n-1)éme* opération regroupe l'ensemble des éléments au sein d'une classe unique.

Chacun des regroupements effectués en suivant cette méthode s'appelle un *noeud*. L'ensemble des éléments terminaux rassemblés dans un noeud s'appelle une *classe***.** Les deux éléments (ou groupes d'éléments) agrégés, l'*aîné* et le *benjamin* de ce noeud.

Ce principe est celui de la classification *ascendante* hiérarchique, famille de méthodes la plus répandue. Il faut noter qu'il existe aussi de méthodes *descendantes*, qui opèrent par dichotomies successives de toute la population.[2]

4.1.1 Le dendrogramme

La classification ainsi obtenue peut être représentée de plusieurs manières différentes. La représentation sous forme d'arbre hiérarchique ou *dendrogramme* constitue sans doute la représentation la plus parlante.

Les regroupements effectués à chaque pas de l'algorithme de classification hiérarchique rassemblent des éléments qui sont plus ou moins proches entre eux. Plus on avance dans le regroupement (plus on se rapproche du sommet de l'arbre), plus le nombre de points déjà agrégés est important et plus la distance minimale entre les classes qu'il reste à agréger est importante. On peut associer à chacun des noeuds de l'arbre cette "plus petite distance".

Cette manière de procéder permet de mettre en évidence, comme sur la figure 4.1 ci-dessous, les noeuds qui correspondent à des augmentations importantes de cette distance minimale et qui rassemblent donc des composants nettement moins homogènes que leur réunion.
On lit sur la figure 4.1 que les éléments C et D, très proches, se sont agrégés dès la première itération, et que les deux blocs [A, B, C, D] d'une part, et [E, F, G, H] d'autre part, qui s'agrègent en dernier, constituent probablement les deux classes d'une bonne partition de la population (i.e. des 8 éléments terminaux) vis-à-vis de la distance utilisée.

---

[1] Dans la pratique il existe un grand nombre de façons de procéder qui correspondent à cette définition, ce qui explique la grande variété des méthodes de classification automatique. Comme nous l'avons signalé plus haut, nous n'utiliserons ici qu'une seule de ces variantes, basée sur l'utilisation de la distance du chi-2 et la maximisation du moment d'ordre 2 d'une partition (Benzécri, 1973), appelée parfois critère de Ward généralisé.

[2] Une méthode de ce type a été préconisée par Reinert (1983) dans le domaine textuel pour classer des *unités de contexte*

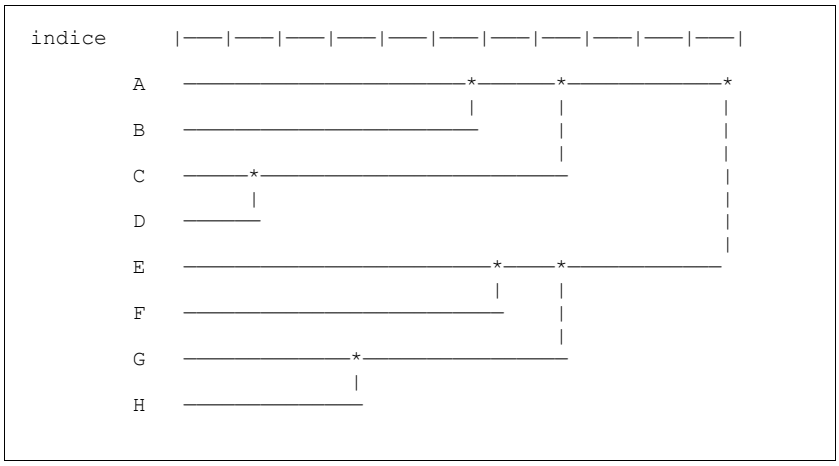


Figure 4.1

Dendrogramme représentant une classification sur un ensemble de 8 éléments

La représentation sous forme de dendrogramme d'une classification hiérarchique matérialise bien le fait que les classes formées au cours du processus de classification constituent une *hiérarchie indicée* de classes partiellement emboîtées les unes dans les autres.

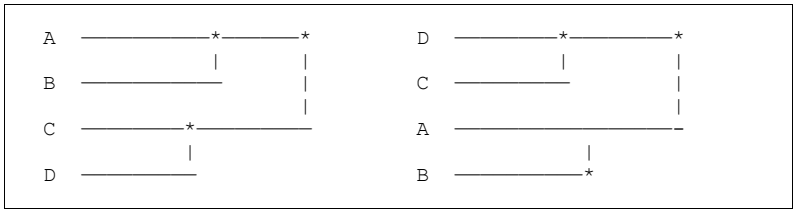


Figure 4.2

Dendrogrammes équivalents

L'interprétation de cette hiérarchie s'appuie sur l'analyse des seules distances entre éléments ou classes faisant l'objet d'un même noeud. En effet, pour une classification construite sur un ensemble d'éléments [A, B, C, D], la représentation sous forme de dendrogramme n'est pas unique
puisque chaque permutation de l'aîné et du benjamin d'un même noeud amène une représentation équivalente de la hiérarchie des classes.
Ainsi, les deux représentations de la figure 4.2 correspondent-elle à une même classification des quatre éléments en une hiérarchie comportant, en dehors des éléments terminaux, les trois classes [C,D], [A,B], [A,B,C,D].

4.1.2 Coupures du dendrogramme

Si le nombre des éléments à classer est important, la représentation complète de l'arbre de classification devient difficile à étudier. Une solution pratique consiste à définir un *niveau de coupure* de l'arbre (qui correspond précisément à un intervalle entre deux des itérations du processus de classification ascendante). Une telle coupure permet de considérer une classification résumée aux seules classes supérieures de la hiérarchie.

On peut définir la coupure du dendrogramme en déterminant à l'avance le nombre des classes dans lesquelles on désire répartir l'ensemble des éléments à classer, ou encore, ce qui revient au même, le nombre de noeuds supérieurs que l'on retiendra pour la classification (le nombre des classes retenues est alors égal au nombre des noeuds augmenté d'une unité).

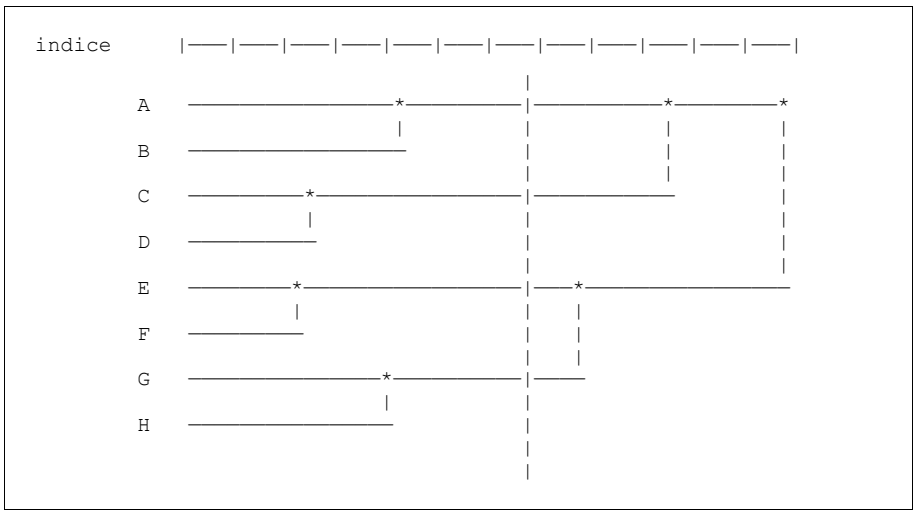


Figure 4.3

Coupure d'un dendrogramme réalisant 4 classes

La coupure pratiquée sur le dendrogramme représenté sur la figure 4.3 opère sur l'ensemble des éléments de départ une partition[1] en quatre classes [A, B], [C, D], [E, F], [G, H], dont la réunion correspond à l'ensemble soumis à la classification.

Chacune des classes est décrite exhaustivement par la liste des éléments qu'elle contient. Le choix d'une coupure dans l'arbre de classification est une opération qui doit en premier lieu s'appuyer sur les valeurs de l'indice : il faut dans la mesure du possible que cette coupure corresponde à un saut de l'indice. Avec les conventions des figures 4.1 à 4.3, les valeurs de l'indice à gauche de la coupure doivent être faibles, et celles à droite de la coupure forte : de cette

[1] Il s'agit ici de la partition de l'ensemble des éléments à classer. Cette notion, courante dans le domaine de la classification automatique doit être distinguée de celle de partition d'un corpus de textes (i.e. la division d'un corpus de textes en parties).

façon, les éléments sont proches à l'intérieur des classes définies par la coupure, et éloignés lorsqu'ils appartiennent à des classes différentes. On reviendra sur ce point au paragraphe 4.3.1 dévolu aux méthodes de classification mixtes.

4.1.3 Adjonction d'éléments supplémentaires

A partir d'une classification hiérarchique réalisée sur un ensemble d'éléments, il est possible de répartir à l'intérieur des classes de la hiérarchie ainsi créée un ensemble d'éléments supplémentaires ou illustratifs.[1]

Pour chacun des éléments supplémentaires, l'algorithme d'adjonction procède de manière particulièrement simple. On commence par rechercher parmi les éléments de l'ensemble de base celui dont l'élément supplémentaire à classer est le plus proche. On affecte ensuite cet élément supplémentaire à toutes les classes de la hiérarchie qui contiennent l'élément de base.

Cette méthode peut être utilisée, comme dans le cas de l'analyse des correspondances, pour illustrer par des données segmentales la classification obtenue à partir des formes. Ici encore, la méthode utilisée permet de classer un nombre important d'éléments supplémentaires sans perturber la classification obtenue sur l'ensemble de base.

4.1.4 Filtrage sur les premiers facteurs

Comme on l'a signalé plus haut, la classification hiérarchique se construit à partir d'une distance calculée entre les couples d'éléments de l'ensemble de base pris deux à deux. Dans le cas général, on mesure les distances entre les éléments soumis à la classification à l'aide de la distance du chi-2 entre les colonnes du tableau, distance qui est également à la base de l'analyse des correspondances.
Rappelons cependant que le but de l'analyse des correspondances est précisément l'extraction, lorsqu'elle est possible, de sous-espaces résumant au mieux l'information contenue dans le tableau de départ. L'interprétation s'appuie en général sur l'hypothèse que les facteurs qui correspondent aux valeurs propres les plus faibles ne constituent qu'un "bruit" non susceptible d'interprétation.

Si l'on suit cette logique, on peut opérer, à partir de l'ensemble des éléments de base, des classifications qui font intervenir les distances mesurées dans le seul

[1] On utilise ici cette notion dans le sens précis de la définition des éléments supplémentaires introduite au chapitre précédent.

espace des premiers facteurs, réputés les plus significatifs au sens statistique du terme.

L'opération qui transforme les distances mesurées entre les éléments de base en les réduisant à leur projection dans l'espace des premiers facteurs constitue un *filtrage* des distances sur les premiers facteurs. Elle permettra, par une réduction sélective de l'information, de classer des milliers de formes ou d'individus.[1]

Cette propriété de filtrage commune à la plupart des méthodes fondées sur la recherche d'axes principaux (décomposition aux valeurs singulières, analyse en composantes principales, analyse des correspondances simples et multiples) sera à nouveau évoquée au paragraphe 7.4 (chapitre 7). Elle est aussi utilisée en *discrimination textuelle* (chapitre 8) et en recherche documentaire dans l'approche désignée par l'expression *latent semantic analysis* (Deerwester et al., 1990 ; Bartell et al., 1992).

4.2 Classification des lignes et des colonnes d'un tableau lexical

Comme au chapitre précédent, on présentera le fonctionnement de la méthode sur un exemple de dimensions réduites, en rappelant que cet exercice pédagogique qui doit mettre en évidence le mécanisme des opérations ne parviendra pas à donner une idée exacte de la valeur heuristique de l'outil. Seront successivement examinées la classification des lignes (formes) et la classification des colonnes (parties).

4.2.1 Classification des formes

On reprendra l'exemple du chapitre 3 croisant 14 formes graphiques et 5 catégories de répondants à une enquête. La classification s'opérera sur les lignes du tableau 4.1 (lequel reproduit le tableau 3.2 du chapitre précédent). Les lignes de ce tableau sont les profils-lignes entre lesquels sont calculés les distances qui nous intéressent.
La première colonne de ce tableau contient un numéro d'ordre qui servira par la suite à décrire les regroupements. Les effectifs originaux de chaque ligne figurent dans la dernière colonne. Le tableau 4.2 décrit les étapes et les paramètres du fonctionnement de l'algorithme.

[1] Une analyse des correspondances préalable procure deux autres avantages : elle fournit une description des données complémentaire, parce que reposant sur des principes différents ; elle peut permettre des économies considérables lors des calculs de distances dans le cas de grands tableaux (en recherche documentaire par exemple).

Tableau 4.1

Rappel des profils-lignes du tableau 3.2 (chapitre 3)

		SDipl (1)	CEP (2)	BEPC (3)	Bacc (4)	Univ (5)	Total	(Eff)
1	*Argent*	26.4	33.2	16.6	15.0	8.8	100.	(193)
2	*Avenir*	16.7	28.3	24.5	23.6	6.9	100.	(318)
3	*Chômage*	25.1	39.2	17.7	14.1	3.9	100.	(283)
4	*Conjoncture*	4.5	31.8	22.7	22.7	18.2	100.	(22)
5	*Difficile*	25.9	40.7	14.8	11.1	7.4	100.	(27)
6	*Économique*	13.0	24.1	22.2	20.4	20.4	100.	(54)
7	*Égoïsme*	19.6	34.6	13.1	24.3	8.4	100.	(107)
8	*Emploi*	15.2	44.3	24.1	7.6	8.9	100.	(79)
9	*Finances*	35.7	25.0	25.0	10.7	3.6	100.	(28)
10	*Guerre*	15.4	26.9	26.9	23.1	7.7	100.	(26)
11	*Logement*	15.4	42.3	13.5	19.2	9.6	100.	(52)
12	*Peur*	15.7	28.3	23.9	23.9	8.2	100.	(159)
13	*Santé*	19.4	29.0	21.5	20.4	9.7	100.	(93)
14	*Travail*	23.2	40.4	19.2	9.3	7.9	100.	(151)
	Total	20.3	33.7	20.2	17.9	7.9	100.	(1592)

Lecture du tableau 4.2

Première étape : Il y a 14 éléments à classer.

La première ligne du tableau 4.2 indique que le premier élément artificiel (*noeud*) est obtenu par fusion des éléments 2 et 10 (*aîné* et *benjamin*) qui sont les formes *avenir* et *guerre*. Ce nouvel élément portera le numéro 15. Il sera caractérisé par le profil moyen de ses deux constituants. La valeur de l'indice correspondante (0.00008) décrit la plus petite distance correspondante, et la masse du nouvel élément (344) est bien la somme des effectifs (344 = 318 + 26).

Seconde étape : 13 éléments à classer (seconde ligne du tableau 4.2).

Les deux éléments les plus proches, qui vont constituer l'élément 16, sont les éléments 15 (*avenir* + *guerre*) et 12 (*peur*).

Troisième étape : 12 éléments à classer (troisième ligne du tableau 4.2)

L'élément 17 est ensuite formé de l'union des éléments 14 et 5 (*travail*, *difficile*), l'élément 18 rassemble le couple (*conjoncture, économique*). Puis s'unissent (*égoïsme*, *logement*), ([*avenir* + *guerre* + *peur*], et *santé*), etc.

Le processus se termine lorsqu'il ne reste plus qu'un seul élément.

Tableau 4.2

Classification hiérarchique des formes
(description des noeuds)

NUM.	AINE	BENJ	EFF.	POIDS	INDICE	HISTOGRAMME DES INDICES DE NIVEAU
15	2	10	2	344.00	.00008	*
16	15	12	3	503.00	.00018	*
17	14	5	2	178.00	.00022	*
18	6	4	2	76.00	.00061	**
19	7	11	2	159.00	.00094	***
20	16	13	4	596.00	.00107	***
21	8	17	3	257.00	.00198	******
22	9	1	2	221.00	.00219	******
23	3	22	3	504.00	.00321	*********
24	21	23	6	761.00	.00523	**************
25	20	19	6	755.00	.00654	*****************
26	25	18	8	831.00	.00890	************************
27	26	24	14	1592.00	.03091	***************************/ /*****

SOMME DES INDICES DE NIVEAU = .06206

On note que la somme des indices (0.062) est, elle aussi, égale à la somme des valeurs propres calculée au chapitre 3. Cette quantité est, on l'a vue, proportionnelle au chi-2 (ou : χ^2) calculé sur la table de contingence.

Analyse des correspondances et classification décomposent donc de deux façons différentes la même quantité (chi-2 classique), qui mesure un écart entre la situation observée et l'hypothèse d'indépendance des lignes et des colonnes de la table.

La représentation graphique de ce dendrogramme va illustrer de façon plus suggestive ce processus d'agrégation. Elle sera confrontée à la représentation graphique obtenue précédemment par analyse des correspondances, de façon à mettre en évidence l'originalité de chacun des points de vue.

La figure 4.4 ci-après représente sous la forme d'un dendrogramme les regroupements successifs du tableau 4.2 : la longueur des branches de l'arbre est proportionnelle aux valeurs de l'indice.

Lecture de la figure 4.4

On lit sur cette figure que les formes *peur, guerre, avenir* sont agrégées très tôt, mais également que le "paquet" *(peur, guerre, avenir, santé)* ne rejoint le couple *(logement, égoïsme)* que beaucoup plus tard.

La comparaison avec la figure 3.1 du chapitre précédent est intéressante. Les deux branches principales du dendrogramme, opposant les 8 premières lignes (de *peur* à *économique*) aux six dernières (de *difficile* à *chômage*) traduisent la principale opposition visible sur le premier axe factoriel (horizontal).

Les principaux regroupements observables dans le plan factoriel de la figure 3.1, sont ceux que décrit le processus d'agrégation hiérarchique, avec cependant quelques différences qui méritent notre attention.

Les points *santé* et *égoïsme* sont proches sur la figure 3.1 (il est vrai qu'ils sont au voisinage de l'origine, et que les proximités en analyse des correspondances sont d'autant plus fiables que les points occupent des positions périphériques...).

Le dendrogramme montre au contraire que le point *santé* est plus proche de la constellation (*peur, guerre, avenir*) que du point *égoïsme*. De la même façon, ce dendrogramme nous montre que le point *égoïsme* est plus proche de *logement* que de *santé*, contrairement à ce que pourrait laisser penser le plan factoriel de la figure 3.1.

Il ne faudrait pas en déduire que la classification donne des résultats plus précis que l'analyse des correspondances. Le bas de l'arbre (en fait, la partie gauche, pour les figures 4.1 à 4.4) donne effectivement des idées plus précises sur les distances locales ; mais, on l'a vu, les branches peuvent pivoter sur elles-mêmes, et la mise à plat de l'arbre ne donne que peu d'information sur les dispositions relatives des constellations plus importantes.

4.2.2 Classification des textes

La classification des colonnes de la table de contingence 3.1 du chapitre précédent se fait à partir des profils colonnes du tableau 3.3 que nous n'avons pas reproduit ici.

Le principe de l'agrégation est en tout point identique à ce qui vient d'être montré sur les lignes, et nous serons donc plus brefs dans les commentaires du tableau 4.3 et de la figure 4.5, homologues des tableaux 4.2 et de la figure 4.4.

(indices en % de la somme S des indices : S = 0.06206, minimum = 0.12%, maximum = 49.81%)

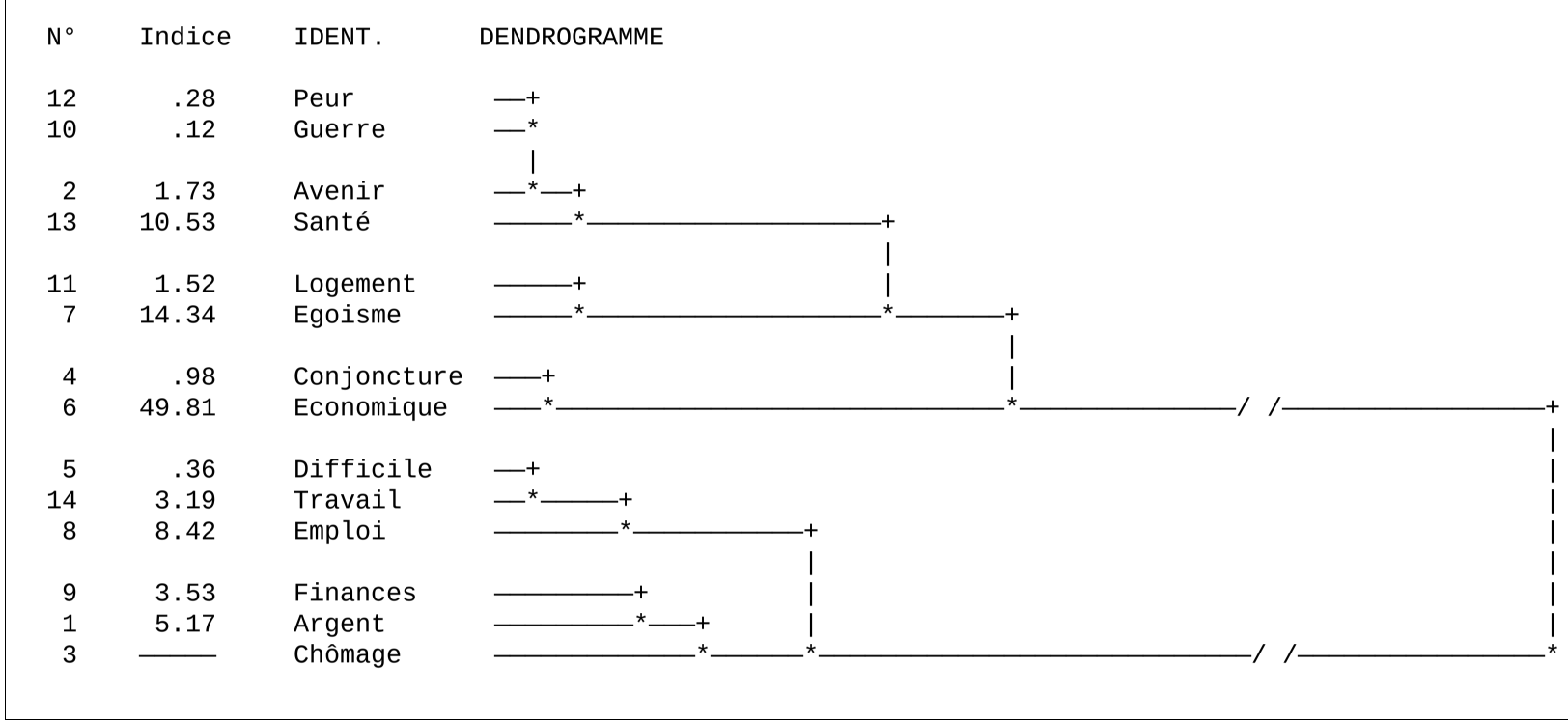


Figure 4.4

Dendrogramme décrivant les proximités entre lignes (formes) du tableau 4.1

Le tableau 4.3 décrit de façon similaire les étapes du fonctionnement de l'algorithme.

Tableau 4.3

Agrégation hiérarchique des colonnes : niveaux de diplômes.

NUM.	AINE	BENJ	EFF.	POIDS	INDICE	HISTOGRAMME DES INDICES DE NIVEAU
6	2	1	2	860.00	.00816	**********
7	4	3	2	607.00	.00950	**********
8	7	5	3	732.00	.01171	**********
9	8	6	5	1592.00	.03269	**********/ /*****

SOMME DES INDICES DE NIVEAU = .06206

Il y a 5 éléments à classer, et le premier noeud, qui porte le numéro 6, est formé des catégories 2 et 1 (numéros des colonnes sur le tableau 4.1) (*Sans diplôme* et *CEP*) qui totalisent 860 occurrences. Ce sont ensuite les catégories *Bacc* et *BEPC*, qui sont rejoint par la catégorie *Université*...

La figure 4.5 schématise le processus, en montrant comme précédemment que les deux branches principales de l'arbre correspondent à une opposition sur le premier axe factoriel de la figure 3.1 du chapitre précédent.

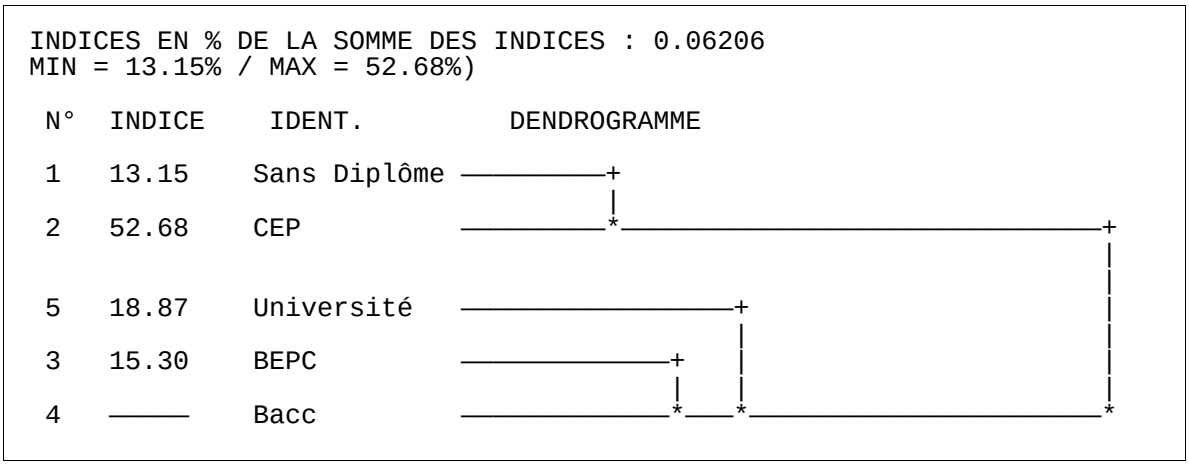


Figure 4.5

Dendrogramme décrivant les proximités entre colonnes du tableau 3.1

On remarque que la somme des indices a toujours la même valeur de 0.062, ce qui était attendu, car cette quantité fait intervenir de façon symétrique les lignes et les colonnes de la table. Toutefois, contrairement à ce qui se passe en analyse des correspondances, la classification des colonnes ne se déduit pas de façon simple de celle des lignes. Il n'existe pas d'équivalent des formules de transition.

4.2.3 Remarques sur les classifications de formes

Bien qu'une certaine symétrie semble relier sur le plan formel les classifications effectuées sur l'ensemble des parties et celles réalisées à partir de l'ensemble des formes, la pratique de la classification sur les tableaux lexicaux montre que les deux types d'analyse répondent à des besoins distincts qui entraînent, dans les deux cas, des utilisations différentes de la méthode.

Lorsqu'il s'agit d'étudier des textes (littéraires, politiques, historiques) les classifications portant sur les formes d'un corpus concernent en général des ensembles dont la dimension dépasse très largement celle de l'ensemble des parties. L'arbre de classification réalisé à partir d'un tel ensemble se présente sous une forme relativement volumineuse qui complique considérablement toute synthèse globale.

Dans la pratique, on abordera l'étude de la classification ainsi réalisée en considérant par priorité les associations qui se réalisent aux deux extrémités du dendrogramme :

a) les classes du niveau inférieur de la hiérarchie constituées par des agrégations de formes correspondant à un indice très bas (i.e. les classes agrégées dès le début de la classification)

b) les classes supérieures, souvent constituées de nombreuses formes, que l'on étudiera comme des entités. On retrouve en général au niveau des classes supérieures les principales oppositions observables dans les premiers plans factoriels.

Associations du niveau inférieur de la hiérarchie

Intéressons-nous tout d'abord aux associations réalisées aux tout premiers niveaux de la classification. Par construction, ces associations regroupent des ensembles de formes dont les profils de répartition sont très similaires (et parfois mêmes identiques) dans les parties du corpus.

Comme on va le voir, seul le retour systématique au contexte permet de distinguer parmi ces associations celles qui proviennent essentiellement de la reprise de segments plus ou moins longs, celles qui sont générées par les cooccurrences répétées de plusieurs formes à l'intérieur de mêmes phrases ou de mêmes paragraphes et celles des associations qui résultent de l'identité plus ou moins fortuite de la ventilation de certaines formes.

Un exemple de "quasi-segments"[1]

Dans un corpus de textes syndicaux, composé de neuf textes relatifs à quatre centrales syndicales[2], on a observé une identité parfaite de la ventilation des formes *pèse* et *lourdement*, identité qui a entraîné l'agrégation de ces deux formes dès le début du processus de classification.

Tableau 4.4

Distribution de deux formes dans neuf textes

	DT1	DT2	DT3	TC	GT1	GT2	GT3	FO1	FO2
lourdement	0	0	0	3	4	2	1	1	0
pèse	0	0	0	3	4	2	1	1	0

Cette identité laisse deviner la présence, dans différentes parties du corpus, des occurrences du stéréotype *"peser lourdement (sur quelque chose ou quelqu'un)*".

Le retour au contexte permet tout à la fois de vérifier l'existence de cette locution dans plusieurs des parties du corpus et de constater que les cooccurrences de ces deux formes ne se produisent pas exclusivement dans ce type de contexte.

L'examen des ventilations qui correspondent à ces deux formes montre tout d'abord qu'elles sont employées, dans ce corpus du moins, dans des contextes du type suivant :

> [CFTC] /.../ la prise en charge par les collectivités locales de l'entretien des espaces verts qui **pèse lourdement** sur les bénéficiaires des logements sociaux.
>
> [CGT1] une nouvelle orientation de la fiscalité /.../ est indispensable pour s'attaquer à l'inflation qui **pèse lourdement** sur les travailleurs et les titulaires de petits revenus et qui a également des effets négatifs sur les échanges extérieurs.

A côté de ces contextes, les deux formes apparaissent dans des contextes légèrement modifiés par rapport au segment considéré ci-dessus.

---

1 Cf. également sur ce point Becue (1993) à qui nous empruntons cette dénomination.

2 Ce corpus, réuni par J. Lefèvre et M. Tournier est composé de textes votés en congrès par les principales centrales syndicales françaises entre 1971 et 1976. Les parties sont notées : DT (CFDT), TC (CFTC), GT (CGT), FO (CGT-Force-Ouvrière). On trouvera des développements qui concernent ce problème dans Salem (1993).

> [CFTC] le comité national de la c-f-t-c-, /.../, constate que les prix de détail continuent à monter et que l'inflation **pèse** de plus en plus **lourdement** sur toutes les catégories sociales à revenu modeste, surtout les personnes âgées et les chargés de familles.

Enfin, chacune des deux formes peut apparaître dans des contextes libres des occurrences de l'autre forme : on retrouve parfois d'autres formes liées à cette même locution qui relativisent l'absence de la forme *lourdement.* Tel est le cas par exemple pour le contexte de la forme *pèse* :

> [CFTC] pour que l'europe **pèse** de **tout son poids** en face des deux grands blocs économiques

ou encore [CGT1] :

> l'économie américaine **pèse** d'un **poids déterminant.**

Cette identité de ventilation dans les parties du corpus des deux formes appelle un double commentaire. Tout d'abord, la ressemblance globale des profils résulte essentiellement de l'existence d'une expression récurrente, plus ou moins figée, qui contient les deux formes.

Cependant, l'identité stricte des deux ventilations est, dans le cas considéré, relativement fortuite, puisqu'elle résulte de l'addition des occurrences des deux formes liées à cette expression avec d'autres occurrences des deux formes présentes dans des contextes indépendants.

Les méthodes de la classification permettent de généraliser la recherche des cooccurrences de couples de formes à l'intérieur d'une même phrase au repérage de cooccurrences à l'intérieur de contextes plus étendu pouvant concerner plusieurs formes.

Classes de ventilations homogènes

On obtient une classification de l'ensemble des formes en un petit nombre de classes en procédant à une coupure de l'arbre correspondant à un niveau suffisamment élevé dans la hiérarchie indicée des classes.

Cette manière de procéder permet de considérer des classes qui regroupent, par construction, des formes dont la ventilation est relativement homogène. L'analyse du contenu de ces classes se fait en retournant fréquemment au contexte pour tenter de trouver les raisons profondes des similitudes que l'on constate entre les profils des formes qui appartiennent à une même classe.
Cette démarche permet également de vérifier que les occurrences des formes graphiques qui correspondent aux différentes flexions d'un même lemme se retrouvent affectées à une même classe, ce qui sera le signe d'une ventilation analogue dans les parties du corpus.

Classes périphériques

Toutes les classes découpées par l'algorithme de classification ne présentent pas le même intérêt. En projetant les centres de gravité de chacune des classes sur les premiers plans factoriels issus de l'analyse des correspondances[1], il est possible de repérer la correspondance de certaines des classes avec les types extrêmes dégagés par l'analyse des correspondances et par là même de sélectionner des ensembles de formes qui illustrent tout particulièrement ces grands types.

On considérera avec une attention toute particulière les groupes de formes qui s'associent très bas lors de la construction de l'arbre de classification et qui se rattachent relativement haut au reste des autres formes.

Documentation par les données segmentales

Ici encore, la lisibilité des classifications obtenues sur les formes s'accroît notablement si l'on prend soin de situer les ventilations relatives aux segments répétés parmi les résultats similaires obtenus sur les formes graphiques.

Comme dans le cas de l'analyse des correspondances, cette implication des segments peut se réaliser de deux manières différentes : analyse directe sur le tableau des formes et des segments répétés pour laquelle tous les éléments sont impliqués en qualité d'éléments principaux, ou implication des segments en qualité d'éléments supplémentaires (ou illustratifs) sur la base d'une classification préalable opérée sur les seules formes graphiques.

L'algorithme d'adjonction d'éléments supplémentaires à une classification (cf. Jambu, 1978) permet de rattacher un par un aux différentes classes de la hiérarchie indicée un ensemble d'individus statistiques n'ayant pas participé à la classification sur les éléments de base sans que cette dernière en soit aucunement affectée. Ici encore nous utiliserons massivement cette propriété pour documenter la classification par des calculs relatifs aux segments répétés. Comme on l'a annoncé plus haut, on peut également soumettre à la classification hiérarchique l'ensemble constitué par la réunion des termes (formes et segments répétés) dont la fréquence dépasse un certain seuil dans le corpus. Bien que cette seconde manière de procéder modifie les résultats

[1] On fait ici allusion aux techniques de calcul des contributions des classes aux facteurs d'une analyse des correspondances Jambu (1978).

obtenus sur les formes, l'expérience montre que les deux procédures ne conduisent pas, en règle générale, à des résultats trop éloignés.

Les nouvelles classes contiennent alors des formes et des segments dont la ventilation est relativement homogène dans les parties du corpus. L'étude de la disposition relative d'une forme et des segments qu'elle contient se trouve grandement facilitée par cette procédure.

Influence de la partition du corpus

On a vu plus haut que les distances mutuelles entre les différentes formes de l'ensemble soumis à la classification sont calculées à partir de leur ventilation dans les parties du corpus. Il s'ensuit que le découpage du corpus en parties revêt une importance primordiale pour la construction d'une classification à partir de l'ensemble des formes d'un même corpus car les variations dans la partition du corpus ont pour effet de rapprocher certains couples de formes et d'en éloigner d'autres ce qui influe forcément sur les classes de la hiérarchie.

Avec l'évolution croissante des capacités des ordinateurs, qui concerne tant les capacités mémoire que la vitesse d'exécution des calculs, les méthodes de la classification ascendante hiérarchique trouveront des applications dans le domaine de la recherche des cooccurrences textuelles à l'intérieur de phrases, de segments de texte de longueur fixe ou de paragraphes. Les algorithmes utilisés ici devraient permettre d'étendre les recherches que nous avons décrites plus haut en s'appuyant sur des partitions plus fines d'un même corpus correspondant à chacune de ces unités.

4.3 La Classification des fichiers d'enquête

Dans le cas du traitement statistique des fichiers d'enquêtes en vraie grandeur, les inconvénients théoriques et pratiques d'une démarche limitée aux méthodes de type factoriel sont particulièrement importants : ils concernent à la fois la nature même des résultats, et la gestion du volume de ces résultats. On insistera ici sur deux points :

a) Les visualisations sont souvent limitées à un petit nombre de dimensions, deux le plus souvent (premier plan factoriel), alors que la dimension réelle du phénomène étudié peut être bien supérieure (cette dimension est mesurée par le nombre d'axes significatifs[1]).

[1] On appelle ici *axe significatif,* au sens statistique du terme, un axe qui explique une part de dispersion ne pouvant être imputée au hasard. On peut déterminer le nombre d'axes significatifs en utilisant une procédure de simulation.

b) Ces visualisations peuvent inclure des centaines de points, et donner lieu à des graphiques chargés ou illisibles, à des listages de coordonnées encombrants, peu synthétiques.

Il faut donc à ce stade faire appel de nouveau aux capacités de gestion et de calcul de l'ordinateur pour compléter, alléger et clarifier la présentation des résultats. L'utilisation conjointe de méthodes de classification automatique adaptées à ce type de problème et des analyses de correspondances permet de remédier à ces lacunes.

Lorsqu'il y a trop de points sur un graphique, il paraît utile de procéder à des regroupements en familles homogènes. Mais les algorithmes utilisés pour ces regroupements fonctionnent de la même façon, que les points soient situés dans un espace à deux ou à dix dimensions.

Autrement dit, l'opération va présenter un double intérêt : allégement des sorties graphiques d'une part, prise en compte de la dimension réelle du nuage de points d'autre part.

Une fois les individus regroupés en classes, il est facile d'obtenir une description automatique de ces classes : on peut en effet, pour les variables numériques comme pour les variables nominales, calculer des statistiques d'écarts entre les valeurs internes à la classe et les valeurs globales ; on peut également convertir ces statistiques en *valeurs-test* et opérer un tri sur ces valeurs-test. On obtient finalement, pour chaque classe, les modalités et les variables les plus caractéristiques.

Après quelques mots sur les algorithmes de classification adaptés aux grands ensembles de données, on présentera un exemple d'application qui prolongera l'analyse des correspondances multiples présentée au chapitre précédent (paragraphe 4.3), et qui illustrera cette nouvelle présentation compacte de l'information.

4.3.1 Les algorithmes de classification mixte

L'algorithme de classification qui nous paraît le plus adapté au partitionnement d'un ensemble comprenant des milliers d'individus est un *algorithme mixte* procédant en trois phases :

a) *Partitionnement initial* en quelques dizaines de classes par une technique du type "nuées dynamiques" ou "k-means", Diday (1971).
On peut résumer sans trop les trahir ces techniques en quelques phrases. On commence par tirer au hasard des individus qui seront des centres provisoires de classes. Puis, on affecte tous les individus au centre provisoire le plus proche (au sens d'une distance telle que la distance du

chi-2 définie au chapitre précédent). On construit ainsi une partition de l'ensemble des individus. On calcule de nouveau des centres provisoires, qui sont maintenant les "centres" (points moyens par exemple) des classes qui viennent d'être obtenues, et on réitère le processus, autrement dit on affecte de nouveau tous les individus à ces centres, ce qui induit une nouvelle partition, etc. Le processus se stabilise nécessairement, mais la partition obtenue dépendra en général du choix initial des centres.[1]

b) *Agrégation hiérarchique des classes obtenues* : Les techniques de classification hiérarchique présentées dans ce chapitre sont assez coûteuses si elles s'appliquent à des milliers d'éléments, c'est pourquoi il faut réduire la dimension du problème en opérant un regroupement préalable en quelques dizaines de classes.

L'intérêt du dendrogramme obtenu est qu'il peut donner une idée du nombre de classes existant effectivement dans la population. Chaque coupure d'un tel arbre va donner une partition, ayant d'autant moins de classes que l'on coupe près du sommet.

c) *Coupure de l'arbre* (en général après une inspection visuelle)

Plus on agrège de points, autrement dit plus on se rapproche du sommet de l'arbre, plus la distance entre les deux classes les plus proches est grande. On a vu (paragraphe 4.1.2) qu'en coupant l'arbre au niveau d'un saut important de l'indice, on peut espérer obtenir une partition de bonne qualité, car les individus regroupés auparavant étaient proches, et ceux regroupés après la coupure seront nécessairement éloignés, ce qui est la définition d'une bonne partition.

d) *Optimisation de la partition obtenue par réaffectations*

Une partition obtenue par coupure n'est pas la meilleure possible, car l'algorithme de classification hiérarchique n'a malheureusement pas la propriété de donner à chaque étape une partition optimisée. On peut encore améliorer la partition obtenue par réaffectation des individus comme indiqué au paragraphe a) ci-dessus. Malgré la relative complexité de la procédure, on ne peut être assuré d'avoir trouvé la "meilleure partition en k classes".

[1] La recherche d'une partition optimale vis-à-vis d'un critère tels que ceux que l'on utilise habituellement en statistique (par exemple maximisation du quotient variance externe / variance interne) se heurte encore à l'*infini combinatoire* et sa mise en évidence n'est actuellement pas possible même sur les ordinateurs les plus puissants.

4.3.2 Séquence des opérations en traitements d'enquêtes

Les analyses complètes incluant la phase de filtrage par analyse des correspondances se dérouleront en définitive de la façon suivante :

a) Choix des éléments actifs (qui correspond au choix d'un point de vue). On peut décrire les individus du point de vue de leurs caractéristiques de base, mais aussi à partir d'un thème particulier : habitudes de consommation, opinions politiques, etc.

b) Analyse des correspondances simples ou multiples à partir de ces éléments actifs.

c) Positionnement des éléments illustratifs. On projettera ainsi toute l'information disponible de nature à comprendre ou à interpréter la typologie induite par les éléments actifs.

d) Examen des premiers graphiques de plans factoriels en général limités aux points occupant les positions les plus significatives.

e) Partition de l'ensemble des individus ou observations selon la méthode qui vient d'être exposée.

f) Position sur les graphiques précédents des centres des principales classes (une partition définit toujours une variable nominale particulière).

g) Description des classes par les modalités et les variables les plus caractéristiques.

4.3.3 Exemple d'application : Situations-types ou noyaux factuels

Cet exemple prolonge l'analyse des correspondances multiples du chapitre précédent. Les 144 individus enquêtés sont décrits par le même ensemble de variables actives. On veut maintenant obtenir un petit nombre de groupes d'individus les plus homogènes possible vis-à-vis de leurs caractéristiques de base. L'intérêt de tels regroupements apparaîtra au chapitre 5, lorsqu'il s'agira d'agréger les réponses libres sans privilégier de critères particuliers.

On aimerait pouvoir croiser des caractéristiques telles que l'âge, le sexe, la profession, le niveau d'instruction, de façon à étudier des groupes d'individus tout à fait comparables entre eux du point de vue de leur situation objective, c'est-à-dire de réaliser, dans les limites du possible, le *toutes choses égales par ailleurs,* situation idéale, mais hors de portée. En pratique, de tels croisements conduisent vite à des milliers de modalités, dont on ne sait que faire lorsque l'on étudie un échantillon de l'ordre de 1 000 individus. De plus, les croisements ne tiennent pas compte du réseau d'interrelations existant entre

ces caractéristiques : certaines sont évidentes a priori (il n'y a pas de jeunes retraités), d'autres sont également connues a priori, mais peuvent souffrir des exceptions (il y a peu d'étudiants veufs, d'ouvriers diplômés d'université), d'autres enfin ont un caractère plus statistique (il y a plus de femmes dans les catégories employés et veufs).

En pratique, on peut espérer trouver des regroupements opératoires en une vingtaine de classes pour un échantillon de l'ordre de 2 000 individus. On verra qu'un regroupement plus sommaire en cinq grandes classes n'est pas totalement dépourvu d'intérêt, en restant compatible avec le volume restreint d'un exemple pédagogique. De telles classes sont appelées *situations-types* ou encore *noyaux factuels*. Elles seront utilisées au chapitre 5 pour procéder à des regroupements de réponses libres (paragraphe 5.2).

Application à l'exemple du paragraphe 3.3.1 (chapitre 3)

La procédure de classification mixte décrite ci-dessus a été appliquée aux 144 individus partir des cinq variables actives décrites au paragraphe 3.3.1. Quatre classes ont été obtenues par coupure du dendrogramme, puis optimisation. On présentera seulement ici la phase nouvelle de description automatique des classes.

Le tableau 4.5 va en effet décrire les classes de façon précise, en comparant les pourcentages de réponses internes aux classes aux pourcentages globaux, puis en sélectionnant les modalités les plus caractéristiques (cf. guide de lecture du tableau 4.5). On note que parmi les modalités les plus caractéristiques d'une classe figurent également des modalités illustratives, qui n'ont pas participé à la fabrication des classes. Tels sont les cas des modalités décrivant des appartenances professionnelles.

On voit assez bien les avantages que présentent les partitions pour décrire des ensembles multidimensionnels :

La notion de classe est intuitive (groupes d'individus les plus semblables possible). La description des classes fait appel à des classements de libellés complets et donc facile à lire, ces classements étant fondés sur de simples comparaisons de pourcentages. Il est donc plus facile de décrire des classes qu'un espace continu.

Tableau 4.5

Exemple de description automatique de la partition en 4 noyaux factuels de l'exemple du paragraphe 3.3.1 (chapitre 3)

MODALITES CARACTERISTIQUES		POURCENTAGES CLA/MOD	POURCENTAGES MOD/CLA	POURCENTAGES GLOBAL	POIDS	V.TEST	PROBA
- CLASSE 1 / 4				45.83	66		
Niveau de diplôme	Baccalauréat	93.02	60.61	29.86	43	7.61	.000
Niveau de diplôme	BEPC	72.22	39.39	25.00	36	3.50	.000
- CLASSE 2 / 4				13.89	20		
Niveau de diplôme	Sans diplôme ou CEP	74.07	100.00	18.75	27	8.82	.000
Profession	EMPLOYE	42.86	30.00	9.72	14	2.58	.005
Age en 3 classes	Entre 30 et 50 ans	24.49	60.00	34.03	49	2.33	.010
- CLASSE 3 / 4				26.39	38		
Niveau de diplôme	Université	100.00	100.00	26.39	38	12.42	.000
Age en 3 classes	Moins de 30 ans	36.49	71.05	51.39	74	2.66	.004
Profession	CADRESUP, PROF LIB	75.00	15.79	5.56	8	2.62	.004
Profession	ETUDIANT	40.43	50.00	32.64	47	2.42	.008
- CLASSE 4 / 4				13.89	20		
Age en 3 classes	Plus de 50 ans	95.24	100.00	14.58	21	9.94	.000
Avez-vous eu des enfants	oui	28.85	75.00	36.11	52	3.59	.000
Etes vous actuellement	Veuf(ve)	100.00	20.00	2.78	4	3.45	.000
Sexe	femme	21.62	80.00	51.39	74	2.57	.005
Profession	MENAGERE	57.14	20.00	4.86	7	2.44	.007
Niveau de diplôme	BEPC	27.78	50.00	25.00	36	2.39	.008
Profession	ARTISANTS, COMMERC.	75.00	15.00	2.78	4	2.39	.009

Mais l'analyse des correspondances permet de visualiser les positions relatives des classes dans l'espace, et aussi de mettre en évidence certaines variations continues ou dérives dans cet espace qui auraient pu être masquées par la discontinuité des classes. Les deux techniques sont donc complémentaires, et se valident mutuellement.

Guide de lecture du tableau 4.5

Nous commencerons par les deux dernières colonnes, qui sont en fait les plus importantes puisqu'elles permettent de sélectionner et de classer les modalités caractéristiques de chaque classe.

Pour chacune des classes, les modalités les plus caractéristiques sont rangées suivant les valeurs décroissantes de la valeur-test V.TEST (avant dernière colonne du tableau), ou, ce qui revient au même, selon les valeurs croissantes de la probabilité PROBA (dernière colonne).

La valeur-test V.TEST est, brièvement, l'analogue de la valeur absolue d'une variable normale centrée réduite, qui est significative au seuil 5% bilatéral si elle dépasse la valeur 1.96, ce qui est le cas pour toutes les modalités retenues ici.

La première colonne numérique CLA/MOD donne les pourcentages de chaque classe dans les modalités : ainsi, tous les veufs (ves) de l'échantillon appartiennent à la classe 4 (CLA/MOD = 100).

La seconde colonne numérique donne le pourcentage interne MOD/CLA de chaque modalité dans la classe. Ainsi, il y a 20% de veufs(ves) dans la classe 4 (MOD/CLA = 20). Ici, ce pourcentage interne est nécessairement plus élevé que le pourcentage global (troisième colonne : GLOBAL) puisque les modalités sélectionnées sont caractéristiques des classes. C'est précisément la différence entre le pourcentage interne et le pourcentage global qui est à la base du calcul de la valeur-test.

Ainsi, la classe 3 contient 38 individus, soit 26% de la population. La modalité "Cadre Sup.", qui concerne 5.56 % de l'échantillon, concerne près de 15.79% des personnes de cette classe. Avec une valeur-test de 2.62, elle vient en quatrième rang pour caractériser la classe.

La troisième colonne, GLOBAL, donne, on l'a vu, le pourcentage de chaque modalité dans la population globale. Il s'obtient simplement en divisant la colonne poids par 144, effectif global de l'échantillon.

La quatrième colonne, POIDS, donne les effectifs bruts des classes et des modalités.

Chapitre 5

Typologies, visualisations

Comment appliquer dans des situations réelles les techniques multidimensionnelles définies sur des exemples d'école aux chapitres précédents ? La complexité de l'information et la multiplicité des points de vue possibles ne permettent pas de proposer une voie unique menant du problème à une solution définitive. Dans ce chapitre, nous allons au contraire tenter de reconnaître divers chemins permettant de reculer quelque peu le moment de la nécessaire intervention interprétative de l'utilisateur. En somme, on ambitionne d'étendre le domaine de l'analyse contrôlable et reproductible, ces deux vocables peu élégants étant peut-être moins polémiques que les qualificatifs *objectif* et *automatique*...

Il s'agira ici d'aider à la lecture d'une série de textes, qu'il s'agisse de textes littéraires, de documents, ou de réponses à des questions ouvertes regroupées en textes artificiels (regroupement par classes d'âge, par profession, par niveau d'instruction ou tout autre critère pouvant présenter de l'intérêt vis-à-vis du phénomène étudié).

Quels sont les textes les plus semblables en ce qui concerne le vocabulaire et la fréquence des formes utilisées (autrement dit, quels sont les textes dont les profils lexicaux sont similaires) ? Quelles sont les formes qui caractérisent chaque texte, par leur présence ou leur absence ? On reconnaît là les questions auxquelles permet de répondre l'analyse des correspondances du *tableau lexical entier* (tableau croisant formes graphiques et textes). Ce type de traitement fera l'objet du paragraphe 5.1 intitulé : *Analyse des correspondances sur tableau lexical*.

Pour le cas des réponses libres dans les enquêtes, l'approche que nous proposons présuppose que les réponses ont été préalablement regroupées. Or les critères de regroupement les plus pertinents ne sont pas connus a priori. Il n'est d'autre part pas toujours possible, compte tenu du nombre de variables pouvant servir de critère de regroupement, et donc du nombre encore plus

grand de croisements de ces variables, d'essayer toutes les combinaisons possibles.

On proposera en fait trois stratégies d'analyse lorsqu'il n'existe pas de variable privilégiée pour regrouper les textes :

a) L'utilisation d'une partition de synthèse. Une technique de classification automatique déjà esquissée au chapitre 4 qui permet de résumer en une seule partition de synthèse les différentes caractéristiques des personnes interrogées. Cette technique, dite de "partition en noyaux factuels" est exposée sur un exemple au paragraphe 5.2 intitulé : *Les noyaux factuels*.

b) Une analyse directe des réponses non regroupées. Si les réponses paraissent suffisamment riches pour être traitées isolément, une analyse directe du tableau lexical entier croisant formes graphiques et réponses peut être opérée. Une telle analyse produira une typologie des réponses, en général assez grossière, et, de façon duale, une typologie portant sur les formes. Les réponses émanent d'individus, dont les caractéristiques sont connues par leurs réponses aux autres questions posées lors de l'enquête. Il sera donc possible d'illustrer ces typologies par des caractéristiques qui auront le statut de variables supplémentaires ou illustratives. Ce traitement direct des réponses pourra conduire à la réalisation d'un post-codage partiellement automatisé. Les problèmes et les éléments de solutions relatifs à cette approche directe seront abordés au paragraphe 5.3, intitulé : *Analyse directe des réponses ou documents.*

c) Une analyse juxtaposant des tables de contingence. Dans ce cas, pour un même vocabulaire V sélectionné, on construit autant de tableaux lexicaux qu'il y a de variables nominales, et l'on soumet à l'analyse la juxtaposition de ces tableaux. Cette méthode, d'interprétation plus délicate, est esquissée à partir d'un exemple au paragraphe 5.4, intitulé : *Analyse des correspondances à partir d'une juxtaposition de tableaux lexicaux*.

Enfin, toutes ces approches qui font intervenir des tableaux impliquant des formes graphiques peuvent aussi être réalisées à partir des unités statistiques que sont les segments répétés. Le paragraphe 5.5, intitulé *Analyse des correspondances à partir du tableau des segments répétés*, reprendra l'optique du premier paragraphe en substituant aux formes les segments. Ces divers traitements seront présentés à partir d'exemples. La plupart d'entre eux se réfèrent à un même corpus de réponses à une question ouverte.

5.1 Analyse des correspondances sur tableau lexical

On a vu au cours des chapitres précédents comment les réponses pouvaient être numérisées de façon complètement "transparente" pour l'utilisateur. Le résultat de cette numérisation peut prendre deux formes différentes, matérialisées par deux tableaux **R** et **T**.

5.1.1 Les tableaux lexicaux de base

Le tableau **R** a autant de lignes qu'il existe de réponses. On note en général ce nombre de lignes par k, nombre d'individus (il peut y avoir des réponses vides, mais il est commode de réserver une ligne pour chaque individu interrogé, de façon à assurer une correspondance aisée avec les autres informations disponibles sur ces individus).

Le tableau **R** a d'autre part un nombre de colonnes égal à la longueur de la plus longue réponse (nombre d'occurrences dans cette réponse). Pour un individu i, la ligne i du tableau **R** contient les adresses des formes graphiques qui composent sa réponse, en respectant l'ordre et les éventuelles répétitions de ces formes. Ces adresses renvoient au vocabulaire propre à la réponse. Le tableau **R** permet donc de restituer intégralement les réponses originales.

En pratique, le tableau **R** n'est pas rectangulaire, car chacune de ses lignes a une longueur variable. Les nombres entiers qui composent le tableau **R** ne peuvent dépasser *V*, étendue du vocabulaire.

Le tableau **T** a le même nombre de lignes que le tableau **R**, mais il possède autant de colonnes qu'il y a de formes graphiques utilisées par l'ensemble des individus, c'est-à-dire *V* colonnes. A l'intersection de la ligne i et de la colonne j de **T** figure le nombre de fois où la forme j a été utilisée par l'individu i dans sa réponse. Il s'agit donc d'une table de contingence (réponses ∞ formes), c'est-à-dire d'un tableau lexical entier.

Le tableau **T** peut être aisément construit à partir du tableau **R**, mais la réciproque n'est pas vraie : l'information relative à l'ordre des formes dans chaque réponse est perdue dans le tableau **T.**

En fait, le tableau **R** est beaucoup plus compact que le tableau **T** : ainsi, une réponse contenant 20 occurrences (pour un lexique de 1 000 formes) correspond à une ligne de longueur 20 du tableau **R** et à une ligne de longueur 1 000 du tableau **T** (cette dernière ligne comprenant au moins 980 zéros...). Les traitements statistiques et algorithmiques qui mettront en jeu le

tableau **T** seront en réalité programmés à l'aide du tableau **R**, moins gourmand en mémoire d'ordinateur.

5.1.2 Les tableaux lexicaux agrégés

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

Reprenant les notations du chapitre 3, on désignera par **Z** le tableau disjonctif complet à *k* lignes et *p* colonnes décrivant les réponses de *k* individus à une question fermée comportant *p* modalités de réponse possibles, ces réponses étant mutuellement exclusives. Autrement dit, une ligne de **Z** ne comportera ici qu'un seul "1", et *(p-1)* "0".

Contrairement au tableau de l'exemple du chapitre 3 (tableau 3.9), le tableau **Z** ne concerne qu'une seule question, et comporte donc un bloc unique. Chaque question fermée de ce type permet de définir une partition des individus interrogés.

Le tableau **C**, obtenu par le produit matriciel:

$$\mathbf{C} = \mathbf{T}' \mathbf{Z}$$

est un tableau à *V* lignes (*V* est, rappelons-le, le nombre total de formes distinctes) et *p* colonnes (*p* est le nombre de modalités de réponses à la question fermée considérée) dont le terme général c_{ij} n'est autre que le nombre de fois où la forme i a été utilisée par l'ensemble des individus ayant choisi la réponse j.

Un exemple de tableau **C**, de dimension modeste, est fourni par le tableau 3.1 du chapitre 3, avec $V = 14$, et $p = 5$. La question fermée était dans ce cas : *Quel est le diplôme d'enseignement général le plus élevé que vous avez obtenu ?*

Il est donc aisé, pour toute question fermée dont les réponses sont codées dans un tableau $\mathbf{Z}_q$, de calculer le tableau lexical agrégé $\mathbf{C}_q$ par la formule :

$$\mathbf{C}_q = \mathbf{T}' \mathbf{Z}_q$$

et donc de comparer les profils lexicaux de différentes catégories de population. Chaque tableau $\mathbf{C}_q$ donne un point de vue différent (le point de vue de la question fermée q) sur la dispersion des profils lexicaux des réponses à la question ouverte étudiée.

5.1.3 Seuil de fréquence pour les formes

Ces comparaisons de profils lexicaux n'ont de sens, d'un point de vue statistique, que si les formes apparaissent avec une certaine fréquence : les hapax ou même les formes rares seront écartés de la phase de comparaisons de fréquences. Ceci a pour effet de réduire considérablement la taille du vocabulaire effectivement pris en compte.

Pour une question ouverte posée à 2 000 personnes, compte tenu de la structure de la gamme des fréquences du vocabulaire, une sélection des formes apparaissant au moins 10 fois peut, dans bien des cas, ramener la valeur de V de 1 000 (pour fixer les idées) à 100...

5.1.4 Présentation de l'exemple

On reprendra, en vraie grandeur, l'exemple de réponses à une question ouverte qui a été présenté dans une version simplifiée au chapitre 3 et que nous avons appelé : la question ***Enfants***.

Numérisation du texte

Le bilan de la première phase de numérisation du texte réalisée par le programme de traitement est le suivant :

Sur 2 000 réponses, on relève 15 457 occurrences, avec 1 305 formes graphiques distinctes. Si l'on sélectionne les formes apparaissant au moins 20 fois, il reste 12 051 occurrences de ces formes, qui sont au nombre de 117.

Le tableau 5.1 donne la liste alphabétique de ces formes les plus fréquentes ; celles-ci sont ensuite classées par ordre de fréquences décroissantes dans le tableau 5.2. Cette "mise à plat" de l'information de base peut paraître surprenante. En fait, il ne s'agit que d'une étape intermédiaire, et les traitements ultérieurs vont redonner forme à cet inventaire.

5.1.5 Construction du tableau lexical agrégé

Les réponses à la question ***Enfants,*** dont on peut trouver quelques exemples au paragraphe 2.2 du chapitre 2 sont en général assez lapidaires. Dans ces conditions, deux réponses qui semblent au premier abord traduire des préoccupations similaires, telles par exemple : *manque d'argent* et *problèmes financiers* peuvent n'avoir aucune forme graphique en commun, et donc ne pas être reconnues comme proches.

Tableau 5.1

Vocabulaire des réponses à la question *Enfants* (Ordre alphabétique)

NUM.	MOTS EMPLOYÉS	FREQUENCES	NUM.	MOTS EMPLOYES	FREQUENCES
1	a	218	60	l	674
2	actuelle	83	61	la	666
3	argent	196	62	le	474
4	au	44	63	les	442
5	aucune	33	64	leur	71
6	aussi	22	65	liberte	52
7	avec	21	66	logement	52
8	avenir	318	67	maladie	38
9	avoir	77	68	manque	160
10	beaucoup	23	69	materielle	27
11	bien	21	70	materielles	57
12	c	98	71	monde	21
13	ca	41	72	moyens	44
14	ce	34	73	n	94
15	cela	23	74	ne	193
16	chomage	285	75	on	84
17	conditions	48	76	ont	24
18	conjoncture	22	77	ou	58
19	couple	95	78	par	23
20	crainte	24	79	parce	22
21	crise	25	80	pas	325
22	d	332	81	peur	160
23	dans	67	82	peut	54
24	de	915	83	plus	87
25	des	208	84	pour	161
26	deux	26	85	pouvoir	41
27	difficile	28	86	probleme	37
28	difficultes	82	87	problemes	108
29	du	155	88	professionnelle	22
30	economique	54	89	qu	51
31	egoisme	109	90	quand	28
32	elever	51	91	que	78
33	emploi	79	92	question	48
34	en	116	93	questions	23
35	enfant	148	94	qui	76
36	enfants	148	95	raison	41
37	entente	22	96	raisons	176
38	est	190	97	responsabilite	21
39	et	204	98	responsabilites	21
40	etre	69	99	ressources	49
41	faire	22	100	s	55
42	fait	25	101	sais	25
43	faut	35	102	sante	95
44	femme	60	103	se	36
45	finances	28	104	si	56
46	financier	24	105	situation	176
47	financiere	91	106	son	28
48	financieres	173	107	sont	44
49	financiers	86	108	sur	29
50	gens	25	109	surtout	34
51	guerre	27	110	tout	41
52	il	105	111	travail	152
53	ils	79	112	trop	52
54	incertain	45	113	un	172
55	incertitude	29	114	une	120
56	independance	23	115	veulent	42
57	insecurite	62	116	vie	179
58	instabilite	21	117	y	56
59	je	62			

Il s'agit là d'un problème important sur lequel on reviendra à plusieurs reprises dans ce chapitre et dans ceux qui suivent. En bref, tout dépend du degré de finesse que l'on assigne à l'analyse, et aussi de la taille de l'échantillon, indissolublement liée à cette exigence.

Tableau 5.2

Vocabulaire de la question *Enfants* (Ordre lexicométrique)

NUM.	MOTS	EMPLOYES	FREQUENCES	NUM.	MOTS	EMPLOYES	FREQUENCES
33	de		915	42	economique		54
76	l		674	82	logement		52
77	la		666	81	liberte		52
78	le		474	145	trop		52
79	les		442	109	qu		51
31	d		332	45	elever		51
98	pas		325	119	ressources		49
11	avenir		318	22	conditions		48
21	chomage		285	112	question		48
1	a		218	69	incertain		45
34	des		208	6	au		44
53	et		204	90	moyens		44
3	argent		196	135	sont		44
92	ne		193	148	veulent		42
52	est		190	115	raison		41
150	vie		179	16	ca		41
131	situation		176	140	tout		41
116	raisons		176	105	pouvoir		41
63	financieres		173	84	maladie		38
146	un		172	106	probleme		37
104	pour		161	128	se		36
85	manque		160	58	faut		35
100	peur		160	18	ce		34
41	du		155	138	surtout		34
141	travail		152	7	aucune		33
49	enfant		148	137	sur		29
50	enfants		148	70	incertitude		29
147	une		120	60	finances		28
48	en		116	37	difficile		28
44	egoisme		109	110	quand		28
107	problemes		108	134	son		28
67	il		105	66	guerre		27
15	c		98	86	materielle		27
127	sante		95	36	deux		26
26	couple		95	30	crise		25
91	n		94	56	fait		25
62	financiere		91	124	sais		25
103	plus		87	65	gens		25
64	financiers		86	29	crainte		24
93	on		84	61	financier		24
2	actuelle		83	94	ont		24
39	difficultes		82	13	beaucoup		23
68	ils		79	113	questions		23
47	emploi		79	71	independance		23
111	que		78	96	par		23
12	avoir		77	19	cela		23
114	qui		76	108	professionnelle		22
80	leur		71	51	entente		22
54	etre		69	55	faire		22
32	dans		67	24	conjoncture		22
74	je		62	97	parce		22
72	insecurite		62	8	aussi		22
59	femme		60	73	instabilite		21
95	ou		58	89	monde		21
87	materielles		57	14	bien		21
154	y		56	118	responsabilites		21
130	si		56	117	responsabilite		21
123	s		55	10	avec		21
101	peut		54				

Une analyse de contenu rapide sur petit échantillon se heurtera à des problèmes de dispersion des formes relatives à un même lemme, et aussi aux problèmes de synonymies. Il n'en est pas de même si l'on s'astreint à traiter des ensembles importants de réponses. Dans ce cas, les fréquences permettent en quelque sorte de consolider les emplois de tel terme ou

locution, et parfois d'identifier des catégories de locuteurs qui les utilisent de manière privilégiée.

Il convient donc, dans un premier temps, de trouver des regroupements d'individus pertinents vis-à-vis du phénomène étudié. On reviendra sur la façon de choisir un tel regroupement. Dans le cas particulier de l'exemple traité, les individus sont regroupés en 9 classes, qui diffèrent soit par l'âge, soit par le niveau d'instruction générale. Cette partition en 9 groupes est obtenue en croisant deux partitions élémentaires qui contiennent chacune 3 classes[1].

Tableau 5.3
Tête de liste du tableau lexical entier
croisant les 117 formes de fréquence supérieure à 20
avec la partition *âge-diplôme* en 9 groupes

	A -30	B -30	S -30	A -50	B -50	S -50	A +50	B +50	S +50
a	14	17	25	34	10	29	61	16	12
actuelle	7	6	7	18	8	5	20	8	4
argent	20	18	22	38	11	17	57	4	9
au	1	2	6	6	3	7	14	0	5
aucune	0	4	2	7	2	0	12	5	1
aussi	3	2	1	4	4	3	3	0	2
avec	2	2	2	4	1	4	6	0	0
avenir	32	40	43	58	21	38	53	17	16
avoir	12	5	11	8	2	5	25	7	2
beaucoup	0	1	1	8	2	1	10	0	0
bien	3	2	2	4	1	2	6	1	0
. . . . .									

Le tableau 5.3 est en tout point analogue au tableau 3.1 du chapitre 3, à ceci près qu'il comporte l'intégralité des formes graphiques de fréquences supérieures à 20 (117 au lieu de 14) et une partition des individus en 9 classes intégrant l'âge des répondants (au lieu d'une partition en 5 classes par niveau d'instruction).

[1] Première partition : trois classes d'âge : moins de 30 ans (notée -30), entre 30 et 50 ans (-50), plus de 50 ans (+50). Seconde partition : trois classes de niveaux d'instruction: Aucun diplôme ou Certificat d'études primaires (notée A), BEPC ou équivalent (notée B), Bac et études Supérieures (notée S).

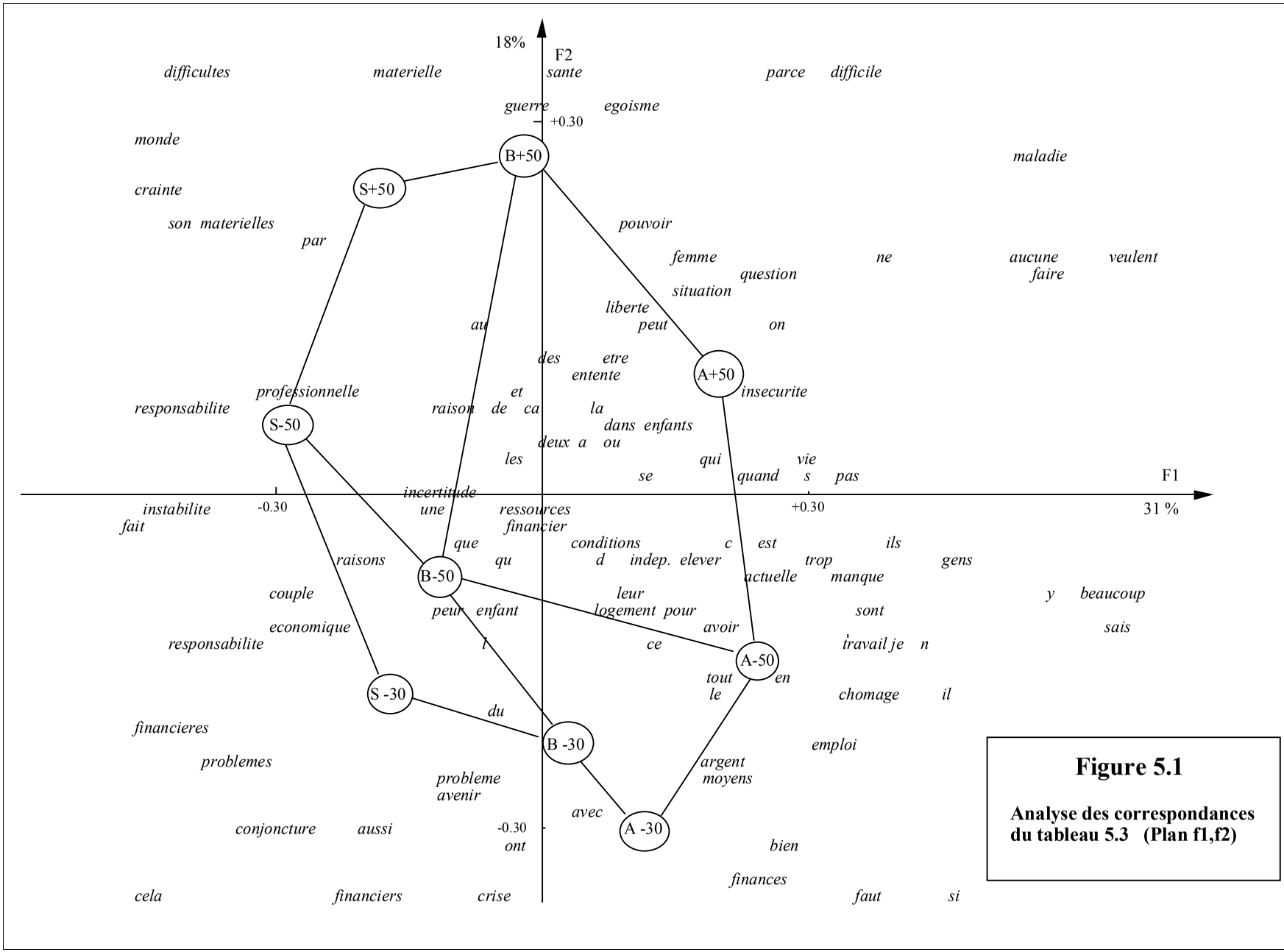


Figure 5.1

Analyse des correspondances du tableau 5.3 (Plan f1,f2)

Pour lire efficacement l'information contenue dans ce tableau, il faut calculer les tableaux de profils-lignes et de profils-colonnes, et représenter les distances entre formes d'une part, entre catégories d'âge et d'instruction, d'autre part. C'est précisément la vocation de l'analyse des correspondances que de procéder à cette double description.

5.1.6 Analyse et interprétation du tableau lexical

La figure 5.1 représente le premier plan factoriel (c'est-à-dire le plan des deux premiers facteurs) de l'analyse des correspondances du tableau 5.3. Les deux premières valeurs propres valent respectivement 0.035 et 0.021, et correspondent à 31% et 18% de la trace (ou encore inertie totale, ou variance totale).

La disposition des points-colonnes est d'une régularité assez étonnante : à partir d'une information purement lexicale (les éléments des profils-colonnes), on retrouve le caractère composite de la partition des individus en 9 classes. A âge égal, les individus sont d'autant plus instruits qu'ils sont situés vers la gauche du graphique ; à niveau d'instruction égal, ils sont d'autant plus âgés qu'ils occupent une position élevée sur l'axe vertical.

Ainsi donc, ces vecteurs décrivant pour chaque catégorie la fréquence de 117 formes (choisies à partir d'un simple critère de fréquence) permettent de reconstituer simultanément la gradation des âges, et celle du niveau de diplôme à l'intérieur de chaque classe d'âge.

L'examen des positions des points représentant les formes reste assez frustrant, à ce niveau d'analyse limitée aux seules formes, en raison de l'absence de contexte.

On devine des *je ne sais pas, je ne vois pas* du côté des personnes peu instruites et/ou âgées. Les formes *monde*, *responsabilités*, *conjoncture* sont invoquées par les personnes les plus instruites. On remarque que la forme *avenir* est résolument du côté des plus jeunes, alors que *égoïsme,* mais aussi *santé*, *maladie* sont du côté des réponses fournies par les personnes les plus âgées.

On note d'ailleurs que les trois classes "jeunes" sont plus resserrées que celles de leurs aînés, phénomène que l'on retrouve souvent dans un cadre socio-économique plus général : à l'intérieur de la classe des personnes âgées, on trouve en effet un éventail de situations (niveau de vie, niveau d'instruction) plus large que chez les jeunes.

Interprétation

On peut faire quelques remarques à ce stade:

a) L'indexation automatique des formes et les calculs de fréquences qu'elle permet ignorent délibérément de nombreuses informations de type sémantique ou syntaxique qui sont accessibles à tout lecteur des textes considérés. La synonymie n'est pas prise en compte, pas plus que l'homonymie.

   La pratique de ce type d'analyse sur des échantillons importants (ici : 2 000 réponses) montre que ces objections peuvent être assez facilement levées dans le cas de discours artificiels construits par juxtaposition de réponses et pour lesquels on recherche surtout des éléments de répétition.

   Dans ce contexte statistique, les dépouillements en formes graphiques se révèlent souvent plus intéressants que les dépouillements en lemmes. Les formes *problème* au singulier et au pluriel occupent des positions similaires sur la figure 5.1 (partie inférieure gauche), ce qui prouve qu'il n'était pas gênant de distinguer les deux formes. Le seuil de fréquence choisi (formes apparaissant plus de 20 fois) a éliminé la forme *difficultés* qui apparaît exactement 20 fois, et n'a laissé que le singulier. L'analyse avec un seuil inférieur, ou le positionnement du pluriel de ce mot en élément supplémentaire montre que les deux formes sont au contraire très distantes, (en bas à droite au singulier, en haut à gauche au pluriel) ce qui n'est pas moins intéressant, car cette opposition renvoie à des contextes d'utilisation très différents (brièvement : mentions de *difficultés matérielles* pour les personnes plutôt âgées et instruites, contre *difficulté de trouver du travail* pour des répondants sensiblement plus jeunes et moins instruits).

   A propos d'autres exemples, on a pu noter que la présence de flexions distinctes d'un même verbe (comme: *peuvent, peut, pouvoir, puisse*) et de synonymes aide à confirmer l'interprétation de certaines zones du plan factoriel : le regroupement de formes graphiques différentes possédant des affinités au plan sémantique est au contraire un critère de validation supplémentaire des résultats empiriques.

b) La prise en compte aveugle de mots-outil (comme : *de, les, quand, que, pour, ...*) n'alourdit en rien l'analyse. Ces formes n'apparaîtront dans une analyse factorielle lexicale que si leur répartition n'est pas uniforme dans les textes, autrement dit que si elles caractérisent électivement

certains discours : dans ce cas, elles méritent effectivement d'être situées par rapport aux autres formes.[1]

c) L'ordre des formes graphiques dans les réponses n'est pas pris en compte : chaque discours est pour les programmes de calcul un "sac de mots", dont seul le profil de fréquences est actuellement exploité. Cette objection est sérieuse, encore qu'un profil de fréquences se révèle à l'expérience beaucoup plus riche en information qu'on ne l'imagine a priori.

Si dans l'absolu, un profil lexical, c'est-à-dire ici une suite de 117 sous-fréquences, n'a pas grand sens, la confrontation de plusieurs profils lexicaux est au contraire riche d'enseignements dans l'optique comparative choisie ici.

Toujours dans une optique fréquentielle, la recherche des *segments répétés* permet de prendre en compte les occurrences d'unités plus riches au plan sémantique que les formes isolées. La sélection des *réponses modales* qui sera évoquée plus bas répond également à plusieurs des objections précédentes en mettant en évidence les contextes les plus fréquents de certaines de ces formes.

Stabilité vis-à-vis d'une lemmatisation interne

Les remarques a) et b) du paragraphe précédent posent la question de la stabilité des résultats en cas de suppression d'une "liste de mots-outils" et/ou de remplacement systématique des formes graphiques par les lemmes auxquelles on peut les rattacher[2].

Le temps d'une expérience, une procédure de lemmatisation interne au sous-corpus utilisé après intervention du seuil permettra de vérifier la stabilité des structures obtenues.

Il s'agit en quelque sorte de vérifier si la structure observée[3] (disposition relative des neuf points catégories sur la figure 5.1) n'est pas un artefact dû à la présence de formes grammaticales particulières, auquel cas les catégories se distingueraient avant tout par la forme de leur discours, et non par le contenu des réponses dont on peut supposer qu'il est plutôt lié aux mots pleins.
Dans la liste de formes du tableau 5.1, on a supprimé les formes suivantes :

---

[1] Une étude sur l'image de l'institution du mariage (Lebart, 1982b) a montré, par exemple, que l'adverbe *quand* est surtout utilisé lors de réponses traditionalistes (*quand on se marie, c'est pour la vie*, ou encore : *quand on se marie, c'est pour avoir des enfants).*

[2] Nous verrons au chapitre 7 une étude en grandeur réelle sur ce même sujet.

[3] Le mot anglais *pattern* serait encore plus précis.

a, au, c, ca, ce, cela, d, dans, de, des, du, en, et, l, la, le, les, n, ne, on, ou, par, parce, quand, que, qu, qui, s, se, si, sur, un, une, y

De plus, on a rassemblé en de mêmes unités les ensembles de formes suivants (la première forme de chaque ligne subsiste et remplace la ou les suivantes).

enfant	*enfants*
avoir	*ont*
finances	*financier, financière, financiers , financières*
problème	*problèmes*
matérielle	*matérielles*
raison	*raisons*
question	*questions*
responsabilité	*responsabilités*
être	*est , sont*
pouvoir	*peut*
faire	*fait*

Il s'agit en quelque sorte d'une lemmatisation interne à l'ensemble des formes sélectionné, (après constitution d'un sous-corpus par seuillage sur la fréquence des formes) et non d'une vraie lemmatisation comme celle présentée au chapitre 2 (paragraphe 2.8).

Il est clair que lorsque la lemmatisation intervient sur la totalité du texte original, elle influe également sur la constitution de l'ensemble des formes que l'on sélectionne par seuillage. En particulier, les verbes, beaucoup plus dispersés en français que les noms et les adjectifs, sont défavorisés par une sélection à partir des fréquences de formes, puisque de nombreuses flexions d'un même verbe peuvent apparaître dans le texte de départ avec une fréquence inférieure au seuil de fréquence retenu.

Néanmoins, l'optique prise pour cette expérience est celle d'un calcul de perturbation et de stabilité interne au sous-corpus sélectionné. Le vocabulaire est ici réduit de 117 formes à 68 "pseudo-lemmes". Pour cet exemple particulier, le graphique 5.1 n'est pas profondément modifié par cette diminution considérable du nombre de formes.

On a pu vérifier que les points-catégories sont disposés d'une façon qui respecte l'ordre des classes d'âge et des niveaux de diplôme. Ce qui est une présomption supplémentaire (et non une preuve) du lien entre les catégories et le contenu des réponses.

5.2 Les noyaux factuels

On a vu qu'il était souvent nécessaire de regrouper les réponses pour pouvoir procéder à des analyses de type statistique. Les profils lexicaux d'agrégats de réponses ont plus de régularité et s'interprètent plus facilement que les réponses isolées. Les exemples précédents ont montré comment ce regroupement a priori pouvait être réalisé à partir des variables retenues en fonction d'un choix d'hypothèses de départ. Mais ceci suppose une bonne connaissance a priori du phénomène étudié, situation qui n'est en général pas réalisée dans les études dites exploratoires.

La technique dite des *noyaux factuels* ou des *situations-types* brièvement exposée au chapitre 4 (exemple du paragraphe 4.3.2 relatif aux techniques de classification) va permettre de donner des éléments de réponse à ce problème.

Etant donnée une liste de descripteurs ou de variables caractérisant les individus, on se pose maintenant le problème de répartir ces individus en groupes les plus homogènes possible vis-à-vis des caractéristiques retenues... sans en privilégier aucune a priori. Comme on l'a dit au chapitre 4, il s'agit en quelque sorte d'approcher à l'intérieur des classes le *toutes choses égales par ailleurs,* si difficile à réaliser en sciences sociales. C'est précisément le type d'opération que permet d'obtenir, autant que faire se peut, un algorithme de classification, selon les modalités définies lors de l'exemple qui clôt le chapitre 4.

Un exemple

L'exemple qui suit concerne une question ouverte posée à l'issue des interviews de l'enquête déjà citée sur les conditions de vie et aspirations des Français (en 1980 et 1981), et pour laquelle on dispose de 4 000 réponses libres. Le questionnaire abordait successivement les thèmes suivants : Famille et politique familiale, Environnement, Niveau de vie, Vie au travail, quelques thèmes généraux (Science, Justice), Vie sociale, Loisirs (cf. Chapitre 2, paragraphe 2.2).

Le libellé de cette question ouverte était le suivant :

Vous venez d'être interrogés longuement sur vos conditions de vie et votre environnement. Peut-être auriez-vous aimé donner votre avis sur certains points non prévus dans ce questionnaire rigide. Avez-vous des remarques à formuler ?

On désire avoir une vue d'ensemble des réponses, mais la variété des thèmes abordés dans cette enquête, et donc l'étendue du contenu potentiel des

réponses ne permettent pas de savoir a priori quels peuvent être les critères les plus pertinents de regroupements des réponses.

A partir d'une batterie de 13 descripteurs, dont la liste figure ci-dessous, on regroupe les 4 000 individus enquêtés en 22 classes (dont une classe "résiduelle", représentant moins de 3% de l'échantillon, qui contient des questionnaires incomplets, ou un ensemble hétérogène de combinaisons de caractéristiques peu fréquentes).

Les deux dernières variables sélectionnées constituaient, à l'époque de l'enquête, d'assez bons indicateurs d'équipement pouvant remédier à certaines lacunes ou à certains biais de la variable "revenu".

Liste des descripteurs *(variables actives de la classification)*

	Nombre de modalités
Catégorie socio-professionnelle	15
Statut matrimonial	6
Diplôme d'enseignement général le plus élevé	8
Situation actuelle de la personne interrogée	8
Présence ou non d'enfants < 16 ans au foyer	2
Statut d'occupation du logement	5
Croisement sexe-activité	4
Taille d'agglomération	5
Nombre de personnes vivant actuellement au foyer	5
Croisement sexe-âge de l'enquêté	8
Revenu global du foyer	7
Possession d'un lave-vaisselle	2
Possession de plusieurs voitures	2

La figure 5.2 représente la coupure de l'arbre hiérarchique correspondant aux 22 classes finalement retenues et donne une description sommaire de la composition de chacune de ces classes.

La figure 5.3 donne une esquisse du plan factoriel que l'on obtient en soumettant à l'analyse des correspondances le tableau lexical entier croisant, en lignes les 150 formes graphiques les plus fréquentes, et en colonnes les 22 classes (ou noyaux factuels). Seules quelques formes ont été représentées sur cette figure. Le caractère composite et systématique de la partition met en évidence des particularités ou des oppositions qui auraient pu être omises à partir de regroupements plus sommaires. Les grands traits de la description fournie par cette figure auraient peut-être été saisis par des regroupements question par question.

Ainsi les personnes instruites manifestent des opinions critiques vis-à-vis du questionnaire, alors que les personnes plus âgées évoquent, en moyenne, des problèmes qui les concernent particulièrement (*retraite, impôts*, etc).

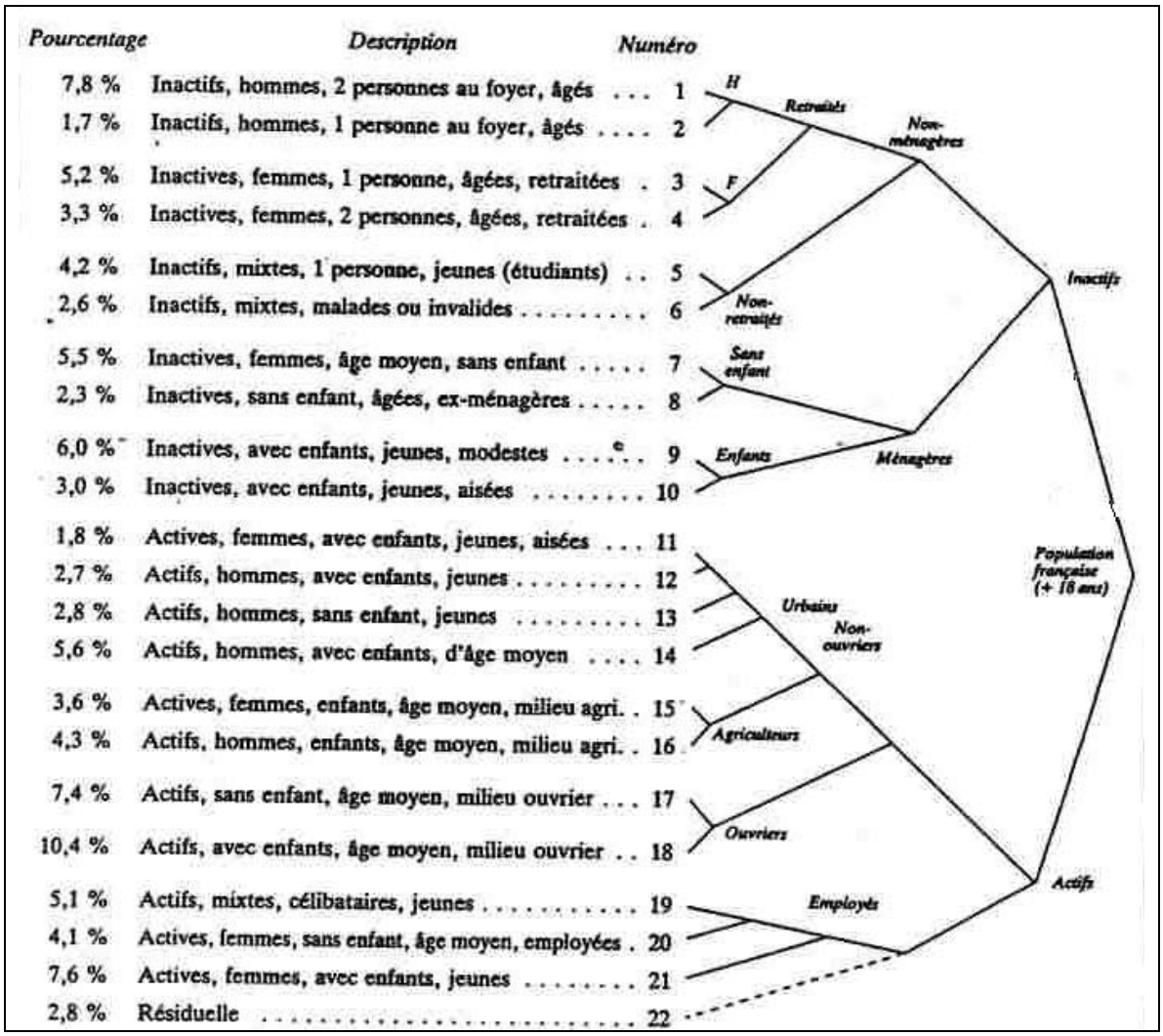


Figure 5.2

Partition en 22 noyaux factuels d'un échantillon de 4 000 personnes

Mais des nuances plus difficiles à saisir peuvent apparaître. On note ainsi, par exemple, une différence considérable de profils lexicaux entre les ouvriers de sexe masculin selon qu'ils sont sans enfant (classe 17), ou avec enfants (classe 18).

Ces derniers,visiblement intéressés et motivés, posent des problèmes non abordés dans le questionnaire, ou prolongent des débats amorcés lors de l'interview, alors que les premiers se contentent de quelques remarques assez laconiques sur le questionnaire.

Ce sont les regroupements des 22 classes observables sur la figure 5.3 qui donnent l'idée des variables sous-jacentes. Les noyaux 1, 2, 3, 4, 8 en bas à

droite sont tous composés de personnes âgées, alors que les noyaux 12 et 13 sont formés de jeunes hommes actifs.

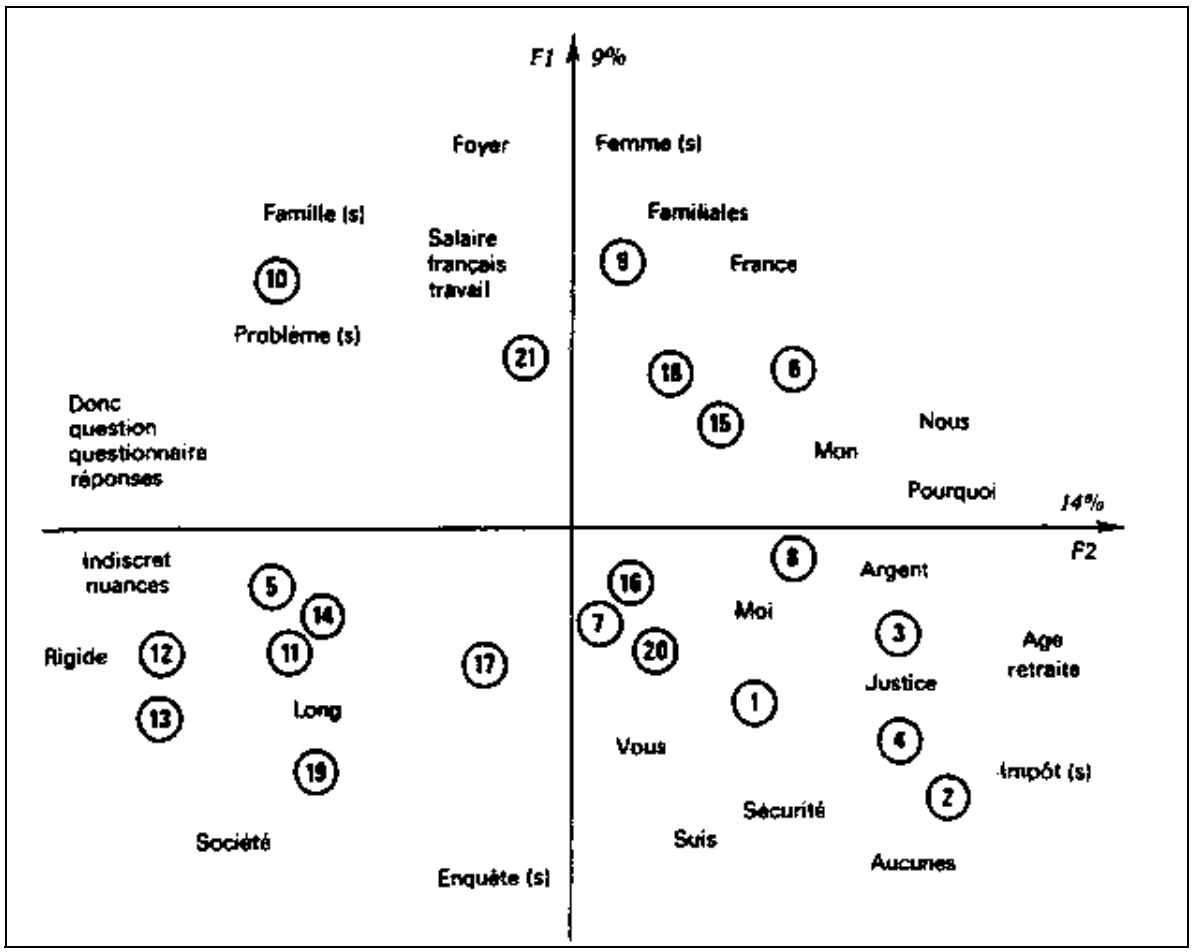


Figure 5.3

Proximités entre les 22 noyaux factuels et quelques formes (Question: *remarques à l'issue de l'interview"*)

Les classes 9, 10 et 21, constituées de femmes, évoquent souvent les lacunes du questionnaire sur certains problèmes de politique familiale. A propos de cette même question ouverte, la technique des réponses modales présentée au chapitre 6 va permettre de situer dans leurs contextes les formes représentées sur cette figure, et d'enrichir considérablement l'interprétation de ces classes.

5.3 Analyse directe des réponses ou documents

Une part importante de notre travail a été consacrée jusqu'ici à l'analyse des profils de fréquences de pour des textes relativement importants du point de vue de leur longueur.
Pour obtenir de tels textes à partir des fichiers de réponses libres, nous avons dû procéder à des regroupements a priori de ces réponses.

Mais certaines des méthodes d'analyse statistique exploratoire présentées aux chapitres 3 et 4 peuvent être appliquées avec profit aux réponses individuelles. Ce traitement direct des données individuelles est recommandé dans les deux cas suivants :

a) Lorsque les réponses sont suffisamment riches du point de vue lexical pour que les profils de fréquence puissent être comparés avec profit, ce qui peut être le cas dans les interviews prolongées, dans le domaine psychologique ou médical par exemple, ou encore lors du traitement d'unités textuelles présentant un certain volume dans le domaine de l'étude des textes socio-politiques.

b) Lorsque l'on veut procéder à un travail préliminaire de description, de regroupement et de mise en ordre.

Il est clair que le premier cas se situe dans le cadre des méthodes préconisées aux paragraphes précédents : une description directe des réponses est maintenant possible. Elle ne ferme d'ailleurs pas la porte définitivement à des regroupements ultérieurs, si cela peut aider l'interprétation ou permettre d'éprouver certaines hypothèses.

Le second cas est sensiblement différent : la notion de profil n'a plus le même sens. En termes statistiques, la variance inter-individus n'a plus le même statut que la variance inter-catégories. Les réponses se distinguent alors plus par la présence ou l'absence de formes que par de véritables variations entre profils de fréquences.

5.3.1 Comment interpréter les distances ?

On commencera par prendre un exemple simple de réponses libres à une question sur la sécurité routière, dont le libellé est le suivant :

Après une question fermée préliminaire :

A votre avis, est-il possible de diminuer fortement le nombre des tués et des blessés dans les accidents de la route ? (réponses: *oui/non*), on demande à ceux qui ont répondu *oui* (environ 80% des enquêtés en 1982) : *Que faut-il faire pour cela ?*
On peut alors trouver des réponses du type[1] :

— *développer l'usage des transports en commun*

ou encore :

— *inciter les gens à se servir le plus possible du train, du bus.*

[1] Cf. Boscher, (1984).

qui sont deux réponses ayant respectivement 7 et 13 occurrences de formes, sans aucune forme graphique commune, et dont les contenus sont pourtant assez voisins.

Inversement, les deux réponses suivantes à la même question :

— *respecter les limitations de vitesse*
— *faire respecter les limitations de vitesse*

ne se distinguent que par une seule forme graphique et ont cependant des contenus sensiblement différents.

Cet exemple correspond à une situation réelle, mais plutôt exceptionnelle. Lors des traitements de tableaux lexicaux agrégés, les réponses de ce type sont en général noyées dans des classes dont le profil lexical moyen possède, lui, une certaine régularité. Quoi qu'il en soit, ces exemples montrent que les distances entre réponses individuelles ne pourront pas être interprétées facilement.

Notre préoccupation n'est pas cependant de procéder à un décodage exhaustif de l'information, mais d'utiliser les redondances et répétitions, lorsqu'elles existent, pour simplifier cette information.

Il est clair qu'une analyse directe, dans ces conditions, pourra au moins regrouper les réponses identiques ou similaires, en laissant dans un premier temps "non-classées" les réponses que distingue l'originalité de leur forme. Il s'agit en somme d'une aide au post-codage, opération manuelle évoquée au chapitre 1, paragraphe 1.4.4.

5.3.2 Analyse du tableau clairsemé T

Cette analyse directe des réponses revient à soumettre à une analyse descriptive le tableau **T** défini plus haut (paragraphe 5.1), dont les k lignes sont les réponses et les V colonnes correspondent à un ensemble de formes.[1] Ceci appelle plusieurs remarques.

a) La proximité entre deux formes, c'est-à-dire entre deux colonnes du tableau **T** sera d'autant plus grande que ces formes apparaîtront souvent dans les mêmes réponses (et non plus seulement dans les mêmes textes ou agrégats de réponses), ce qui permettra de mieux appréhender les

[1] Le traitement d'un tableau aussi grand impliquera en général la mise en oeuvre d'algorithmes de calcul particuliers, utilisant le tableau réduit **R** au lieu du tableau **T,** et évitant le calcul et le stockage d'une matrice à diagonaliser d'ordre (V,V) (cf. par exemple, Lebart, 1982a).

voisinages syntagmatiques. L'analyse directe rendra mieux compte des contextes.

b) Les formules de transition du chapitre 3 (paragraphe 3.2) donnent aux représentations factorielles issues de l'analyse des correspondances de la matrice **T** une interprétation assez simple : les *k* individus seront approximativement aux centres de gravité des formes qu'ils auront utilisées, et réciproquement, ces formes seront approximativement positionnées aux centres de gravité des individus qui les auront choisies pour s'exprimer (nous disons approximativement pour rappeler la présence d'un coefficient dilatateur β pour chaque axe, qui déplace la position des centres de gravités...).

c) Les *k* lignes du tableau **T** représentent les réponses ou les individus interrogés. Les réponses aux questions fermées du questionnaire des mêmes *k* individus peuvent constituer les colonnes d'un tableau $\mathbf{T}^{+}$, et donc être positionnées comme éléments illustratifs (ou supplémentaires) sur les plans factoriels issus de l'analyse des correspondances de **T**. Ceci permet une visualisation très rapide des proximités entre les mots et les caractéristiques des répondants. En pratique, cela suggère des critères de regroupement.

Les modalités pratiques de ce type d'approche exploratoire seront exposées à partir d'un exemple.

5.3.3 Exemple d'application

Ici encore, on reprendra l'ensemble des 2 000 réponses à la question ouverte ***Enfants***.

Le tableau **T** compte ici *k*= 2 000 lignes et 117 colonnes (liste des formes figurant dans le tableau 5.1 au début de ce chapitre). Notons qu'un tel tableau, appelé *tableau clairsemé*, contient 95% de valeurs nulles, puisque les quelques 12 000 occurrences se répartissent dans plus de 234 000 cases !
La figure 5.4 représente le plan des deux premiers facteurs issus de l'analyse des correspondances du tableau **T**.
Les deux mille points-lignes correspondant à des individus anonymes ne figurent pas sur ce graphique, mais le tableau 5.4 nous donne les coordonnées, sur les mêmes axes factoriels, de certaines caractéristiques des répondants. C'est la figure 5.5 qui donnera, pour ce même plan factoriel, les positions de ces caractéristiques, projetées a posteriori en tant que modalités illustratives.

Tableau 5.4

Positionnement des modalités illustratives dans le plan de l'analyse directe des réponses

MODALITES	*EFF.*	*DISTO*	*COORDONNEES*		*VALEURS-TEST*	
			1	*2*	*1*	*2*
Genre (Sexe)						
homme	963	1.20	.07	.04	6.1	3.8
femme	1037	.83	-.05	-.03	-6.1	-3.8
statut matrimonial						
célibataire	348	5.21	.06	.13	2.8	5.6
marié(e)	1191	.59	-.02	-.04	-3.3	-5.2
cohabitation mar.	127	19.02	.27	.10	6.3	2.3
divorcé(e)	119	15.82	.24	-.02	6.3	-.5
veuf/ve	215	9.05	-.22	.01	-7.6	.3
instruction						
sans diplôme	435	4.62	-.20	-.01	-9.3	-.5
cep	642	1.96	-.09	-.07	-6.4	-5.1
bepc	396	4.01	.13	.02	6.5	1.0
bacc	339	4.73	.18	.09	8.6	4.4
université	180	8.29	.08	.05	2.7	1.7
autres diplômes	8	476.82	-.50	-.02	-2.3	-.1
age en 3 classes						
moins de 30 ans	509	3.02	.08	.15	4.9	9.0
de 30 à 50 ans	705	1.80	.08	-.04	5.9	-2.7
plus de 50 ans	786	1.54	-.12	-.06	-10.2	-5.3

Il convient cependant d'effectuer au préalable une préparation du tableau lexical clairsemé. Certaines réponses ne contiennent aucune des formes sélectionnées au seuil de fréquence que l'on a fixé, et s'éliminent d'elles-mêmes ; d'autres ne contiennent qu'une seule de ces formes : l'analyse des correspondances a alors pour propriété de les isoler sur un axe qui aura une valeur propre voisine de 1, voire égale à 1 si cette forme apparaît exclusivement dans des réponses où elle est isolée. On restreint dans la pratique la typologie aux réponses ayant au moins deux formes sélectionnées.

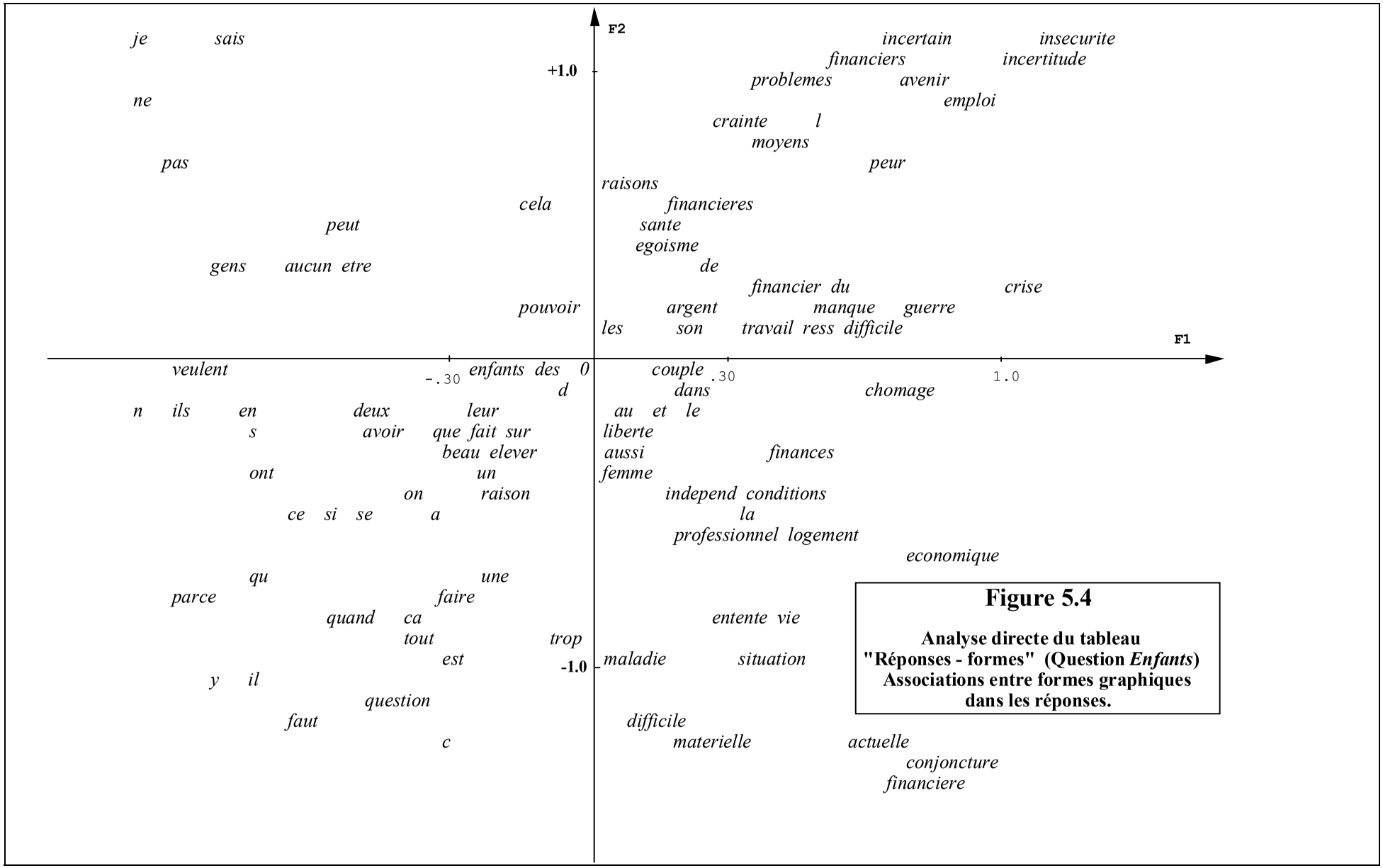


Figure 5.4

Analyse directe du tableau "Réponses - formes" (Question *Enfants*) Associations entre formes graphiques dans les réponses.

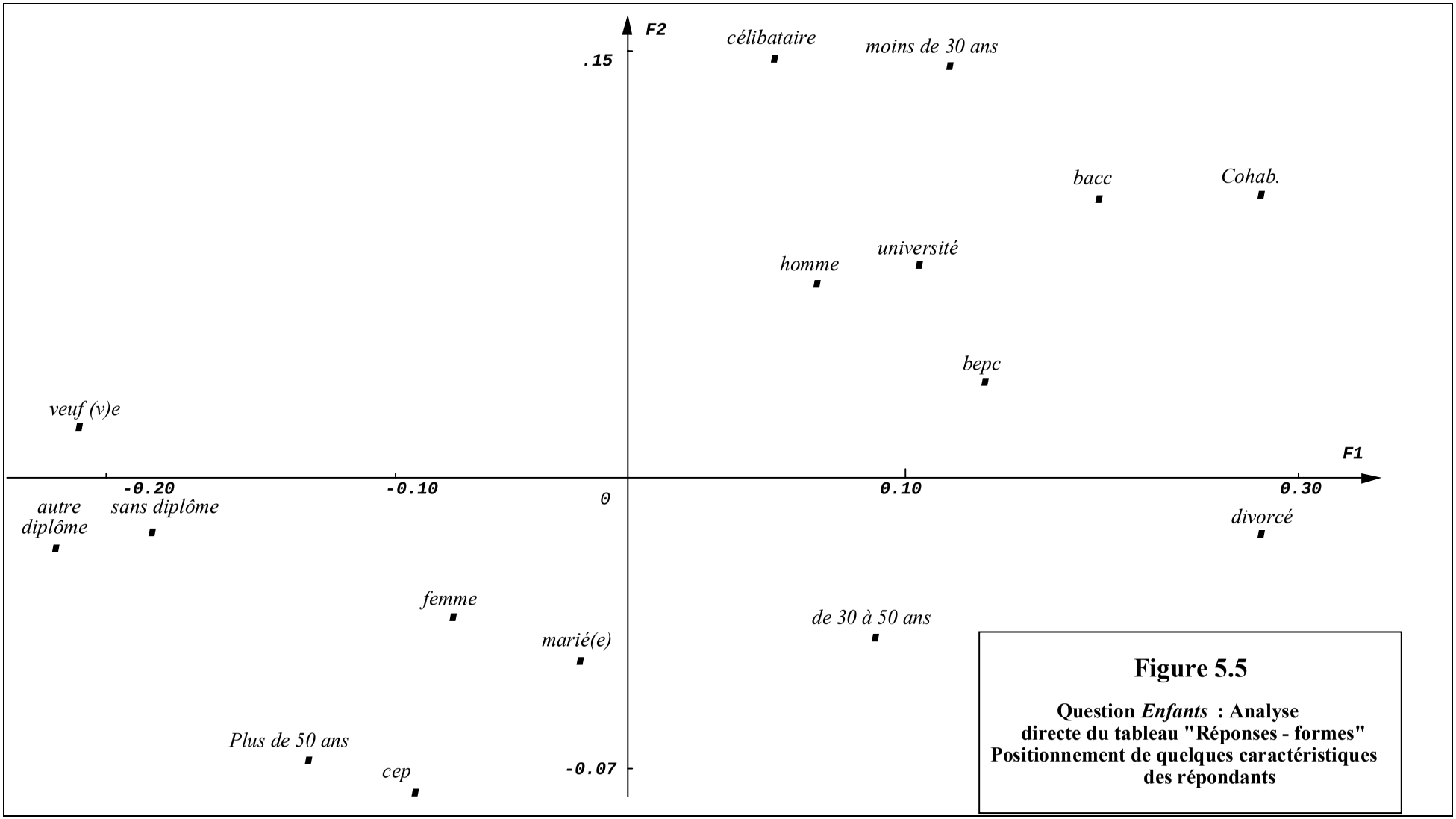


Figure 5.5

Question *Enfants* : Analyse directe du tableau "Réponses - formes" Positionnement de quelques caractéristiques des répondants

Lecture de la figure 5.4

Les proximités entre formes traduisent maintenant des cooccurrences de ces formes dans les réponses (et non plus dans des parties de textes ou dans des agrégats de réponses) ; on ne s'étonnera donc pas de voir grossièrement reconstitués certains éléments de phrases sur le graphique.

On lit ainsi sur la partie supérieure gauche de la figure 5.4 : *je ne sais pas,* ou encore, plus bas *ils ne veulent pas*, ou *ils veulent* et/ou peut-être *ils peuvent*.

On note en effet une certaine concentration de "formes à caractère grammatical" (appelées parfois mots-outil) dans la partie gauche du graphique, qui apparaît comme plus caractéristique des personnes âgées peu instruites, avec, comme conséquence de la structure démographique de la population interrogée, plus de femmes (figure 5.5). Cette concentration semble traduire des difficultés d'expression, des hésitations ou des réticences à répondre, plus fréquentes dans ces dernières catégories.

On note aussi qu'une partie des liaisons observées traduisent simplement le fait que les réponses sont soumises aux règles de la syntaxe, (accords entre substantifs et adjectifs au singuliers ou au pluriel, par exemple), et donc que les proximités entre formes ne dépendent pas seulement du contenu des réponses. Mais ceci n'est que partiellement vrai, car le tableau 5.4 et la figure 5.5 montrent une très forte polarisation socio-démographique de ce plan factoriel, et donc que, comme c'est souvent le cas lors de l'étude quantitative des matériaux textuels, la forme et le fond ne peuvent être facilement dissociés au niveau statistique.

En fait, dans l'analyse d'un tableau clairsemé comme le tableau **T,** les premiers axes factoriels n'épuisent qu'une faible part de l'information contenue dans le tableau de départ.

Il s'agit là d'un résultat général pour ce type de codage. Il faut cependant se garder d'interpréter avec pessimisme les faibles taux d'inertie obtenus sur les premiers axes car l'inertie totale, dans ce cas, ne constitue pas la seule information de référence. Elle contient un "bruit" incompressible (au sens de la théorie du signal), imputable au caractère clairsemé du tableau.

La décroissance des valeurs propres est très faible ($\lambda_1 = 0.418$, $\lambda_2 = 0.310$, $\lambda_3 = 0.291$, $\lambda_4 = 0.287$, etc.

Les pourcentages d'inertie correspondant valent $\tau_1 = 3.23$, $\tau_2 = 2.40$, $\tau_3 = 2.25$, $\tau_4 = 2.22$, ...). Les dix premiers axes ne restituent que 21% de l'inertie totale.

On n'obtient donc pas, dans ce cas, une synthèse visuelle, comme dans celui des tableaux agrégés (cas des figures 5.1, 5.3) mais un "épluchage" progressif de l'information.

On notera que l'échelle de la figure 5.5 est près de dix fois plus petite que celle de la figure 5.4 : autrement dit, portées sur la figure 5.4, ces caractéristiques seraient toutes concentrées près de l'origine des axes.

L'explication de ce phénomène est simple : les réponses du type *je ne sais pas*, par exemple, sont souvent le fait de personnes âgées, mais il s'en faut de beaucoup que toutes les réponses de ces personnes se réduisent à ces quelques formes. D'où une position beaucoup moins excentrée mais néanmoins très significative, du point qui représente cette catégorie.

Il est facile de sélectionner les modalités supplémentaires (ou illustratives) les plus significatives, au sens statistique de ce terme. Prenons l'exemple du graphique (non représenté ici, puisque les répondants sont anonymes) des 2 000 points-individus ou points-lignes sous-jacents à la figure 5.4. On sait que le point représentatif de la catégorie supplémentaire "personnes de plus de 50 ans" est le centre de gravité (dilaté) des points-individus concernés.

Les coordonnées de ce pseudo-centre de gravité figurent sur la dernière ligne du tableau 5.4. Pour plus de clarté, ces centres de gravité ont été portés sur une figure séparée, la figure 5.5. Dans l'hypothèse où ces individus pourraient être considérés comme tirés au hasard parmi les 2 000 personnes enquêtées (hypothèse qui dénoterait ici un vocabulaire indifférencié pour cette catégorie), ce centre de gravité partiel ne doit pas trop s'éloigner du centre de gravité du nuage (origine des axes factoriels).

On peut convertir cette distance au centre de gravité en "valeur-test", qui sera alors la réalisation d'une variable normale centrée réduite (deux dernières colonnes du tableau 5.4). Autrement dit, dans l'hypothèse d'un tirage au hasard, la valeur-test d'une catégorie supplémentaire a 95 chances sur 100 d'être comprise entre -1.96 et +1.96 . Comme on le lit sur le tableau 5.4, la valeur-test du point "plus de 50 ans" sur l'axe horizontal est de -10.2 ! C'est ce qu'il est convenu d'appeler une modalité hautement significative. Toutes les modalités ont d'ailleurs, pour ce qui concerne notre exemple, des valeurs-test significatives sur le premier axe, et presque toutes sur le second.

La consultation séquentielle, pour les couples d'axes factoriels successifs, des graphiques analogues des figures 5.4 et 5.5 permet donc de dégager systématiquement les formes ou groupes de formes choisis par les mêmes répondants, et d'identifier ces répondants par leurs caractéristiques. Les caractéristiques ainsi mises en évidence permettront ensuite de procéder à des regroupements et d'obtenir des visualisations plus synthétiques.

Une application directe des techniques de classification sera également très profitable à ce stade exploratoire. Elle peut alors constituer la première étape d'un *post-codage assisté*, qui consiste à regrouper les réponses similaires de façon à réduire les dimensions du problème.

Mais il sera indispensable de prendre alors en compte l'information de type segmentale à laquelle sera consacré le paragraphe 5.5.

5.4 Analyse des correspondances à partir d'une juxtaposition de tableaux lexicaux

Rappelons les différents points de vue adoptés jusqu'ici pour les analyses de réponses à des questions ouvertes ou plus généralement de textes courts et nombreux à propos desquels existe une information externe.

La première étape (paragraphe 5.1) a permis de procéder à un regroupement a priori, à partir d'une partition de l'ensemble des textes considérée comme privilégiée.

La seconde (paragraphe 5.2) suggère de construire une partition synthétique et polyvalente à partir des variables externes disponibles, en l'absence de variable privilégiée a priori.

La troisième (paragraphe 5.3), toujours à cause de l'absence d'un critère de regroupement privilégié, montre qu'une analyse directe des réponses, malgré le *bruit* considérable qu'elles contiennent, et le peu de signification au plan statistique des distances inter-individuelles, permet de mettre en évidence certains traits structuraux, et de sélectionner pour un éventuel regroupement ultérieur des variables externes qui leur sont probablement liées.

La quatrième étape, abordée dans le présent paragraphe, sera consacrée à l'analyse, non plus d'une table de contingence, mais d'une juxtaposition de tables de contingences. Cette juxtaposition est obtenue de la façon suivante : en lignes figurent toujours les unités statistiques de base (formes, segments, ou lemmes), en colonne figurent les partitions juxtaposées correspondant à différentes variables.

Il ne s'agit pas ici d'une partition de synthèse car il y a simplement juxtaposition et non croisement. Les distances entre formes sont donc des distances moyennes, pour lesquelles chacune des partitions a la même importance.[1]

[1] Cette stratégie d'analyse proposée par Benzécri, a été implémentée dans le logiciel SPAD (1984) (procédure TALEX), puis dans SPADT (1988). Son intérêt dans le cas ou

Il faut donc que ces partitions ne soient pas trop hétérogènes, pour que l'interprétation des proximités entre formes reste possible

Comme dans les sections précédentes 5.1 et 5.3, on présentera un exemple d'application relatif à la même question et la même enquête, de façon à permettre des comparaisons avec dans un cadre maintenant familier pour le lecteur. Le seuil de fréquence sera plus bas (14 au lieu de 20) ce qui donne 154 formes au lieu des 117 sélectionnées dans les tableaux 5.1 et 5.2. En effet, comme les catégories sont juxtaposées et non croisées, les effectifs des regroupements sont importants, et il est alors possible de travailler sur des fréquences plus faibles.

Les questions qui définiront les partitions à juxtaposer seront précisément celles qui figurent en lignes du tableau 5.4 : *genre, statut matrimonial, niveau d'instruction, âge en trois classes*. Bien entendu, il est possible de croiser certaines d'entre elles si l'on remarque lors d'une phase de dépouillement préalable que les interactions correspondantes présentent un intérêt particulier.

Tableau 5.5

Premières lignes de la juxtaposition de tables de contingence
***En ligne* : les 16 premières formes**
***En colonne* : 4 partitions de l'échantillon**

	Genre		Statut Matrimonial				Instruction /Diplômes						Age		
	H	F	CE	MA	CO	DI	VE	SA	CE	BE	BA	UN	AU	–30	–50	+50
a	99	119	39	134	12	9	24	46	63	43	46	19	1	56	73	89
actuelle	32	51	15	54	2	5	7	15	30	22	14	2	0	20	31	32
argent	95	101	45	103	11	15	22	51	64	33	29	17	2	60	66	70
au	19	25	11	26	4	1	2	8	13	5	11	7	0	9	16	19
aucune	20	13	2	26	1	0	4	8	11	11	2	1	0	6	9	18
aussi	12	10	4	13	4	0	1	2	8	6	4	2	0	6	11	5
avec	12	9	3	16	0	0	2	1	11	3	1	5	0	6	9	6
avenir	154	164	66	192	19	23	18	53	90	78	75	22	0	115	117	86
avoir	27	50	17	50	1	0	9	12	33	14	12	6	0	28	15	34
beaucoup	9	14	3	17	0	0	3	5	13	3	0	2	0	2	11	10
bien	7	14	4	11	1	1	4	5	8	4	4	0	0	7	7	7
c	44	54	15	61	2	6	14	23	35	17	14	8	1	24	34	40
ca	23	18	7	24	3	3	4	8	15	4	11	3	0	8	17	16
ce	13	21	4	22	2	2	4	3	16	6	3	6	0	8	14	12
cela	11	12	9	14	0	0	0	2	5	5	4	7	0	11	7	5
chomage	125	160	41	184	14	16	30	72	111	50	40	11	1	79	88	118
.....																

les partitions sont constituées par de nombreuses questions fermées a été souligné par Cibois (1990).

Le tableau 5.5 est un extrait (16 premières lignes sur 154) de la table qui sera soumise à l'analyse des correspondances. Chaque occurrence d'une forme apparaît quatre fois (une fois dans chaque sous-tableau). On voit que ce type d'analyse peut être intéressant lorsque l'on a affaire à des petits échantillons, car les profils-colonnes restent fondés sur des comptages importants.

Enfin, la Figure 5.6 donne le plan principal de visualisation obtenu par l'analyse de ce tableau composite. Le premier axe correspond à 34% de l'inertie, le second à 13%.

Lecture de la figure 5.6

Les résultats présentés ici contiennent moins de détails que les résultats correspondants représentés par la figure 5.1, en ce qui concerne les effets séparés de l'âge et du niveau d'instruction, mais le graphique est cependant très riche : la présence du statut matrimonial et du genre (sexe) est une information intéressante.

Le fait que le niveau d'instruction soit plus détaillé est également appréciable, car les diplômes *Bacc* et *Université*, regroupés précédemment, sont distants sur la figure 5.6, comme d'ailleurs les rubriques *Cep* et *Sans diplôme*.

On remarque par exemple que les cinq formes construites à partir du vocable *finance* attestées dans le corpus (finances, financier, financiers, financières, financière) se regroupent dans le même quadrant, avec la constellation (bacc, célibataire, cohabitation, moins de 30 ans, divorcé(e)), mais également avec les couples singulier/pluriel (*problème, problèmes* et *responsabilité, responsabilités*). Ces formes étaient plus dispersées précédemment.

On retrouve l'expression particulière des personnes âgées et peu instruites, et la plupart des grandes oppositions observables lors des visualisations précédentes.

On peut aussi noter la stabilité de certains traits au travers de différentes enquêtes (ici par exemple le fait que l'emploi de la forme *enfant* au singulier correspond à un niveau d'instruction plus élevé en moyenne que l'emploi du pluriel *enfants*).

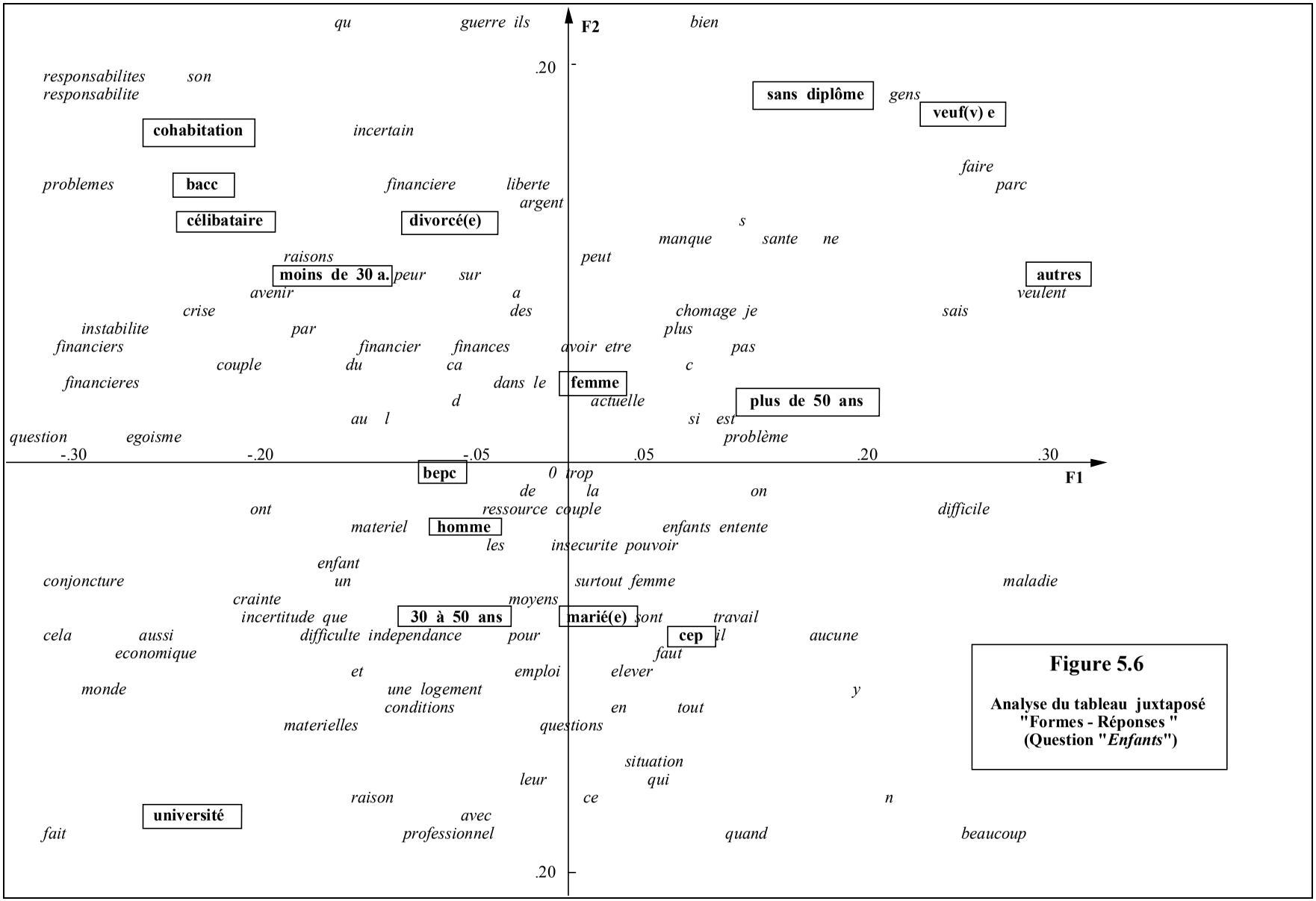


Figure 5.6

Analyse du tableau juxtaposé "Formes - Réponses " (Question "*Enfants*")

5.5 Analyse des correspondances à partir du tableau des segments répétés.

Cette section recoupe en fait les quatre sections précédentes ; les opérations effectuées à partir des formes pouvant être généralisées aux segments. Il est possible d'utiliser directement les données segmentales dans le cas d'une partition privilégiée, qui a fait l'objet du paragraphe 5.1, mais aussi dans le cas d'une partition obtenue à partir des noyaux factuels. (paragraphe 5.2).
Dans le cas d'une analyse directe des réponses, cela reste possible si les réponses sont longues et riches lexicalement (paragraphe 5.3). Enfin, cela est également possible dans le cas de juxtaposition de tableaux de contingence (paragraphe 5.4).

On a choisi de présenter l'utilisation de segments dans les phases de visualisation réalisées à partir du premier exemple traité dans la section 5.1. On reprendra donc la même question ouverte, et la même partition en 9 classes croisant les variables Age et Niveau de diplôme.

Les tableaux segmentaux sont en général assez volumineux. A titre d'exemple, on notera que pour les 2 000 réponses à la question ouverte de référence ***Enfants***, alors que 249 formes graphiques ont plus de 6 occurrences, on relève 330 segments de longueur 2 apparaissant au moins 6 fois, 133 segments de longueur 3, 39 de longueur 4, 18 de longueur 5, un seul de longueur 6 (tous ces segments étant assujettis à apparaître au moins 6 fois). On a donc un tableau de 523 segments distincts dont l'effectif total est égal à 8 670.

Comme les formes, les segments sont évidemment des éléments de description des réponses, et plus généralement, des parties de corpus. Mais la notion de profil segmental n'est pas aussi pure que celle, largement utilisée aux chapitres précédents, de profil lexical. Un tableau lexical entier comme le tableau 5.3 est en effet homogène parce que doublement additif : pour une ligne, la somme des termes est la fréquence globale de la forme, pour une colonne, la somme est la longueur de la partie correspondante du corpus.

La somme des colonnes d'un tableau analogue dont les lignes seraient des segments n'a plus une signification aussi naturelle : les segments ne sont pas indépendants et constituent des éléments d'information largement redondants. Doit-on utiliser des seuils de fréquences différents selon la longueur des segments? Doit-on utiliser simultanément les segments et les formes?

Toutes ces interrogations sont justifiées, mais l'expérience concrète des traitements statistiques d'informations segmentales montre en fait une grande robustesse des résultats vis-à-vis de ces choix. En bref, on ne

bouleverse que rarement le système des distances entre catégories ou parties en complétant les profils lexicaux des catégories ou parties par leurs profils segmentaux. Mais la présence des segments sur les graphiques et les listages va faciliter grandement leurs interprétations en contribuant à lever l'équivoque du contexte immédiat.

Un exemple

Toujours pour la question ***Enfants***, on soumet à l'analyse des correspondances le tableau à 523 lignes et 9 colonnes croisant l'ensemble des 523 segments apparaissant plus de 6 fois avec les 9 modalités de la variable "âge-éducation".
On retrouve alors dans le plan des deux premiers facteurs le réseau régulier de proximités entre les 9 catégories déjà observé sur la figure 5.1, obtenue directement à partir des formes lexicales. Pour alléger la présentation, on publiera ci-dessous une analyse portant sur un extrait des 523 segments précédents obtenu de la façon (automatique...) suivante : on retient les segments de longueur 2 apparaissant plus de 50 fois, ceux de longueur 3 ou plus s'ils apparaissent plus de 6 fois.
On obtient 111 segments, qui suffiront eux-aussi, comme les formes lexicales originales, et comme l'intégralité des 423 segments, à reconstituer le réseau régulier "âge-éducation" dans le plan des deux premiers facteurs.

Le tableau 5.6 donne la liste intégrale de ces 111 segments avec leurs fréquences. Le tableau **T** soumis à l'analyse compte donc 111 lignes et 9 colonnes. L'effectif total du tableau ainsi constitué s'élève cette fois à 2 647 occurrences de segments. Rappelons que les segments participent cette fois à l'analyse en qualité d'éléments actifs.

La figure 5.7 représente le plan principal de visualisation issu de l' analyse des correspondances de la table de contingence d'ordre (111 x 9) croisant en ligne les 111 segments du tableau 5.6 avec les neuf catégories de la variable croisée Sexe-Age utilisée au paragraphe 5.1. Les deux premières valeurs propres valent respectivement 0.13 et 0.07, et correspondent respectivement à 33% et 18% de la trace. Pour les deux premiers facteurs, les pourcentages d'inertie sont donc très semblables aux pourcentages obtenus sur les formes simples[1]. La figure 5.7 qui représente les deux premiers axes factoriels de l'analyse de **T** rappelle évidemment la figure 5.1, à ceci près que la structure est moins régulière, et a subi une légère rotation.

[1] Les valeurs propres sont plus grandes, ce qui était prévisible, car la table de contingence segmentale est beaucoup moins pleine (2 647 occurrences au lieu de 12 051 pour la table de contingence 5.3 qui concerne les occurrences de formes). Le coefficient 12 051/2 647 = 4.55 est approximativement un facteur d'échelle entre les deux ensembles de valeurs propres.

Tableau 5.6

Inventaire partiel des segments répétés *(Question "Enfants")*
Seuil pour les segments de longueur 2 : 50 ; de longueur 3 ou plus : 6

Freq	Long.	Segment
8	3	a pas de
7	3	avec un enfant
17	3	conditions de vie
8	3	crainte de l'avenir
10	3	d'avoir un enfant
7	3	dans la famille
12	3	dans le couple
113	2	de l'avenir
117	2	de la
53	2	de travail
9	3	de l'avenir pour
21	3	de la femme
7	3	de la situation
49	3	de la vie
26	3	de ne pas
8	3	de vie actuelle
9	4	de ne pas pouvoir
8	4	de plus en plus
55	2	des enfants
7	3	des raisons de
8	3	des raisons financieres
11	4	difficultes de la vie
11	3	il n'y a
22	3	il y a
10	4	il n'y a pas
7	4	il y a des
7	4	il y en a
8	4	je n'en vois pas
20	4	je ne sais pas
17	3	l'avenir de l'enfant
12	3	l'avenir des enfants
7	3	l'avenir le chomage
13	3	l'insecurite de l'emploi
51	2	la femme
79	2	la peur
87	2	la situation
127	2	la vie
11	3	la conjoncture actuelle
8	3	la crainte de
9	3	la femme travaille
44	3	la peur de
10	3	la peur des
15	3	la peur du
8	3	la situation actuelle
11	3	la situation economique
22	3	la situation financiere
7	3	la situation materielle
23	3	la vie actuelle
7	3	la vie chere
17	3	la vie est
32	4	la peur de l'avenir
7	4	la peur des responsabilites
152	2	le chomage
11	3	le chomage le
9	3	le chomage les
9	3	le fait de
31	3	le manque d'argent
35	3	le manque de
10	3	le travail de
7	4	le manque de moyens
12	4	le manque de travail
7	5	le fait de ne pas
9	5	le travail de la femme
59	2	les enfants
19	3	les conditions de
9	3	les conditions materielles
10	3	les difficultes de
18	3	les difficultes financieres
15	3	les moyens financiers
7	3	les problemes d'argent
9	3	les problemes financiers
12	3	les raisons financieres
15	4	les conditions de vie
8	5	les difficultes de la vie
53	2	manque d'argent
88	2	manque de
8	3	manque de liberte
8	3	manque de moyens
13	3	manque de ressources
23	3	manque de travail
11	3	n'y a pas
7	4	n'y a pas de
12	3	ne pas avoir
13	3	ne pas pouvoir
9	3	ne sont pas
14	3	ne veulent pas
7	3	pas de travail
7	3	pas les enfants
103	2	peur de
65	3	peur de l'avenir
8	3	peur de la
15	3	peur de ne
14	3	peur du chomage
7	4	peur de l'avenir pour
12	4	peur de ne pas
8	5	peur de ne pas pouvoir
22	3	pour les enfants
51	2	problemes financiers
8	3	qui n'est pas
7	3	raison de sante
93	2	raisons financieres
15	3	raisons de sante
9	3	raisons financieres et
12	4	travail de la femme
52	2	un enfant
7	3	une question de

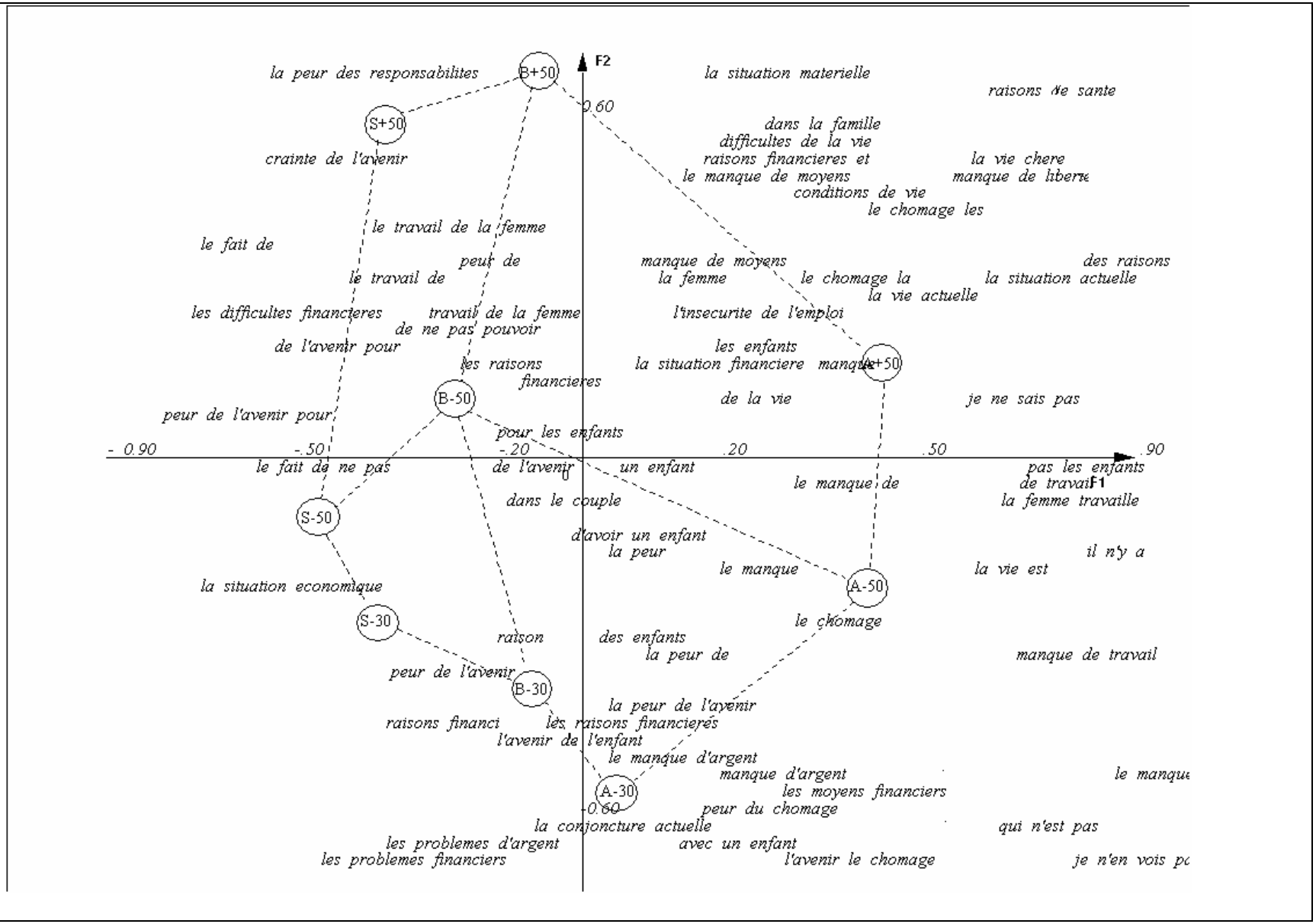


Figure 5.7 : Analyse des correspondances du tableau segmental - Question Enfants. Plan (f_1,f_2)

Rappelons que les coordonnées d'un point-catégorie sont les composantes de son profil segmental. Il n'était donc pas du tout évident a priori de retrouver simultanément l'ordre des classes d'âge et l'ordre des niveaux de diplôme. Il y a donc variation continue des segments utilisés avec l'âge (à niveau de diplôme constant) et avec le niveau de diplôme (à âge constant).

Les positions des points-segments vont permettre de compléter les interprétations que suscitaient les positions des points-formes de la figure 5.1.

La figure 5.7 suscite quelques remarques spécifiques:

a) La première de ces remarques porte sur la similitude, en ce qui concerne les positions des points-catégories, des figures 5.1 et 5.7. Il eût été équivalent, dans ce cas particulier caractérisé par une grande stabilité du "pattern" des catégories, de positionner les points-segments en éléments illustratifs sur la figure 5.1.

b) Le fonctionnement des segments répétés comme "sélecteur automatique de contexte" éclaire incontestablement certains points laissés obscurs par l'analyse directe des formes graphiques.
Examinons par exemple la position de la forme *manque* sur la figure 5.1, près des points *A-50* (aucun diplôme, entre 30 et 50 ans) et *A+50* (aucun diplôme, plus de 50 ans). Il lui correspond les segments *le manque, le manque de, manque de* qui occupent des positions au voisinage des points homologues *A-50* et *A+50* sur la figure 5.7.
Mais les expansions de ces segments (segments plus longs qui les contiennent) sont beaucoup plus dispersées sur cette figure : *le manque de travail* , en bas tout-à-fait à droite, caractérise donc les personnes sans diplôme, *le manque d'argent*, sur l'axe vertical, près du point *A-30* , caractérise plutôt les jeunes, *le manque de moyens*, beaucoup plus haut, entre les points *A+50*. et *B+50* appartient plutôt aux réponses des plus âgés.
Il en va de même pour la forme *peur*, qui est seulement *peur du chômage* chez les plus jeunes peu instruits (*A-30*), et *peur des responsabilités* chez les plus âgés et plus instruits, entre les points *S+50* et *B+50*. La forme synonyme *crainte* n'apparaît, en concurrence avec *peur*, que chez les répondants plus instruits.

c) Comme cela a déjà été signalé plusieurs fois, le contenu des réponses est fortement lié aux termes de l'expression... au point que l'on peut penser que la forme joue tour à tour les rôles d'aide ou d'obstacle dans le déchiffrage de l'information. Quelquefois, une idée qui semble être la même au prime abord est exprimée de façons différentes, avec

pourtant des mots voisins. Ainsi, ce sont surtout les personnes plus âgées qui mentionnent l'activité féminine comme une des raisons pouvant faire hésiter à avoir un enfant. Mais parmi ces personnes, les plus instruites disent : *le travail de la femme* (sous le point *S+50*), alors que les moins instruites disent simplement : *la femme travaille* (entre les points *A+50* et *A-50*). Ces différences de formes (qui traduisent des différences de lieux d'expression, de lieux d'observation), stimulantes dans toutes phases de recherche ou d'exploration, appartiennent à ce que l'on peut appeler l'aspect sociolinguistique du matériau analysé[1]. Il s'agit de nuances qui disparaissent dans une opération de post-codage, ou qui peuvent être masquées par une lemmatisation. Les personnes plus éduquées nominalisent plus facilement et désignent plutôt le phénomène que l'action correspondante[2]. Mais ces différences ne se limitent peut-être pas à la forme, et c'est dans ce sens que ces matériaux bruts, ni codés ni transformés, peuvent stimuler le sociologue : les deux segments renvoient-ils aux mêmes activités ?

d) Une quatrième remarque porte sur l'itinéraire parcouru et sur l'outil dont on dispose en fin de course. A partir d'une saisie aveugle et automatique du texte des réponses et des caractéristiques des répondants, on est en mesure d'obtenir, sans intervention manuelle ni interprétative, des représentations du type de celle présentée à la figure 5.7, plus suggestive que celle de la figure 5.1, elle-même beaucoup plus suggestive que les tableaux de fréquences.
Le seul choix important qui a été effectué est celui de la variable de regroupement (ici la variable "âge-diplôme"), mais il ne s'agit évidemment pas d'un choix unique et définitif, puisque l'analyse peut être refaite pour n'importe quel autre critère. Toutes les remarques faites à ce sujet à propos des analyses de formes s'appliquent. En particulier, une partition en noyaux factuels préliminaire peut aider à trouver le critère ou la combinaison de critères ad hoc.

e) Enfin, il faut insister sur le caractère complémentaire des analyses segmentales : la figure 5.7, volontairement limitée aux seuls segments, ne se substitue pas à la figure 5.1, elle l'enrichit.
Prenons l'exemple de la forme *crise ,* qui apparaît 25 fois dans le corpus (cf. tableau 5.1). Le segment le plus fréquent qui contient cette forme (*la crise*) n'apparaît que 19 fois ; chacun des segments plus

[1] A propos des variations des modes d'expression dans les diverses couches sociales, et plus généralement de la sociologie du langage, on pourra consulter, par exemple, la synthèse de Achard (1993).
[2] Cf. aussi les travaux de Somers (1966), que l'on évoquera à nouveau au chapitre 8.

longs, (*la crise économique, la crise de société, la crise actuelle*), a une fréquence d'apparition inférieure à notre seuil choisi égal à 6, ce qui a pour effet de faire disparaître cette forme de la figure 5.7, ce qui est ici dommageable car les contenus des segments restent proches de ceux de la forme isolée.

Parmi les solutions possibles à ce type de problème, on peut proposer une analyse simultanée des formes et des segments, ou bien adopter une approche dans laquelle soit les formes, soit les segments figurent en éléments supplémentaires.

Chapitre 6

Eléments caractéristiques, réponses ou textes modaux

Les typologies et visualisations du chapitre précédent donnent des panoramas globaux des tableaux lexicaux, que ceux-ci soient agrégés ou non, qu'ils correspondent aux parties naturelles d'un corpus (chapitres, oeuvres d'un même auteur, articles, discours datés, etc.) ou qu'ils soient constitués à partir de regroupements artificiels (regroupement de réponses libres par catégories, de documents par thèmes, etc.).

Les représentations spatiales fournies par l'analyse des correspondances gagnent à être complétées par quelques paramètres d'inspiration plus probabiliste : les *spécificités* ou *formes caractéristiques*, définies et illustrées dans ce chapitre. Il existe en effet des outils statistiques qui permettent de décrire chacune des classes d'une partition en exhibant tout à la fois les unités qu'elle contient en grand nombre par rapport aux autres classes et, au contraire, les unités qu'elle contient en très petit nombre. Dans le cas des unités de base, qu'il s'agisse de formes, de segments ou de lemmes, on parlera d'éléments caractéristiques ou de spécificités (6.1).

De façon plus générale, et inévitablement plus complexe, on peut aussi mettre en évidence, pour chacune des parties du corpus, des associations caractéristiques composées de plusieurs unités. Ainsi, des regroupements de réponses peuvent être caractérisés par quelques réponses dites caractéristiques (ou encore *réponses modales*). Ces réponses, brièvement, contiennent un maximum de formes spécifiques d'un recueil particulier...

Elles permettent de replacer les formes graphiques dans leur contexte et de résumer chaque recueil. Dans le cas de recherche documentaire, il peut s'agir de documents modaux (caractérisant un thème particulier) et on parlera alors de réponses modales, qu'il s'agisse effectivement de réponses à des questions ouvertes, ou de documents (6.2). Dans le cas de textes plus longs, on pourra mettre en évidence des unités de contexte de longueur variable (phrases, paragraphes, etc.).[1]

[1] On trouvera dans Salem (1993) des applications de ce dernier type.

6.1 Formes caractéristiques, spécificités.

L'analyse des correspondances crée une typologie portant à la fois sur l'ensemble des parties du corpus et sur l'ensemble des formes attestées dans celui-ci. Cependant, il peut être utile de compléter ces analyses globales par des calculs probabilistes effectués séparément à partir de chacune des sous-fréquence du tableau lexical.

Différentes méthodes de calcul permettent d'aboutir à ce type de paramètres. Nous avons choisi d'exposer ci-dessous le calcul des éléments caractéristiques ou spécificités, adaptation aux matériaux textuels de tests statistiques classiques[1].

6.1.1 Le calcul des spécificités

Nous noterons, exprimées en nombre d'occurrences de formes simples, les quantités :

- k_{ij} - sous-fréquence de la forme numéro i dans la partie j du corpus;
- $k_{i.}$ - fréquence de la forme numéro i dans l'ensemble du corpus;
- $k_{.j}$ - longueur de la partie numéro j;
- $k_{..}$ - longueur totale du corpus (ou simplement k).

A partir des trois derniers nombres contenus dans le tableau ci-dessus, le calcul de spécificités permet de porter une appréciation sur la sous-fréquence k_{ij} sans recourir à la notion de proportionnalité dont on a expliqué qu'elle était, dans la plupart des cas, mal adaptée à l'étude quantitative des textes, surtout lorsque ceux-ci sont de longueurs très différentes.

Le modèle probabiliste

On commence par imaginer une *population* d'objets d'effectif total k. Parmi tous ces objets, on en suppose $k_{i.}$ portant une "marque" qui les distingue des autres : boules de couleur particulière ou, dans le cas qui nous intéresse, objets correspondant aux occurrences d'une même forme graphique de fréquence totale $k_{i.}$. Les objets restants de notre population sont tous

[1] Cf. Lafon (1980) pour un exposé et d'autres exemples sur la méthode des spécificités.

confondus en un même sous-ensemble et considérés comme "non-marqués". Le nombre des objets non marqués est donc égal à $k - k_{i.}$.
Prélevons maintenant, en pratiquant des tirages aléatoires sans remise, un *échantillon* contenant exactement $k_{.j}$ objets dans notre population. A l'issue de ce tirage, le nombre des objets marqués que contient l'échantillon est noté k_{ij}.

Les nombres k, $k_{i.}$, $k_{.j}$ définis plus haut constituent les paramètres du modèle.

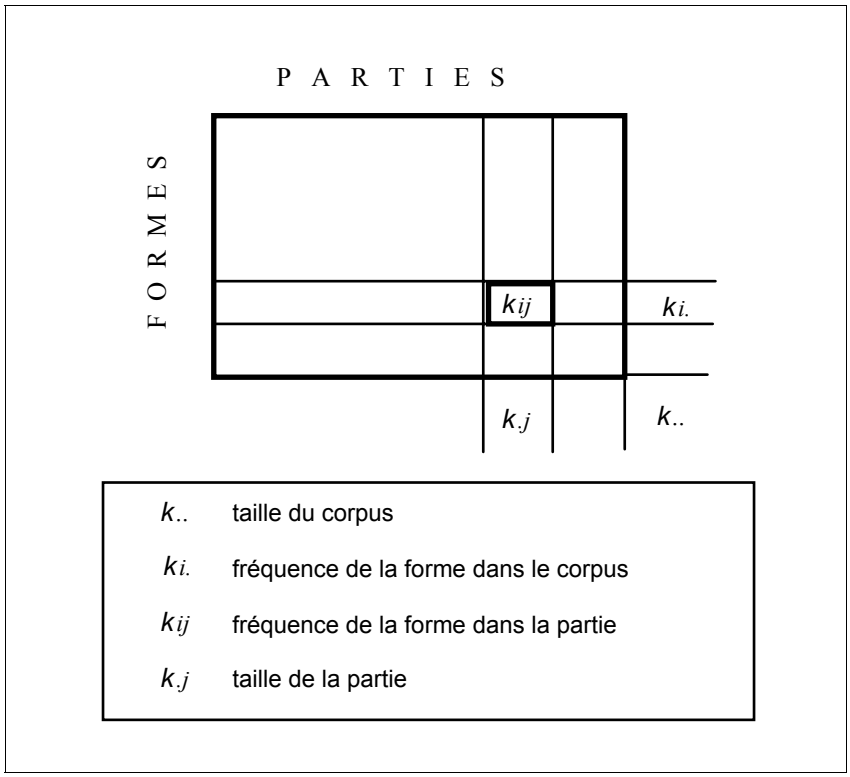


Figure 6.1
Les 4 paramètres du calcul des spécificités

Pour porter un jugement sur le résultat k_{ij}, il nous faut le situer parmi des comptages de même nature qui correspondent à l'ensemble de tous les échantillons, composés de $k_{.j}$ objets, qu'il est possible de prélever à partir de la population de départ.

Pour chaque échantillon de taille $k_{.j}$, le nombre k_{ij} des objets marqués peut prendre une valeur obligatoirement comprise entre 0 et $k_{i.}$, nombre total des objets marqués. Inversement, toujours pour des nombres k, $k_{.j}$ et $k_{i.}$ fixés, et pour chaque nombre n compris entre 0 et $k_{i.}$, il est possible de recenser le

nombre $N(n)$ des échantillons de longueur $k_{.j}$ pour lesquels k_{ij} est strictement égal à n.
Si l'on divise chacun des nombres $N(n)$ par le nombre total des échantillons de longueur $k_{.j}$, on obtient une distribution de probabilité (de paramètres : k, $k_{i.}$, $k_{.j}$) sur l'ensemble des nombres compris entre 0 et $k_{i.}$. La somme de ces nombres est égale à l'unité. La loi de probabilité correspondant à un tirage sans remise dans l'hypothèse d'indépendance est la loi hypergéométrique. Cette loi est voisine de la loi multinomiale lorsque les prélèvements sont petits par rapport à la population (on peut alors assimiler tirages sans remise et tirages avec remise).

Nous noterons :

$$Prob\ (k, k_{i.}, k_{.j}, n\ )$$

la probabilité ainsi calculée de voir apparaître exactement n objets marqués lors d'un tirage sans remise d'un échantillon de longueur $k_{.j}$ parmi une population d'effectif total k, sachant que l'ensemble de la population compte $k_{i.}$ objets marqués.

On peut voir figure 6.2 l'exemple d'une telle distribution calculée pour les paramètres :

k	=	160 000	taille du corpus.
$k_{.j}$	=	20 000	taille de la partie j
$k_{i.}$	=	36	fréquence de la forme i.

Comme on le voit, le *mode* de cette distribution (valeur la plus probable) est égal à 4. Les probabilités décroissent rapidement à mesure que l'on s'éloigne de cette fréquence.

Nous pouvons maintenant utiliser la distribution de probabilité ainsi construite à partir des paramètres k, $k_{.j}$ et $k_{i.}$ pour porter un jugement sur la fréquence absolue n_0 observée lors du tirage de notre échantillon. Pour cela nous commencerons par situer n_0 par rapport au mode de la distribution.

Si cette valeur est très proche du mode, nous ne pourrons pas dire grand chose à propos du résultat observé.

Si en revanche elle lui est nettement supérieure, nous calculerons la quantité $P_{sup}(n_0)$ qui est la probabilité de voir apparaître, toujours sous les hypothèses retenues plus haut, un nombre d'objets marqués égal ou supérieur à n_0 parmi les $k_{.j}$ objets prélevés au hasard.

La probabilité $P_{sup}(n_0)$ est égale à la somme de toutes les probabilités correspondant à des valeurs de n égales ou supérieures à n_0. C'est donc la

somme des probabilités *Prob (k, $k_{i.}$, $k_{.j}$, n)* pour les valeurs de *n* comprises entre n_0 et $k_{i.}$.
Si n_0 est inférieur au mode, nous calculerons de la même manière la quantité $P_{inf}(n_0)$ qui est la probabilité de voir apparaître, toujours sous les mêmes hypothèses, un nombre d'objets marqués égal *n* ou inférieur à n_0.

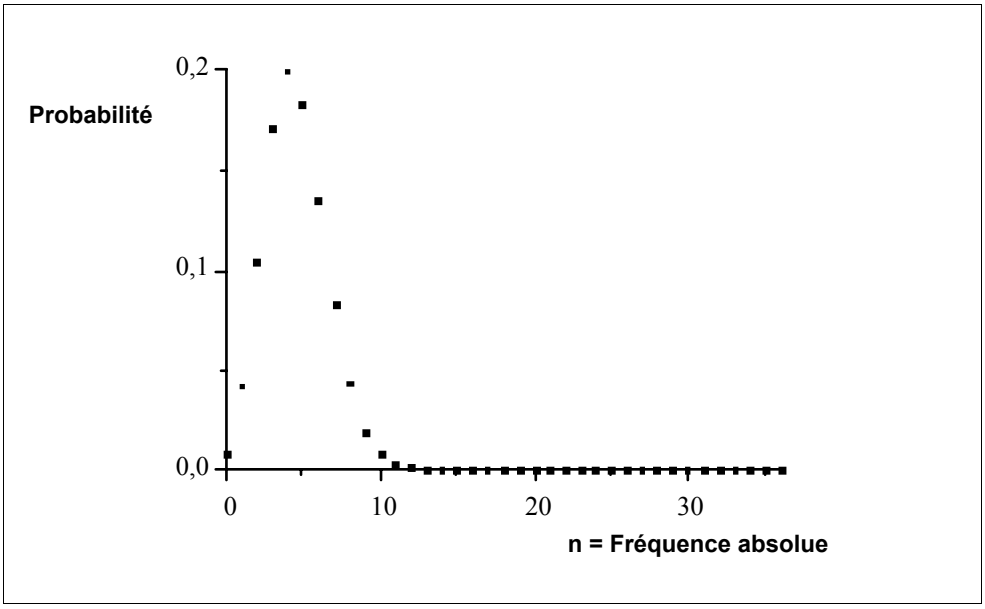


Figure 6.2

Distribution des probabilités hypergéométriques pour les paramètres k =160 000, $k_{.j}$ = 20 000, $k_{i.}$ = 36.

Cette dernière probabilité est égale, quant à elle, à la somme des probabilités *Prob (k, $k_{i.}$, $k_{.j}$, n)* pour les valeurs de *k* comprises entre 0 et n_0.

Si la probabilité $P_{inf}(n_0)$ est très faible pour un échantillon, on en conclura que cet échantillon compte, par rapport à l'ensemble des échantillons de même type, "anormalement" peu d'objets marqués.

Inversement, si la probabilité $P_{sup}(n_0)$ est très faible pour cet échantillon, on conclura que l'échantillon compte un nombre "anormalement" élevé d'objets marqués.

Pratique du calcul des spécificités

Pour chacune des cases du tableau, contenant le nombre k_{ij} des occurrences de la forme numéro i dans la partie numéro j, nous allons apprécier cette

sous-fréquence par rapport aux nombres k, $k_{i.}$, $k_{.j}$, en utilisant le modèle décrit plus haut.
Pour sélectionner les probabilités que l'on considérera comme faibles, et par conséquent dignes d'attention, on fixe arbitrairement au début de l'expérience un *seuil* de probabilité qui ne changera pas au cours du calcul sur l'ensemble des sous-fréquences contenues dans le tableau. On note cette valeur par la lettre ***s***.

On effectue ensuite le même type de calcul pour chacune des sous-fréquences k_{ij}. Les paramètres du modèle étant fixés, on peut alors construire la distribution de probabilités sur les nombres compris entre 0 et $k_{i.}$. En fonction de la position de k_{ij} par rapport au mode de la distribution, on calcule ensuite selon la méthode décrite plus haut, les probabilités : $P_{inf}(k_{ij})$ ou $P_{sup}(k_{ij})$.

Formes spécifiques négatives, banales, spécifiques (caractéristiques)

Si la première de ces probabilités se révèle plus faible que le seuil retenu, on en conclura que la sous-fréquence observée k_{ij} est relativement faible, c'est-à-dire que la forme i est plutôt mal représentée dans la partie j par rapport à ce que le modèle hypergéométrique laissait prévoir. Dans ce cas on dira que la forme numéro i est *spécifique négative* pour la partie numéro j, ce qu'on notera *S*-.

Si c'est, au contraire, la seconde de ces probabilités qui se révèle en dessous du seuil ***s***, on en conclura que la sous-fréquence k_{ij} est relativement élevée, c'est-à-dire que la forme i est plutôt abondante dans la partie j. On dira que la forme i est une forme *spécifique positive* pour la partie considérée. On notera *S*+ cette éventualité. On parle aussi dans ce cas de *forme caractéristique.*

Il se peut également qu'aucune des deux probabilités $P_{sup}(k_{ij})$ et $P_{inf}(k_{ij})$ ne se révèle inférieure au seuil ***s*.** On dira dans ce cas que la forme i est *banale* pour la partie j. Les formes banales pour une partie sont signalées par la lettre *b*.

Si une forme ne présente aucune spécificité (i.e. si elle est banale pour chacune des parties du corpus), on dira que cette forme appartient au *vocabulaire de base* du corpus.

Les diagnostics ainsi obtenus par la méthode des spécificités sur chacune des sous-fréquences figurant dans le tableau analysé peuvent être ensuite réordonnés pour permettre différents types de lecture.

La première de ces lectures s'effectue à partir de listes dans lesquelles les diagnostics de spécificité sont classés en premier lieu d'après les formes qu'ils affectent. On peut alors obtenir une caractérisation de chacune des formes du corpus par ses spécificités positives ou négatives dans les différentes parties du corpus.

A l'inverse, on peut pratiquer un autre type de lecture de ces résultats en ordonnant d'abord les diagnostics de spécificité en fonction des parties du corpus.

Ce second type de regroupement permet de rassembler, pour chaque partie, les formes qu'elle sur-emploie (les formes *S*+ de cette partie) ainsi que les formes qu'elle sous-emploie (les formes *S*-). Cette seconde lecture des résultats de la méthode permet de caractériser chacune des parties du corpus tout à la fois par les formes qu'elle utilise plus particulièrement et par celles qu'elle a tendance à éviter.

Dans chacun des états on ordonne les diagnostics de spécificité par ordre de probabilité croissante, c'est-à-dire du plus spécifique au plus banal.

Fréquences minimales, seuil d'absence spécifique

Les formes de faible fréquence, et parmi elles les hapax, constituent un cas limite. En effet, dans le cas d'une forme de fréquence égale à 1, les paramètres : k (longueur du corpus), $k_{i.}$ (fréquence totale de la forme), étant fixés, le diagnostic porté sur la sous-fréquence observée (laquelle ne peut être égale, pour ce qui la concerne, qu'à 0 ou 1) dépend alors essentiellement de $k_{.j}$, la longueur de la partie j. Dans ce cas les résultats du calcul perdent de leur intérêt. Dans la pratique, on se borne donc à calculer les spécificités pour les seules formes dont la fréquence dépasse un certain seuil que l'on appellera la *fréquence minimale retenue* et que l'on note $freq_{min}$.

Pour un seuil de probabilité donné, il existe pour chaque partie j, de longueur $k_{.j}$, une fréquence n_j telle que : toute forme de fréquence égale ou supérieure à n_j dans le corpus et absente de la partie est spécifique négative pour cette partie. C'est le *seuil d'absence spécifique* (négative) pour la partie j.

6.1.2 Un exemple de calcul des spécificités

Reprenons l'exemple des réponses à la question ***Enfants*** introduit au début de ce chapitre. On peut voir (tableau 6.1, qui reprend la ligne correspondante du tableau 5.3) la ventilation de la forme *chômage* dans les neuf parties constituées à partir du croisement des 3 classes d'âge avec les 3 catégories de

diplômes définies plus haut (cf. chapitre 5). Cette forme possède 285 occurrences dans le corpus.

Tableau 6.1

Ventilation de la forme *chômage* dans les réponses à la question *Enfants*

forme : *chômage* (285 occurrences)

Diplôme	A	A	A	B	B	B	S	S	S
Age	-30	-50	+50	-30	-50	+50	-30	-50	+50
Fréquence	35	55	93	25	15	10	19	18	15

Partant de cette ventilation, et compte tenu des longueurs respectives de chacune des neuf parties, le calcul des spécificités fournit (tableau 6.2) au seuil de 0.05, quatre diagnostics (positifs ou négatifs).

Tableau 6.2

Spécificité, au seuil de 0.05, sur la ventilation de la forme *chômage* dans les réponses à la question *Enfants*.

	partie	*fréq totale*	*fréq partie*	*Probabilité*
A -30	285	35	*S+*	0.0006
A +50	285	93	*S+*	0.0168
S -30	285	19	*S-*	0.0159
S -50	285	18	*S-*	0.0023

Comme on le constate, les diagnostics fournis par la méthode ne sont pas tous au même niveau. Certaines des probabilités calculées sont beaucoup plus faibles que les autres. Cependant, si l'on ne retient de ces résultats que le caractère qualitatif spécifique/non-spécifique, on peut résumer, comme au tableau 6.3, l'emploi de la forme par chacune des parties.

Tableau 6.3

Spécificités de la forme *chômage*

Diplôme	A	A	A	B	B	B	S	S	S
Age	-30	-50	+50	-30	-50	+50	-30	-50	+50
Fréquence	35	55	93	25	15	10	19	18	15
Spécificité	*S+*	*b*	*S+*	*b*	*b*	*b*	*S-*	*S-*	*b*

Certaines des formes ne trouvent de spécificité dans aucune des parties du corpus. Le tableau 6.4 donne quelques-unes de ces formes les plus fréquentes, que l'on a appelé formes banales.

Tableau 6.4

Quelques formes banales.

Diplôme	A	A	A	B	B	B	S	S	S	
Age	-30	-50	+50	-30	-50	+50	-30	-50	+50	Tot.
enfants	10	7	15	24	9	20	47	9	7	148
emploi	7	9	5	19	6	5	21	4	3	79
insécurité	3	3	5	14	3	8	18	3	5	62
femme	2	9	18	4	4	5	5	9	4	60

Les résultats de spécificités obtenus à partir des ventilations des formes dans les parties du corpus peuvent être, on l'a vu, ordonnés de différentes manières. On peut rassembler comme nous l'avons fait plus haut les diagnostics relatifs à une même forme graphique en les ordonnant d'après les parties qu'ils affectent. Inversement, on peut établir, pour chacune des parties, la liste des formes qui trouvent des spécificités (positives et négatives) dans cette partie.

Sur le tableau, 6.5 on peut voir, toujours au seuil de 0.05, la liste des formes spécifiques de la partie A-30 (jeunes peu diplômés).

Tableau 6.5

Les diagnostics de spécificité, au seuil de 0.05, pour la partie A-30 dans les réponses à la question *Enfants*.

partie : A-30 (Aucun diplôme - moins de 30 ans)

forme	*fréq.*	*fréq. totale*		*Probabilité*
chômage	285	35	*S+*	0.0006
le	474	50	*S+*	0.0016
avoir	7	12	*S+*	0.0060
avenir	318	32	*S+*	0.0194
très	18	4	*S+*	0.0313
argent	196	20	*S+*	0.0493
et	204	4	*S-*	0.0013
peut	54	0	*S-*	0.0212
que	78	1	*S-*	0.0259
situation	176	6	*S-*	0.0376

Le tableau s'ouvre sur les formes spécifiques positives triées de la plus spécifique (i.e. celle à laquelle se trouve affectée la plus petite probabilité) à la plus "banale". Viennent ensuite les formes spécifiques négatives triées selon le même critère.

Il est également possible de calculer des spécificités à partir de décomptes dans les parties, ou groupes de parties d'un corpus, des occurrences de toute autre unité constituée à partir de regroupements, ou de subdivisions opérées sur la base des décomptes en formes graphiques. C'est à ce stade que l'on pourra si on le désire regrouper le pluriel et le singulier d'une même forme, les différentes flexions d'un même verbe, afin de comparer les diagnostics obtenus sur ces regroupements à ceux obtenus lors de l'analyse des formes simples.

6.1.3 Liste des formes spécifiques ou caractéristiques

Aux formes caractéristiques sont attachées, dans les listages produits par les logiciels, des *valeurs-test,* qui sont d'autant plus grandes que les probabilités ci-dessus sont petites. Ces valeurs-tests mesurent l'écart existant entre la fréquence relative d'une forme dans une classe avec sa fréquence relative globale calculée sur l'ensemble des réponses ou individus.[1]

Cet écart est normé de façon à pouvoir être considéré comme une réalisation de variable normale centrée réduite, dans l'hypothèse de répartition aléatoire de la forme étudiée dans les classes. Dans cette hypothèse, la valeur-test a *95* chances sur *100* d'être comprise entre les valeurs *-1.96* et *+ 1.96* .[2] Mais ce calcul reposant sur une approximation normale de la loi hypergéométrique n'est utilisé que lorsque les effectifs concernés ne sont pas trop faibles.

Le tableau 6.6 présente pour les quatre catégories "extrêmes" (correspondant aux deux classes d'âge et aux deux niveaux de diplômes extrêmes) les 5 premières spécificités positives et les cinq dernières spécificités négatives.

Les valeurs-test servent de critère de classement des formes à l'intérieur de chaque classe (ou partie). On retrouve les cinq premières formes du tableau 6.5 pour la classe "aucun diplôme-moins de 30 ans", à une interversion près des formes *très* et *avenir*. Cette interversion est due à des bases de calcul différentes.

---

[1] Des logiciels comme SPAD.T et Lexico1 (brièvement décrits en annexe) produisent, pour chaque partition des individus, un bilan exhaustif des formes et segments caractéristiques (spécificités positives et négatives) correspondant à chaque classe (ou partie).

[2] Les valeurs-test ont déjà été évoquées, dans un cadre différent, au chapitre 4, à propos de la description automatique des classes d'une partition par le tableau 4.5.

Tableau 6.6

Formes graphiques caractéristiques (question *Enfants*) pour les 4 classes extrêmes de la partition "âge-diplôme"

LIBELLE DE LA FORME		POURCENTAGE INTERNE	POURCENTAGE GLOBAL	FREQUENCE INTERNE	FREQUENCE GLOBALE	V.TEST	PROB.
TEXTE NUMERO 1	(A—30 = Moins de 30 ans, Aucun diplôme ou cep)						
1	chomage	3.90	2.25	35.	285.	3.092	.001
2	le	5.57	3.74	50.	474.	2.756	.003
3	avoir	1.34	.61	12.	77.	2.427	.008
4	avenir	3.57	2.51	32.	318.	1.912	.028
5	tres	.45	.14	4.	18.	1.819	.034
5	plus	.22	.69	2.	87.	—1.660	.048
4	situation	.67	1.39	6.	176.	—1.884	.030
3	que	.11	.62	1.	78.	—2.012	.022
2	peut	.00	.43	0.	54.	—2.092	.018
1	et	.45	1.61	4.	204.	—3.122	.001
TEXTE NUMERO 3	S—30 = (Moins de 30 ans, bacc ou université)						
1	financieres	2.62	1.37	34.	173.	3.619	.000
2	couple	1.46	.75	19.	95.	2.718	.003
3	vis	.46	.13	6.	16.	2.674	.004
4	responsabilites	.54	.17	7.	21.	2.672	.004
5	que	1.23	.62	16.	78.	2.565	.005
5	finances	.00	.22	0.	28.	—1.665	.048
4	travail	.69	1.20	9.	152.	—1.718	.043
3	manque	.69	1.26	9.	160.	—1.917	.028
2	sante	.31	.75	4.	95.	—1.920	.027
1	chomage	1.46	2.25	19.	285.	—2.010	.022
TEXTE NUMERO 7	A+50 = (Plus de 50, aucun diplôme ou cep)						
1	ne	2.48	1.52	85.	193.	5.038	.000
2	pas	3.56	2.57	122.	325.	4.123	.000
3	ils	1.05	.62	36.	79.	3.428	.000
4	veulent	.64	.33	22.	42.	3.335	.000
5	sais	.41	.20	14.	25.	2.866	.002
5	problemes	.50	.85	17.	108.	—2.665	.004
4	financiers	.32	.68	11.	86.	—3.051	.001
3	l	4.15	5.32	142.	674.	—3.627	.000
2	avenir	1.55	2.51	53.	318.	—4.348	.000
1	financieres	.55	1.37	19.	173.	—5.108	.000
TEXTE NUMERO 9	S+50 = (Plus de 50ans , Bacc ou université)						
1	sante	1.78	.75	15.	95.	2.969	.001
2	difficultes	1.54	.65	13.	82.	2.759	.003
3	materielles	1.18	.45	10.	57.	2.645	.004
4	assurer	.59	.14	5.	18.	2.549	.005
5	les	5.09	3.49	43.	442.	2.406	.008
5	n	.24	.74	2.	94.	—1.697	.045
4	en	.36	.92	3.	116.	—1.703	.044
3	trop	.00	.41	0.	52.	—1.933	.027
2	si	.00	.44	0.	56.	—2.051	.020
1	y	.00	.44	0.	56.	—2.051	.020

Les spécificités du tableau 6.5 sont calculées sur l'ensemble du corpus, conformément à une tradition bien établie en lexicométrie. Le tableau 6.6, comme les analyses des correspondances du chapitre précédent, est calculé d'après la table de contingence esquissée au tableau 5.3, c'est à dire après intervention du seuil de fréquence des formes.

Dans ce second cas, le choix de ce seuil de fréquence de sélection des formes, qui modifie le nombre total des occurrences du texte partiel ainsi que le nombre d'occurrences pour chaque partie du texte, aura une influence sur les probabilités calculées. Les calculs du tableau 6.6 utilisent de plus une approximation normale de la loi hypergéométrique. On voit cependant que les résultats sont robustes au regard de ces options de calculs.

Le tableau 6.6 a été établi avec un seuil minimal de 14 occurrences par forme, ce qui correspond à 154 formes, et à 12 661 occurrences, sur 15 457.

Le tableau 6.7 reprend le premier texte (Jeunes sans diplômes) dans le cas d'une sélection avec un seuil minimal de 21 occurrences (cas des tableaux et graphiques du chapitre 5), d'où un texte de 12 051 occurrences.

Tableau 6.7

Formes graphiques caractéristiques (question *Enfants*)
Réponses de la classe 1 de la partition "âge-diplôme"
Spécificités positives, seuil de fréq. = 20

	forme	pourcentage interne	pourcentage global	frequence interne	frequence globale	v.test	prob.
TEXTE NUMERO 1	A—30 = (Moins de 30 ans, Aucun diplôme ou cep)						
1	chomage	4.08	2.36	35.	285.	3.069	.001
2	le	5.83	3.93	50.	474.	2.727	.003
3	avoir	1.40	.64	12.	77.	2.414	.008
4	avenir	3.73	2.64	32.	318.	1.887	.030
5	argent	2.33	1.63	20.	196.	1.509	.066
5	plus	.23	.72	2.	87.	—1.671	.047
4	situation	.70	1.46	6.	176.	—1.901	.029
3	que	.12	.65	1.	78.	—2.023	.022
2	peut	.00	.45	0.	54.	—2.100	.018
1	et	.47	1.69	4.	204.	—3.140	.001

On constate que la sélection des formes conservées après seuillage est inchangée (la forme *très* ne fait plus partie du corpus partiel, puisqu'elle n'a que 18 occurrences dans le corpus) mais que les fréquences et les indices de probabilité sont parfois légèrement modifiées.
D'une façon générale, il est recommandé, et cela est conforme à la pratique de la lexicométrie, de calculer les formes caractéristiques sur l'ensemble du corpus, avant toute sélection par seuil. Toutefois, lorsqu'il s'agit de compléter

une visualisation factorielle qui, elle, intervient nécessairement après un seuil, il est loisible de travailler sur le corpus partiel, afin d'assurer une compatibilité des tableaux lexicaux utilisés.

Formes caractéristiques et visualisation factorielle

On peut confronter les résultats du tableau 6.6 aux proximités observables sur la figure 5.1 du chapitre 5 ; il va sans dire que les résultats sont plus complémentaires que concurrents. L'approximation plane de la figure 5.1 fait perdre une certaine information, mais décèle un ordre et des régularités invisibles sur le tableau 6.6.

Une représentation utile consiste à ne garder sur la figure 5.1 que les points les plus caractéristiques, et à les joindre aux points-catégories qu'ils caractérisent.

Les points anti-caractéristiques (spécificités négatives) sont beaucoup plus difficiles à identifier sur les graphiques factoriels. Ainsi, par exemple, la forme *chômage* est la forme la plus *anti-caractéristique* de la classe 3 (S-30 = jeunes instruits), alors que ce n'est pas, loin s'en faut, la forme la plus éloignée (sur la figure 5.1 du chapitre 5) du point S-30.

Attention aux comparaisons multiples !

Le calcul simultané de plusieurs valeurs-test ou de plusieurs seuils de probabilités se heurte à l'écueil des *comparaisons multiples,* bien connu des statisticiens.

Supposons que les parties de texte soient parfaitement homogènes et donc que l'hypothèse d'indépendance entre les formes et les parties soit réalisée. Les valeurs-test attachées aux spécificités, pour une partie donnée, sont alors toutes des réalisations de variables aléatoires normales centrées réduites indépendantes. Dans ces conditions, *en moyenne*, sur 100 spécificités calculées, 5 seront en dehors de l'intervalle [-1.96 , +1.96], et 5 dépasseront la valeur 1.65 (test unilatéral). Le seuil de 5% n'a de sens en fait que pour un seul test, et non pour des tests multiples.

Autrement dit, l'utilisateur non averti trouvera presque toujours "de quoi s'étonner" au seuil de 5%... On résout de façon pragmatique cette difficulté en choisissant un seuil plus sévère. On doit garder à l'esprit les ordres de grandeurs suivants : le seuil 1% correspond à une valeur supérieure à 2.33, et le seuil 1 pour 1 000 à une valeur supérieure à 3.09. Avec ces seuils, beaucoup des spécificités calculées au tableau 6.6 restent significatives.
On se référera à nouveau au tableau 6.6 lors de la recherche des "réponses modales" (ou réponses caractéristiques) de chaque classe. Les réponses

modales, étudiées au paragraphe suivant, permettront en effet d'atténuer l'impression de morcellement que peut susciter l'examen des formes isolées (tableau 6.6) en situant ces formes caractéristiques dans leurs contextes.

6.2 Les réponses modales

Une technique simple à mettre en œuvre va permettre de répondre simultanément à plusieurs des objections qui ont été soulevées concernant le caractère fragmentaire et désarticulé de toute étude limitée aux formes isolées de leur contexte immédiat : il s'agit de la sélection automatique des *réponses modales* (appelées encore *phrases caractéristiques* ou *documents caractéristiques*, pour d'autres types d'applications).

Les réponses modales ne seront pas des réponses artificielles dressant un portrait de chaque groupement, mais des réponses authentiques, effectivement fournies par les personnes interrogées, choisies en raison de leur caractère représentatif, pour une catégorie donnée d'individus.

6.2.1 La sélection des réponses modales

Il existe deux modes de sélection de ces réponses modales : un premier mode est fondé sur des calculs de distances selon des critères géométriques simples (sélection suivant la distance du chi-2). Un second mode s'appuie sur les calculs de spécificité.

Sélection suivant la distance du chi-2

Le principe de cette sélection est schématiquement le suivant : une fois décodées par l'ordinateur (phase de numérisation), les réponses libres de k individus (k=1 000, pour fixer les idées) donnent lieu à la construction d'un tableau rectangulaire **T** ayant k lignes (individus ou réponses) et autant de colonnes que de formes graphiques sélectionnées (V= 400, par exemple).

Une réponse est alors une ligne du tableau, donc un vecteur à 400 composantes. Si cette réponse est formée de 25 formes différentes, seulement 25 de ces composantes seront différentes de zéro.

Un groupement de réponses (un discours artificiel) est un ensemble de vecteurs-lignes, et le profil lexical moyen de ce groupement est obtenu en calculant la moyenne des vecteurs-lignes de cet ensemble. Si ce regroupement s'opère selon les modalités d'une question fermée dont les

réponses sont codées dans un tableau **Z**, on a vu au paragraphe 5.1 que le tableau lexical agrégé **C** se calcule par la formule:

$$\mathbf{C = T'Z}$$

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

Ces distances expriment l'écart entre le profil d'une réponse et le profil moyen de la classe à laquelle cette réponse appartient. La distance choisie entre ces profils de fréquences sera la distance du chi-2, en raison de ses propriétés distributionnelles soulignées au chapitre 3.

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

$$d^2(i,m) = \sum_{j=1}^{p} (t../t.j) \left(t_{ij} / t_{i}. - c_{jm} / c.m\right)^2$$

avec les notations usuelles :

t.. désigne la somme globale des éléments du tableau **T**, c'est-à-dire le nombre total d'occurrences.

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

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

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

On peut, pour chaque regroupement, classer ces distances par ordre croissant, et donc sélectionner les réponses les plus représentatives au sens du profil lexical, qui correspondront aux plus petites distances. Ce mode de sélection des réponses modales est appelé dans ce qui suit : sélection selon le critère du chi-2.

Sélection par les formes caractéristiques

Un autre mode de calcul de réponses modales est le suivant : une réponse modale d'un groupement est une réponse qui contient, autant que faire se peut, les formes les plus caractéristiques de ce groupement (spécificités). Pour chaque groupement, ces formes sont classées par ordre de

significativité (cf. tableau 6.6 par exemple) : le rang d'une forme dans ce classement est d'autant plus petit qu'elle est caractéristique.

Une formule empirique simple consiste à associer à chaque réponse le rang moyen des formes qu'elle contient : si ce rang moyen est petit, cela signifie que la réponse ne contient que des formes très caractéristiques du groupement.

Mieux que les rangs moyens, on peut prendre les valeurs-test correspondantes (au plus petit rang moyen correspond alors la plus grande valeur-test moyenne). On conçoit facilement que certains arbitrages soient nécessaires : entre une réponse courte formée de peu de formes très caractéristiques et une réponse plus longue, qui contient nécessairement des formes moins caractéristiques...

Ce mode de calcul (critère des formes caractéristiques) a la particularité de favoriser les réponses courtes, alors que le critère du chi-2 a au contraire tendance à favoriser les réponses longues.

Quel que soit le mode de calcul, on imprimera en fait plusieurs réponses caractéristiques pour chaque groupement, car il est extrêmement improbable qu'il existe parmi les réponses originales une réponse qui résume à elle seule toutes les particularités d'une catégorie. Dans les listages de sortie, on pourra trouver, par exemple, six réponses par classe (trois réponses pour chacun des deux modes de calcul...).

6.2.2 Mise en oeuvre et exemples

Pour une question ouverte (ici encore, la question ***Enfants*** nous servira d'exemple de référence), et pour une partition de la population (la partition en catégories *âge* - *diplôme*), on obtient donc, sans codification préalable ni médiation :

- Une visualisation des proximités entre formes et catégories (chapitre 5).
- Les spécificités ou formes caractéristiques de chaque catégorie (paragraphe 6.1 de ce chapitre).
- Les réponses modales de chaque catégorie.

Reprise de l'exemple du chapitre 5

On donnera tout d'abord un exemple de réponses modales correspondant aux quatre classes extrêmes de la partition en neuf classes : les moins de 30 ans

sans diplôme, les moins de 30 ans de niveau d'instruction "Bac et plus", les plus de 50 ans sans diplôme, les plus de 50 ans "Bac et plus". On abandonne donc simultanément la classe d'âge et le niveau de diplôme intermédiaires.

Pour chacune de ces quatre classes, on donnera les trois premières réponses modales, pour chacun des deux critères de sélection. Les tableaux 6.8a et 6.8b présentent, pour chacune des quatre classes extrêmes (pour l'âge et le niveau d'instruction) les réponses sélectionnées d'après le critère des formes caractéristiques

Les réponses sont très laconiques : puisque la forme *chômage* est la plus caractéristique de la première classe (tableau 6.6), et qu'il existe des réponses formées de ce simple mot, il est naturel qu'il s'agisse, avec ce critère, de la réponse la plus caractéristique. Le critère de sélection correspondant est alors égal à la valeur test 3.09.

L'article *le* fait légèrement baisser la valeur du critère. On peut vérifier (tableau 6.6) que la valeur de ce critère pour la réponse *le chômage* est la moyenne arithmétique des critères des deux formes qu'elle contient.

Le caractère presque caricatural des réponses sélectionnées par ce premier critère est frappant si l'on compare les tableaux 6.8 et 6.9. L'utilisateur consultera d'abord les premières réponses, qui constituent un résumé, un squelette sémantique, que la lecture des réponses choisies selon le second critère permettra de nuancer et compléter.

Tableau 6.8 a

Critère des formes caractéristiques
Réponses modales pour la question *Enfants*
(Catégories *moins de 30 ans*)

```
CRITERE DE            REPONSES OU INDIVIDUS CARACTERISTIQUES
CLASSEMENT
_____________________________________________________________________
TEXTE NUMERO  1  (A—30  = Moins de 30 ans, Aucun diplôme ou cep)
_____________________________________________________________________
   3.092 ——    1   chomage
   3.092 ——    2   chomage
   3.092 ——    3   chomage
   3.092 ——    4   chomage
   2.924 ——    5   le chomage
_____________________________________________________________________
   TEXTE NUMERO   3  (S—30 = (Moins de 30 ans, bacc/université)
_____________________________________________________________________
   2.361 ——    1   raisons financieres
   2.361 ——    2   raisons financieres
   2.361 ——    3   raisons financieres
   2.361 ——    4   raisons financieres
   2.361 ——    5   raisons financieres
```

Tableau 6.8 b

Critère des formes caractéristiques
Réponses modales pour la question *Enfants*
(Catégories *plus de 50 ans*)

```
CRITERE DE           REPONSES OU INDIVIDUS CARACTERISTIQUES
CLASSEMENT
_____________________________________________________________________
   TEXTE NUMERO  7   A+50  = (Plus de 50, aucun diplôme ou cep)
_____________________________________________________________________
   3.544 ——    1   je ne sais pas
   3.544 ——    2   je ne sais pas
   3.053 ——    3   ne sait pas
   3.053 ——    4   ne sait pas
   3.053 ——    5   ne sait pas
_____________________________________________________________________
   TEXTE NUMERO  9   S+50  =  (Plus de 50ans , Bacc/université)
_____________________________________________________________________
   2.969 ——    1   sante
   1.890 ——    2   egoisme
   1.843 ——    3   les revenus
   1.823 ——    4   l'egoisme,difficultes materielles
   1.721 ——    5   les difficultes financieres
```

Remarques

Cet exemple appelle plusieurs remarques, qui concernent aussi bien l'application elle-même que la méthode.

a) La première est un constat élémentaire, mais qu'il faut faire à ce stade : on a obtenu effectivement un résultat ayant une certaine cohérence, une sorte de résumé de l'information initiale, sans codification ni pré-interprétation du texte.

b) Les réponses ci-dessus sont des réponses originales intégrales, choisies parmi 2 000 réponses saisies... elles permettent de rétablir les formes dans leur contexte, lorsqu'il existe. Elles pourraient d'ailleurs être positionnées sur la figure 6.1, en occupant chacune le centre de gravité des formes graphiques qu'elles contiennent. Mais cela encombrerait une représentation déjà très chargée.

c) Les différences d'expression sont assez considérables, mais ne semblent pas un obstacle à l'analyse de l'information, bien au contraire : Peut-on vraiment considérer, par exemple, que *manque d'argent* invoqué par les jeunes sans diplôme, est tout à fait synonyme des *raisons financières*, mentionnées par les jeunes diplômés ? Bien entendu, un post-codage aurait classé ces deux réponses sous une même rubrique.

Tableau 6.9

Critère de la distance du chi-2. Réponses modales pour la question *Enfants*

```
CRITERE DE            REPONSES OU INDIVIDUS CARACTERISTIQUES
CLASSEMENT
-------------------------------------------------------------------
TEXTE NUMERO  1  A-30  = moins de 30 ans, aucun diplôme ou cep)
-------------------------------------------------------------------
 .861 --  1   manque d'argent,l'avenir,le chomage
 .876 --  2   le chomage,le manque d'argent de toutes manieres
 .887 --  3   la peur de l'avenir le chomage
 .887 --  4   la peur de l'avenir,le chomage

-------------------------------------------------------------------
TEXTE NUMERO  3  S-30  = moins de 30 ans, bacc ou université
-------------------------------------------------------------------
 .924 --  1   l'instabilite des relations du couple,la situation
               financiere encore plus et la peur d'un avenir    precaire
             pour les enfants
 .925 --  2     raisons financieres,avenir de l'enfant
 .935 --  3     problemes financiers,l'engagement,le contrat a long terme que
             cela suppose
 .936 --  4   je ne suis pas peut etre les obligations que cela  impliquent
             et les depenses en plus

-------------------------------------------------------------------
TEXTE NUMERO  7  A+50  = plus de 50, aucun diplôme ou cep
-------------------------------------------------------------------
 .865 --  1   je ne sais pas si on en a pas envie ce n'est pas la
               peine,il y a des gens qui n'aiment pas les enfants
 .909 --  2   je ne peux pas vous dire.. le genre de caractere de la personne
 .914 --  3   la mesententе c'est le seul probleme un couple qui
               ne s'entend pas ils vont pas avoir d'enfants parce
               qu'apres c'est des problemes plus graves
 .918 --  4   les difficultees financieres,il y a des gens qui   n'aiment pas
             les enfants,il vaut mieux ne pas avoir d'enfants que de les
             rendre malheureux

-------------------------------------------------------------------
TEXTE NUMERO  9  S+50  =  plus de 50ans , bacc/université
-------------------------------------------------------------------
 .894 --  1   l'instabilite economique et sociale du pays et plus
               concretement les difficultes financieres de chacun
               d'entre nous,la base du pouvoir d'achat et l' insecurite de
             l'emploi
 .908 --  2   une question de logement, ensuite les ressources financieres,
             ensuite  la situation geographique des epoux apres ca peut etre
             un certain egoisme,des gens qui limitent les naissances dans le
             desir de mieux elever les enfants qu'ils auront acceptes
 .913 --  3   la sante de l'epouse,les conditions materielles de
               vie et les causes ideologiques,crainte du futur le
               chomage,la guerre
 .921 --  4   les conditions de vie et les difficultes d'embauche
```

Mais le fait que la différence de formes soit liée à une différence de situations socio-économiques pose à nouveau le problème du *contenu* de chacune de ces réponses. N'y a-t-il pas quasi-impossibilité dans un cas et difficulté dans l'autre ? Comme c'est souvent le cas lors de l'utilisation de techniques exploratoires, on est conduit inéluctablement à critiquer l'information de base et même les concepts qui président au recueil de cette information.

d) Il peut arriver, bien que cela ne soit pas le cas ici, que certaines formes rares apparaissent dans les réponses caractéristiques. Il faut noter que pour les deux critères, la sélection des réponses ne se fait que sur les formes retenues, après intervention du seuil de sélection des formes selon leur fréquence (ce seuil vaut 13 dans notre exemple). Mais l'édition des réponses respecte leur libellé original. Il peut donc arriver, dans des cas plutôt exceptionnels, qu'une réponse ne contienne que des formes dont la fréquence est inférieure au seuil, à l'exception d'une seule, qui soit un mot-outil spécifique de la catégorie. Avec le critère de fréquence lexicale (mais pas avec le critère du chi-2), cette réponse pourra apparaître comme caractéristique.

Un exemple en est donné par une des réponses modales de la classe 9 (personnes instruites de plus de 50 ans) : parmi les formes spécifiques de cette classe figurait la conjonction *et*, qui traduit d'ailleurs le fait que les personnes de cette classe sont également les plus disertes (la forme *et* ne figure pas dans le tableau 6.6 car elle occupe le rang 6, avec une valeur-test de 2.3). On obtient comme réponse modale, à un niveau voisin de celles citées plus haut : "*Elles aiment mieux batifoler et rigoler ensemble*", réponse qui ne doit son bon classement qu'à la présence de la forme *et* et à l'absence des autres formes fréquentes du corpus. Le critère de fréquence lexicale n'est donc à utiliser qu'avec circonspection. Le seuil de fréquence doit être fixé suffisamment bas[1].

6.2.3 Autres exemples

Pour donner une idée à la fois de la diversité des réponses modales pour d'autres questions ouvertes et d'autres catégories, et des interprétations auxquelles elles conduisent, on donnera ci-dessous quelques exemples complémentaires (il s'agit de sélections réalisées uniquement selon le critère du chi-2). Ces exemples sont extraits de l'enquête sur les conditions de vie et aspirations des Français (cf. chapitre 3, paragraphe 3.2.2).

Le premier exemple (a) concerne des *inquiétudes générales*, étudiées selon la catégorie socioprofessionnelle des répondants.

Le second exemple (b) concerne un exemple d'explicitation de réponse, du type *"pourquoi ?"*, à une question fermée (cf. chapitre 1, paragraphe 1.4.3).

[1] Notons, toujours dans le cadre empirique de cet exemple, que le fait d'abaisser le seuil de 13 à 6 fait effectivement disparaître cette réponse modale "parasite", mais modifie très peu l'ensemble des autres réponses modales. Une seule nouvelle réponse modale apparaît, (toujours avec le critère de la fréquence lexicale, et toujours pour la classe 4):" *maladies héréditaires* "(*maladies*: fréquence 9, *héréditaires*, fréquence 7).

Le troisième et dernier exemple reprend la question plus méthodologique posée à l'issue de chaque entrevue de cette même enquête (question déjà étudiée au chapitre 5, paragraphe 5.2).

a) Inquiétudes par catégorie socioprofessionnelle

Les réponses modales concernent la question *"Qu'est-ce qui vous inquiète le plus en ce qui concerne votre avenir ?"* pour quelques catégories socio-professionnelles :

Agriculteurs :

— *De pouvoir continuer notre métier jusqu'à la fin de la vie; l'avenir des jeunes dans l'agriculture, problèmes du foncier et de l'équipement lourd en agriculture.*
— *Les problèmes d'énergie, le fuel est trop cher, on en a besoin pour les tracteurs.*
— *Il faudrait que les prix augmentent à la production, sinon l'exploitation ne sera plus rentable.*

Ouvriers :

— *Garantie de l'emploi jusqu'à la retraite, toucher une retraite convenable.*
— *Ne pas disposer de suffisamment d'années de vie pour profiter de la retraite.*
— *L'emploi et le chômage.*

Ménagères, femmes au foyer :

— *Je m'inquiète plus pour l'avenir de mes enfants que pour le mien.*
— *Rien pour moi, mais pour mes enfants: leur orientation, leur formation, leur travail.*
— *Pas de travail pour les jeunes.*

Cadres moyens et supérieurs :

— *Cette crise économique nous entraîne vers une catastrophe, la crise politique vers la guerre.*
— *L'impact de la crise sur mon activité professionnelle par réduction des crédits et du recrutement.*
— *Montée du chômage, elle aura peut-être des conséquences très graves sur le plan psychologique.*

Etudiants :

— *La peur de ne pas trouver un travail intéressant, le ronronnement progressif, la passivité, la démission collective devant les grands problèmes de notre temps.*

Retraités :

— *Tout ce que je demande, c'est de pouvoir vivre comme maintenant, et surtout de garder une bonne santé, le reste, je ne m'en inquiète pas.*

On obtient donc un résumé assez vivant des réponses et de leur dispersion dans différentes catégories sans avoir à lire l'ensemble des deux mille réponses qui correspondent en fait à un texte de près de cent pages.

Bien évidemment, la partition en catégories socioprofessionnelles ne permet pas d'épuiser toutes les formes de réponses : ainsi, si l'on distingue de l'ensemble des "retraités" la catégorie des "retraités avec une seule personne au foyer", on obtient pour cette catégorie les réponses modales suivantes :

— *J'ai peur de tomber malade et d'être seule, le reste ne me fait pas peur.*
— *Etre malade et être seule, ne pas pouvoir rester dans ma maison jusqu'au bout.*
— *La solitude, la santé.*

Ces différences de contenu et de tonalité entre les retraités en général et les retraités vivant seuls illustrent bien l'importance du choix des regroupements à opérer sur les réponses. Aurait-on pu saisir par un post-codage la véhémence des étudiants, l'angoisse des personnes âgées seules, l'attitude oblative des ménagères, le pragmatisme à court et moyen terme des agriculteurs, la globalité des points de vue des cadres ?

Probablement, si le codage est réalisé par un sociologue chevronné, et si l'exploitation statistique des données codées permet de prendre en compte les cooccurrences de ces items ; mais la procédure proposée ici nous paraît moins coûteuse et plus réaliste.

b) Opinions sur le mariage pour une classe d'âge, par sexe

La question ouverte est maintenant la question "*Pouvez-vous dire pourquoi ?"*, à la suite de la question fermée suivante, qui concerne l'institution du mariage (cf. Tabard, 1974) :

Parmi ces opinions, laquelle se rapproche le plus de la vôtre?
Le mariage est:

- *Une union indissoluble.*
- *Une union qui peut être dissoute dans les cas graves.*
- *Une union qui peut être dissoute par simple accord des deux parties.*

On se limite ici à l'examen de deux classes d'une partition croisant le genre et l'âge des répondants (6 000 individus correspondant à trois phases consécutives de l'enquête précitée, de 1978 à 1980) : les hommes de plus de 60 ans et les femmes de plus de 60 ans.

Ces classes d'âges sont toutes deux "traditionalistes" : à la question fermée préliminaire, 48% des hommes et 44% des femmes ont répondu qu'ils considéraient le mariage comme une union indissoluble, alors que le pourcentage moyen obtenu par cette réponse dans l'ensemble de la population (résidents métropolitains âgés de 18 ans et plus) est de 28%.

Les réponses modales des hommes de plus de 60 ans sont :

— *Quand on se marie, on ne fait pas n'importe quoi.*

— *On se marie pour toujours.*

— *Si on se marie, c'est pour toujours, autrement, il vaut mieux rester célibataire.*

— *Je suis catholique, dans ma religion, on ne divorce pas.*

— *Quand on est marié, on doit rester ensemble.*

Les réponses modales des femmes de la même classe d'âge sont plus longues et plus nuancées :

— *De mon temps, c'était comme ça, bien sûr, ce n'est pas interdit de divorcer, mais il vaut mieux passer bien des choses et rester ensemble.*

— *Parce que je suis catholique pratiquante, et que quand on est marié, on l'est.*

— *Plutôt que d'être malheureux toute sa vie, il vaut mieux divorcer, mais avant d'en arriver là, il vaut mieux supporter beaucoup de choses.*

— *Si le mari est saoul du matin au soir, violent, il vaut mieux se séparer, les enfants sont trop malheureux.*

— *En principe, je suis contre le divorce, sauf dans les cas graves (alcoolisme, mauvais traitements infligés aux enfants et au conjoint).*

On relève une assez nette différence d'expression et d'argumentation entre ces deux catégories de personnes âgées.

La forme *malheureux* est trois fois plus fréquente chez les femmes de cette classe d'âge que dans l'ensemble de la population, la forme *supporter* y est quatre fois plus fréquente. Plus circonstanciées, les réponses des femmes sont en retrait par rapport à celles des hommes.

Il est très probable qu'un post-codage "a priori" aurait désincarné les réponses et donc gommé ces différences qui tiennent à la fois à la tonalité globale de la réponse, à la prolixité du répondant, aux nuances de l'argumentation.

C'est en ce sens que l'exploitation des questions ouvertes constitue un prolongement du questionnaire susceptible de ménager des surprises, mais surtout fournissant les moyens de juger la pertinence du libellé des questions et parfois même des hypothèses sous-jacentes au questionnement.

c) Exemple à partir de noyaux factuels

On reprend l'exemple des 22 noyaux factuels donné au chapitre précédent, section 5.2. Sans donner une liste complète de réponses modales pour l'ensemble des 22 classes, on trouvera, à titre d'exemple, des réponses modales relatives à sept d'entre elles :

Classe 17 (hommes ouvriers sans enfant)
Classe 18 (hommes ouvriers avec enfants)
Classe 9 (femmes au foyer de niveau de vie modeste avec enfants)
Classe 10 (femmes au foyer aisées, avec enfants)
Classe 7 (femmes au foyer sans enfant)
Classe 21 (femmes actives de condition modeste avec enfants)
Classe 11 (femmes actives aisées avec enfants)

Examinons tout d'abord les deux catégories d'*hommes ouvriers* :

Classe 17 (hommes ouvriers sans enfants)

— *Non, le questionnaire est très complet.*
— *Je ne connais pas l'utilité de votre questionnaire, mais je trouve qu'il est pas mal fait.*
— *Questionnaire ridicule qui ne peut servir à rien si ce n'est être utilisé par la police ou les impôts.*

Classe 18 (hommes ouvriers avec enfants)

— *Pourquoi faire des différences dans les familles entre deux et trois enfants au niveau des allocations. Ce n'est pas notre faute si nous n'avons pas pu avoir plus de deux enfants, et c'est injuste de nous pénaliser.*
— *Il y a trop d'injustice au niveau de la répartition des allocations familiales, pour un enfant, on devrait avoir la même chose.*
— *Pour les allocations familiales, trop d'écart entre deux et trois enfants, chaque enfant devrait avoir une somme égale, ce sont tous les mêmes Français.*

— *Pourquoi, avec un petit salaire comme le nôtre avons-nous si peu d'allocations familiales, alors que nous avons deux enfants en étude et ceux-ci étant grands nous coûtent plus cher que des enfants en bas-âge.*

Ces dernières interventions en forme de doléances, souvent longues, contrastent avec les réponses libres des ouvriers sans enfants au foyer. On voit que les différences de forme et de contenu des réponses de ces deux groupes pourtant voisins de personnes enquêtées sont importantes.

La partition en noyaux factuels nous donne trois classes de *femmes au foyer* (classes 7, 9, et 10, selon la figure 5.2 du chapitre 5), dont les réponses seront très diversifiées selon la présence ou l'absence d'enfants au foyer, et également selon le niveau de vie.

Pour la classe 9 (*ménagères de niveau de vie modeste, jeunes, avec enfants*), il y a peu de critiques de l'enquête mais une avalanche de remarques et de revendications, comme le suggèrent les réponses modales suivantes.

Classe 9 (femmes au foyer de niveau de vie modeste avec enfants)

— *On aurait pu aborder le sujet de la femme au foyer pour la prendre plus en considération, on ne parle que des femmes qui travaillent*
— *J'aurais aimé que l'on demande aux mères au foyer, et même aux femmes en général si elles accepteraient mieux de rester chez elles si elles touchaient un certain salaire.*
— *La femme au foyer travaille et n'a pas de salaire ni de retraite en rapport avec son travail, elle est handicapée par rapport à la femme qui travaille*

Pour la classe 10, l'attitude est tout à fait différente, comme le laisse prévoir la position du point correspondant sur la figure 5.3 du chapitre 5, plus à gauche que celui de la classe 9, et donc plus proche des remarques et critiques sur l'enquête.

Classe 10 (femmes au foyer aisées, avec enfants)

— *Les réponses pour certaines questions sont trop rigides, pas assez nuancées.*
— *Le questionnaire tourne en rond sur les questions de salaire, on répète plusieurs fois les mêmes questions.*

Enfin, les réponses de la classe 7 sont assez variées, comme peut le laisser prévoir la position du point correspondant sur la figure 5.3.

Classe 7 (femmes au foyer sans enfant)

— *Le libellé des questions limite trop les réponses et empêche de dire ce qu'on pense.*
— *C'est vraiment parce que l'enquêtrice était sympathique que j'ai répondu jusqu'au bout à votre questionnaire, qui, lui, n'est pas très passionnant.*

— *Sur la sécurité sociale, c'est inadmissible de faire attendre si longtemps les gens sans argent pour liquider un dossier.*

Il est intéressant de constater que les réponses des femmes actives urbaines (classes 11, 20, 21) diffèrent assez peu des réponses des femmes au foyer, le principal critère discriminant étant, comme nous l'avons signalé plus haut, la présence d'enfants au foyer, avec également un rôle important dévolu au niveau de vie.

Classe 21 (femmes actives de condition modeste avec enfants)

— *N'a pas été posée la question du salaire de la femme au foyer, suffisant pour éviter que la femme au foyer ne soit obligée de travailler.*
— *Le questionnaire ne parle pas du tout du travail à mi-temps.*
— *On ferait bien de donner un salaire à la femme au foyer afin de permettre qu'elle reste plus chez elle avec ses enfants, ce qui permettrait d'avoir plus d'enfants et de libérer des emplois.*

Alors que les centres d'intérêt et la tonalité générale des réponses fournies par les femmes de la classe 11 sont fort différents :

Classe 11 (femmes actives aisées avec enfants)

— *Trop long, trop détaillé, je n'aime pas parler de ma vie privée.*
— *Non, sinon j'en reviens à la formation, à diplômes égaux, les hommes sont plus favorisés que les femmes, il y a aussi le problème de la régionalisation dans ce domaine, et cela, on en tient absolument pas compte.*

Ces derniers exemples de classes et de réponses montrent que les partitions en noyaux factuels peuvent mettre en évidence une certaine combinatoire de situations qui permettent de détecter, grâce aux réponses modales, des familles de préoccupations fort différentes. L'intérêt heuristique de la méthode des noyaux factuels tient à la mise en évidence de croisements ou d'interactions qui aurait pu échapper à une sélection faite "a priori" par l'utilisateur.

D'une manière générale, la procédure de sélection des réponses modales apporte des nuances et des compléments importants à l'information fournie par les réponses aux questions fermées. La personne en charge d'analyser l'enquête a sous les yeux la diversité et la complexité de l'individu "répondant", habituellement dissimulées sous des pourcentages qui ne sont neutres qu'en apparence. L'interprétation des réponses donnée par les répondants eux-mêmes mérite aussi d'être prise en considération.

Chapitre 7

Partitions longitudinales, contiguïté

Les textes rassemblés en corpus constituent souvent un ensemble sur lequel nous possédons de nombreuses informations externes (auteur, genre, date de rédaction, dans le cas des corpus rassemblant des textes dus à différents auteurs, catégories socioprofessionnelles, classes d'âge, degré d'instruction dans le cas des agrégats de réponses libres). Ces informations nous permettent d'établir, avant toute expérience de caractère quantitatif visant à rapprocher certaines parties sur la base du stock lexical qu'elles utilisent, une série de relations (même auteur, dates voisines, même genre, etc.).

Le présent chapitre introduit des méthodes qui permettront de confronter analyses lexicales et informations a priori pour un corpus dont les parties sont munies de telles relations : il ne s'agit plus de découvrir ou de redécouvrir des structures, mais d'éprouver la réalité de structures connues, et d'évaluer le niveau de dépendance des typologies obtenues à partir des textes vis-à-vis de ces informations a priori.

7.1 Les trois structures de base

On peut voir sur la figure 7.1 des représentations qui correspondent à trois structures distinctes mises en relation avec un même corpus de textes muni d'une partition en neuf parties. Ces représentations seront formalisées en *graphes de contiguïté* ultérieurement dans ce chapitre.

Les schémas 1 et 2 peuvent correspondre à une situation dans laquelle trois éditorialistes (A, B et C) d'un même quotidien ont écrit à tour de rôle un éditorial pendant neuf jours de parution du journal.

Ces textes ont été rassemblés ensuite en un même corpus lexicométrique qui compte neuf parties. On peut désigner la suite de ces articles par :

A1, B1, C1, A2, B2, C2, A3, B3, C3

Le corpus ainsi constitué peut être examiné par rapport à deux structures a priori, notées S1, et S2.

La première de ces structures, S1, tient compte exclusivement de la variable *auteur de l'article*. C'est-à-dire que l'on considère comme contigus pour cette structure, les articles dûs à un même auteur.

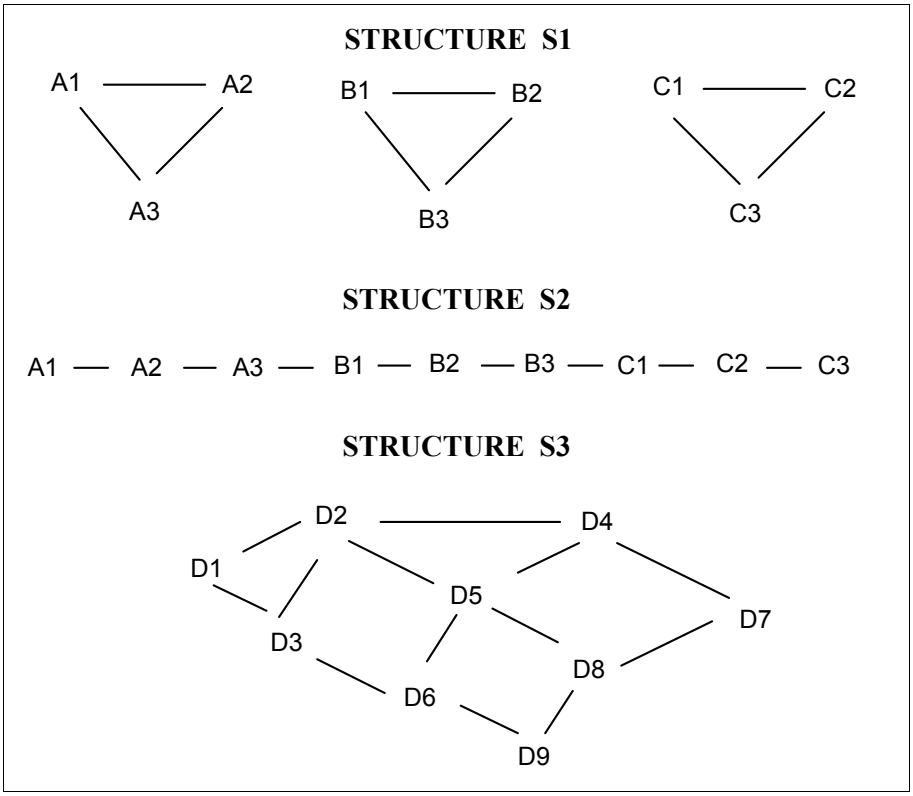


Figure 7.1

Graphes correspondant à trois types de structures courantes : Partition, chronologie, graphe non orienté.

La structure S2 rend compte de la variable *consécutivité dans le temps*. Sont considérés comme contigus, dans ce second cas, les articles rédigés consécutivement sans considération d'auteur.

On rencontre des structures plus générales, représentées par le schéma S3 qui correspond au cas d'un corpus entre les parties duquel existeraient des relations de contiguïté (thématiques, appartenance politique, documents relatifs à des régions limitrophes, etc.).

Les structures de *partition* (type S1) sont d'un usage courant en statistique. L'outil privilégié dans le cas où les variables étudiées sont numériques est

l'*analyse de la variance*, technique classique qui permet de tester l'hétérogénéité de certains regroupements de variables.

Les structures de *séries chronologiques* (type S2) sont étudiées par les méthodes d'une autre branche importante de la statistique : les *séries temporelles et processus stochastiques.*

Les structures de *graphe* (type S3), plus générales, relèvent de techniques moins largement répandues et moins développées que les précédentes, les *analyses statistiques de contiguïté*. Cependant, ces techniques permettent de procéder facilement aux inférences correspondant aux situations les plus courantes. Dans la mesure où elles embrassent comme cas particuliers les deux types S1 et S2, ces techniques fournissent un outil unique pour la prise en compte de la plupart des structures a priori susceptibles d'être observées par le praticien de la statistique textuelle.

Dans la phase délicate de l'interprétation des résultats, les méthodes de l'analyse statistique de la contiguïté interviennent comme complément des analyses purement exploratoires. Elles permettront en fait de prolonger un temps la démarche formelle et de réduire encore la part de subjectivité inhérente à tout commentaire.

Ce chapitre se poursuit par un rappel qui concerne l'analyse statistique de la contiguïté dans le cas, le plus simple, de l'analyse des valeurs d'une seule variable (7.2). La démarche est ensuite étendue à l'étude des facteurs issus d'une analyse des correspondances réalisée à partir d'un tableau lexical (7.3). Le paragraphe (7.4) est consacré à l'application des méthodes présentées à un corpus expérimental composé d'agrégats de réponses libres à une question ouverte. On introduit ensuite (7.5) la problématique particulière des structures longitudinales et l'on présente une application particulièrement importante des analyses de données longitudinales dans le domaine textuel : les séries textuelles chronologiques (7.6). Enfin, une application de l'analyse de la contiguïté au problème de la *recherche en homogénéité d'auteurs* terminera ce chapitre (7.7).

7.2 Homogénéité des valeurs d'une variable par rapport à une partition.

Commençons par le cas le plus simple d'une variable prenant des valeurs numériques sur les éléments d'un ensemble. Par la suite, les notions introduites seront utilisées pour étudier l'ensemble des facteurs sur l'ensemble des parties issus de l'analyse des correspondances d'un tableau lexical. Dans ce qui suit, J symbolise un ensemble d'éléments, *p* le nombre

de ces éléments. Il est possible de définir sur l'ensemble (J ∞ J) une relation dite : *relation de contiguïté*. Nous noterons cette relation R(j,j'). Cette relation est symétrique (si R(j,j'), alors R(j',j)). Nous supposerons ici de plus que cette relation n'est pas réflexive (R(j,j) n'est jamais vrai) ce qui revient à poser, de façon conventionnelle, que chaque partie j n'est pas contiguë à elle même.[1]

7.2.1 Graphe associé à une structure de contiguïté

Cette relation de contiguïté peut être représentée de différentes façons. Une des représentations traditionnelles s'effectue à l'aide d'un graphe. Dans ce qui suit nous désignerons ces *graphes de contiguïté* par les symboles G, éventuellement indicés (G_1, G_2, G_3...).

Pour cette représentation, chaque élément j de l'ensemble J correspond à un *sommet* du graphe. Chacun des couples de sommets du graphe liés par la relation de contiguïté R(j,j') est relié par une *arête*. Le nombre des arêtes adjacentes à un même sommet j est appelé le *degré* de ce sommet. Nous noterons ce nombre m_j. Si tous les sommets sont reliés par une arête, le graphe est dit *complet*. Un tel graphe possède *p(p-1)* arêtes.

Il est usuel de distinguer les arêtes selon leur sens. Ainsi, l'arête qui va du sommet j au sommet j' est, en principe, distincte de celle qui joint j' à j. On note *m* le nombre des arêtes du graphe.

En fait on se limitera ici, pour les relations a priori qui nous intéressent, aux graphes symétriques (tels que R(j,j') □ R(j',j)), que l'on peut aussi considérer comme des graphes non-orientés avec *m/2* arêtes.

7.2.2 Matrices de contiguïté

On peut également représenter la relation de contiguïté R(j,j') par le biais d'une matrice carrée **M** dite *matrice de contiguïté*. Cette matrice dans laquelle chaque élément vaut 0 ou 1 possède autant de lignes et de colonnes que l'ensemble J compte d'éléments.

Son terme général $m_{jj'}$ vaut :

[1] Dans la littérature sur le sujet on définit parfois des structures de contiguïté qui incluent des proximités à distance 1, 2, ...n , cf. par exemple Lebart (1969). Nous nous limiterons ici aux structures de contiguïté pour lesquelles deux parties sont immédiatement contiguës (distance 1) ou disjointes bien que les résultats soient également généralisables à des structures de contiguïté plus complexes.

$m_{jj'} = 1$ si j et j' sont contigus ;
$m_{jj'} = 0$ dans le cas contraire.

On voit que cette matrice est symétrique du fait de la symétrie de la relation de contiguïté. Selon la convention posée plus haut, les termes situés sur la diagonale principale de la matrice **M** sont tous nuls.

Tableau 7.1 **Matrice de contiguïté relative à la structure S1**

	A1	B1	C1	A2	B2	C2	A3	B3	C3
A1	0	0	0	1	0	0	1	0	0
B1	0	0	0	0	1	0	0	1	0
C1	0	0	0	0	0	1	0	0	1
A2	1	0	0	0	0	0	1	0	0
B2	0	1	0	0	0	0	0	1	0
C2	0	0	1	0	0	0	0	0	1
A3	1	0	0	1	0	0	0	0	0
B3	0	1	0	0	1	0	0	0	0
C3	0	0	1	0	0	1	0	0	0

Tableau 7.2 **Matrice de contiguïté relative à la structure S2**

	A1	B1	C1	A2	B2	C2	A3	B3	C3
A1	0	1	0	0	0	0	0	0	0
B1	1	0	1	0	0	0	0	0	0
C1	0	1	0	1	0	0	0	0	0
A2	0	0	1	0	1	0	0	0	0
B2	0	0	0	1	0	1	0	0	0
C2	0	0	0	0	1	0	1	0	0
A3	0	0	0	0	0	1	0	1	0
B3	0	0	0	0	0	0	1	0	1
C3	0	0	0	0	0	0	0	1	0

Les relations suivantes lient le nombre d'arêtes m, les degrés m_j , et les éléments $m_{jj'}$ de la matrice **M** :

$$m = \sum_{j=1}^{p} m_j = \sum_{j=1}^{p} \sum_{j'=1}^{p} m_{jj'},$$

Les tableaux 7.1 et 7.2 donnent les matrices de contiguïté correspondant aux structures S1 et S2.

Si la matrice du tableau 7.1 voit ses lignes et ses colonnes réordonnées simultanément, de façon à regrouper les symboles par leurs premières lettres A, B ou C, on voit mieux apparaître les trois groupes qui constituent cette structure (tableau 7.3).

Tableau 7.3

Matrice de contiguïté relative à la structure S1, après reclassement des lignes et des colonnes

	A1	A2	A3	B1	B2	B3	C1	C2	C3
A1	**0**	**1**	**1**	0	0	0	0	0	0
A2	**1**	**0**	**1**	0	0	0	0	0	0
A3	**1**	**1**	**0**	0	0	0	0	0	0
B1	0	0	0	**0**	**1**	**1**	0	0	0
B2	0	0	0	**1**	**0**	**1**	0	0	0
B3	0	0	0	**1**	**1**	**0**	0	0	0
C1	0	0	0	0	0	0	**0**	**1**	**1**
C2	0	0	0	0	0	0	**1**	**0**	**1**
C3	0	0	0	0	0	0	**1**	**1**	**0**

7.2.3 Le coefficient de contiguïté

Imaginons maintenant une variable *Z* prenant des valeurs z_j sur chacun des sommets du graphe G défini plus haut. Ces valeurs peuvent correspondre à des mesures faites sur chacune des parties du corpus (la longueur moyenne des phrases dans chaque partie par exemple ou toute autre grandeur ne dépendant pas directement de la longueur de chacune des parties[1]).

Nous commencerons par calculer la moyenne $\bar{z}$ des valeurs de cette variable sur les sommets du graphe G.

$$\bar{z} = \frac{1}{p} \sum_{j=1}^{p} z_j$$

La *variance empirique* de la variable Z vaut :

$$v(Z) = \frac{1}{2p(p-1)} \sum_{j'=1}^{p} \sum_{j=1}^{p} (z_j - z_{j'})^2 \qquad (1)$$

que l'on notera aussi :

$$v(Z) = \frac{1}{2p(p-1)} \Sigma^{jj'} (z_j - z_{j'})^2$$

On retrouve en développant la partie droite, la formule plus classique :

[1] En effet, si la variable z dépend directement de la longueur des parties du corpus l'étude de sa variation au fil des parties risque fort de se confondre avec l'étude de la différence de longueur entre les parties.

$$v(Z) = \frac{1}{(p-1)} \sum_{j=1}^{p} (z_j - \bar{z})^2 ,$$

La variance *v(Z)*, que l'on appelle ici variance totale, mesure la dispersion des valeurs z_j autour de leur moyenne $\bar{z}$ sur l'ensemble des sommets du graphe G.

On calcule la *variance locale v*(Z)* de la variable Z sur les seuls sommets contigus du graphe G au moyen d'une formule analogue à la formule (1) mais pour laquelle la sommation s'effectue cette fois pour les seuls couples j et j' qui sont reliés par une arête du graphe (au lieu de la sommation sur tous les j et j'). La sommation selon ce dernier critère est représentée par le symbole Σ^*.

$$v^*(Z) = \frac{1}{2m} \Sigma^* (z_j - z_{j'})^2$$

ou de façon équivalente :

$$v^*(Z) = \frac{1}{2m} \sum_{j'=1}^{p} \sum_{j=1}^{p} m_{jj'} (z_j - z_{j'})^2 ,$$

que l'on note aussi :

$$v^*(Z) = \frac{1}{2m} \Sigma^{jj'} m_{jj'} (z_j - z_{j'})^2,$$

Rappelons que dans ces formules[1]:

— $m = \sum_{j=1}^{p} m_j = \Sigma^{jj'} m_{jj'}$, est le nombre total des connexions.

— Σ^* indique une sommation faite pour l'ensemble des sommets j et j' tels que j et j' soient contigus.

— $\Sigma^{jj'}$ indique une sommation sur l'ensemble des sommets j et j' pris deux à deux.

— $\bar{z} = (1/p) \sum_{j=1}^{p} z_j$, est la moyenne des z_j.

[1] Il faut remarquer que, compte tenu des notations que nous avons adoptées, la distinction (ou la non-distinction) des arêtes selon leur orientation amène un calcul légèrement différent pour un résultat identique.

Le coefficient de contiguïté $c(Z)$ (Geary, 1954) se calcule alors comme :

$$c(Z) = \frac{v^*(Z)}{v(Z)} = \frac{p(p-1)\sum^{*}(z_j - z_{j'})^2}{m\sum^{jj'}(z_j - z_{j'})^2}$$

Ce rapport est proche de l'unité si la variance locale (la variance mesurée sur les sommets contigus) est proche de la variance mesurée sur l'ensemble des sommets du graphe, ce qui indique une présomption de répartition aléatoire sur le graphe. Plus il est proche de zéro, plus on peut dire que la variable z prend sur les sommets contigus des valeurs voisines par rapport aux sommets pris deux à deux de façon quelconque. Si le coefficient est nettement supérieur à l'unité on en conclura que les valeurs de la variable *z* sont, au contraire, "anormalement" différentes sur les sommets voisins.

7.2.4 Calcul des moments du coefficient de contiguïté

Sous l'hypothèse selon laquelle les valeurs z_j peuvent être considérées comme des réalisations de variables aléatoires normales indépendantes, on peut calculer les quatre premiers moments du coefficient $c(Z)$[1].

A partir de ces moments, on détermine μ_1, μ_2, μ_3 et μ_4, moments centrés d'ordre 2, 3 et 4 pour cette même distribution.

Le coefficient d'asymétrie de Fisher :

$$\gamma_1 = \mu_3 / \sigma^3 = \mu_3 / \mu_2^{3/2}$$

et le coefficient d'aplatissement :

$$\gamma_2 = \mu_4 / \sigma^4 - 3 = \mu_4 / \mu_2^2 - 3$$

permettent de comparer la distribution empirique à une loi normale. Si γ_1 et γ_2 sont proches de zéro on appliquera, en première approximation, le test de l'écart réduit pour juger de l'écart du coefficient $c(Z)$ par rapport à l'unité.

7.2.5 Un cas particulier : les séries temporelles

Le coefficient $c(Z)$ constitue une généralisation du coefficient de Von Neumann (1941). En effet, dans le cas où la relation de contiguïté se réduit à une relation de consécutivité comme c'est le cas, par exemple, pour la structure S2 présentée plus haut, le numérateur du coefficient de contiguïté peut s'écrire :

[1] Pour un exposé plus complet cf. Lebart (1969). Pour d'autres applications de la notion de contiguïté, cf. Aluja Banet et al. (1984), Burtschy et al. (1991).

$$c = \frac{1}{2(p-1)} \frac{\sum_{j=2}^{p} (z_j - z_{j-1})^2}{v(Z)}$$

puisque le nombre total des connexions dans une structure de consécutivité avec p éléments est égal à *p-1*.

Dans ce dernier cas, le coefficient de Geary est une des formes du coefficient (plus classique) de Von Neumann qui mesure l'autocorrélation d'une suite de valeurs z_j.

7.2.6 Utilisation du coefficient *c*

Revenons à l'exemple évoqué plus haut et imaginons une variable *Z* prenant des valeurs sur les neuf parties d'un corpus (par exemple, la fréquence relative des occurrences de la forme *de*). Nous commencerons par calculer le coefficient *c(Z)* pour les structures de contiguïté S_1 et S_2. Notons ces nombres respectivement $c(Z \mid S_1)$ et $c(Z \mid S_2)$.

Si le coefficient $c(Z \mid S_1)$ est très proche de l'unité, nous dirons que la variable z_j ne subit pas de variation notable liée à la différence d'auteur. Dans le cas où le coefficient $c(Z \mid S_2)$ serait très proche de l'unité nous dirions que la variable z_j ne subit pas d'effet chronologique.

La proximité à l'unité s'analyse dans la pratique en fonction des moments calculés au paragraphe 7.2.4. Dans tous les cas cependant, si les valeurs z_j varient moins pour les textes produits par un même auteur qu'elles ne varient par rapport au temps, le calcul du coefficient $c(Z \mid S_1)$ amènera une valeur nettement plus petite que la valeur $c(Z \mid S_2)$. Dans le cas contraire, d'un coefficient $c(Z \mid S_2)$ plus petit nous dirons que c'est l'effet chronologique qui prime au contraire sur l'effet "auteur".

7.3 Homogénéité des facteurs d'une analyse des correspondances en fonction d'une structure a priori

Comme plus haut, on supposera que l'on est en présence d'un corpus de textes divisé en *p* parties par la partition P. De plus, on dispose d'une structure S1, structure a priori de contiguïté sur l'ensemble des parties.

A cette structure correspond la matrice de contiguïté **M**. A partir de ce corpus, nous avons construit la table de contingence (formes x parties) correspondant à cette partition. Ce tableau qui compte n lignes et p colonnes

a ensuite été soumis à une analyse des correspondances afin d'en extraire *(p-1)* facteurs.

Les facteurs sur l'ensemble des parties sont notés $F_\alpha(j)$ où l'indice α indique le numéro du facteur et varie de 1 à *(p-1)*. L'indice j indique ici le numéro de partie et varie entre 1 et *p*.

7.3.1 Homogénéité d'un facteur par rapport une structure.

Le calcul du coefficient de contiguïté pour la suite des valeurs $F_\alpha(j)$ permet de mesurer le lien entre ce facteur et la structure de contiguïté S1. Pour chaque facteur ce coefficient se calcule selon la formule employée pour les variables numériques :

$$c\,(F_\alpha) \;=\; \frac{(p-1)\ \sum^{*}\,(F_\alpha(j) - F_\alpha(j'))^2}{2m\ \sum^{j}\,(F_\alpha(j) - F_\alpha(.))^2} \qquad (2)$$

Dans ce qui suit nous noterons ce coefficient c_α.

On a noté $F_\alpha(.) = \frac{1}{p}\sum_{j=1}^{p} F_\alpha(j)$ la moyenne des valeurs du facteur. Cette quantité est nulle si toutes les parties ont le même poids.

7.3.2 Homogénéité dans l'espace des k premiers facteurs.

L'analyse réalisée pour chacun des facteurs pris isolément par rapport à une structure de contiguïté peut être heureusement complétée par une analyse portant sur la distance entre les éléments de l'ensemble J dans l'espace engendré par les *k* premiers facteurs.

Le carré de la distance du chi-2 entre deux points j et j' s'écrit en fonction du tableau des profils :

$$d^2(j,j') \;=\; (1/f_i)\sum_{i=1}^{N} (f_{ij}/f_{.j} - f_{ij'}/f_{.j'})^2$$

Cette même quantité s'écrit :

$$d^2(j,j') \;=\; \sum_{\alpha=1}^{n_f} \left(F_\alpha(j) - F_\alpha(j')\right)^2$$

en fonction des n_f facteurs issus de l'analyse des correspondances.

Le carré de la distance entre deux points j et j' dans l'espace engendré par les k premiers facteurs s'écrit :

$$d_k^2(j,j') = \sum_{\alpha=1}^{k} \left(F_\alpha(j) - F_\alpha(j')\right)^2$$

La quantité :

$$G_k = \frac{p(p-1) \sum^{*} d_k^2(j,j')}{m \sum^{jj'} d_k^2(j,j')} \qquad (3)$$

constitue donc un coefficient *c* calculé à partir des distances $d_k^2(j,j')$ prises cette fois dans le sous-espace des *k* premiers facteurs.

Comme on le voit d'après les formules (2) et (3), dans le cas ou *k* est égal à 1 on a l'égalité $G_1 = c\ (F_1)$. Lorsque *k* est égal à *(p-1)* nombre total des facteurs issus de l'analyse des correspondances, le coefficient G_k rend compte du rapport entre la variance "sur la structure" et la variance totale pour la distance du chi-2 calculée sur le tableau des fréquences. Les valeurs intermédiaires de *k* correspondent à des distances filtrées par les premiers facteurs.

Dans la mesure où les parties n'ont pas toutes la même importance numérique, on calculera dans le cas de corpus dont les parties sont de taille très dissemblable la quantité :

$$G_k = \frac{p(p-1) \sum^{*} p_{.j}\, p_{.j'}\, d_k^2(j,j')}{m \sum^{jj'} p_{.j}\, p_{.j'}\, d_k^2(j,j')} \qquad (4)$$

pour laquelle les distances $d_k^2(j,j')$ sont pondérées par le poids relatif de chacune des parties.

7.4 Les agrégats de réponse et l'analyse de la contiguïté

Les coefficients que nous venons d'introduire vont nous permettre de formaliser le dépouillement des résultats de l'analyse des correspondances sur le tableau des agrégats présenté au chapitre 5 (table 5.3, puis figure 5.1). Dans le cas des agrégats de réponses constitués à partir de la variable

nominale composite (âge x diplôme), la matrice de contiguïté $\mathbf{S_{ad}}$ (ad : comme âge x diplôme) peut être représentée par le graphe de la figure 7.2. Pour chacun des axes factoriels issus de l'analyse du tableau (formes x agrégats), on calcule alors les coefficients c_α et G_α

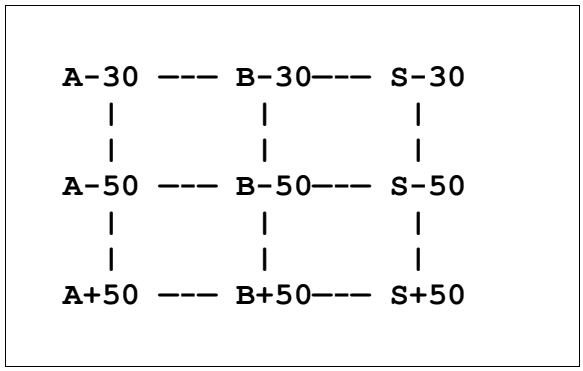


Figure 7.2

Graphe de contiguïté décrivant la structure des colonnes de la table 5.3 (chapitre 5)

Structure et filtrage

L'exemple qui suit concerne l'analyse d'un tableau dont les lignes correspondent aux 321 formes les plus fréquentes pour la question ***Enfant*** (au lieu de 117 pour l'exemple du chapitre 5) et les colonnes correspondent à la même partition en 9 classes croisant *âge* et *niveau de diplôme*.

Tableau 7.4

Valeurs propres issues de l'analyse d'un tableau (321 formes x 9 agrégats) et coefficients c_α et G_α calculés par rapport à la structure de contiguïté S_{ad} pour chacun des facteurs

Axe	*Valeur propre*	c_α	G_α
1	0.057	0.397	0.397
2	0.037	0.400	0.398
3	0.028	1.489	0.698
4	0.027	1.071	0.761
5	0.025	1.190	0.849
6	0.022	1.120	0.873
7	0.021	1.206	0.909
8	0.017	1.093	0.926

Comme on pouvait le prévoir au vu des résultats de l'analyse de ce tableau (figure 5.1) les coefficients c_α et G_α qui correspondent aux deux premiers facteurs ont ici des valeurs faibles, signe d'une bonne représentation de la structure $\mathbf{S_{ad}}$ par le plan des deux premiers axes. Le calcul de ces mêmes coefficients pour les axes suivants permet de vérifier que ces axes ne possèdent pas cette propriété constatée à propos des deux premiers.

Dans le cas particulier du corpus ***Enfant*** muni de la partition (âge x diplôme) que nous venons d'étudier, les calculs de contiguïté (tableau 7.4) confirment un résultat aisément perceptible lors de l'examen du plan des deux premiers facteurs (figure 5.1, du chapitre 5). Mais ils montrent surtout que ces *deux premiers facteurs* ont d'une certaine façon *l'exclusivité* de cette structure. La croissance du coefficient G_α montre également que cette structure est de plus en plus floue lorsque la dimension augmente : le coefficient de contiguïté vaut 0.926 pour les distances calculées dans l'espace à 8 dimensions.

Autrement dit, le lien entre la structure a priori et les distances (du chi-2) globales est faible. La propriété de *filtrage* des premiers axes de l'analyse des correspondances, si souvent remarquée par les praticiens, et cependant si difficile à définir et étayer théoriquement, est encore vérifiée sur cet exemple. Le squelette schématisé par la figure 7.2 est à peine décelable dans l'espace, et pourtant représenté sans interversion par le premier plan factoriel (figure 5.1). On emploie ici à dessein le mot squelette, car l'analyse des correspondances a été souvent comparée à un appareil radiographique.

7.5 Partitions longitudinales d'un corpus

La relation de consécutivité constitue le cas le plus simple de relation de contiguïté sur les parties d'un corpus. On obtient des partitions de ce type, lorsque l'on construit, par exemple, des agrégats de réponses à une question ouverte à partir des valeurs d'une même variable numérique (âge, revenu mensuel, ou nombre d'enfants) ou encore à partir d'une variable dont les différentes modalités peuvent être rangées dans un ordre qui correspond à un degré d'intensité pour cette variable (niveau d'instruction par exemple). Tel est le cas encore lorsqu'on range les parties d'un corpus de textes selon un ordre chronologique (rédaction, publication, etc.)

Dans ce qui suit on appellera ce type de partition : des *partitions longitudinales* d'un corpus en fonction d'une variable. On examinera successivement le cas de réponses à une question ouverte regroupées selon les modalités d'une variable ordonnée (classes d'âges), puis, dans le

paragraphe suivant, dans le cadre plus général des séries textuelles chronologiques, le cas d'une série de discours politiques.

7.5.1 Exemple de partition longitudinale

On désigne par *corpus* ***Life*** un ensemble des réponses fournies par 1 000 personnes de langue anglaise à une question ouverte portant les choses qu'elles jugent personnellement importantes pour elles dans la vie[1]. Le libellé de la question était : *"What is the single most important thing in the life for you ? "* suivie d'une relance *"What other things are very important to you ?"*.

En constituant des agrégats de réponses fondés sur l'appartenance des répondants à une même classe d'âge, nous pouvons espérer mettre en évidence des variations liées à la variable "âge du répondant".

Tableau 7.5

Découpage en 6 classes du corpus des réponses à la question *Life* d'après l'âge des répondants.

code	val.	occ.	formes	hapax	fmax	
Ag1	18-24	2003	429	247	112	my
Ag2	25-34	2453	512	295	148	my
Ag3	35-44	2529	510	290	151	family
Ag4	45-54	2519	493	258	159	my
Ag5	55-64	1864	459	264	110	my
Ag6	65-++	3326	623	333	144	health

Le tableau 7.5 montre les bornes retenues pour effectuer le découpage des agrégats en 6 classes notées (Ag1, Ag2, …, Ag6) en fonction de l'âge. Ce découpage amène, on le voit, des parties dont la taille est comparable. Remarquons que le simple calcul de la fréquence maximale suffit à faire apparaître une préoccupation *(health)* propre aux personnes les personnes les plus âgées.

[1] Cette question est extraite d'une enquête par sondage réalisée à la fin des années quatre-vingt au Royaume Uni. Il s'agit du volet britannique d'une enquête comparative internationale réalisée auprès de cinq pays (Japon, Allemagne, France, Grande Bretagne et USA,) sous la direction des Pr. C. Hayashi, M. Sasaki et T. Suzuki. Pour la série de travaux dans lesquels s'inscrit cette opération, cf. Hayashi et al.(1992). Cf. aussi Sasaki et Suzuki (1989), Suzuki (1989). Cette enquête se situe elle-même dans la lignée d'une enquête permanente sur le *Japanese National Character* réalisée régulièrement depuis 35 ans au Japon (Hayashi, 1987). Les réponses des individus à la question et à la relance ont été regroupées ici. Les différences dans l'ordre des réponses fournies par les sujets interrogés font par ailleurs l'objet d'une étude particulière.

Classification sur les agrégats

La figure 7.3 montre le résultat d'une classification ascendante hiérarchique (selon le critère de Ward généralisé, comme les méthodes présentées au chapitre 4) obtenue à partir de cette partition. Le dendrogramme obtenu à partir de l'analyse du tableau (6 classes d'âge, 483 formes de fréquence supérieure ou égale à 3) possède une propriété particulière : ses noeuds rassemblent toujours des classes d'âge consécutives.

Les classes extrêmes (Ag1 et Ag6) s'agrègent très haut, dans l'arbre de classification, à l'ensemble constitué par les autres classes, signe d'une particularité plus marquée par rapport aux classes intermédiaires. Comme plus haut, nous interpréterons cette circonstance en concluant que les répondants qui appartiennent à des classes d'âge voisines produisent des réponses plus semblables entre elles que les répondants qui appartiennent à des classes d'âge plus distantes.

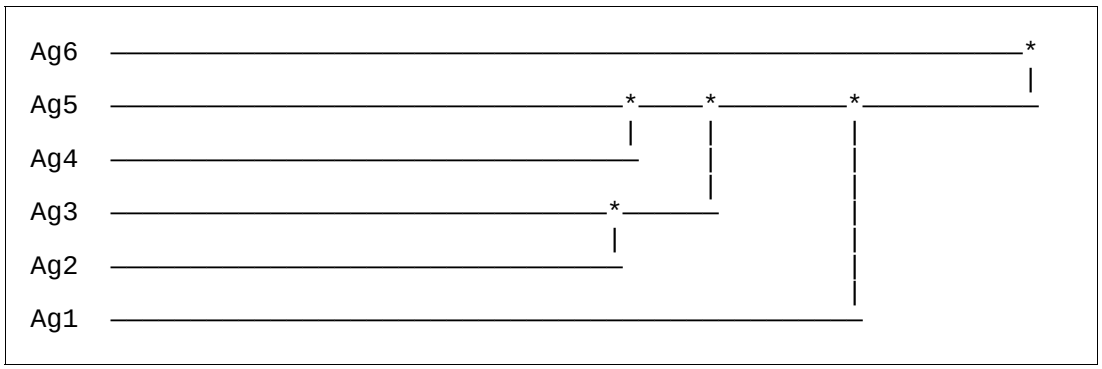


Figure 7.3

Classification ascendante hiérarchique réalisée à partir d'une partition en 6 classes d'âge de l'ensemble des réponses à la question *Life*.

7.5.2 Analyse de la gradation "classe d'âge"

L'analyse des correspondances réalisée à partir d'un tableau, pour lequel l'ensemble des distances entre agrégats est dominé par l'existence d'une gradation, produit des résultats d'un type particulier.

Dans le cas ou la gradation est absolument régulière, la représentation des éléments sur les plans engendrés par les premiers facteurs acquiert une forme très caractéristique. Les différents facteurs à partir du second se révèlent des fonctions du premier (quadratique pour le facteur numéro 2, cubique pour le facteur numéro 3, etc.). L'analyse des correspondances représente de manière

plutôt complexe (en introduisant des facteurs qui sont des fonctions polynomiales du premier) un phénomène qui est fondamentalement de nature unidimensionnelle. Dans la littérature statistique ce phénomène est connu sous le nom d'*effet Guttman*, du nom du statisticien qui étudia de manière systématique[1] les tableaux de données correspondants.

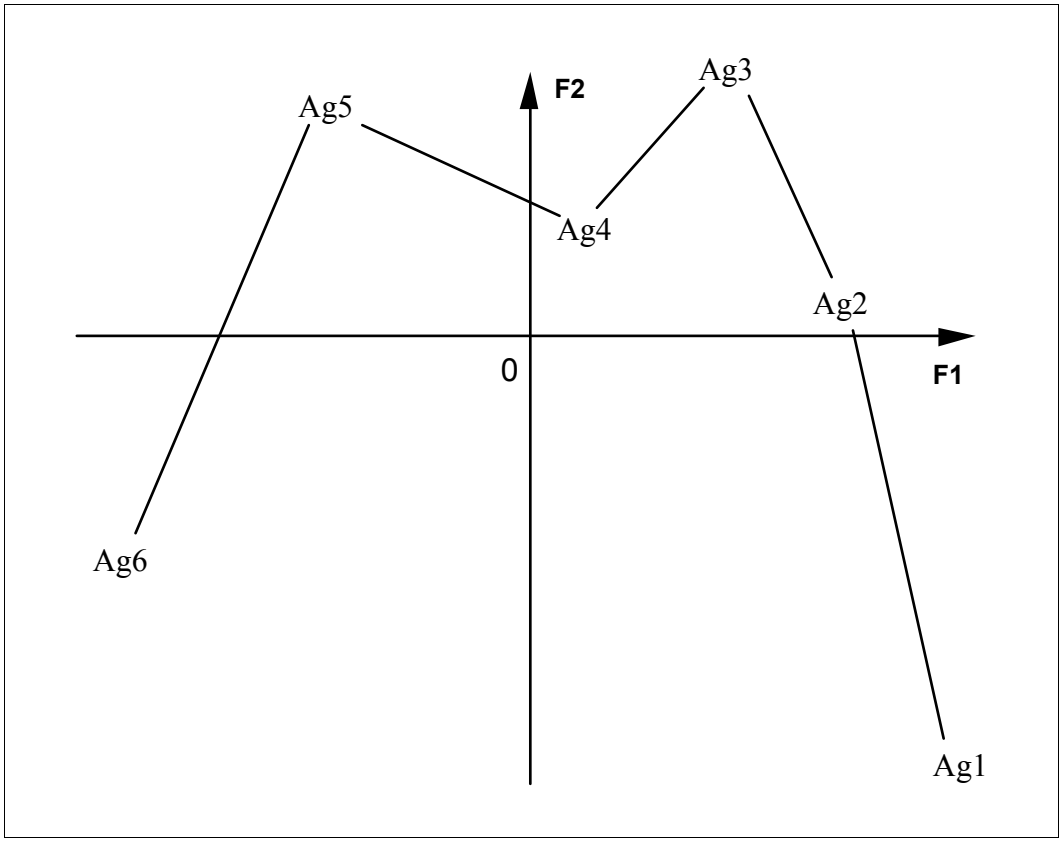


Figure 7.4

L'effet Guttman : représentation schématique du plan factoriel 1 x 2 tableau (6 classes d'âges x 483 formes).
Question ouverte *Life*

Devant des résultats de ce type, on évitera donc, en général, de commenter séparément les oppositions constatées sur chacun des axes factoriels. Ici, le schéma classique de l'interprétation d'une typologie que l'on affine au fur et à mesure par la prise en compte de nouveaux axes factoriels doit faire place à la reconnaissance d'une situation caractéristique globale liée à l'existence et à la dominance d'une gradation.
Lors de l'analyse des correspondances du tableau des distances calculées sur les stocks lexicaux correspondant à chacune des parties du corpus des

[1] Cf. Benzécri (1973) Tome IIb, chapitre 7, Cf. aussi Guttman (1941) et Van Rijckevorsel (1987).

réponses à la question ***Life*** - figure 7.4 - les points correspondants aux classes d'âge (Ag1, ... , Ag6) dessinent sur le plan des deux premiers facteurs une courbe incurvée en son centre qui rappelle le schéma théorique mentionné plus haut. L'examen des facteurs suivant confirme l'impression que l'analyse est bien dominée par l'existence d'une gradation.

Les formes et les segments pourvus d'une coordonnée factorielle extrême correspondent dans ce cas à des termes utilisés par les plus jeunes ou les plus âgés des répondants. Cependant, pour mettre en évidence les termes qui sont responsables, au plan quantitatif, de la gradation d'ensemble constatée sur les premiers facteurs, il faudra mettre en oeuvre des méthodes plus spécifiques.

7.5.3 Spécificités connexes

Pour répondre à cette dernière préoccupation, on va construire des indicateurs permettant de repérer les termes, formes ou segments, dont la ventilation présente des caractéristiques qui peuvent être mises en relation avec la gradation d'ensemble. Considérons un terme de fréquence totale F dont la ventilation dans les n parties d'un corpus est donnée par la liste des sous-fréquences :

$$f_1,\ f_2,\ f_3,\ \ldots,\ f_n.$$

Comme toujours, T désigne la longueur totale du corpus et la liste : $t_1,\ t_2,\ t_3,\ \ldots,\ t_n$, donne la longueur de chacune des parties.

On a vu au chapitre 6 que ces éléments permettent de calculer, pour chacune des sous-fréquences f_i, et pour un seuil de probabilité s fixé, un diagnostic de spécificité S_i qui sera symbolisé par "S+", "S-" ou "b" suivant que nous aurons affaire à une spécificité positive, négative ou à une forme banale pour la partie i.

Convenons encore d'affecter à ces diagnostics un indice supérieur, égal à 1, qui indiquera que nous avons affaire à des spécificités de premier niveau, c'est-à-dire des spécificités portant sur des périodes élémentaires du corpus prises isolément.

Dans ces notations, la liste des spécificités de premier niveau s'écrit donc :

$$S^1_1,\ S^1_2,\ S^1_3,\ \ldots\ S^1_n,$$

On appellera spécificités de niveau 2 les spécificités calculées de la même manière pour chaque couple de parties consécutives.

Tableau 7.6

Spécificités longitudinales majeures pour le corpus des réponses *Life* découpé en 6 classes d'après l'âge des répondants.

terme	F	fp	sp	code
to do	25	13	+E06	Ag1-
being	116	32	+E05	Ag1-
friends	116	32	+E05	Ag1-
good job	13	8	+E05	Ag1-
having	70	20	+E04	Ag1-
a good job	6	5	+E04	Ag1-
a good	54	18	+E04	Ag1-
a	300	62	+E04	Ag1-
money	170	80	+E06	Ag1-Ag2
career	11	9	+E04	Ag1-Ag2
be happy	41	23	+E04	Ag1-Ag2
to get a	6	6	+E04	Ag1-Ag2
being happy	30	19	+E04	Ag1-Ag2
my job	37	29	+E04	Ag1-Ag3
education	25	20	+E04	Ag1-Ag3
job	143	128	+E12	Ag1-Ag4
work	118	100	+E07	Ag1-Ag4
family	686	582	+E07	Ag1-Ag5
my	809	670	+E05	Ag1-Ag5
my family	225	194	+E04	Ag1-Ag5
happiness	227	198	+E04	Ag1-Ag5
the children	25	21	+E07	Ag2-Ag3
children	125	66	+E05	Ag2-Ag3
wealth	11	10	+E04	Ag2-Ag3
of	312	193	+E05	Ag2-Ag4
can	35	8	-E04	Ag2-Ag4
do	47	17	-E05	Ag2-Ag5
security	40	37	+E05	Ag2-Ag5
want	31	10	-E04	Ag2-Ag5
to keep	27	8	-E04	Ag2-Ag5
keep	50	19	-E04	Ag2-Ag5
health	611	564	+E06	Ag2-Ag6
welfare	22	11	+E04	Ag3-
welfare of	13	8	+E04	Ag3-
the	331	155	+E06	Ag3-Ag4
the family	78	47	+E06	Ag3-Ag4
need	10	9	+E04	Ag3-Ag4
to	522	203	-E05	Ag3-Ag5
are	65	57	+E04	Ag3-Ag6
me	33	31	+E04	Ag3-Ag6
be	137	10	-E04	Ag4-
grandchildren	30	30	+E09	Ag4-Ag6
I	248	158	+E04	Ag4-Ag6
they	24	21	+E04	Ag4-Ag6
religion	13	13	+E04	Ag4-Ag6
good health	176	117	+E04	Ag4-Ag6
as	83	44	+E04	Ag5-Ag6
worry	8	7	+E04	Ag6-

Guide de lecture du tableau 7.6

Les spécificités calculées concernent les sous-fréquences des termes (formes ou segments) dont la fréquence dépasse 5 occurrences dans le corpus. Seuls les diagnostics les plus importants ont été retenus dans ce tableau (indice de probabilité inférieur à 10^{-4}).

Les diagnostics retenus ont été reclassés en fonction des catégories caractéristiques sélectionnées par la méthode (des plus jeunes aux plus âgés).

La colonne F indique la fréquence du terme dans l'ensemble du corpus.

La colonne fp indique la fréquence partielle du terme dans la partie ou le groupe de parties consécutives pour lesquelles le diagnostic de spécificité a été établi.

Le diagnostic de spécificité est constitué par un signe et un exposant (+Exx renvoie à une spécificité positive à laquelle est attachée une probabilité d'ordre 1/10 à la puissance xx).

La partie, ou le groupe de parties consécutives, du corpus pour laquelle la spécificité positive à été établie apparaît entre deux barres verticales.

Dans le cas des spécificités de niveau 1, la comparaison porte, pour un terme i et une partie j donnés, sur les quatre nombres $T, t_i, F, f_i,$ la spécificité de niveau 2 se calcule donc à partir des nombres : $T, t_i+t_{i+1}, F, f_i+f_{i+1}$. Les spécificités de niveau 2 sont au nombre de $n-1$. Elles sont notées :

$$S^2_1, S^2_2, S^2_3, \ldots S^2_{n-1},$$

Comme plus haut, l'indice supérieur indique le niveau des spécificités auxquelles nous avons affaire. L'indice inférieur indique, de son côté, le numéro de la première des deux parties consécutives sur lesquelles porte le calcul.

On définit de la même manière, les spécificités de niveau k, c'est-à-dire les n-k+1 spécificités calculées pour des séquences de k parties consécutives qui s'écrivent :

$$S^k_1, S^k_2, S^k_3, \ldots S^k_{n-k+1},$$

Nous conviendrons que le niveau supérieur est celui qui regroupe les deux spécificités de niveau *n-1*. Remarquons que les deux nombres obtenus sont respectivement égaux aux spécificités de niveau 1 calculées pour la dernière et la première partie du corpus.

En effet, les calculs, effectués séparément sur les sous-fréquences d'une forme dans deux sous-ensembles créés par la division du corpus en deux parties connexes dont la réunion est égale à l'ensemble, conduisent nécessairement au même résultat puisque la connaissance de la sous-fréquence f dans l'une des parties permet de calculer la sous-fréquence *(F-f)* dans la partie complémentaire. Ce qui entraîne :

$$S_1^k = S_{k+1}^{n-k}$$

Les spécificités connexes, ainsi calculées, permettent de repérer certaines des irrégularités liées au temps dans la ventilation d'un terme dans les parties d'un corpus.

Application à la question Life

Les diagnostics de spécificités rassemblés au tableau 7.6 concernent des termes dont la fréquence est supérieure à 5 occurrences dans l'ensemble du corpus ***Life***. Pour chaque terme on a commencé par calculer les spécificités simples (cf. chapitre 6) qui concernent la ventilation du terme dans chacune des parties du corpus. Dans un deuxième temps, on a calculé les spécificités qui concernent les parties consécutives (niveau 2, 3, etc.). A la fin de ce processus, on a retenu, pour chaque terme, le diagnostic correspondant à la probabilité la plus faible.

Les diagnostics présentés au tableau 7.6 correspondent donc aux faits de répartition les plus remarquables du point de vue de leur distorsion par rapport à une ventilation qui résulterait d'un tirage au sort.

Les spécificités longitudinales ainsi calculées mélangent des constats que l'on pouvait aisément prévoir et des découvertes plus inattendues. Les formes *job* et *work* sont rapprochées par leur répartition dans les catégories d'âge (plutôt fréquentes dans les catégories 1 à 4).

Good job, *career* et *money* mais aussi *friends* et *being happy* sont plutôt mentionnées par des catégories de répondants plus jeunes (Ag1 et Ag2). Comme dans l'enquête ***Enfants***, et pour des raisons que l'on peut comprendre, le terme *health* est majoritairement mentionné par les catégories les plus âgées

La présence importante dans les réponses des classes Ag3 et Ag4 de l'article *the* est liée au fait que ceux-ci, dont le niveau d'instruction est supérieur en moyenne, font le plus souvent des phrases plus longues, à la syntaxe plus

précise, avec aussi un degré de nominalisation supérieur (cf. au chapitre 5, les remarques qui suivent le paragraphe 5.5).

7.6 Séries textuelles chronologiques

Dans le domaine des analyses textuelles, et dans le domaine socio-politique en particulier, les analyses de presse, les études à caractère historique, conduisent souvent à la constitution d'un type de corpus réalisé par l'échantillonnage au cours du temps d'une même source textuelle sur une période plus ou moins longue.

Les textes ainsi réunis en corpus sont souvent produits dans des conditions d'énonciation très proches, parfois par le même locuteur. Leur étalement dans le temps doit permettre de les comparer avec profit, de mettre en évidence ce qui varie au cours du temps.

Nous appelons *séries textuelles chronologiques* ces corpus homogènes constitués par des textes produits dans des situations d'énonciation similaires, si possible par un même locuteur, individuel ou collectif, et présentant des caractéristiques lexicométriques comparables.

Selon les cas, l'étude de cette dimension peut se révéler plus ou moins intéressante. Ainsi, on peut décider d'entreprendre l'étude de l'évolution du vocabulaire d'un même auteur durant toute la période qui l'a vu produire son oeuvre, en s'appuyant dans un premier temps, quitte à la remettre en cause par la suite, sur la date de rédaction présumée de chacune de ses productions.[1]

Les études réalisées sur des séries textuelles chronologiques ont mis en évidence l'importance d'un même phénomène lié à l'évolution d'ensemble du vocabulaire au fil du temps : *le Temps lexical*. Ce renouvellement constitue, la plupart du temps, la caractéristique lexicométrique fondamentale d'une série chronologique[2]. Tout émetteur produisant des textes sur une période de temps assez longue utilise sans cesse de nouvelles formes de vocabulaire qui

[1] C'est ce qu'a fait Ch. Muller (1967) dans son étude sur le vocabulaire du théâtre de Pierre Corneille.

[2] Après les recherches, désormais classiques, de Ch. Muller sur l'évolution du lexique de Pierre Corneille et celles d'Etienne Brunet sur les données du Trésor de la Langue Française, cf. Brunet (1981). Citons parmi des études plus récentes : Habert, Tournier (1987) et Salem (1987), sur un corpus de textes syndicaux français (1971-1976), Peschanski (1981) sur un corpus d'éditoriaux de *l'Humanité* (1934-1936), Romeu (1992) sur des éditoriaux parus dans la presse espagnole entre 1939 et 1945, Gobin, Deroubaix (1987) sur des déclarations gouvernementales en Belgique (1961-1985), Bonnafous (1991) sur 11 ans d'éditoriaux de presse française (1974-1984) sur le thème de l'immigration, Labbé (1990) sur le premier septennat de F. Mitterrand.

viennent supplanter, du point de vue fréquentiel, d'autres formes dont l'usage se raréfie.

Il s'ensuit que les vocabulaires des parties correspondant à des périodes consécutives dans le temps présentent en général plus de similitudes entre eux que ceux qui correspondent à des périodes séparées par un intervalle de temps plus long. Bien entendu, il peut arriver que des clivages lexicaux plus violents relèguent au second plan cette évolution chronologique. Cependant, sa mise en évidence lors de plusieurs études portant sur des corpus chronologiques nous incite à lui prêter un certain caractère de généralité et à penser que sa connaissance est indispensable pour chaque étude de ce type.

7.6.1 La série chronologique *Discours*

Le corpus ***Discours***, présenté au chapitre 2, rassemble, comme nous l'avons signalé plus haut, les textes de 68 interventions radiotélévisées de F. Mitterrand survenues entre juillet 1981 et mars 1988. Ce corpus constitue un bon exemple de *série textuelle chronologique*.

Tableau 7.7

Les 7 périodes du corpus *Discours*

	code	*années*	*occurrences*	*formes*	*hapax*	*fmax*
1	Mit1	(81-82)	30867	4296	2298	1388
2	Mit2	(82-83)	41847	5077	2611	1807
3	Mit3	(83-84)	53050	5622	2818	2040
4	Mit4	(84-85)	36453	4601	2362	1285
5	Mit5	(85-86)	53106	5509	2726	1975
6	Mit6	(86-87)	41205	4920	2499	1523
7	Mit7	(87-88)	40730	4748	2417	1526

Partition en 7 années

On obtient une partition chronologique du corpus en regroupant l'ensemble des interventions en 7 parties qui correspondent chacune à une période d'un an (du 21 mai au 20 mai de l'année suivante).

On trouve au tableau 7.7 les principales caractéristiques lexicométriques des parties ainsi découpées.

Comme le montre le tableau 7.7, ce découpage produit une partition relativement équilibrée du corpus tant du point de vue du nombre des occurrences affectées à chacune des parties que de celui du nombre des formes.

La sélection des 1 397 formes dont la fréquence est supérieure à 20 occurrences représente 256 855 occurrences soit 86% des occurrences du corpus.

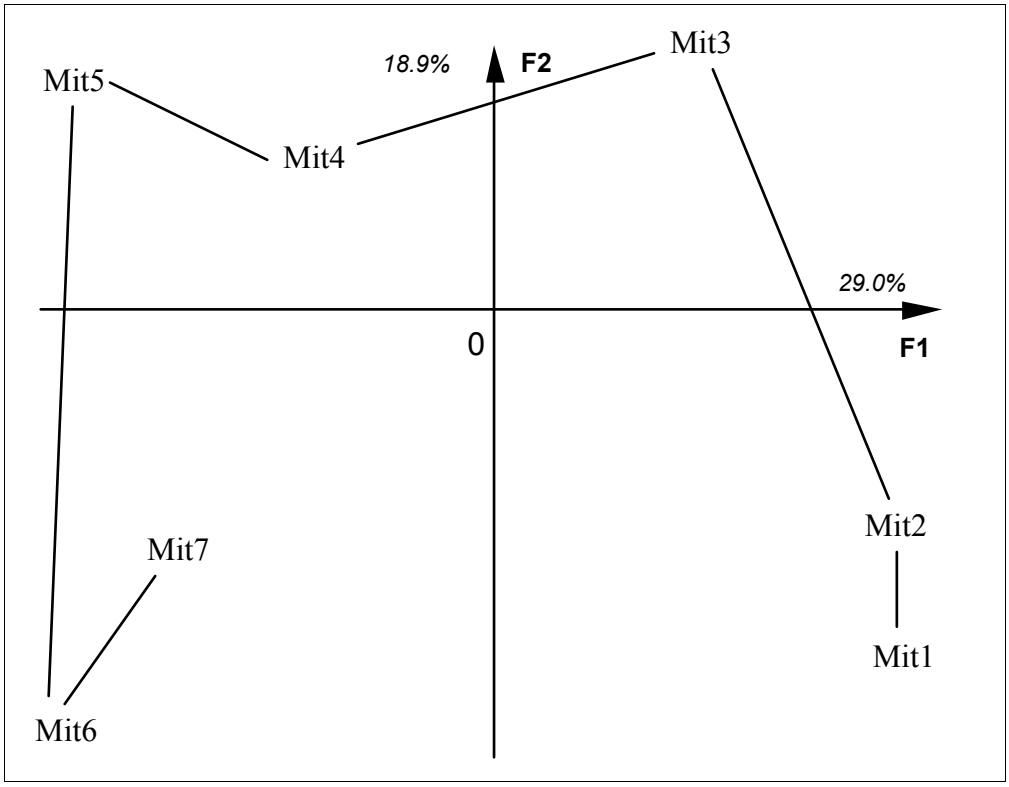


Figure 7.5

Corpus *Discours* : Esquisse des facteurs 1 et 2 issus de l'analyse des correspondances du tableau [1 397 formes (F≥ 20) ∞ 7 périodes]

Les résultats de l'analyse des correspondances du tableau (1 397 formes x 7 périodes) en ce qui concerne les différentes périodes sont résumés par la figure 7.5

L'examen des valeurs propres relatives à chacun des axes montre que trois axes se détachent de l'ensemble (λ_1=0.026, τ_1=29.0%, λ_2=0.017, τ_2=18.9%, λ_3=0.014, τ_3=16.2%)

On vérifie immédiatement que la disposition des périodes sur les premiers axes factoriels obéit dans les grandes lignes au schéma du temps lexical

exposé plus haut. Le décalage le plus marquant par rapport à ce schéma vient de la position de la dernière période sur le second axe.

En effet, la période Mit7 prend sur cet axe une position plus centrale que celle qui correspondrait à une évolution lexicale homogène par rapport à l'évolution d'ensemble. A partir de ce constat, on étudiera l'hypothèse d'une inflexion particulière du discours présidentiel dans la dernière année du septennat.

On remarque par ailleurs que l'évolution est loin d'être homogène dans le temps entre les différentes périodes. Ainsi les périodes Mit1 et Mit2 sont-elles rapprochées sur le plan des deux premiers axes factoriels alors que les périodes Mit5 et Mit6 sont au contraire très distantes. Les périodes Mit3 et Mit4 se signalent par le fait qu'elles rompent l'alignement de l'ensemble Mit1-Mit6.

7.6.2 Spécificités chronologiques

L'analyse des spécificités chronologiques (spécificités longitudinales dans le cas d'un corpus chronologique), permet de mettre en évidence les termes particulièrement employés (et les termes sous-employés) au cours d'une période ou d'un groupe de périodes consécutives dans le temps.

Le tableau 7.8, donne la liste des spécificités chronologiques majeures du corpus ***Discours***. Seules quelques formes extrêmement significatives d'un point de vue statistique[1] ont été sélectionnées pour ce tableau. La liste s'ouvre sur un diagnostic de spécificités relatif à la forme *nous* dont la méthode nous signale l'abondance relative dans les premières parties. On note, qu'a l'inverse, la fréquence de la forme *je*, moins forte dans les premières périodes à tendance à augmenter avec le temps.

La période numéro 1 est caractérisée par la présence des termes *nationalisations, reformes, relance*, lesquels disparaissent par la suite comme on peut le constater.

Un examen plus détaillé de cette liste permettra de dégager des thèmes relatifs à chacune des périodes et de suivre leurs fluctuations dans le temps. Le classement en probabilité permet comme toujours de hiérarchiser les écarts.

Comme plus haut, ces listes, dont on pourra régler le volume global en faisant varier le seuil de sélection des termes, peuvent être triées en fonction

[1] Il s'agit de formes dont l'indice de probabilité est inférieur à $1/10^{-14}$.

des périodes du corpus si l'on désire illustrer les fluctuations du vocabulaire au fil des périodes du corpus.

Tableau 7.8

Spécificités chronologiques majeures du corpus *Discours*
(cf. guide de lecture du tableau 7.6)

terme	F	fp	sp	code
nous	2059	1148	+E35	01-03
l' iran	50	41	+E27	07-
nationalisations	42	31	+E22	01-
étudiants	28	27	+E22	06-
chaîne	39	34	+E21	05-
israël	71	54	+E20	01-02
majorité	212	130	+E19	05-06
réformes	39	27	+E18	01-
a	2863	1875	+E18	04-07
relance	32	24	+E17	01-
avons	523	458	+E17	01-05
de	11544	3195	+E17	01-02
c	3240	2643	+E17	03-07
je ne	794	567	+E16	04-07
nous avons	413	366	+E16	01-05
pour 100	204	193	+E16	01-05
cela	1636	1096	+E15	04-07
pas	4067	2092	+E15	05-07
pays	748	358	-E15	03-06
cinquième	35	27	+E14	05-
je	5024	3156	+E14	04-07

7.6.3 Les accroissements spécifiques

Pour compléter l'étude de l'évolution d'un vocabulaire au fil du temps, il reste encore à construire des outils qui permettront de signaler des changements brusques dans l'utilisation d'un terme lors d'une période donnée par rapport aux périodes précédentes - et non à l'ensemble du corpus comme c'est le cas dans les utilisations que nous venons de décrire.

Calcul des accroissements spécifiques

Dans ce qui suit, nous sommes toujours confrontés à un tableau lexical du type de ceux que nous avons étudiés plus haut. Cependant, la fréquence f_{ij} de la forme i dans la partie j va être appréciée cette fois en fonction de quantités qui ne correspondent pas à celles que nous avons employées pour le calcul des spécificités simples.
En effet, si l'on se fixe une période j (j varie entre 2 et p - nombre total des périodes du corpus, puisqu'il s'agit d'accroissement), il est possible pour un

terme donné de comparer la fréquence f_{ij} à la fréquence de ce même terme dans l'ensemble des parties précédentes.

Pour effectuer cette comparaison, on commencera par calculer les quatre paramètres du modèle hypergéométrique. Deux de ces paramètres sont identiques par rapport au calcul des spécificités simples, ce sont les paramètres:

f_{ij} — fréquence du terme i dans la partie j;
t_j — nombre des occurrences de la partie j;

Les deux autres sont des analogues des paramètres employés lors du calcul des spécificités simples pour lesquels la sommation s'effectue cette fois non plus pour l'ensemble des parties mais pour les j premières parties.

T^j — le nombre des occurrences du corpus réduit aux seules j premières parties du corpus;
F^j_i — la fréquence du terme i dans ce même sous-corpus.

Guide de lecture du tableau 7.9

Les spécificités calculées concernent les sous-fréquences des termes (formes ou segments) dont la fréquence dépasse 20 occurrences dans le corpus. Seuls les diagnostics dont l'indice de probabilité est inférieur à 10^{-6} ont été retenus.

La colonne Fx indique la fréquence du terme dans les x premières parties du corpus. Cas particulier, pour la dernière période du corpus Fx=F ; fréquence totale de la forme dans le corpus.

La colonne f indique la fréquence partielle du terme dans la partie pour lesquelles le diagnostic d'accroissement spécifique a été établi (ici la dernière partie).

Le diagnostic d'accroissement spécifique affecté d'un exposant

– / Exx renvoie à un accroissement spécifique positif, sur-emploi spécifique par rapport aux parties précédentes, à laquelle est attachée une probabilité d'ordre 1/10 à la puissance xx.
– \ Exx renvoie à un accroissement spécifique négatif, sous-emploi spécifique par rapport aux parties précédentes, à laquelle est attachée une probabilité d'ordre 1/10 à la puissance xx.

Les diagnostics retenus ont été classés en fonction des spécificités positives ou négatives mises en évidences puis en fonction de l'indice probabilité.

Tableau 7.9

Corpus *Discours* accroissements spécifiques majeurs de la partie Mit7 (période : 12/04/87 - 4/03/88)

terme	F7	f	spec.
l iran	50	41	/E27
iran	53	41	/E25
arabe	34	23	/E13
monde arabe	21	17	/E12
d instruction	20	16	/E11
instruction	23	17	/E11
l irak	29	18	/E09
irak	32	18	/E08
élection	35	18	/E07
président	303	73	/E07
d armes	27	15	/E07
un président	28	15	/E07
politiques	105	34	/E07
armes	93	32	/E07
juge	35	17	/E07
pays	748	151	/E07
manière	25	13	/E06
oui	256	63	/E06
candidat	26	14	/E06
y	1749	305	/E06
vraiment	143	40	/E06
il y	1014	191	/E06
désarmement	38	17	/E06
nous avons	413	27	\E06
inflation	83	0	\E06
avons	523	35	\E07
jeunes	134	2	\E07
nous	2059	182	\E12

Ces deux derniers paramètres ont été pourvus d'un indice supérieur qui rappelle que la sommation est une sommation partielle, limitée aux premières périodes du corpus.

Remarquons tout de suite que pour la dernière des parties du corpus, et pour elle seulement :

$$T^p = T; \quad F^p{}_i = F_i \, .$$

En d'autres termes, pour la dernière des parties le calcul des accroissements spécifiques amène des résultats identiques au calcul des spécificités simples pour cette partie. Pour chaque période donnée, on peut alors obtenir une liste des termes dont l'emploi a notablement augmenté ainsi que celle des termes brutalement disparus.

L'étude des accroissements spécifiques majeurs du corpus permet de localiser les changements les plus brusques sur l'ensemble du corpus. On a rassemblé au tableau 7.9 les diagnostics les plus importants qui concernent la seule partie Mit7 laquelle, comme nous l'avons vu sur le premier plan factoriel issu de l'analyse du tableau (1 397 formes x 7 parties) manifeste une position particulière par rapport à l'évolution d'ensemble.

Ce tableau illustre très nettement l'émergence au cours de l'année 1987-1988 d'un vocabulaire lié à des conflits internationaux du moment *(iran, monde arabe, irak, armes)* et aux affaires juridiques elles-mêmes liées aux attentats terroristes *(juge, instruction)* alors que refluent de manière spectaculaire les préoccupations liées à l'*inflation* et aux *jeunes*. On note également la spécificité négative de formes liées à la première personne du pluriel (*nous, avons*). En ne retenant que quelques formes liées aux diagnostics les plus marqués d'accroissement spécifique, nous nous sommes bornés à une illustration de la méthode. Il va sans dire qu'une sélection plus large de termes (que l'on obtient sans peine en choisissant un seuil de probabilité plus élevé) permet de remarquer des accroissements dont l'évidence est moins nette pour le politologue.

7.6.4 Étude parallèle sur un corpus lemmatisé

Nous avons vu au chapitre 2, à propos de la série chronologique ***Discours***, comment les décomptes parallèles réalisés d'une part en formes graphiques et d'autre part sur des unités lemmatisées amenaient des résultats très éloignés les uns des autres en ce qui concerne les principales caractéristiques lexicométriques du corpus : nombre des formes, des hapax, fréquence maximale, etc.[1]

Le présent paragraphe nous permettra d'évaluer l'incidence de l'opération de lemmatisation sur les résultats auxquels aboutissent les analyses multidimensionnelles présentées dans les chapitres qui précèdent.

Pour cette comparaison effectuée à partir du décompte des occurrences de chacun des lemmes dans les 7 périodes du corpus, nous avons conservé les seuils de sélection établis lors de l'analyse effectuée sur les formes graphiques. Nous avons ensuite soumis aux diverses méthodes d'analyses multidimensionnelles un tableau croisant les 1 312 lemmes dont la fréquence est au moins égale à 20 occurrences avec les sept périodes considérées dans l'expérience relatée plus haut.

[1] Rappelons que la série chronologique ***Discours*** que nous considèrons actuellement à fait l'objet d'une lemmatisation, partiellement assistée par ordinateur, réalisée par D. Labbé. Cf. Chapitre 2 du présent ouvrage, et Labbé (1990). On trouvera un commentaire plus développé de l'ensemble de ces résultats dans Salem (1993).

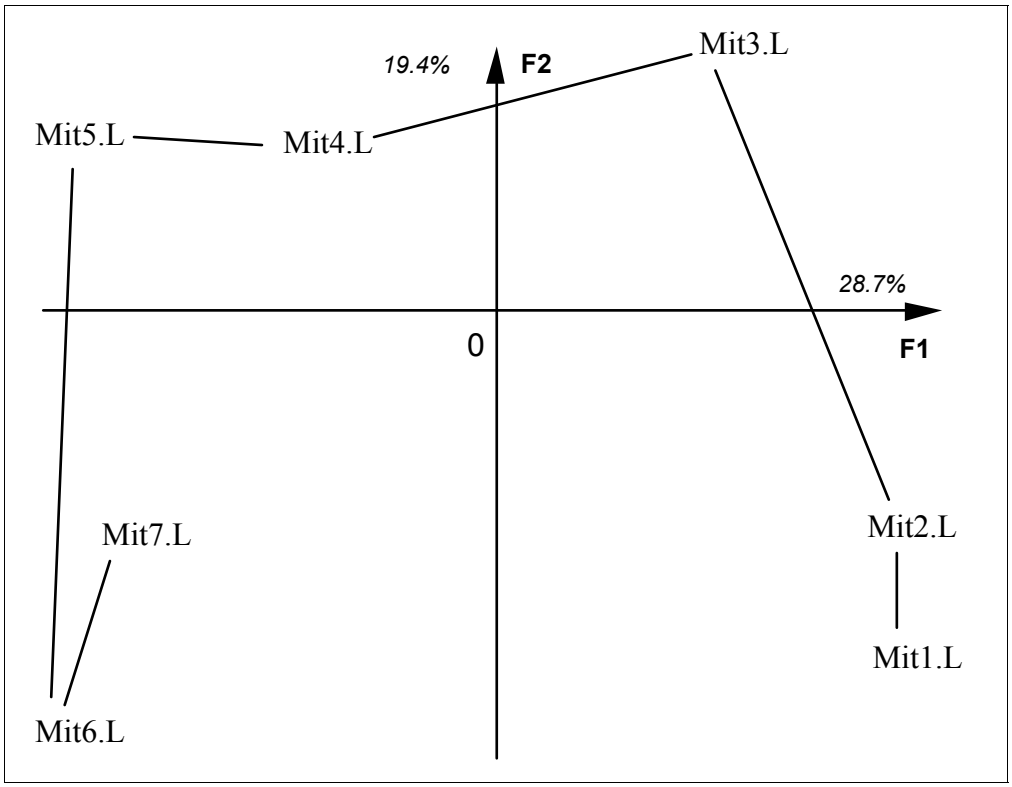


Figure 7.6

Corpus *Discours lemmatisé*
Facteurs 1 et 2 issus de l'analyse des correspondances
du tableau (1 312 lemmes (F≥ 20) x 7 périodes)

Les résultats de l'analyse des correspondances de ce dernier tableau sont résumés, pour ce qui concerne les différentes périodes, par la figure 7.6. Les symboles qui désignent les 7 périodes du corpus (Mit1.L, Mit2.L, ..., Mit7.L) ont été affectés de la lettre "L" qui rappelle que les décomptes sont effectués cette fois ci à partir de lemmes.
Ici encore trois axes se détachent de l'ensemble auxquels correspondent des valeurs propres et des pourcentages d'inertie très voisins de ceux obtenus lors de l'analyse sur les formes graphiques (λ_1=0.024, τ_1=28.7%, λ_2=0.016, τ_2=19.4%, λ_3=0.014, τ_3=16.6%)

L'examen de la disposition relative des périodes sur le plan des premiers axes factoriels confirme l'impression que nous nous trouvons bien devant des résultats extrêmement proches de ceux obtenus à partir des formes graphiques (cf. fig 7.5 du présent chapitre).

La comparaison des listes d'unités les plus contributives pour chacun des ensembles d'unités de décompte (lemmes et formes graphiques) qui servent de base à la typologie dans l'une et l'autre des analyses montre que les différences sont peu importantes dans l'ensemble. De la même manière la

liste des diagnostics de spécificité chronologique et d'accroissements spécifiques les plus forts, calculés à partir des unités lemmatisées recoupent dans une grande mesure les diagnostics obtenus à partir des formes graphiques.[1]

Ainsi, à partir de tableaux qui présentent des différences importantes dans la mesure ou ils résultent de comptages effectués sur des unités elles mêmes très différentes, les méthodes de la statistique textuelle (analyse des correspondances, classification hiérarchique, spécificités chronologiques) produisent des résultats très proches.

L'expérience présentée ci-dessus ne constitue évidemment pas à elle seule une preuve définitive de ce que les analyses multidimensionnelles mettent en évidence des contrastes entre les textes peu dépendants du choix des unités de décompte. Ici encore c'est la finalité de l'étude entreprise qui soufflera les choix à opérer.

La décision est aussi, si l'on peut dire, d'ordre économique. Il est dans l'absolu toujours préférable de disposer d'un double réseau de décomptes (en formes graphiques et en lemmes).[2] Une *lemmatisation complète*, sur un corpus important, reste une opération coûteuse. Indispensable dans un travail de recherche, elle est beaucoup moins justifiée s'il s'agit d'obtenir rapidement des visualisations ou des typologies de parties de corpus d'une certaine richesse lexicale. On peut alors travailler sur la base des formes graphiques sans crainte de passer à côté de l'essentiel. En revanche, si l'on a affaire à des textes courts non regroupés, les analyses sur texte lemmatisé donneront un point de vue parfois différent, souvent complémentaire.

7.7 Recherches en homogénéité d'auteur

Application au problème du livre d'Isaïe.[3]

Dans ce paragraphe, on appliquera les méthodes de l'analyse de la contiguïté, à un problème philologique ancien : celui de l'homogénéité du livre de la Bible que l'on nomme : *Le livre d'Isaïe* (cf. encadré ci-dessous). Sans insister sur les particularités lexicométriques de l'hébreu biblique, langue dans laquelle le texte est parvenu jusqu'à nous, on se consacrera essentiellement à l'éclairage qu'apportent à ce problème les typologies empiriques réalisées sur la base des comptages de formes.[4]

---

[1] Cf. Salem (1993) p 740-746.

[2] Mais il ne faut pas perdre de vue que la lemmatisation d'un corpus de plusieurs centaines de milliers d'occurrences est une opération de longue haleine.

[3] Pour plus de détails, cf. Salem (1979). Les résultats des analyses quantitatives ont été repris sous leur forme originale ; l'exposé méthodologique a été réactualisé.

[4] Rappelons cependant que cette langue, comme la plupart des langues de la famille sémitique, s'est donné une écriture consonantique (i.e. on ne transcrit que les consonnes

Un problème vieux de deux mille ans . . .

Combien d'auteurs ont-ils participé à la rédaction du livre de la Bible attribué au prophète Isaïe? Depuis longtemps les différentes écoles de la critique biblique divergent. Pour les uns, le livre entier doit être attribué à un même auteur. D'autres avancent des découpages en deux, trois, six, voire dix-sept parties qui selon eux doivent chacune être attribuées à un auteur différent.

Si l'oeuvre dont tous les commentateurs s'accordent à reconnaître la grande qualité littéraire, se présente comme rédigée avant l'exil de Babylone (environ 600 av J.C.), la mention du roi Cyrus (Is.Ch44.28) qui vécut quelques cent cinquante ans plus tard, laisse planer un doute sur l'éventualité d'un auteur unique pour l'ensemble de l'oeuvre.

A partir de remarques de ce type, l'hypothèse d'une oeuvre constituée par compilation de plusieurs textes d'origines différentes investit rapidement le milieu des études bibliques et donne lieu à une inflation sur le nombre possible des auteurs. Avec le développement des études bibliques, le livre se trouve fragmenté en parcelles de plus en plus fines dont les critiques affirment que chacune d'entre elles doit être attribuée à un auteur différent.

Signalons enfin que parmi les *Manuscrits de la Mer Morte* exhumés en 1947, les archéologues ont découvert une copie du livre d'Isaïe qui se trouve, du point de vue de l'établissement du texte, dans un état très proche de celui qui est parvenu jusqu'à nous à travers les manuscrits de l'époque médiévale. Avant cette découverte le manuscrit le plus ancien de la Bible était un manuscrit daté de l'an 1009 (après JC). Cette découverte signifiant pour nous que s'il y a eu compilation, elle a été réalisée à une date très ancienne qui ne peut être postérieure à 150 avant J.C.

Recherches textuelles en homogénéité d'auteur.

De nombreuses études "textuelles", c'est à dire se fondant exclusivement sur le matériau textuel du livre et non sur les recoupements d'ordre historique que l'on peut faire à partir du texte, ont été publiées à propos de cette oeuvre. La plupart de ces études s'appuie sur des comptages réalisés manuellement ou automatiquement. Ces comptages portent sur des unités

de la chaine vocalique laissant le loisir au lecteur de rétablir les voyelles ; opération que le lecteur expérimenté effectue sans grande difficulté). En hébreu, on a coutume de diviser les mots du discours en trois catégories : noms, verbes et particules libres. La base des verbes est constituée par des racines trilitères. Aux mots, peuvent s'adjoindre différentes particules proclitiques ou enclitiques. Dans les années 70, le Centre Analyse et de Traitement Automatique de la Bible (CATAB), dirigé par G.E.Weil, a établi une transcription sur cartes perforées de l'ensemble de la Bible hébraïque.

très différentes dont les auteurs affirment, dans chacun des cas, qu'elles permettent l'identification d'un auteur de manière privilégiée[1].

Nous évoquerons également au chapitre 8, dévolu à l'analyse discriminante textuelle, les problèmes d'attribution d'auteurs, en insistant sur certains modèles statistiques sous-jacents. On réservera la terminologie *homogénéité d'auteur* aux cas pour lesquels il n'y a pas plusieurs auteurs potentiels identifiés, mais simplement incertitude sur l'unicité de l'auteur d'une série de textes.

Cette tradition de recherche en homogénéité d'auteur dans le domaine biblique prend largement ses sources dans les recherches analogues réalisées dans d'autres domaines parfois très éloignés[2].

Au plan méthodologique, le schéma qui régit la plupart de ces études peut-être ramené à trois phases principales :

a) promotion au rang de variable discriminante, pour l'attribution d'un texte à un auteur donné, d'un type de comptage particulier portant sur les occurrences d'un phénomène textuel de préférence assez fréquent, (occurrences d'une particule, d'une tournure syntaxique, proportion d'emploi des voix du verbe, etc.).
b) discussion d'exemples montrant que, sur un corpus composé de groupes de textes dont on sait avec certitude que chaque groupe est composé de textes écrits par un auteur différent, la variable privilégiée prend effectivement des valeurs qui discriminent les groupes.
c) application du critère ainsi validé à un problème n'ayant pas de solution connue.

La difficulté majeure créée par ce genre d'étude vient précisément de la diversité des critères d'homogénéité proposés. Le point le plus discutable est que ces critères n'ont été, en général, élaborés et utilisés qu'en vue de résoudre un seul problème. On est en droit de se poser la question : pour justifier une partition quelconque d'une oeuvre n'est-il pas possible de

[1] Cf. par exemple une étude portant sur une codification de la morphologie du verbe dans le livre d'Isaïe, Kasher (1972), ou encore une étude portant sur la répartition dans ce même livre d'une particule que l'on nomme "waw conversif", Radday (1974).

[2] Radday (1974) signale d'ailleurs que le critère portant sur les occurrences du waw conversif lui a été inspiré par une étude du Rev. A.Q. Morton : *The authorship of the Pauline corpus* (Morton, 1963), sur des textes évangéliques en grec ancien. Morton estime en effet que la variation du nombre des occurrences de la particule και dans différents textes des évangiles permet de conclure que certains des textes attribués à Paul n'ont pu être écrits par ce dernier, d'où la transposition proposée dans une langue complètement différente.

trouver un critère discriminant, parmi la masse des critères imaginables, et ce, quelle que soit cette partition ?

Pour sortir de ces difficultés, il est donc important de ramener l'ensemble de ces problèmes à l'intérieur du cadre plus général du décompte exhaustif des formes graphiques d'un texte soumis à l'analyse multidimensionnelle des tableaux ainsi obtenus. La statistique textuelle ne peut prétendre trancher de manière indiscutable ce débat entre érudits. Cependant les méthodes exposées plus haut peuvent apporter sur ces questions un éclairage quantitatif dont on a plusieurs fois prouvé la pertinence.

7.7.1 Le corpus informatisé.

Le texte sur lequel nous avons travaillé compte 16 931 occurrences. Plusieurs possibilités de découpage du corpus en unités de base se présentent alors. Plutôt que de choisir une partition en tranches de longueur fixe (1 000 mots, 100 versets, etc.) nous avons opté pour la partition du livre en 66 chapitres[1] que l'on désignera par la suite (Is1, Is2, ..., Is66) partition presque universellement reçue de nos jours. C'est dans ce langage des chapitres que s'expriment les partitions en fragments homogènes proposés par la plupart des critiques.

Dans la mesure où nous envisageons de faire apparaître, à l'aide de méthodes quantitatives, des fragments homogènes à l'intérieur du livre, nous éviterons par tous les moyens d'introduire en qualité d'hypothèses les découpages de la critique biblique que nous nous proposons de mettre à l'épreuve[2].

Au cours de l'étude, l'unité de chacun des 66 chapitres ne sera plus remise en cause. Cette partition servira ensuite à construire des groupes de chapitres homogènes au plan statistique.

---

[1] Signalons que cette partition en chapitres n'a été établie que vers le début du 13ème siècle par Stephen Langton, futur archevêque de Canterbury. Elle est utilisée aujourd'hui par la plupart des spécialistes de la Bible.

[2] On trouvera une description des différentes partitions de ce livre proposées au cours des sciècles par la critique dans l'article (Weil et al. 1976). Il est facile de constater que les chapitres Is36 à Is39 constituent un groupe très particulier puisque le texte de chacun de ces chapitres correspond, souvent mot pour mot, à un passage du Deuxième livre des Rois (autre livre de la Bible). Certains des critiques considèrent que ce groupe de chapitres ne fait pas réellement partie du livre d'Isaïe et l'écartent purement et simplement de leur partition de l'oeuvre en fragments homogènes. Nous n'écarterons pas d'emblée ces chapitres de nos analyses ; leur présence nous permettra au contraire de juger, à travers les classements qui leur seront attribués, de la pertinence des méthodes utilisées.

En effet, pour être utile, la *partition de travail* qui sert de base aux regroupements futurs doit répondre à plusieurs critères :

a) les éléments de base doivent être choisis suffisamment ténus pour que la majorité des parties ne chevauchent pas les frontières, inconnues a priori, entre ce qui constitue des fragments homogènes au plan statistique.

b) cette partition ne doit pas introduire subrepticement une partition a priori, résultant de considérations extra-textuelles, puisque le but final de l'analyse est précisément de dégager une telle partition avec un minimum de considérations a priori.

c) enfin, les éléments de base doivent être choisis suffisamment longs pour que leurs profils lexicaux représentent bien des moyennes statistiques susceptibles d'être rapprochées par des algorithmes de classement.

Pour cette division en chapitres, qui respecte l'unité thématique de petits fragments, le chapitre le plus long, Is37, compte 566 occurrences, le plus court Is12 n'en compte que 62. Cependant, la grande majorité des chapitres oscillent entre 150 et 400 occurrences soit environ du simple au triple.

A partir du décompte des occurrences de chacune des formes les plus fréquentes dans chacun des 66 chapitres du livre, on construit le tableau à double entrée de dimensions (89 formes x 66 chapitres).[1]

L'objectif principal de cette étude étant d'aboutir à un découpage en parties homogènes, on se gardera d'introduire trop tôt, même à des fins de validation, des structures de contiguïtés par trop inspirées des découpages traditionnels de la critique biblique, puisqu'il s'agit précisément de mettre ces découpages à l'épreuve. Il est néanmoins possible d'introduire une hypothèse de contiguïté particulière qui servira de point d'appui pour mesurer la pertinence des typologies obtenues.

L'hypothèse d'homogénéité globale des chapitres consécutifs

Le découpage en unités relativement fines que sont les 66 chapitres permet en effet de supposer que la grande majorité des coupures entre les chapitres ne constituent pas également des coupures entre fragments dus à des auteurs différents.

[1] La liste des 100 formes les plus fréquentes nous avait été fournie par l'équipe du CATAB. A l'époque nous en avions écarté d'emblée une liste de 11 noms propres, ne souhaitant pas y trouver un appui pour le découpage recherché, précaution que nous jugerions superflue aujourdhui.

Sous cette hypothèse, les mesures de contiguïté, nous permettent de dire, pour toute variable numérique prenant ses valeurs sur l'ensemble des chapitres, si cette variable prend des valeurs voisines sur les chapitres consécutifs.

Considérons en effet, une variable z_j prenant ses valeurs sur l'ensemble J des chapitres (cette variable peut être le résultat du décompte au fil des chapitres d'une unité textuelle, un facteur, ou encore toute autre fonction numérique). Si l'on stipule l'hypothèse d'homogénéité globale des chapitres consécutifs, c'est à dire si l'on accepte, dans notre cas, que deux chapitres consécutifs ont de fortes chances d'appartenir au même fragment-auteur, il s'ensuit que pour toute variable z_j prenant ses valeurs sur l'ensemble J des chapitres, les différences du type $z_j - z_{j-1}$ sont, pour la plupart d'entre elles, des différences entre chapitres appartenant à un même fragment.

Dans le cas où la variable z_j discrimine réellement les fragments-auteur, c'est à dire dans le cas ou elle prend des valeurs similaires pour les chapitres j et j - 1 situés à l'intérieur d'un même fragment-auteur, et des valeurs différentes pour les fragments-auteur distincts, le coefficient de contiguïté calculé sur les chapitres consécutifs doit être nettement inférieur à l'unité.

Commençons par appliquer le calcul de ces coefficients à l'étude de la variation au fil des chapitres de la variable *D*, mesurée pour chacun des chapitres et définie comme suit :

$$D_j = (V_j - P_j) * 100 / t_j$$

où :

— V_j est le nombre des occurrences de formes verbales du chapitre j.

— P_j est le nombre des occurrences de particules dans ce même chapitre.

— t_j est le nombre des occurrences dans le chapitre.

En d'autres termes D_j représente la différence entre le nombre des verbes et celui des particules rapportée au nombre des occurrences dans chacun des chapitres.[1]

[1] Une étude parallèle (Salem 1979) a montré que les livres réputés anciens comportaient, proportionnellement, plus de particules par rapport aux verbes que les livres plus récents. Ces proportions se modifiant au fur et à mesure de l'évolution de la langue en raison vraisemblablement de la formation de verbes nouveaux par l'intégration en qualité d'affixes de particules précédemment libres aux racines anciennes. Cette circonstance permet, sous certaines conditions, de considérer la variable *VP* comme un critère de datation dont l'efficacité varie avec la longueur des fragments considérés. Ce critère ne tient pas compte de la possibilité de réécriture à une date plus rapprochée de nous de textes beaucoup plus anciens; procédé dont on peut supposer qu'il a été fréquemment utilisé avant la fixation définitive des textes que nous considérons.

La figure 7.7 montre comment la quantité D_j varie pour les 66 chapitres du livre d'Isaïe et pour les 40 chapitres que contient un autre livre de la Bible, le livre de l'Exode, dont on a de nombreuses raisons de croire qu'il est beaucoup plus ancien.

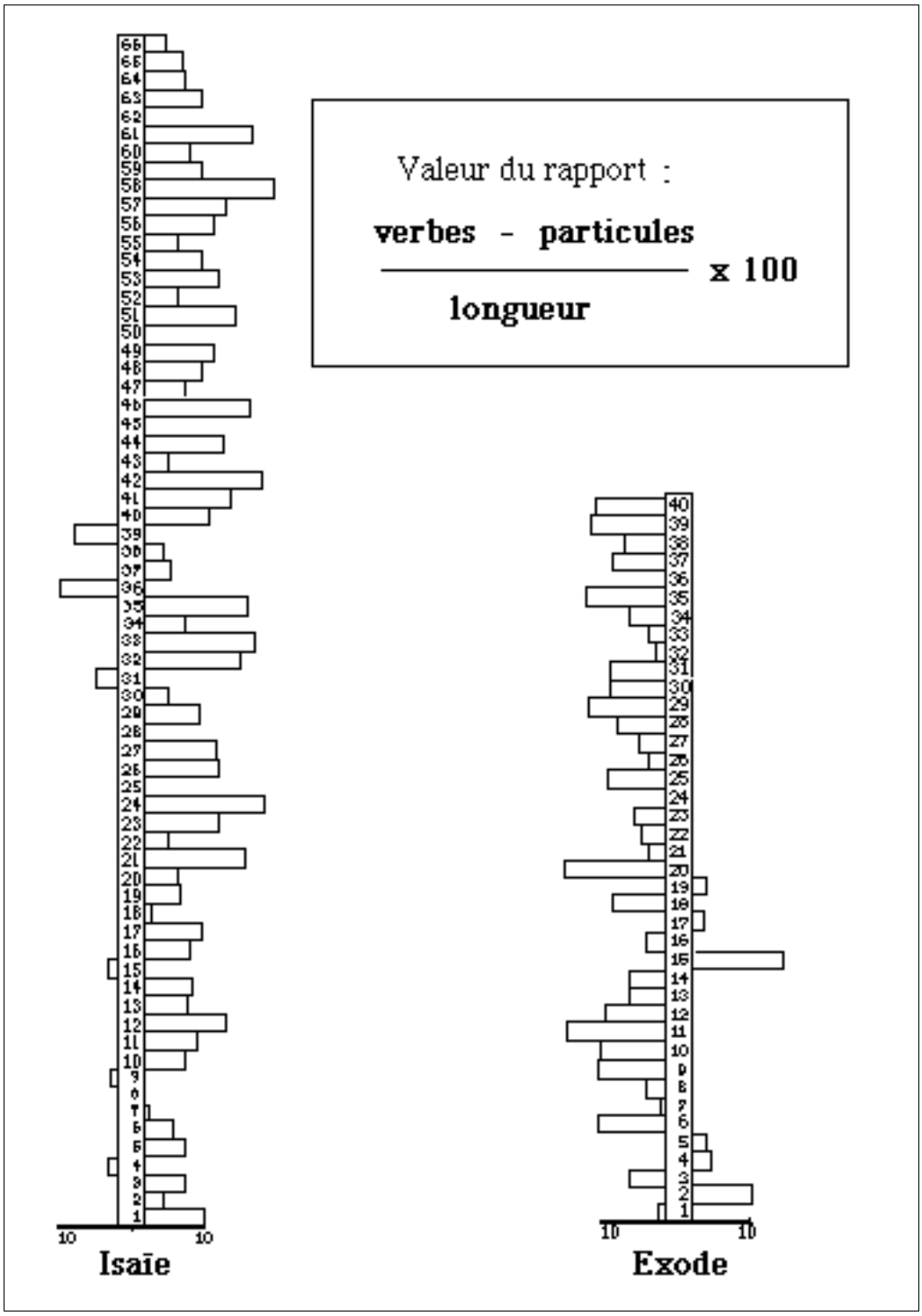


Figure 7.7

Etude d'homogénéité pour la différence *Verbes-Particules* pour les chapitres du livre d'Isaïe et du livre de l'Exode.

Guide de lecture de la figure 7.7

Les chapitres d'un même livre sont représentés par ordre croissant de bas en haut. Pour chacun des chapitres, les écarts de la variable D_j sont représentés par un rectangle qui s'écarte sur la droite du numéro correspondant au chapitre si le nombre des verbes excède celui des particules, et sur la gauche de ce numéro dans le cas inverse.

On note que les chapitres du livre d'Isaïe comptent dans l'ensemble plutôt plus de verbes, à l'inverse des chapitres du livre de l'Exode dans lesquels les particules sont plutôt plus abondantes. Ce qui confirmerait, si besoin était, les présomptions d'antériorité du livre de l'Exode.

Pour le livre d'Isaïe, les exceptions les plus marquantes sont constituées par les chapitres Is36 et Is39 dont on a déjà signalé la particularité à l'intérieur d'un groupe Is36-Is39. Pour ces chapitres, l'excédent de particules par rapport aux verbes pourrait conforter l'hypothèse d'une écriture antérieure aux autres chapitres du livre. Les chapitres Is37 et Is38 qui appartiennent visiblement au même groupe thématique que Is36 et Is39 ne présentent pas la même répartition que ces derniers.

Pour chacun des deux livres, on a calculé le coefficient de contiguïté pour les valeurs de la variable D_j :

pour les 66 chapitres du livre d'Isaïe, $c(D_j.) = 0.86$,

pour les 40 chapitres du livre de l'Exode, $c(D_j.) = 0.93$.

Dans les deux cas le coefficient est très faiblement inférieur à l'unité. L'examen des tables de l'écart type pour la loi de ce coefficient, en fonction du nombre des parties, conduit à la conclusion que le coefficient calculé pour les deux livres s'écarte de l'unité de manière non significative.

La conclusion qui s'impose est donc que la variable D_j distingue les deux séries de chapitres en raison, pouvons-nous supposer, de leur date d'écriture différente. Par contre, cette variable ne varie pas de manière très homogène au fil des chapitres consécutifs. Il semble donc difficile d'utiliser la variable D_j. pour délimiter des fragments homogènes à l'intérieur de chacun des livres.

Analyse sur les chapitres

L'examen de l'histogramme des valeurs propres, figure 7.8, montre que les deux premières valeurs propres se détachent très nettement. Les valeurs propres correspondant aux deux axes suivants (3 et 4) se détachent ensuite

des valeurs propres qui correspondent au reste des facteurs lesquelles décroissent ensuite régulièrement.

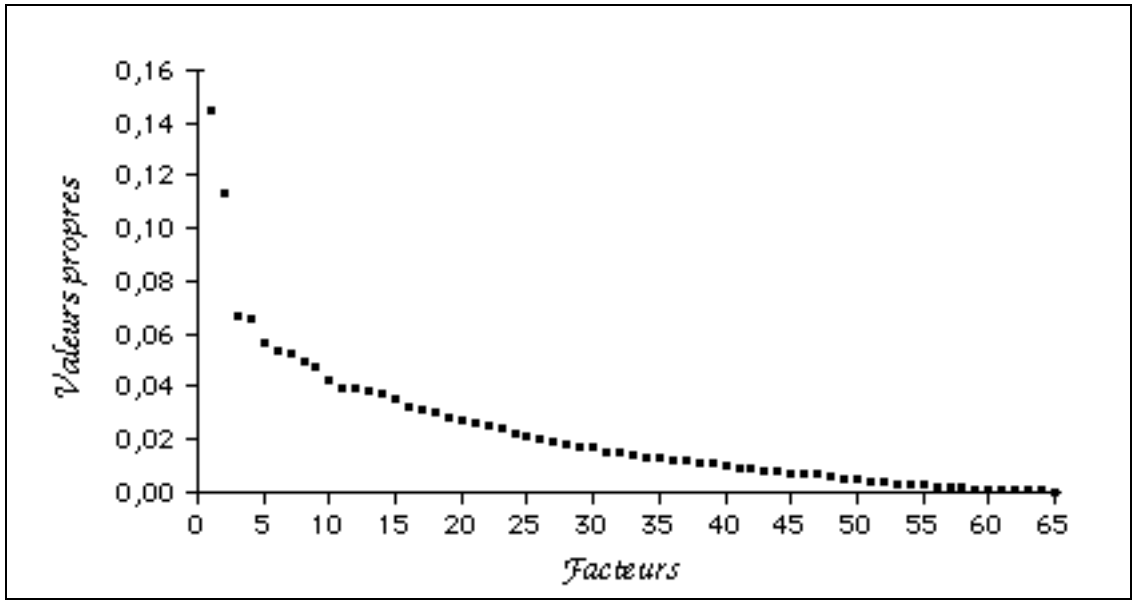


Figure 7.8

Isaïe : Les valeurs propres issues de l'analyse du tableau (89 formes x 66 chapitres)

En considérant les facteurs sur l'ensemble des chapitres on s'aperçoit vite que les chapitres consécutifs occupent sur ces facteurs des positions similaires.

7.7.2 Validation des résultats

Pour chacun des 65 facteurs issus de l'analyse des correspondances du tableau (89 formes x 66 chapitres), le coefficient de contiguïté va nous permettre d'apprécier dans quelle mesure le facteur possède ou non la propriété de "proximité des valeurs sur les chapitres contigus".

Le coefficient à été calculé à partir des 66 coordonnées factorielles relatives à chacun des chapitres. On trouve sur la figure 7.9 le résultat de ces calculs.

La figure 7.10 fournit les valeurs du coefficient G_α (coefficient de contiguïté mesuré sur les α premiers facteurs, cf. paragraphe 7.3.2).

Comme le montrent nettement ces figures, les premiers facteurs issus de l'analyse des correspondances possèdent à un très haut degré la propriété d'homogénéité sur les chapitres consécutifs que nous avons étudiée plus haut (en plus de leur propriété spécifique qui est de maximiser l'inertie du nuage).

Cette propriété est particulièrement frappante pour ce qui concerne les deux premiers facteurs.

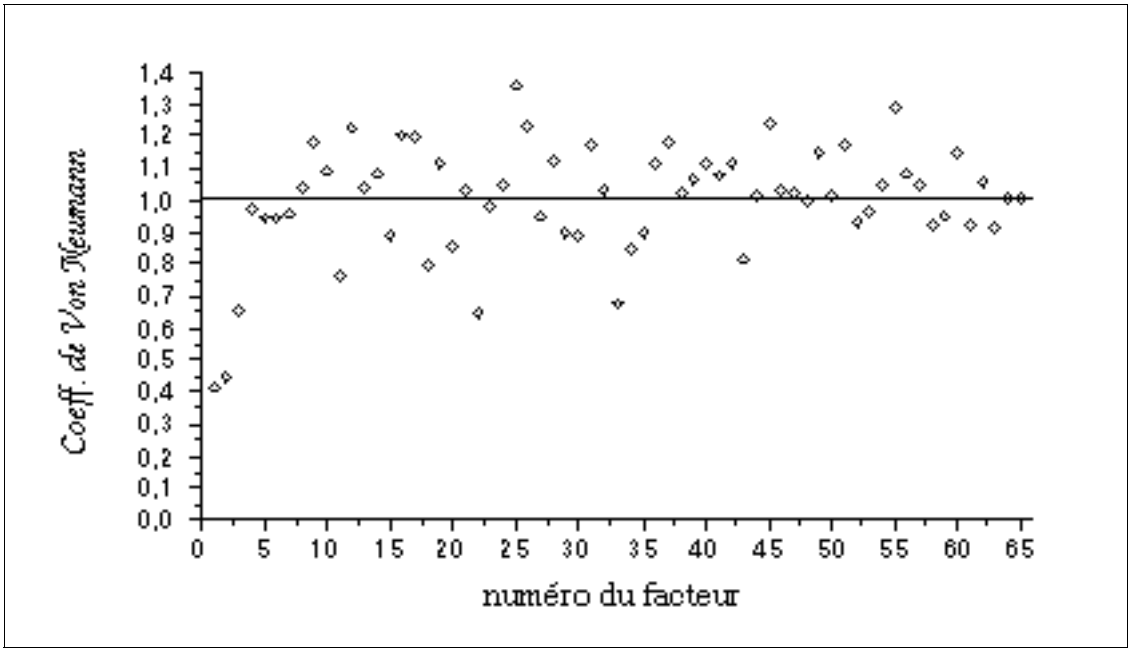


Figure 7.9

***Isaïe* : Le calcul du coefficient de Von Neumann pour les 65 facteurs issus de l'analyse des correspondances du tableau (89 formes x 66 chapitres)**

Les deux méthodes proposées ci-dessous permettent de préciser le diagnostic que l'on peut faire à partir de ces données.

Un test d'autocorrélation

L'écart type du coefficient de contiguïté vaut $\sigma = 0.122$. Il nous permet de définir un intervalle de confiance approximatif $1 \pm 2\,\sigma$ qui vaut dans notre cas [0.756, 1.244].

Comme on le voit en se reportant à la figure 7.9 , un très petit nombre de coefficients ainsi calculés à partir des facteurs se trouve à l'extérieur de cet intervalle[1]. Ces coefficients sont, pour ce qui concerne les valeurs inférieures à l'unité : $c_1 = 0.410$, $c_2 = 0.449$, $c_3 = 0.654$, $c_{22} = 0.642$, $c_{33} = 0.677$ et pour ce qui concerne les valeurs supérieures à l'unité : $c_{55} = 1.290$.

Ici encore, la distribution de l'écart n'est pas symétrique dans ce sens que l'on ne rencontre que très peu de facteurs ayant la propriété de prendre sur les chapitres consécutifs des valeurs fortement différentes.

[1] On se trouve ici encore devant un problème de comparaisons multiples (cf. chapitre 6, paragraphe 6.1.3). L'intervalle de confiance approximatif pour les coefficients c est donc trop large, ce qui incite à penser que seuls les deux premiers sont nettement significatifs.

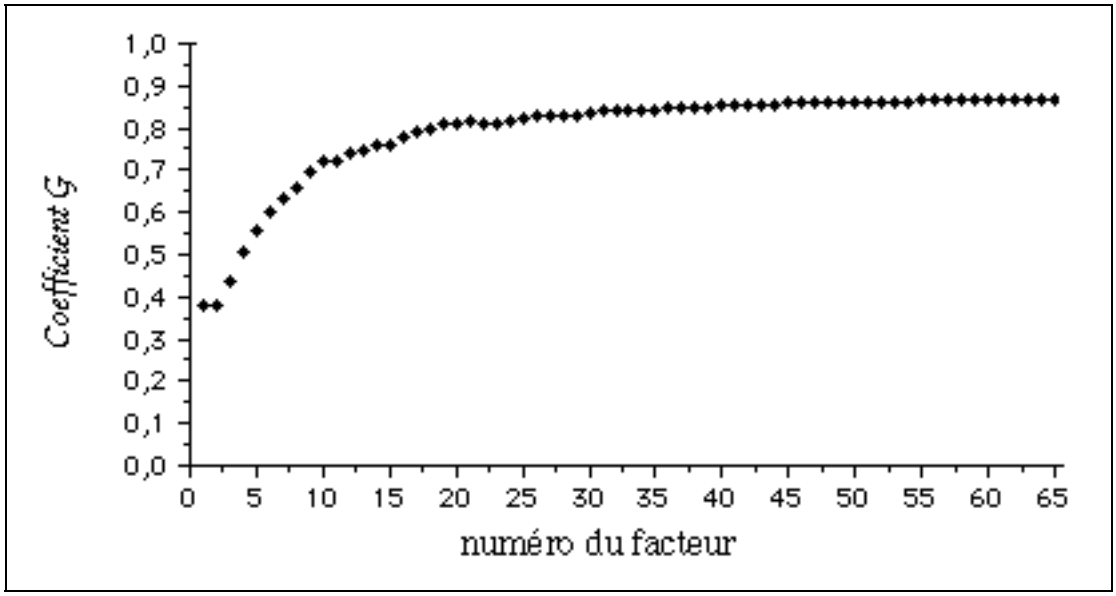


Figure 7.10

***Isaïe* : Le calcul du coefficient G_α pour les 65 facteurs issus de l'analyse des correspondances du tableau (89 formes x 66 chapitres)**

De tout ce qui précède, on retiendra que les deux premiers facteurs, outre le fait qu'ils correspondent aux plus fortes valeurs propres, possèdent plus que les autres la propriété de rapprocher les chapitres consécutifs.

7.7.3 Fragments de *n* chapitres consécutifs.

L'analyse du tableau (89 formes x 66 chapitres) produit plusieurs facteurs (les deux premiers de façon incontestable) qui possèdent globalement, et à un très haut degré, la propriété de rapprocher les chapitres consécutifs.

Cependant, certains des chapitres s'écartent sensiblement des chapitres qui leur sont voisins dans le texte sans que l'on puisse savoir si cet écart doit être mis au compte de ce que l'on pourrait appeler des fluctuations d'échantillonage ou, au contraire, d'une différence profonde dans le stock lexical employé.

Pour tenter d'avancer sur ce point, on procédera à des analyses du même type sur des fragments plus importants. Pour ne pas introduire des découpages qui peuvent nous être suggérés par la tradition de la critique biblique, nous nous sommes orientés vers des découpages en unités sensiblement égales constituées par la réunion d'un même nombre de chapitres consécutifs.

On présentera ici un découpage en 13 fragments de 5 chapitres consécutifs[1]. A l'issue de cette étape qui doit fournir une typologie plus nette, l'unité des fragments constitués de manière aussi brutale sera remise en cause, lors d'une phase consacrée tout particulièrement à l'affinage du découpage.

Tableau 7.10

***Isaïe* : Le regroupement des 66 chapitres en 13 fragments de cinq chapitres consécutifs**

1	*Chap 1 - 5*	*8*	*Chap 36 - 40*
2	*Chap 6 - 10*	*9*	*Chap 41 - 45*
3	*Chap 11 - 15*	*10*	*Chap 46 - 50*
4	*Chap 16 - 20*	*11*	*Chap 51 - 55*
5	*Chap 21 - 25*	*12*	*Chap 56 - 60*
6	*Chap 26 - 30*	*13*	*Chap 51 - 66*
7	*Chap 31 - 35*		

Pour ce dernier découpage, le poids des occurrences décomptées pour l'ensemble des 89 formes varie du simple au triple.

Ici encore, trois axes se détachent nettement avec des pourcentages d'inertie respectivement égaux à : τ_1=29.3 %, τ_2=17.6 % et τ_3=10.0 %.

On retrouve sur le premier axe l'opposition entre les premiers et les derniers fragments. Du côté négatif, les 5 derniers fragments s'opposent fortement au reste. Comme dans la première analyse, les contributions les plus fortes sont dues aux fragments 9 et 10 (Chap 40 à 50) qui portent à eux seuls 60% de l'inertie de l'axe.

Le second axe est constitué par l'opposition du fragment 8 (Chap 36 à 40) à tous les autres. Ici encore, ces fragments sont très bien expliqués par le plan des axes 1 et 2.

Comparaison des deux analyses

En comparant sur l'ensemble des formes les facteurs issus de l'analyse (89 formes x 13 fragments) avec les facteurs correspondants issus de l'analyse des correspondances du tableau (89 formes x 66 chapitres) on s'aperçoit que les deux ensembles présentent de très nombreuses similitudes. Cette circonstance n'est pas surprenante dans la mesure ou ces fragments ont été constitués en opérant, par rapport à la première analyse, des regroupements

[1] Le dernier fragment 13 compte 6 chapitres afin que la partition couvre l'ensemble du livre.

de chapitres consécutifs dont les profils se ressemblent. Dans le cas d'une égalité parfaite des profils à l'intérieur de chaque groupe, le principe d'équivalence distributionnelle (Chapitre 3) nous aurait assuré une totale similitude des deux analyses.

7.7.4 Classification des chapitres

Les résultats des classifications hiérarchiques sont souvent sensibles aux fluctuations dans les données initiales car celles-ci se répercutent à tous les niveaux de la hiérarchie contrairement aux premiers facteurs des analyses des correspondances qui présentent quant à eux, toutes conditions égales conservées, plus de stabilité en général.

La classification ascendante hiérarchique portant sur l'ensemble des 66 chapitres nous a fourni des résultats très médiocres : si nous avons retrouvé les grandes classes dégagées par l'analyse des correspondances, le nombre des exceptions devenait beaucoup trop important pour que nous puissions effectuer une synthèse à partir des classes obtenues.

Dans un second temps on a classé les fragments obtenus plus haut à partir des regroupements de chapitres consécutifs.

On examine le tracé de l'arbre - fig 7.11 - en se souvenant que les seules proximités à interpréter sont celles qui font l'objet de la formation d'une classe.

Ainsi, le fait que les fragments 9 et 10 soient réunis très bas dans l'arbre témoigne d'une proximité relative importante entre eux. Cette classe ne sera réunie au reste des fragments qu'à la dernière étape de l'algorithme d'agrégation.

Par ailleurs, on constate que l'algorithme de classification agrège dans la majorité des cas des fragments consécutifs. La seule exception à cette règle est fournie par la classe 15 qui agrège les fragments 3 et 5 auxquels le fragment 4 vient presque immédiatement se joindre pour former avec eux la classe 17 qui contient les fragments 3, 4 et 5.

Le fait que les classes de la hiérarchie agrègent des fragments consécutifs nous amène à supposer que des parties homogènes, au moins du point de vue de la description à laquelle nous nous livrons dans ce chapitre, existent dans le livre dont la taille dépasse les fragments que nous avons constitués jusqu'à présent.

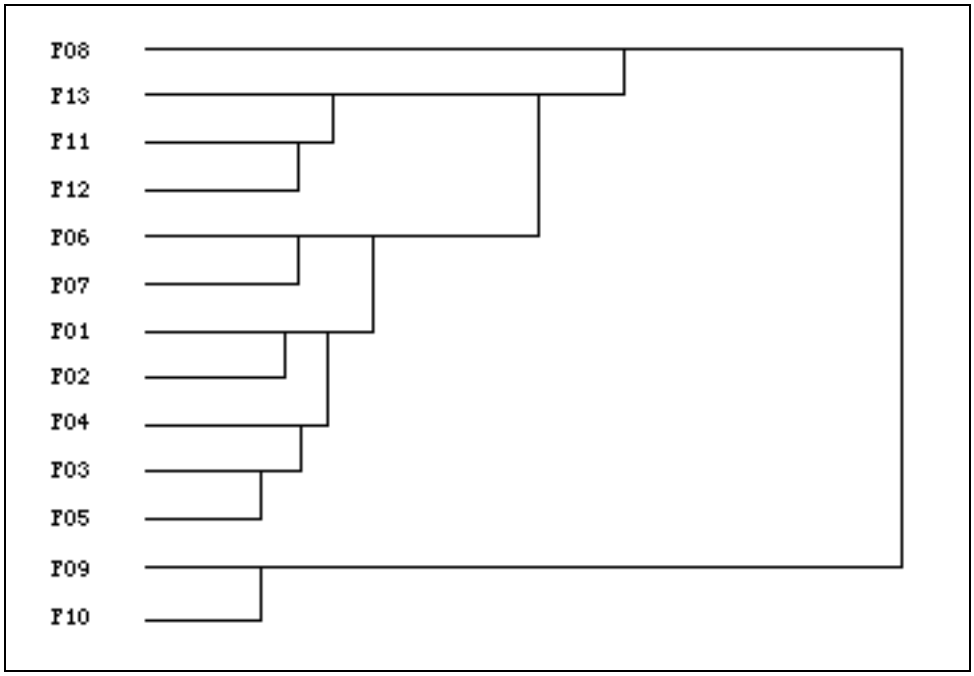


Figure 7.11

***Isaïe* : Classification ascendante hiérarchique à partir du tableau (89 formes x 13 fragments)**

Il est donc possible d'avancer d'un pas et d'étudier l'hypothèse d'une quadripartition provisoire de l'oeuvre selon le schéma présenté au tableau 7.11. Bien entendu les frontières entre les parties ainsi découpées, si elles ne peuvent être précisées, dans cette expérience, au-delà de la frontière de chapitre, peuvent être appréhendées avec plus de précision en analysant les caractéristiques de chacun des chapitres par rapport au fragment auquel il participe.

Tableau 7.11

***Isaïe* : Partition empirique du livre en 4 parties à partir du regroupement de fragments de chapitres consécutifs**

1	Fragments	1 - 7	—	Chap	1 - 35
2	Fragment	8	—	Chap	36 - 40
3	Fragments	9 -10	—	Chap	41 - 50
4	Fragments	11-13	—	Chap	51 - 66

A l'issue de ces expériences, le problème de la détermination du nombre d'auteurs qui ont participé à la rédaction du livre d'Isaïe n'est certes pas résolu. Les facteurs mis en évidence par l'analyse des correspondances à partir du tableau (89 formes x 66 chapitres) possèdent la propriété très

intéressante de rapprocher les chapitres consécutifs et plus généralement les chapitres qui se trouvent être proches dans le texte.

Incontestablement, à l'intérieur de ces grandes classes les chapitres consécutifs présentent des similitudes quant à l'emploi des formes retenues pour l'analyse.

Chapitre 8

Analyse discriminante textuelle

Les méthodes statistiques évoquées dans les chapitres précédents s'appliquent pour la plupart dans le cadre d'une démarche exploratoire, ou, si l'on préfère une terminologie plus traditionnelle, d'une démarche descriptive. Plus dynamique et plus interactive que la simple description, *l'exploration* recourt à la statistique multidimensionnelle pour obtenir des visualisations ou des regroupements d'éléments qui peuvent être soit des textes, soit des unités décomptées à l'intérieur de ces textes : c'est une recherche d'organisations, de traits structuraux, de résumés suggestifs.

Ces méthodes complètent une panoplie (beaucoup plus étendue) de techniques dévolues à des démarches que l'on qualifie souvent de décisionnelles, encore qu'il ne s'agisse en réalité que d'aide à la décision. Dans le domaine de la statistique textuelle, l'aide à la décision pourra concerner l'attribution d'un texte à un auteur ou à une période, la réponse à une requête sous forme de choix d'un document dans une base de donnée, la codification d'une information exprimée en langage naturel.

Les procédures statistiques qui permettent de réaliser ces attributions, ces choix, ces codifications, relèvent de l'analyse discriminante. Elles visent classiquement à prédire l'appartenance d'un "individu" à une classe ou une catégorie (parmi plusieurs possibles) à partir de variables mesurées sur cet individu.[1] Cette prédiction est rendue possible par une phase d'apprentissage réalisée sur un ensemble d'individus pour lesquels les variables et les catégories sont simultanément connues (*learning or training sample*). Nous

[1] Les premières analyses discriminantes furent réalisées à propos de mesures biométriques et anthropométriques par les statisticiens Fisher et Mahanalobis, qui tentaient de prévoir une appartenance ethnique à partir de mensurations effectuées sur le squelette (Fisher, 1936; Mahalanobis, 1936). Ils furent les premiers à utiliser la technique que l'on désigne parfois sous le nom d'analyse factorielle discriminante, ou encore de discrimination linéaire : c'est la méthode la plus ancienne, et c'est encore une des méthodes les plus utilisées actuellement.

avons signalé (Chapitre 2) que les notions de variable et d'individu sont moins évidentes pour les chercheurs qui étudient les matériaux textuels qu'elles ne le sont pour un biométricien ou un démographe. Le découpage et parfois l'élaboration des unités statistiques constituent, on l'a vu, une phase importante de la recherche. L'analyse discriminante textuelle fera grand cas de ces étapes préparatoires.

On distinguera au paragraphe 8.1 deux grandes familles de préoccupations qui s'apparentent, dans le domaine textuel, aux procédures statistiques de discrimination et de reconnaissance des formes. La première famille, fait abstraction du contenu des textes et s'intéresse essentiellement à la forme. Elle relève du domaine de la stylométrie et sera étudiée aux paragraphes 8.2 (brève revue des unités et indices de la stylométrie) et 8.3 (un exemple de modèle statistique en stylométrie, avec suggestion de méthodes alternatives). La seconde famille que l'on désigne ici par *discrimination globale*, parce qu'elle prend en compte le contenu, et parfois le contenu et la forme des textes, est présentée au paragraphe 8.4, puis développée avec le support d'un exemple au paragraphe 8.5.

8.1 Deux grandes familles de problèmes dans le domaine textuel

Parmi les applications les plus courantes des techniques qui s'apparentent à la discrimination, on peut distinguer deux grandes séries de préoccupations :

- Les applications à des corpus littéraires (attributions d'auteurs, datation, par exemple) cherchent à s'affranchir du contenu pour saisir des caractéristiques de *forme* (souvent : de style) à partir des distributions statistiques de vocabulaire, d'indices ou de ratios. Il s'agit de saisir des "invariants" d'un auteur ou d'une époque, dissimulés ou peu apparents, comme des habitudes ou des tics insignifiants de prime abord permettent aux détectives de Simenon ou de Conan Doyle d'identifier un individu à son insu.

- A l'opposé, les applications réalisées en recherche documentaire, en codification automatique, dans le traitement des réponses à des questions ouvertes, visent essentiellement le *contenu,* le sens, la substance des textes. La façon dont une réponse est formulée importe moins que son classement dans un groupe de documents présentant une certaine homogénéité au plan du contenu. Nous appellerons *discrimination globale* cette seconde famille d'application.

Dans la pratique, comme nous l'avons signalé plus haut, il n'est pas toujours utile de chercher à maintenir cette distinction. Lors du traitement statistique

de réponses à des questions ouvertes, la forme des réponses, les connotations véhiculées par des termes apparemment équivalent constituent répétons-le une information très riche qui permet souvent de nuancer et d'infléchir les contenus plus manifestes que l'on peut repérer.

8.1.1 Discrimination à partir de la forme : la stylométrie

Dans de nombreuses applications à des corpus littéraires, politiques, historiques, tel le problème d'attribution que nous avons vu au chapitre précédent, les analyses discriminantes ont été utilisées pour attribuer à des auteurs connus des textes (poésies, pièces de théâtre, romans, textes religieux ou sacrés, etc.) d'origine mal établie, ou pour déterminer la date ou l'époque de leur écriture. Ces procédures d'attribution n'ont pas toujours fait appel à l'analyse discriminante des biométriciens, comme en témoigne le travail de pionnier de Yule (1944).

C. R. Rao (1989) dans son ouvrage de réflexion générale *Statistics and Truth* cite les travaux plus récents de Thisted et Efron (1987) à propos de l'attribution à Shakespeare d'un poème découvert en 1985, travaux sur lesquels nous reviendrons plus loin. Il cite également l'imposant travail de Mosteller et Wallace (1964) sur l'attribution des "Federalist papers".

Parmi ces 77 textes politiques, publiés anonymement à New York à la fin du dix-huitième siècle, 12 textes n'ont pu être attribués à un auteur parmi deux possibles. Le traitement statistique permet de désigner l'auteur le plus probable de ces 12 textes litigieux. On pourra aussi consulter sur un thème voisin un travail plus récent de Holmes (1992) sur l'homogénéité des "Mormon Scriptures", et une revue assez complète de ce même auteur sur les méthodes de la *stylométrie* (1985).[1]

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

L'utilisation systématique de l'analyse des correspondances et des techniques de classification automatique (cf. Benzécri et al., 1981; 1992) a considérablement enrichi ces approches.

[1] La seconde version du travail de Mosteller et Wallace (1984) contient également un panorama général des tentatives d'attribution d'auteur.

8.1.2 Discrimination globale : recherche documentaire, codification, validation

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

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

Dans ce cadre particulier, il est possible de faire appel à plusieurs sources extérieures d'information pour résoudre les problèmes de classement : des analyseurs syntaxiques, premiers pas vers une *compréhension* de la requête, des dictionnaires ou des réseaux sémantiques pour lemmatiser et désambiguïser les requêtes[1], éventuellement à des corpus artificiels faisant appel à des experts.[2]

Mais beaucoup de techniques utilisées, et parmi les plus efficaces si l'on en croit leurs auteurs, ont recours à des outils multidimensionnels très proches de ceux préconisés par Benzécri (1973, 1977, 1981), ou dans le cadre des grands tableaux clairsemés par Lebart (1982a). Deerwester et al. (1990) proposent ainsi sous le nom de *Latent Semantic Analysis* une méthode très semblable à la discrimination d'après les premiers facteurs d'une analyse des correspondances (ces auteurs utilisent en fait une décomposition aux valeurs singulières, technique qui est à la base de l'analyse des correspondances et de l'analyse en composantes principales).

D'autres auteurs insistent sur la complémentarité entre modèles et méthodes descriptives dans la recherche documentaire (cf. Fuhr et al., 1991), et sur l'intérêt de visualisations les plus synthétiques possibles (Fowler et al., 1991). On peut dire que la tendance actuelle est à l'effacement progressif des barrières interdisciplinaires, et que, de plus en plus, les "décisionistes purs et durs" reconnaissent l'importance des phases de description et d'exploration.

Nous allons présenter ces deux volets (stylométrie et discrimination globale) en nous appuyant sur deux exemples de méthodes et d'applications qui

[1] Cf. par exemple, en matière de classification de documents, les travaux de Blosseville et al. (1992), Hebrail et al. (1990, 1993).

[2] Cf. les travaux de pionniers de Palermo et Jenkins (1964). Cf. également Bouroche et Curvalle (1974).

mettent en évidence, pensons-nous, la divergence des démarches, et leurs intérêts respectifs.

Après un bref exposé des spécificités des unités statistiques et indices de la stylométrie, le premier exemple (paragraphe 8.3) reprend et reconsidère le problème stylométrique, évoqué plus haut, d'attribution éventuelle à Shakespeare d'un poème récemment découvert. Le second exemple (paragraphe 8.5) présente pour illustrer l'analyse discriminante globale une application multilingue dans le cadre d'une enquête internationale.

8.2 Les unités et indices de la stylométrie

Ici encore, les unités de bases seront dérivées de la forme graphique, mais il convient de noter que de nombreux travaux stylométriques ont pu utiliser, pour comparer des textes, des auteurs ou des genres, des comptages qui fractionnent encore cette unité[1].

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

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

8.2.1 "Mots-outil", parties du discours

Comme nous l'avons expliqué plus haut, la notion de mot-outil n'est pas une notion de la statistique textuelle dans la mesure ou elle ne se prête à aucune formalisation satisfaisante. De nombreux auteurs utilisent cependant cette notion en s'appuyant sur l'intuition commune que l'on peut dresser, dans chaque langue, une liste de formes que l'on appelle parfois encore "mots-

[1] Hors du cadre de la stylométrie, on trouvera une analyse intéressante des constituants du mot français dans Gruaz (1987).

grammaticaux" et qui ont en commun la propriété d'être moins marqués au plan sémantique[1].

Sur la base d'une telle liste Demonet et al. (1975) ont proposé de mesurer un "taux de fonctionnalité" propre à chaque discours ou à chaque type de discours en recensant le nombre des occurrences qui correspondent à des occurrences d'une liste de "mots-outil" préalablement dressée par eux. Cette propriété a par ailleurs, été largement utilisée, dans les études informatisées, pour écarter de la liste des formes considérées des unités correspondant en grande partie aux formes les plus fréquentes, réputées peu dignes d'intérêt.

A l'inverse, le courant qui s'occupe des problèmes d'attribution d'auteur a longtemps privilégié l'étude de ce type d'unité, posant que leur emploi, moins maîtrisé lors de la rédaction du texte pouvait constituer une marque d'auteur privilégiée[2].

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

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

Une catégorisation des unités graphique du texte (analyse syntaxique qui permet d'affecter à chaque forme une catégorie grammaticale[3]), permet de calculer la proportion de noms, de verbes, d'adjectifs, etc. Ces proportions ont également été utilisées en stylométrie. Ainsi, Somers (1966) affirme que la proportion de substantifs dénote instruction et aisance dans l'expression, Brainerd (1974) étudie le pouvoir discriminant de la proportion d'articles.

---

[1] Ce dernier trait n'exclut pas de s'intéresser à de telles unités dans certains domaines d'étude. Que l'on songe, par exemple au rôle important que peuvent jouer dans l'analyse de textes politiques les mots-outil *pour* et *contre*.

[2] Cf. par exemple les travaux de Radday (1974) et Morton (1963) évoqués au chapitre 7, paragraphe 7.7, à propos de l'homogénéité du livre d'Isaïe.

[3] Aujourd'hui, la réalisation d'une telle opération ne peut être entièrement confiée à un ordinateur. Des progrès importants ont été réalisés dans le domaine de l'analyse syntaxique automatisée des textes, comme en témoigne, par exemple, l'amélioration constante des correcteurs orthographiques que l'on trouve désormais sur la plupart des machines de traitement de texte.

Précisons qu'une telle catégorisation est évidemment nécessaire pour identifier d'éventuels mots-outil.[1]

8.2.2 La richesse du vocabulaire

Pour conclure sur ce bref survol des unités statistiques de la stylométrie, il faut mentionner les indices de richesse de vocabulaire : Si V désigne le nombre de formes différentes, et T le nombre total d'occurrences, pour chaque texte, les premiers indices proposés furent les quotients du nombre de mots distincts V par le nombre total T d'occurrences : $R = V / T$ (*type -token ratio*) ou encore $R' = LogV / LogT$.

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

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

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

$$D = \Sigma_r \, r(r-1)V_r \, / \, T(T-1)$$

où V_r est le nombre de formes distinctes apparaissant exactement r fois dans le texte. C'est donc le quotient du nombre de paires d'occurrences d'une même forme par le nombre total de paires d'occurrences. Le numérateur n'est plus borné, et D dépend beaucoup moins de T que R ou R'. D est simplement la probabilité que deux occurrences prises au hasard dans le corpus correspondent à une même forme.
Cet indice bien connu des statisticiens est très proche dans son principe de la caractéristique K introduite, très antérieurement, par Yule dans le domaine des études stylistiques et qui peut être définie comme :

$$K = 10^4 \; D \, (T-1)/T$$

[1] Notons que si l'isolement de mots-outil demande une catégorisation et une désambiguïsation du texte (cas de la forme *pas*, par exemple) il existe aussi des locutions contenant des mots pleins qui sont des substituts de mots-outil, et qu'une lemmatisation préliminaire pourrait masquer (cf. par exemple Benzécri, 1992a).

8.3 Modèles statistiques en stylométrie : un exemple

On comprend, à la lecture de ce qui précède, que domaines, préoccupations et méthodologie ne sont pas indépendants. Ainsi, l'essentiel des travaux stylométriques met en oeuvre des méthodes statistiques unidimensionnelles. Il s'agit le plus souvent de calculs de paramètres discriminants, à l'exception des travaux, en général plus récents, mettant en oeuvre des vecteurs de mots-outil, ou des combinaisons d'indices stylométriques divers.

A côté des études empiriques utilisant la flore d'indices descriptifs évoquée précédemment, il existe une démarche plus "modélisatrice", moins facile d'accès, mais qui ne manque pas d'intérêt[1].

Nous allons nous appuyer sur le cas déjà cité de l'étude d'attribution d'auteur réalisée à propos du poème (peut-être écrit par Shakespeare) découvert en 1985 (Thisted et Efron, op.cit.) en retraçant brièvement l'histoire de la démarche, les hypothèses, le modèle, les résultats, et en suggérant d'autres approches.

8.3.1 Modélisations de la gamme des fréquences

Nous mentionnerons sous ce chapitre certaines applications au domaine textuel des modèles de capture de l'écologie des années cinquante. Les écologistes disposent des pièges ou des trappes pour capturer des spécimens de diverses espèces.

Si l'on fait l'hypothèse, raisonnable et acceptable empiriquement, que le nombre de spécimens effectivement capturés d'une espèce donnée suit une *loi de Poisson* dont l'unique paramètre dépend de l'espèce, on peut estimer, après un certain temps de capture, à la fois le nombre de spécimens par espèce, et également, ce qui peut paraître surprenant, le nombre probable d'espèces "piègeables" dans la zone de capture...

Avant de préciser ces résultats à l'intention des lecteurs plus mathématiciens (pour lesquels le modèle - non-paramétrique - sera présenté en section 8.3.3) traduisons-les en termes de statistique textuelle.

A l'espèce correspond ici la forme graphique, au spécimen une occurrence de cette forme. Il va de soi qu'il ne s'agit que de modéliser le bilan des

[1] On insistera ici sur les modèles de génération de la gamme des fréquences, et non sur les modèles d'ajustement. A cette dernière catégorie appartient par exemple le modèle de distribution de Waring-Herdan, dont on trouvera par exemple une présentation détaillée dans le manuel classique de Muller (1977).

distributions lexicales d'une oeuvre à un instant donné, et non le processus de création. Nous n'avons pas jusqu'à présent fait d'hypothèse d'indépendance entre les nombres d'occurrences par forme, ce qui eût été difficilement acceptable, même si les notions d'associations et de cooccurrences disparaissent au niveau d'un bilan.

Ce modèle permet donc d'estimer, pour un *texte* donné l'évolution de la gamme des fréquences (nouvelles fréquences des formes déjà utilisées, et nouvelles formes) à mesure que le volume du texte augmente. Etant donné un corpus de textes sur lesquels des comptages ont déjà été effectués et une oeuvre nouvelle, il est donc possible de voir si elle satisfait au canon défini par toutes les oeuvres antérieures, y compris dans l'accroissement du vocabulaire qu'elle implique.

En fait, le modèle n'est valide que si le processus est homogène dans le temps : il ne s'applique pas a priori (sauf vérification empirique) au cas de changement de genre (théâtre en vers versus théâtre en prose, par exemple).

8.3.2 Les matériaux disponibles pour le problème d'attribution

A partir de la somme de Spevack (1968), il a été possible de connaître la distribution du nombre de mots distincts en fonction de leur fréquence d'apparition dans l'oeuvre de Shakespeare (tableau 8.1).

Selon les décomptes effectués par cet auteur, Shakespeare aurait employé 31 534 formes distinctes, pour un total de 884 647 occurrences. On notera que prés de la moitié des mots distincts sont des hapax.[1]

A partir de cette distribution, et du modèle esquissé plus haut et que l'on considérera plus loin en détail, il est possible d'estimer non seulement les fréquences d'apparition de chacune des formes, mais également les fréquences de nouvelles formes. C'est ce qui a conduit Efron et Thisted (1976) à tenter de répondre à la question *"How many words did Shakespeare know?"*, et à estimer à 35 000 le nombre de mots distincts supplémentaires si le processus de "création" s'était poursuivi indéfiniment.

Cette extrême sollicitation du modèle le conduit très certainement en dehors de son domaine de validité, car comme nous l'avons précisé plus haut, l'hypothèse d'homogénéité du processus est fondamentale, et celle-ci est peu vraisemblable sur longue période. De plus, considérer que les mots qui

[1] Bien entendu, cette précision est tout-à-fait illusoire, puisque pour une pièce aussi étudiée que *Hamlet*, il existe des paragraphes entiers absents dans certaines sources, alors que des désaccords sur l'identification des mots subsistent pour les parties de textes communes à toutes les sources.

pourraient être utilisés plus tard sont des mots tous connus actuellement constitue une hypothèse supplémentaire. Mais le plus surprenant, pour un statisticien, est l'ordre de grandeur somme toute raisonnable du résultat obtenu.

Tableau 8.1

Gamme des fréquences dans l'oeuvre de Shakespeare (tableau partiel)

Fréq. f_i	*Effectif V_i*	*Fréq. f_i*	*Effectif V_i*
1	14376	*8*	519
2	4343	*9*	430
3	2202	*10*	364
4	1463	*11*	305
5	1043	*12*	259
6	837	.....	
7	638	*>12*	846

Nombre des formes $= \Sigma_i V_i$ 31 534

Nombre des occurrences $= \Sigma_i f_i V_i$ 884 647

C'est en 1985, neuf années après la parution de l'article précité qu'a été découvert par Gary Taylor un poème formé de neuf strophes et de 258 mots (429 occurrences) pouvant être attribué à Shakespeare. Ce fut l'occasion pour Thisted et Efron (1987) d'appliquer leur modèle sur un accroissement très modeste (t = 0,05 %) de l'oeuvre de cet auteur.

Par prudence, ces auteurs traitent simultanément le cas de sept autres poèmes élisabéthains définitivement attribués, dont quatre sont dus à Shakespeare.

Les huit poèmes élisabéthains :

Ben Jonson	An Elegy
C. Marlowe	Four poems
J.Donne	The Ecstasy

Shakespeare	Cymbeline (extraits)
Shakespeare	A Midsummer Night's Dream (extraits)
Shakespeare	The Phoenix and Turtle
Shakespeare	Sonnets (extraits)
Shakespeare	**(?)** Taylor's Poem

Les textes ont été choisis de façon à être sensiblement équivalents en ce qui concerne la longueur, et comparables du point de vue du genre et des grands thèmes abordés.

Le tableau 8.2 ci-après contient les distributions des formes de ces huit poèmes selon le canon Shakespearien : ainsi, le poème de Taylor contient 9 formes jamais utilisées par Shakespeare, et 140 formes utilisées par lui plus de 100 fois.

Tableau 8.2

Distribution des formes dans les huit poèmes, selon leur fréquence d'apparition dans l'oeuvre de Shakespeare.

Freq.	**BJon**	**Marl**	**Donn**	**Cymb**	**Mids**	**Phoe**	**Sonn**	**Tayl**	***Total***
0	8	10	17	7	1	14	7	9	73
1	2	8	5	4	4	5	8	7	43
2	1	8	6	3	0	5	1	5	29
3-4	6	16	5	5	3	9	5	8	57
5-9	9	22	12	13	9	8	16	11	100
10-19	9	20	17	17	6	18	14	10	111
20-29	12	13	14	9	9	13	12	21	103
30-39	12	9	6	12	4	7	13	16	79
40-59	13	14	12	17	5	13	12	18	104
60-79	10	9	3	4	9	8	13	8	64
80-99	13	5	10	4	3	5	8	5	53
+100	148	138	145	120	103	111	155	140	1060
Total	243	272	252	215	156	216	264	258	1876

Les auteurs ont calculé, pour chaque case de ce tableau, la valeur théorique fournie par le modèle dans l'hypothèse selon laquelle chaque texte serait un texte supplémentaire de Shakespeare. Ainsi, par exemple, pour le poème *Taylor*, le tableau 8.3 donne les valeurs théoriques calculées dans cette hypothèse, selon une formule qui est explicitée dans la section technique 8.3.3 ci-après.

L'adéquation des valeurs observées aux valeurs théoriques est calculée pour les huit poèmes en utilisant une grille plus fine (les 100 valeurs de fréquence de 0 à 99) en utilisant un modèle linéaire généralisé. Les auteurs sont ainsi conduits à rejeter comme non-shakespeariens tous les poèmes connus comme tels, mais aussi *The Phoenix and Turtle*, de Shakespeare.

Le poème *Taylor* n'est pas rejeté. Le test apparaît donc comme plutôt trop sévère, puisqu'il rejette même un poème de Shakespeare, ce qui renforce pour ces auteurs la présomption d'appartenance du nouveau poème au corpus shakespearien.

Tableau 8.3

Valeurs théoriques et observées pour le poème *Taylor*

Fréquences	*Taylor observé*	*Taylor théorique*
0	9	6.97
1	7	4.21
2	5	3.33
3-4	8	5.36
5-9	11	10.24
10-19	10	13.96
20-29	21	10.77
30-39	16	8.87
40-59	18	13.77
60-79	8	9.99
80-99	5	7.48

8.3.3 Précision sur le modèle non-paramétrique d'estimation

A l'usage du lecteur statisticien, on précise dans ce paragraphe le modèle esquissé plus haut[1]. L'hypothèse de base concerne la distribution de la forme s qui est une loi de Poisson de paramètre λ_S. Il y a potentiellement S formes, mais toutes ne sont pas observées.

On désignera par $G(\lambda)$ la fonction de répartition empirique (inconnue) des nombres λ_1 , λ_2 , ... λ_S. Désignons également par n_x le nombre de formes distinctes observées exactement x fois.

L'espérance mathématique η_x du nombre exact de formes distinctes observées exactement x fois dans le passé, représenté ici conventionnellement par l'intervalle [-1, 0] s'écrit :

$$\eta_x = E(n_x) = s\int_0^{\infty} e^{-\lambda} \frac{\lambda^x}{x!} dG(\lambda) \qquad (8.1)$$

Notons que l'espérance E_t du nombre de formes distinctes observées à l'instant t mais non observées auparavant (formes nouvelles) s'écrit de façon simple.

[1] Le lecteur non-matématicien pourra se dispenser de la lecture de ce paragraphe et se reporter directement au paragraphe suivant.

$$E_t = s\int_0^{\infty} e^{-\lambda}(1-e^{-\lambda t})dG(\lambda) \qquad (8.2)$$

En substituant le développement :

$$1-e^{-\lambda t} = \lambda t - \frac{\lambda^2 t^2}{2!} + \frac{\lambda^3 t^3}{3!} - \ldots \qquad (8.3)$$

dans l'équation 8.2, et en tenant compte de 8.1, on trouve que E_t peut s'écrire comme la somme de la série :

$$E_t = \eta_1 t - \eta_2 t^2 + \eta_3 t^3 + . \qquad (8.4)$$

Cette étonnante relation, découverte par Good et Toulmin (1956) donne donc l'espérance du nombre de nouvelles formes pour une proportion supplémentaire *t* de texte, en fonction des espérances des nombres de formes η_x observées *x* fois au cours du passé, sans qu'il n'ait été fait aucune hypothèse sur la forme de *G(λ),* qui n'intervient pas dans le calcul.

Cette formule a permis de calculer le premier terme du tableau 8.3, en substituant aux valeurs théoriques η_x les valeurs calculées n_x. La valeur de *t*, quotient des nombres d'occurrences, est de :

$$t = 4.85\ 10^{-4}\ \ (= 429 / 884\,647).$$

Avec une valeur si petite, seul le premier terme *14 376 x t (= 6.97)* est non-négligeable.

Ainsi, avec ce modèle, le nombre de formes nouvelles, dans le cas d'un très petit accroissement relatif du volume des écrits, s'obtient par une simple règle de trois à partir du nombre d'hapax antérieurement observé.

Intéressons-nous plus généralement à l'espérance mathématique ν_x du nombre exact de formes distinctes observées exactement *x* fois dans l'intervalle [-1 , 0], et présentes dans l'intervalle [0 , t]. Elle vaut :

$$\nu_x = E(n_x) = s\int_0^{\infty} e^{-\lambda}\frac{\lambda^x}{x!}(1-e^{-\lambda t})dG(\lambda)$$

La même procédure de substitution utilisant le développement 8.3 nous donne cette fois-ci une formule un peu plus complexe :

$$\nu_x = \sum_{k=1}^{\infty} (-1)^{k+1} \binom{x+k}{k} t^k \eta_{x+k}$$

Pour obtenir une estimation de cette dernière quantité, on peut substituer aux valeurs théoriques η_X les valeurs calculées n_X :

$$\nu_x^* = \sum_{k=1}^{\infty} (-1)^{k+1} \binom{x+k}{k} t^k n_{x+k}$$

Il s'agit d'un *estimateur empirique bayésien* de ν_x .

Pour $x = 0$, (formes nouvelles) on retrouve la formule 8.4.

C'est à partir de cette formule que sont calculées les valeurs théoriques telles que celles qui figurent dans le tableau 8.3.

8.3.4 Autres approches du problème

Considérons de nouveau le tableau 8.2 donnant la distribution des formes des huit poèmes selon les classes de fréquences d'apparition des mêmes formes dans l'oeuvre de Shakespeare.

Si la distribution des formes par fréquence d'emploi dans l'oeuvre de Shakespeare peut être considérée comme caractéristique de cet auteur, on doit s'attendre à une certaine hétérogénéité dans les profils des colonnes.

Effectivement, le chi-2 (test d'indépendance classique sur les tables de contingence), calculé (malgré la faiblesse de certains effectifs), vaut, pour 77 degrés de liberté (77 = [8 - 1] x [12 - 1]) , et avec les notations usuelles (voir chapitre 3) :

$$\chi^2 = 1876 \times \Sigma_{ij} (f_{ij} - f_{i.} f_{.j})^2 / f_{i.} f_{.j} = 104.7$$

Cette valeur a sensiblement une chance sur 100 d'être dépassée dans l'hypothèse d'homogénéité, qui correspond ici à l'indépendance des lignes et des colonnes de la table. Cette hypothèse est donc douteuse.

L'un des mérites de l'analyse des correspondances est de nous décrire, dans le cas d'un rejet de l'hypothèse d'indépendance, comment cette hypothèse est rejetée.
La figure 8.1 décrit donc les positions des poèmes et des classes de fréquences dans le plan factoriel. Le poème litigieux *Taylor* est proche de l'origine, donc du profil moyen.

Les poèmes de Shakespeare sont alignés le long d'une droite joignant *The Phoenix* à *Midsummer*.

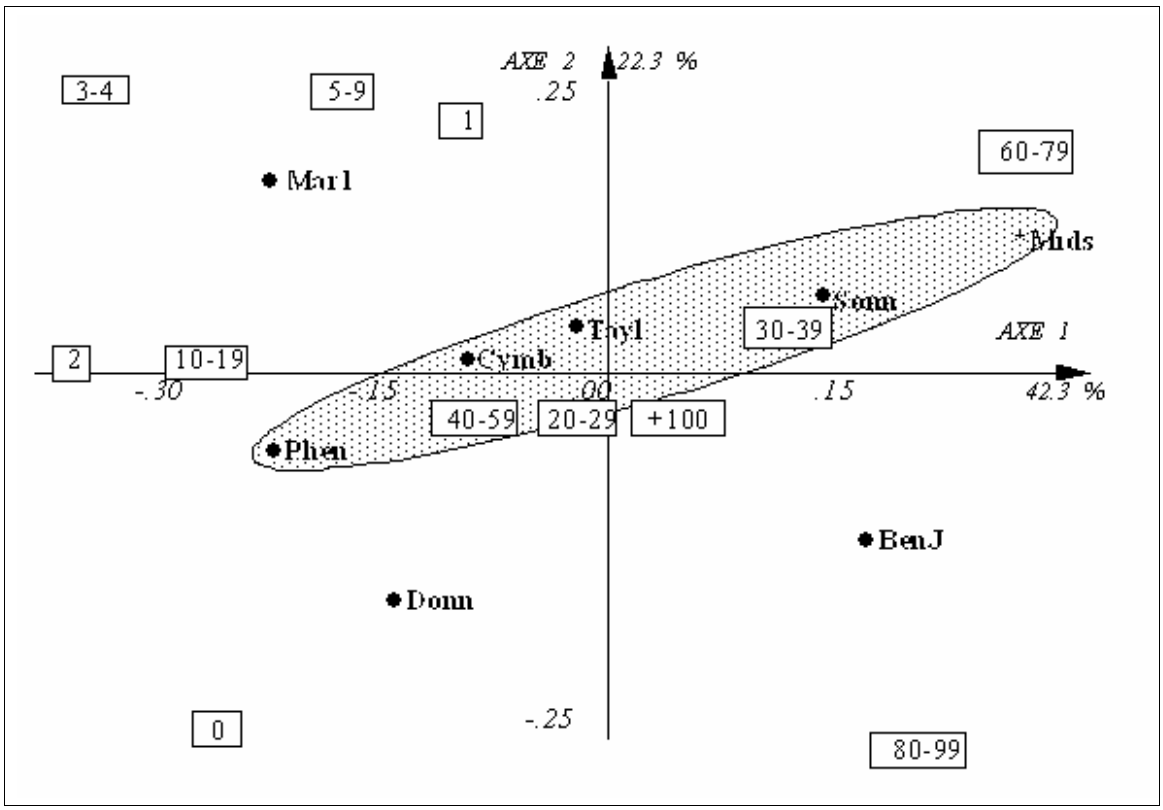


Figure 8.1

Premier plan factoriel de l'A.C. de la table 12 x 8
Classes de fréquence x Poèmes (tableau 8.2)
La zone ombrée contient les poèmes attribués à Shakespeare et le poème *Taylor*

The Phoenix, qui a été rejeté comme non-shakespearien par les tests issus du modèle poissonien, est effectivement assez périphérique, entre *Marlowe* et *Donne*, dans une zone riche en mots nouveaux (fréquence "0" chez Shakespeare) et en fréquences "2" et "10-19". A l'opposé, l'extrait de *Midsummer* est anormalement pauvre en formes nouvelles (exclusives).

Cette visualisation nuancée, fondée sur des comptages n'ayant subi aucune transformation, donne l'idée d'une approche empirique élargie, qui consisterait à multiplier les poèmes analysés (aussi bien de Shakespeare que d'autres auteurs élisabéthains) de façon à pouvoir déterminer des zones de rejets dans un plan factoriel analogue à celui de la figure 8.1. Ces zones seraient fondées sur des variations de densité dans le plan, et non sur un tout petit nombre de jalons.
Il s'agit là d'une approche voisines de l'analyse discriminante telle qu'elle sera évoquée dans la section suivante, à ceci près que les variables de bases utilisées jusqu'à présent (répartition des formes selon des classes de fréquence pour un auteur donné) sont réputées indépendantes du contenu.

8.4 Analyses discriminantes globales

Dans les analyses qui vont suivre, à l'usage de la recherche documentaire ou des analyses ou codifications de réponses libres dans les enquêtes, la discrimination se fera avec l'ambition de prendre en compte forme et contenu, ou parfois contenu seul. Précisons, s'il en est besoin, que prendre en compte n'est pas comprendre.

8.4.1 Principe général

Le principe de base commun aux méthodes d'analyse discriminante les plus répandues est le suivant :

a) on dispose de n points dans un espace de dimension élevée p, et ces n points sont répartis en K catégories, étiquetées de 1 à K.

- *exemple 1* : 65 textes, chacun caractérisé par son profil lexical calculé à partir des 200 formes graphiques les plus fréquentes du corpus global (espace à $p = 200$ dimensions : $\mathbf{R}^{200}$). Les $n = 65$ textes correspondent à $K = 2$ auteurs.

- *exemple 2* : les 1500 points sont des réponses à une question ouverte caractérisées une sélection de 100 formes ou segments (espace à $p = 100$ dimensions : $\mathbf{R}^{100}$). Les répondants ont (ou n'ont pas) acheté un certain produit ($K = 2$).

- *exemple 3* : chaque point, caractérisé par quelques mots-clés choisis parmi les 1 500 d'un thesaurus, correspond à un document dans un ensemble qui en comprend 10 000 (espace à $p = 1\,500$ dimensions : $\mathbf{R}^{1500}$). Ces documents sont répartis en $K = 200$ sous-matières.

b) On dispose également de m points, dans le même espace à p dimensions que les n points précédents (c'est-à-dire décrits par les mêmes variables ou les mêmes profils), mais ces points sont sans étiquette. Et l'on désire affecter à chacun de ces m points l'étiquette la plus probable parmi un ensemble d'étiquettes définies a priori.

L'exemple 1, pour $m = 12$, correspondant au problème cité des *Federalist papers* : il s'agit de savoir lequel de deux auteurs a écrit chacun des 12 textes.

L'exemple 2, pour $m = 1$ est, une tentative de prévision de comportement d'achat à partir d'une réponse libre.

L'exemple 3 est un problème de classement de document.

Une méthode simple consiste à calculer dans chaque cas les K profils moyens dans $\mathbf{R}^p$ et à affecter chaque nouveau point au profil qui lui est le plus proche.

On peut aussi tenir compte de la forme de chacun des K *sous-nuages* de points, car, comme l'illustre la figure 8.2 la proximité au point moyen n'est pas toujours le meilleur indice d'appartenance à la classe.

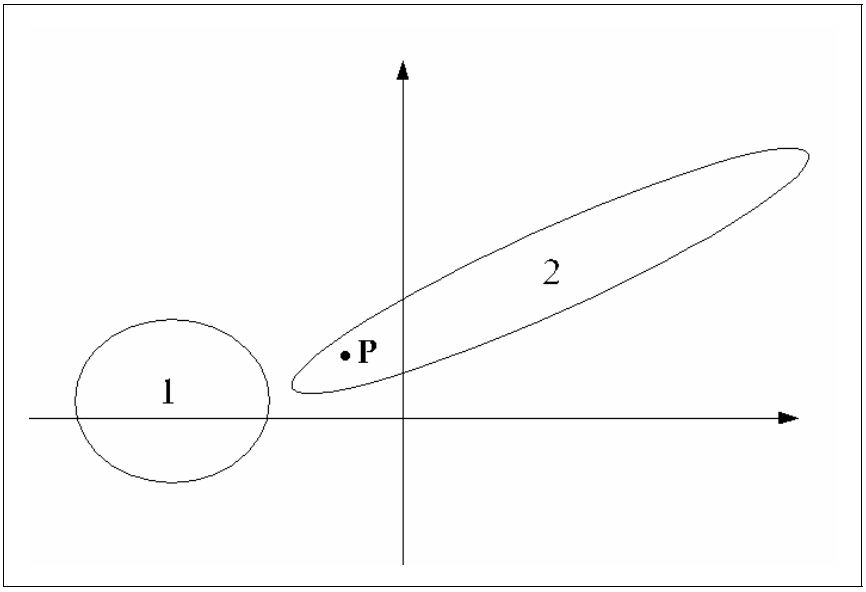


Figure 8.2

Le point P est plus proche du centre 1 que du centre 2, mais appartient plus vraisemblablement à la classe 2, compte tenu des densités observables.

Comme on le pressent, cette notion d'indice d'appartenance n'est pas une chose simple, et de nombreuses variantes techniques seront possibles pour que les distances calculées puissent s'interpréter en termes de probabilités. Ainsi, dans le cas de la figure 8.2, la *distance de Mahalanobis locale*, qui tient compte de la forme des ellipsoïdes[1] correspondant à chaque sous-nuage, rendra le point P plus proche du point 2 que du point 1.
Les principales méthodes de discrimination adaptées aux vastes tableaux de données qualitatives sont :

- la *discrimination sous indépendance conditionnelle* (cf. Goldstein et Dillon, 1978 ; Bochi et al., 1993),

[1] La distance de Mahalanobis locale du point **X** au groupe k, qui est utilisée en *analyse discriminante quadratique*, s'écrit $d_k(\mathbf{X}) = (\mathbf{X}-\mathbf{m}_k)' \mathbf{S}_k^{-1}(\mathbf{X}-\mathbf{m}_k)$ où $\mathbf{S}_k$ est la matrice de covariance interne au groupe **k**, de point moyen (centre de gravité) $\mathbf{m}_k$.

- la *discrimination par estimation directe de densité* (Habbema et al., 1974 ; Aitchison et Aitken, 1976),
- la *discrimination par la méthode des plus proches voisins* (cf. par exemple, Hand, 1986),
- la *discrimination sur coordonnées factorielles* (qui permet d'appliquer aux variables qualitatives toutes les méthodes qui s'appliquent aux variables numériques : analyse factorielle discriminante, discrimination quadratique, etc.) sur laquelle nous reviendrons (Benzécri, 1977 ; Wold, 1976).

On trouvera une revue récente et complète de la plupart de ces méthodes dans l'ouvrage de McLachlan (1992). Mais ces méthodes ne s'appliquent qu'après la détermination des unités statistiques de base, détermination dont on connaît l'importance dans le cas des données textuelles.

8.4.2 Unités pour la discrimination globale

Il est possible de faire varier la nature et la qualité de la discrimination en agissant sur le vocabulaire de base de différentes façons :

a) Les formes peuvent être sélectionnées à l'aide d'un seuil de fréquence préalablement fixé.

b) Les décomptes de formes peuvent être enrichis de comptages portant sur les segments et les quasi-segments,

c) Le texte peut être lemmatisé, (avec ou sans élimination des mots-outil),

d) Seules les formes (segments, lemmes, etc.) caractéristiques des classes à discriminer peuvent être sélectionnées a priori,

A partir du vocabulaire de travail ainsi constitué, il est possible :

e) De procéder à une analyse des correspondances préalable de la table (unités statistiques, individus), et de ne retenir que les premiers axes (filtrage et régularisation par analyse des correspondances),

f) De procéder à une classification des unités statistiques (cf. Hand, 1981 ; Celeux et al., 1991), et de travailler sur des agrégats d'unités.

Toutes ces possibilités impliquent donc que l'utilisateur définisse une stratégie. Même dans le cas d'une discrimination globale, les domaines et les préoccupations vont avoir une influence sur ces transformations préliminaire du vocabulaire de base.

Ainsi, dans le cas de réponses à des questions ouvertes (ou de recherche documentaire ou simplement de textes très courts), le fait de lemmatiser (procédure c) ou de regrouper (procédure f) ou encore de combiner les deux (procédure que l'on note c x f) modifie le résultat des calculs de distances entre textes. Deux textes avec des profils disjoints au départ peuvent acquérir des éléments communs après ces transformations : deux flexions d'une même unité lexicale se rassemblant en un élément unique. Ainsi, lemmatisation et/ou regroupements peuvent rapprocher deux éléments (réponses, documents) n'ayant au départ aucune forme en commun lors de la phase de segmentation automatique du texte.

8.4.3 Discrimination et réponses modales

Le résultat de la numérisation et des étapes de transformations préliminaires conduit à une matrice **T**, d'ordre (*n,V*) analogue à celle qui a été définie au chapitre 5, dont les lignes correspondent aux *n* individus, documents ou textes, et les colonnes aux *V* différentes unités textuelles (formes, segments, lemmes, etc..). Les *K* groupes à discriminer définissent une partition des individus, documents ou textes. Ils peuvent être décrit par une matrice binaire **Z** d'ordre (*n, K*) comme celle qui a été définie pour caractériser une question fermée.

On a vu dans ce chapitre que, pour toute question fermée dont les réponses sont codées dans un tableau **Z**, il est possible de calculer le tableau lexical agrégé **C** par la formule :

$$\mathbf{C = T'Z}$$

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

Plusieurs outils permettent d'aider la lecture des tableaux lexicaux agrégés : l'analyse des correspondances, les techniques de classification qui lui sont complémentaires, les listes de formes caractéristiques, les listes de réponses modales.

Chacun de ces outils va permettre de construire une ou plusieurs techniques de discrimination. En particulier, la procédure de calcul des réponses modales est en elle-même une analyse discriminante.

Réponses modales utilisant les formes caractéristiques

Les spécificités ou formes caractéristiques (cf. chapitre 6) sont, rappelons-le, les formes "anormalement" fréquentes dans une partie du corpus ou dans les réponses d'un groupe d'individus (les formes anti-caractéristiques que l'on a

également appelées spécificités négatives sont au contraire anormalement rares dans ce groupe).

Une *réponse modale* pour une catégorie donnée a été définie comme une réponse d'un individu appartenant à cette catégorie, et contenant des formes qui caractérisent l'ensemble de la catégorie. Ainsi, une réponse contenant beaucoup de formes caractéristiques, et très peu de formes anti-caractéristiques, si elle existe, sera très caractéristique pour la catégorie.

On peut donc associer à toute réponse i du groupe g la somme $s_i(g)$ des valeurs-test relatives aux formes composant la réponse (les formes anti-caractéristiques étant affectées de valeurs-test négatives). Les m réponses modales correspondront aux m individus i associés aux plus grandes valeurs de $s_i(g)$. On peut considérer que $s_i(g)$ définit une *règle de discrimination*, puisque pour un individu donné i, la valeur de g pour laquelle $s_i(g)$ est maximale peut définir le groupe le plus probable d'appartenance de i.

Les réponses modales d'après un critère de distance

Rappelons le principe de ces sélections présenté au chapitre 6. Une réponse (ou un document, un texte) est un point-ligne i de **T**, donc un vecteur à V composantes.

On a appris à calculer des distances entre des réponses et les regroupements de ces réponses, puisque les réponses (lignes de **T**) et les regroupements de réponses (colonnes de **C** , ou lignes de **C**', transposée de **C**) sont tous représentés par des vecteurs d'un même espace. Toujours avec les notations définies au chapitre 6, la distance du chi-2 entre un point-ligne i de **T** et un point-colonne g de **C** est donnée par la formule:

$$d^2(i,g) = \sum_{j=1}^{p} \left(\frac{t_{..}}{t_{.j}}\right) \left(\frac{t_{ij}}{t_{i.}} - \frac{c_{jg}}{c_{.g}} \right)^2$$

Alors que la sélection des réponses caractéristiques se traduit par la recherche, pour un groupe g donné, l'individu i qui rend minimale la distance $d^2(i,g)$, la discrimination (une méthode possible de discrimination) consiste à chercher, de façon inverse, pour un individu i donné, la valeur de g qui rend minimale la quantité $d^2(i,g)$.

8.4.4 Discrimination régularisée par analyse des correspondances préalable

Les analyses des correspondances peuvent décrire les tableaux **C** qui sont des tables de contingence (dont les "individus" sont des occurrences de formes) ou encore les tableaux clairsemés **T**. Elles permettent de visualiser les associations entre formes et groupes ou modalités.

Elles donnent aussi la possibilité de remplacer des variables qualitatives (simple présence ou fréquence d'une forme) par des variables numériques (les valeurs des coordonnées factorielles), et donc d'appliquer les méthodes classiques d'analyse discriminante (linéaire ou quadratique) après changement de coordonnées.

L'analyse des correspondances peut donc servir de "pont" entre les données textuelles (qualitatives, et parfois clairsemées) et les méthodes usuelles d'analyse discriminante.

Mais surtout, comme cela sera illustré lors de l'exemple du paragraphe 8.5, le fait d'abandonner les derniers axes factoriels opère une sorte de filtrage de l'information (cf. : Wold, 1976 ; Benzécri, 1977) qui renforce le pouvoir de prédiction de l'ensemble de la procédure[1]. Ces propriétés sont utilisées en recherche documentaire (cf. Furnas et al., 1988 ; Deerwester et al., 1990).

C'est en ce sens que l'on parle de régularisation de l'analyse discriminante. Les techniques de régularisations (cf. par exemple Friedman,1989) visent en effet à rendre possible des discriminations dans des cas que les statisticiens décrivent comme "pauvrement posés" (à peine plus d'individus que de variables) ou mal posés (moins d'individus que de variables). De tels cas se présentent lorsque les échantillons d'individus sont petits, ou lorsque le nombre de variables est important, ce qui est souvent la situation de la statistique textuelle. Le fait de travailler sur un petit nombre de facteurs est alors un avantage décisif[2].

Dans les exemples qui vont suivre, on procédera à une analyse des correspondances préalable du tableau **T**, et l'on travaillera sur les p premiers axes factoriels. Le calcul du nombre d'axes principaux à retenir peut être déterminé de façon à optimiser les résultats de la discrimination (cf. Lebart,

[1] L'expérience relatée au paragraphe 7.4 du chapitre 7 illustre cette propriété de filtrage de la méthode : la plupart des traits structuraux sont restitués par les premiers axes principaux.

[2] Notons que les distances du chi-2 entre profils utilisées pour calculer les réponses modales deviennent des distances euclidiennes usuelles lorsqu'elles sont calculées à partir des coordonnées factorielles.

1992a), avec les précautions méthodologiques concernant la validité des résultats qui seront évoquées au paragraphe suivant.

8.4.5 Validation d'une discrimination

Les nombreuses possibilités de choix laissées aux utilisateurs (choix et élaboration des unités statistiques, choix de la technique de discrimination) ne doivent ni les rendre perplexes ni les décourager. Car la qualité d'une discrimination peut être évaluée avec des critères précis, plus simples que ceux qui interviennent dans le cas des méthodes exploratoires.

Pour établir la validité d'une discrimination, il est nécessaire de procéder à des contrôles sur une partie de l'échantillon (échantillon-test) qui n'a pas servi à établir la règle de discrimination. Cette règle est établie sur la partie de l'échantillon initial que l'on nomme *échantillon d'apprentissage*. Le critère le plus simple est le "pourcentage de bien classés", qui sera calculé à la fois pour l'échantillon d'apprentissage (ce qui donnera toujours une idée trop optimiste de la qualité de la discrimination) et pour l'échantillon-test.

Cette distinction entre sous-échantillons est fondamentale, et finalement assez intuitive si l'on décrit la phase d'apprentissage avec les termes des procédures d'apprentissage en intelligence artificielle.

La procédure de discrimination, en réalisant un maximum du pourcentage de bien classés sur l'échantillon d'apprentissage, va en fait utiliser toutes les particularités de cet échantillon, et donc des informations accidentelles (du bruit) pour arriver à ses fins... Dans certains cas, si le nombre de paramètres est grand, on peut atteindre un pourcentage de 100% d'individus bien classés sur l'échantillon d'apprentissage, avec un pourcentage très faible sur tout autre échantillon : c'est le phénomène de l'*apprentissage par coeur*. La procédure, pourrait-on dire, a appris sans comprendre, c'est-à-dire ici sans avoir saisi les traits structuraux généraux, sans avoir fait la part du structurel et du contingent dans l'échantillon d'apprentissage. On comprend que la vraie valeur d'une analyse discriminante ne se mesure que sur un, ou mieux sur une série d'échantillons test.

McLachlan (1992) fait une revue intéressante des méthodes de calcul d'erreur en analyse discriminante. Des techniques comme le Jackknife et le Bootstrap (cf. Efron (1982), pour une revue de ces deux familles de méthodes) permettent effectivement d'apprécier la confiance que l'on peut accorder à une procédure de discrimination, comme leurs variantes que sont les techniques de *Leave-one-out* (ou de *cross-validation*) (cf. Lachenbruch and Mickey, 1968 ; Stone, 1974 ; Geisser, 1975), qui consistent à construire n échantillons en supprimant à chaque fois un individu, et à appliquer la règle de

discrimination à cet individu (ces dernières techniques très coûteuses ne sont applicables qu'à de très petits échantillons).
Nous appliquerons lors de l'exemple du paragraphe 8.5 qui va suivre une technique de *leave-x%-out*, beaucoup plus économique dans le cas de grands échantillons, qui revient à tirer plusieurs échantillons test représentant chacun un pourcentage donné de l'échantillon total. Ceci nous permettra de représenter la qualité de la prédiction et les erreurs d'affectation par des *matrices de confusion* croisant les catégories initiales (réelles) avec les catégories construites à l'aide des fonctions discriminantes.

8.5 Discrimination globale et validation d'après un exemple

Dans quelle mesure peut-on prévoir l'appartenance d'un individu à une catégorie (une catégorie démographique, par exemple), à partir de ses réponses à une question ouverte ? Dans ce qui suit, on montrera comment éprouver le pouvoir de prédiction des profils lexicaux construits à partir des réponses aux questions ouvertes d'une enquête.

Dans l'exemple ci-dessous, les catégories démographiques, au nombre de six, sont obtenues par croisement des modalités de la variable genre (masculin, féminin) avec trois classes d'âge. Ces catégories sont communes à trois enquêtes, réalisées dans trois pays, et dans trois langues différentes pour ce qui concerne les questions ouvertes.

Les matrices de confusion calculées sur échantillons-test caractériseront, pour chacun des trois pays, non seulement la qualité de la prédiction, mais aussi la structure des erreurs de prédiction (quelles sont les catégories bien séparées, lesquelles sont confondues, etc.). Elles permettront donc de procéder à des comparaisons entre les trois pays, malgré le caractère très hétérogène (et par conséquent peu comparable a priori) de l'information de base.

8.5.1 L'exemple et le problème

Une enquête internationale[1] a eu lieu en 1989-1990 sur les comportements, habitudes et préférences alimentaires de trois grandes métropoles : Paris, New York et Tokyo (Cf. Akuto, 1992).

Dans chacune de ces trois villes, cette enquête a donné lieu à un millier d'entrevues environ (entretiens face à face).

[1] Cette enquête s'est déroulée à l'initiative de l'Institute of Research on Urban Life (Institut de Recherche Japonais dépendant de la Tokyo Gas Company Ltd) sous la direction du Pr.H. Akuto.

Le questionnaire était composé de nombreuses questions fermées (communes aux trois pays) décrivant de façon détaillée les caractéristiques socio-démographiques des répondants, le rythme, les horaires, les lieux et circonstances des repas ou collations. Le questionnaire comportait en outre deux questions ouvertes, également communes aux trois pays.

Le libellé de ces questions, dans la version française, était le suivant :

- *Quels sont les plats que vous aimez et mangez fréquemment?*
  avec une relance :
  Y-a-t-il d'autres plats que vous aimez et mangez fréquemment?

- *Pouvez-vous donner des exemples de ce que vous considérez comme un "repas idéal" ?* avec la relance :
  Y-a-t-il d'autres choses qui feraient partie de votre "repas idéal" ?

Dans chacune des enquêtes, on a pu constater que la structure syntaxique de la réponse fournie à cette même question variait fortement d'un répondant à l'autre. Si certaines des réponses sont constituées par simple énumération d'une liste d'items (noms de produits alimentaires ou de spécialités culinaires, en l'occurrence), d'autres se présentent sous forme de phrases relativement construites, motivant les choix retenus par des considérations qui peuvent toucher à la diététique ou même à l'art de vivre.

Pour chacun des "corpus de texte" recueillis dans les trois enquêtes, nous avons pu vérifier que les caractéristiques d'ensemble de la gamme des fréquences du vocabulaire ressemblaient fortement à celles que l'on observe lors du dépouillement des questions ouvertes que nous avons traitées antérieurement (présence de quelques formes très fréquentes, profusion des hapax, etc.).

On trouvera ci-dessous quelques exemples de réponses qui relèvent de chacun des types que nous venons de décrire.

Réponses de Paris

— poulet, légumes, viande en sauce / plats surtout équilibrés avec des ingrédients le plus possible naturels.

— boeuf, canard / beaucoup de vitamines moins de matière grasses et d'additifs artificiels.

— sole au vin / repas avec entrées, plats de résistance, dessert, avec des amis dans la bonne humeur.

Réponses de New York

— baked potatoes with veal chops, fried shrimp with yellow rice, steaks, lobster with curry rice and potatoes, french onion.
— collard greens, lamb chops, rice and chicken.

— pot roast,mashed potatoes, lamb chops, salads/cucumber and pepper salad, pot roast with mashed potatoes and gravy, apple pie.

Réponses (romanisées) de Tokyo

— NIMONO / EIYO NO BARANSU GA TORE, MITAME NI UTSUKUSHII KOTO.
— YASAISUPU / ESA DE NAKU SHOKUJI DE ARU KOTO, KAZOKU SOROTTE SHOKUJI O TORU, ANKA DE ARU KOTO.
— NIZAKANA, SARADA, NABERYORI, CHUKARYORI / ANZEN NA ZAIRYO O TSUKAI BARANSU NO YOI OISHIIMONO O KAZOKU YUJIN TO TANOSHIKU.

[— Pot-au-feu / équilibré au plan de la nutrition, beau à regarder.
— Soupe de légume / pas de nourriture préparée n'importe comment, des choses bon-marché.
— Poisson cuit, salade, bouillons, cuisine chinoise / C'est agréable de manger en famille et avec des amis, de bonnes choses équilibrées à base de produits naturels.]

Dans les premières analyses qui sont présentées ici, les réponses à ces deux questions ont été regroupées, et considérées comme une seule réponse. Ce regroupement permet d'avoir des réponses individuelles relativement riches, au prix d'une certaine perte d'homogénéité. Bien entendu, les trois langues seront traitées séparément. Les réponses en japonais ont été romanisées, pour assurer une compatibilité avec les logiciels disponibles.

Certaines des catégories de répondants, communes aux trois enquêtes, comme le genre (sexe) et l'âge, peuvent être caractérisées à l'intérieur de chaque ensemble par des profils lexicaux que l'on comparera ensuite.

Bien que repérés au départ dans des espaces différents, les points qui représentent ces catégories pourront être caractérisés de façon plus intrinsèque par leurs distances respectives, et par les configurations géométriques qui résument ces distances.

L'analyse discriminante permettra d'évaluer, pour une ville donnée, et donc dans notre cas pour une langue donnée, le pouvoir de prédiction de la réponse textuelle sur ces catégories socio-démographiques. Elle permet ainsi de répondre à la question : dans quelle mesure la réponse ouverte d'un individu nous apporte de l'information sur son genre et son âge ? Cette notion de pouvoir de prédiction peut être mesurée de façon intrinsèque, et donner lieu à des comparaisons entre pays.

Nous étudierons ici une variable composite comprenant six catégories (trois catégories d'âge, croisées avec le genre) qui sera notre "grille de passage" entre les trois enquêtes. Les six classes de cette variable sont les suivantes :

1	moins de 30 ans	hommes	4	de 30 à 50 ans	femmes
2	moins de 30 ans	femmes	5	plus de 50 ans	hommes
3	de 30 à 50 ans	hommes	6	plus de 50 ans	femmes

En somme, à partir d'individus inconnus et de mots inconnus, (au moins considérés comme tels dans une première phase), on peut espérer observer et comparer des configurations de catégories connues. On peut donc finalement conduire des comparaisons internationales, comme cet exemple va tenter de le montrer, sans qu'il soit toujours nécessaire de procéder à une traduction pendant certaines phases de traitement statistique.[1]

Sous-échantillon de Tokyo

Le texte total des 1 008 réponses contient 6 219 occurrences de 832 formes graphiques distinctes.

L'analyse utilisera les 139 formes apparaissant au moins 7 fois dans l'échantillon des 1 008 réponses aux deux questions ouvertes. Ces 139 formes donnent lieu à 4 975 occurrences, soit 80% du texte total.

Sous-échantillon de Paris

Le texte total des 1 000 réponses contient maintenant 11 108 occurrences de 1 229 formes distinctes.

Les 112 formes apparaissant au moins 18 fois donnent lieu à 7 806 occurrences, soit 70% du texte total.

Sous-échantillon de New York

Le sous-échantillon des personnes interrogées à New York est sensiblement plus réduit que les précédents (634 personnes). Le corpus des réponses libres contient 6 511 occurrences de 638 formes distinctes.

Sur ces 6 511 occurrences, 5 034 (soit 77%) concernent les 83 formes les plus fréquentes (apparaissant plus de 12 fois), qui serviront de base à notre étude.

8.5.2 Vocabulaire et analyse pour Tokyo

Seule, l'analyse relative au volet japonais de l'enquête sera exposée ici de façon détaillée. Le lecteur pourra se reporter aux travaux cités plus haut pour l'exposé des résultats correspondant aux deux autres enquêtes.

Le tableau 8.4 nous montre les 139 formes apparaissant au moins 7 fois dans l'échantillon des 1 008 réponses aux deux questions ouvertes.

On note la présence de mots-outil (conjonctions, prépositions comme TO, NO, NI, DE), de mots étrangers adoptés par la langue japonaise.

[1] Le détail des analyses relatives aux trois pays est donné dans Akuto (1992) et Akuto et Lebart (1992) ; un point de vue critique sur ces analyses est donné dans Benzécri (1992a).

Tableau 8.4

Réponses libres du sous-échantillon de Tokyo (texte romanisé) Les 139 formes graphiques les plus fréquentes

NUM.	MOTS EMPLOYES	FREQ.	NUM.	MOTS EMPLOYES	FREQ.	NUM.	MOTS EMPLOYES	FREQ.
1	AGEMONO	14	48	KOTO	128	95	SOBA	10
2	AJI	21	49	MENRUI	9	96	SOROTTE	24
3	ANKA	7	50	MISOSHIRU	26	97	SOZAI	20
4	ARI	15	51	MITAME	14	98	SUKI	9
5	ARU	40	52	MO	9	99	SUKINA	47
6	ATTA	11	53	MONO	209	100	SUKIYAKI	50
7	BARANSU	152	54	NABE	14	101	SUNOMONO	9
8	CHAHAN	7	55	NABEMONO	54	102	SUPAGETTEI	42
9	CHIRASHIZUSHI	7	56	NADO	7	103	SURU	24
10	CHUUKA	41	57	NAI	11	104	SUSHI	65
11	CHUUSHIN	9	58	NAKU	7	105	SUTEKI	13
12	DE	127	59	NANDEMO	15	106	TABERARERE BA	8
13	DEKIRU	8	60	NI	97	107	TABERARERU	26
14	DK	14	61	NIHON	44	108	TABERU	127
15	EIYOU	90	62	NIHONSHOKU	21	109	TABETAI	25
16	EIYOUKA	8	63	NIHONSOBA	7	110	TANOSHII	18
17	FUNIKI	10	64	NIKU	36	111	TANOSHIKU	37
18	FURAI	14	65	NIKURUI	10	112	TEMPURA	38
19	FURANSU	10	66	NIMONO	96	113	TENKABUTSU	8
20	GA	106	67	NITSUKE	9	114	TO	92
21	GOHAN	11	68	NIZAKANA	20	115	TOKI	12
22	GURATAN	20	69	NO	300	116	TOMONI	12
23	GYOUZA	15	70	O	223	117	TONKATSU	22
24	HAMBAGA	7	71	ODEN	15	118	TORE	15
25	HAMBAGU	42	72	OHITASHI	9	119	TORETA	73
26	IROIRONO	7	73	OISHII	95	120	TORETEIRU	8
27	ISSHONI	14	74	OISHIKU	75	121	TORI	8
28	ITAMEMONO	14	75	OOI	7	122	TORIIRETA	7
29	JIBUN	27	76	OOKU	16	123	TORU	23
30	JIKAN	12	77	PIZA	11	124	TSUKATTA	8
31	KAISEKI	8	78	RAISU	15	125	TSUKEMONO	14
32	KAKETE	8	79	RAMEN	27	126	UDON	15
33	KAKIFURAI	7	80	RYOURI	258	127	WA	17
34	KARAAGE	11	81	SAKANA	63	128	WASHOKU	87
35	KARADA	15	82	SAKANARUI	14	129	YAKINIKU	62
36	KARE	79	83	SAKE	13	130	YAKITORI	8
37	KARORI	11	84	SANDO	12	131	YAKIZAKANA	57
38	KATEI	14	85	SAPPARISHITA	8	132	YASAI	95
39	KATSU	7	86	SARADA	20	133	YASAIITAME	29
40	KATSUDON	7	87	SASHIMI	94	134	YASUKU	10
41	KAZOKU	64	88	SHABUSHABU	7	135	YOI	62
42	KENKOU	28	89	SHICHUU	21	136	YOKU	42
43	KENKOUTEKI	8	90	SHINSENNA	11	137	YUKKURI	17
44	KENKOUTEKINA	7	91	SHOKUHIN	8	138	YUUJIN	12
45	KI	11	92	SHOKUJI	184	139	ZAIRYOU	16
46	KISETSU	14	93	SHOKUSURU	7			
47	KONOMI	7	94	SHURUI	10			

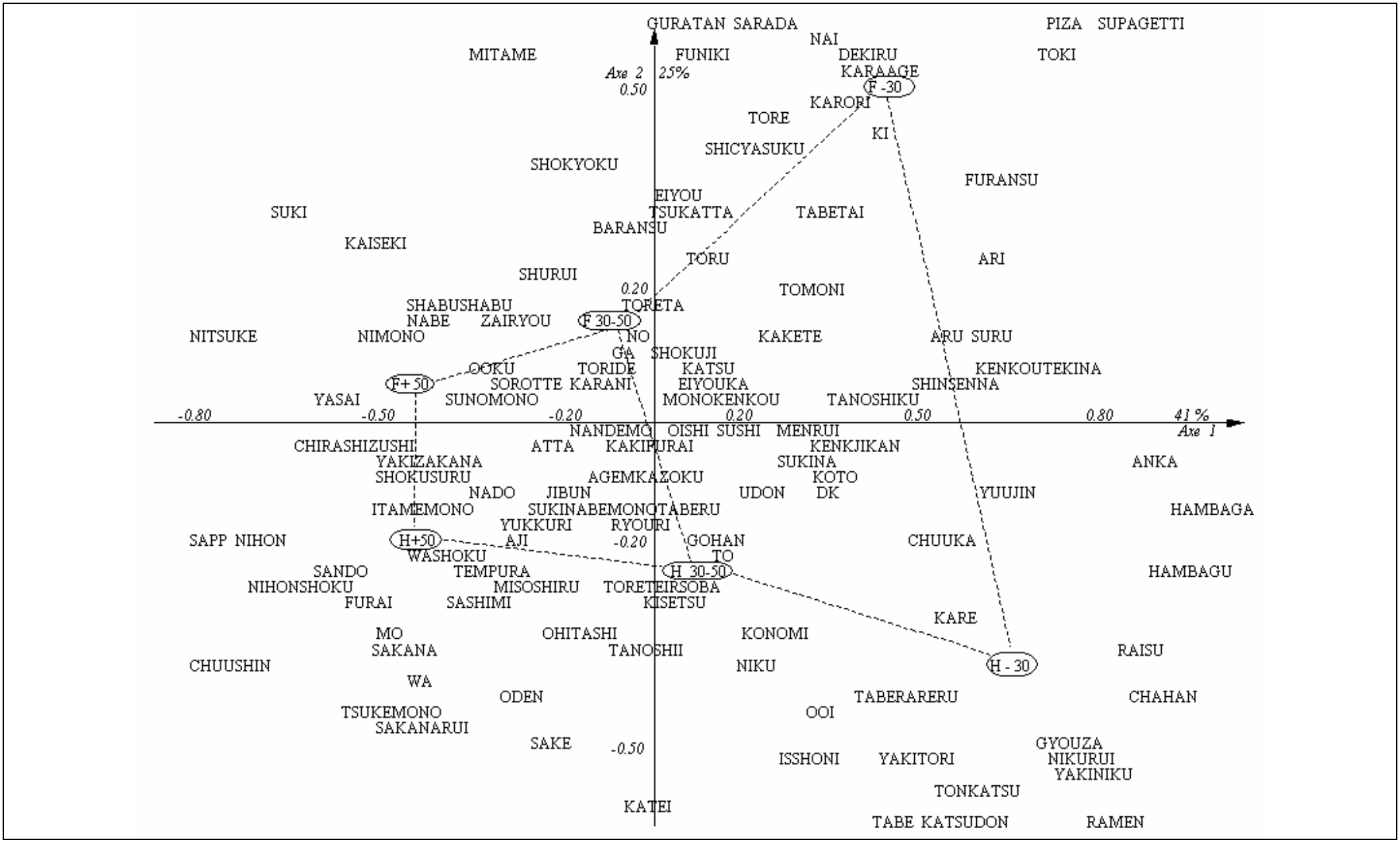


Figure 8.3 Analyse des correspondances de la table Formes- Catégories (Sexe-Age). Enquête alimentaire, Tokyo.

Citons, comme exemples de transcriptions : BARANSU (balanced : équilibré), GURATAN (gratin), HAMBAGU, HAMBAGA (variétés de hamburgers), KARE (curry), SUPAGETTEI, SUTEKI, PIZA (spaghetti, steak, pizza), RAISU (riz), SARADA (salade).

Apparaissent également quelques classiques de la cuisine japonaise, bien connus des occidentaux comme les SASHIMI, SUSHI, SUKIYAKI, YAKITORI, TEMPURA. D'autres formes seront commentées plus loin.[1]

Lecture de la figure 8.3

La figure 8.3 représente le premier plan factoriel issu de l'analyse des correspondances de la table de contingence à 139 lignes et 6 colonnes obtenues en regroupant les 1 008 profils en 6 classes correspondant à la variable composite à six modalités *âge-genre* évoquée plus haut.

On doit tout d'abord remarquer que les profils lexicaux permettent de reconstituer sans interversion les positions relatives des classes : l'âge augmente de la droite vers la gauche, les hommes occupent le bas du graphique, les femmes la partie supérieure. Le premier axe (41% de l'inertie) est donc un axe qui rend compte de l'âge, le second (25% de l'inertie), un axe qui sépare les hommes et les femmes.

Du côté des jeunes, et surtout des garçons, les réponses mentionnent les hamburgers, de la viande grillée, (YAKI = grillé, NIKU = viande, YAKITORI = poulet grillé), du riz, des nouilles (RAMEN). Pour les jeunes filles ou femmes, on peut noter PIZA, SUPPAGETTEI, FURANSU (français). D'une manière générale, les jeunes sont surtout ceux qui s'intéressent à la cuisine étrangère, les filles étant peut-être encore plus sensibles à la cuisine européenne que les garçons.

Du côté des classes les plus âgées, on trouve NIHON (japonais), et les diverses sortes ou préparations de poissons (SAKANA ou ZAKANA), avec mention de la KAISEKI (cuisine Kyotoïte raffinée).

Si les distances représentées dans ce plan factoriel sont une bonne représentation des distances dans l'espace initial à cinq dimensions (il y a six catégories), on constate de plus une convergence progressive des points relatifs aux deux genres pour une même classe d'âge lorsque l'âge augmente, ce qui donne une forme trapézoïdale à la configuration. Nous reviendrons sur cette constatation plus loin.

[1] La romanisation de l'écriture japonaise a introduit une confusion que permettaient de lever les Kanjis chargés de sens, ou même les accents. La forme graphique romanisée SAKE, par exemple, désignera aussi bien dans notre codage le vin de riz que le saumon. Ce type de mutilation de l'information de base n'enlève que peu de chose à la richesse des profils lexicaux agrégés, comme on va pouvoir en juger.

Les explications ne manquent pas à cette observation : traditionalisme, vie commune, et donc homogénéité des modes de vie des personnes plus âgées, ouverture à la fois culturelle (goûts éclectiques) et matérielle (non cohabitation ou cohabitation partielle, fréquentation des restaurants, invitations) des plus jeunes. Une telle ouverture favorise probablement une différentiation des goûts et des habitudes alimentaires.

Tableau 8.5

Sélection des formes caractéristiques (Tokyo, classes d'âge extrêmes).

LIBELLE DE LA FORME GRAPHIQUE	*POURCENTAGE INTERNE*	*GLOBAL*	*FREQUENCE INTERNE*	*GLOBALE*	*V.TEST*	*PROBA*
moins de 30 ans - hommes						
1 YAKINIKU	3.95	1.25	26.	62.	5.509	0.000
2 KARE	3.95	1.59	26.	79.	4.427	0.000
3 HAMBAGU	2.58	0.84	17.	42.	4.268	0.000
4 CHAHAN	0.91	0.14	6.	7.	3.991	0.000
5 RAMEN	1.82	0.54	12.	27.	3.812	0.000
6 RAISU	1.21	0.30	8.	15.	3.483	0.000
plus de 50 ans - hommes						
1 SASHIMI	4.00	1.89	23.	94.	3.421	0.000
2 SAKANA	2.78	1.27	16.	63.	2.938	0.002
3 CHUUSHIN	0.87	0.18	5.	9.	2.926	0.002
4 NIHON	2.09	0.88	12.	44.	2.720	0.003
5 NABEMONO	1.91	1.09	11.	54.	1.731	0.042
6 YASAI	2.96	1.91	17.	95.	1.720	0.043
moins de 30 ans - femmes						
1 SUPAGETTEI	3.57	0.84	24.	42.	6.536	0.000
2 PIZA	1.04	0.22	7.	11.	3.595	0.000
3 SARADA	1.34	0.40	9.	20.	3.239	0.001
4 GURATAN	1.04	0.40	7.	20.	2.237	0.013
5 EIYOU	2.97	1.81	20.	90.	2.160	0.015
6 TOKI	0.74	0.24	5.	12.	2.155	0.016
plus de 50 ans - femmes						
1 YASAI	4.19	1.91	38.	95.	4.910	0.000
2 NIMONO	3.64	1.93	33.	96.	3.712	0.000
3 WASHOKU	3.09	1.75	28.	87.	3.056	0.001
4 NITSUKE	0.66	0.18	6.	9.	2.905	0.002
5 NIHON	1.76	0.88	16.	44.	2.719	0.003
6 JIBUN	1.21	0.54	11.	27.	2.559	0.005

Reste cependant à vérifier la réalité du fait statistique, manifestée par cette forme trapézoïdale dont on observe peut-être une projection déformée. On va tout d'abord consulter les compléments usuels des analyses des correspondances présentés au chapitre 6 : les formes caractéristiques, qui permettent de chiffrer en termes probabilistes les liaisons formes-catégories,

et les réponses modales, qui permettent de situer ces formes dans leur contexte.

Lecture du tableau 8.5 (formes caractéristiques)

Le tableau 8.5 complète la figure 8.3 en faisant apparaître les formes caractéristiques de chaque catégorie. Pour lire ce tableau, prenons par exemple la forme YAKINIKU, qui apparaît caractéristique des hommes de moins de 30 ans.

On lit sur cette table que le pourcentage des occurrences de cette forme chez les hommes de moins de 30 ans (3,95%) est plus de trois fois supérieur à son pourcentage global (1,25%). L'avant dernière colonne donne à cet écart une valeur-test de 5,51. La dernière colonne nous fournit d'ailleurs les premiers chiffres de ce seuil.[1]

On observe que certaines formes, comme GURATAN, caractérisent les femmes des deux premières classes d'âge. En revanche, une forme comme YASAI (légumes) caractérise les personnes âgées des deux sexes, tout en étant beaucoup plus caractéristique des femmes.

Lecture du tableau 8.6 (réponses modales)

Le tableau 8.6 nous donne quelques réponses caractéristiques pour chacune des catégories étudiées. Rappelons que les réponses caractéristiques d'une classe sont des réponses originales sélectionnées comme étant les plus proches du centre de cette classe, (au sens de la distance du chi-2 entre le profil lexical de la réponse et le profil lexical moyen de la classe).[2]

Les réponses caractéristiques sont utiles car elles plongent les formes dans leur contexte réel, et remettent en mémoire la complexité de l'information de base. Ainsi, la forme SUPAGETTEI, très caractéristique des jeunes femmes, se rencontre parfois chez leurs homologues masculins. Si la réponse 3 est quand même typique des garçons, c'est qu'elle est lestée par ailleurs de formes très caractéristiques des hommes jeunes.

---

[1] Rappelons que la valeur-test convertit, pour plus de lisibilité, une probabilité critique en variable normale standardisée : la valeur 1.96 correspond ainsi au seuil 0.025, alors que la valeur 5.51 correspond à une probabilité de l'ordre de 10^{-6}.

[2] Un autre mode de calcul consiste à caractériser une réponse par la valeur-test moyenne des formes qu'elle contient, mais ce critère qui favorise trop les réponses lapidaires, n'a pas été utilisé ici (cf. chapitre 6, paragraphe 6.2).

Tableau 8.6

Sélection des réponses caractéristiques
(Echantillon de Tokyo, classes d'âges extrêmes)

moins de 30 ans - hommes

-- 1 YAKINIKU, KARE RAISU, RAMEN
-- 1 OISHII TO OMOTTE TABERARERU KOTO

-- 2 RAMEN,YAKINIKU,KARE
-- 2 OISHIKUTE RYOU NO ARU KOTO

-- 3 KARE RAISU, HAMBAGU, SUPAGETTEI
-- 3 NIKURUI, ABURAPPOI RYOURI, RAMEN NO YOUNA MONO

plus de 50 ans - hommes

-- 1 NABEMONO, MAGURO NO SASHIMI
-- 1 YASAI TO SAKANA GA ISSHONI NATTA RYOURI

-- 2 WASHOKU
-- 2 SAKANA, YASAI,TOUFU, NIMONO, JUNNIHONTEKINA MONO

-- 3 NABEMONO, TONKATSU, YAKIZAKANA
-- 3 YASAI SAKANA CHUUSHIN NO WASHOKU

moins de 30 ans - femmes

-- 1 KARE, GURATAN, PIZA,UNAGI, SUPAGETTEI
-- 1 EIYOU NO BARANSU TORETA SHOKUJI

-- 2 GURATAN, PIZA, SUPAGETTEI, CHIRASHI, KARE, HAMBAGA,
-- 2 SARADA, OISHIKU EIYOU GA ARI MAINICHI HENKA NI
-- 2 TONDA KONDATE NO SHOKUJI

-- 3 SUPAGETTEI
-- 3 BARANSU TORETA SHOKUJI

plus de 50 ans - femmes

-- 1 YASAI NO NIMONO
-- 1 WASHOKU

-- 2 YASAI NO NIMONO, SASHIMI, NIZAKANA, YAKIZAKANA
-- 2 JIBUN NI ATTA SUKINA MONO O TABERU KOTO

-- 3 YASAI NO NIMONO, KORUDOBIFU, SUTEKI
-- 3 JIBUN DE RYOURISHITA MONO O OISHIKU AJIWAU KOTO GA
-- 3 DEKIREBA RISOU DE ARU

8.5.3 Réalité des configurations

Les analyses opérées sur Paris et New York donnent des *patterns* analogues à ceux observés pour la Figure 8.3. Pour Tokyo et New York, on observe un éloignement marqué des deux catégories les plus jeunes : les garçons et les filles de moins de 30 ans ont, semble-t-il des réponses très différentes. On n'observe pas cette divergence pour Paris (Figure 8.4).

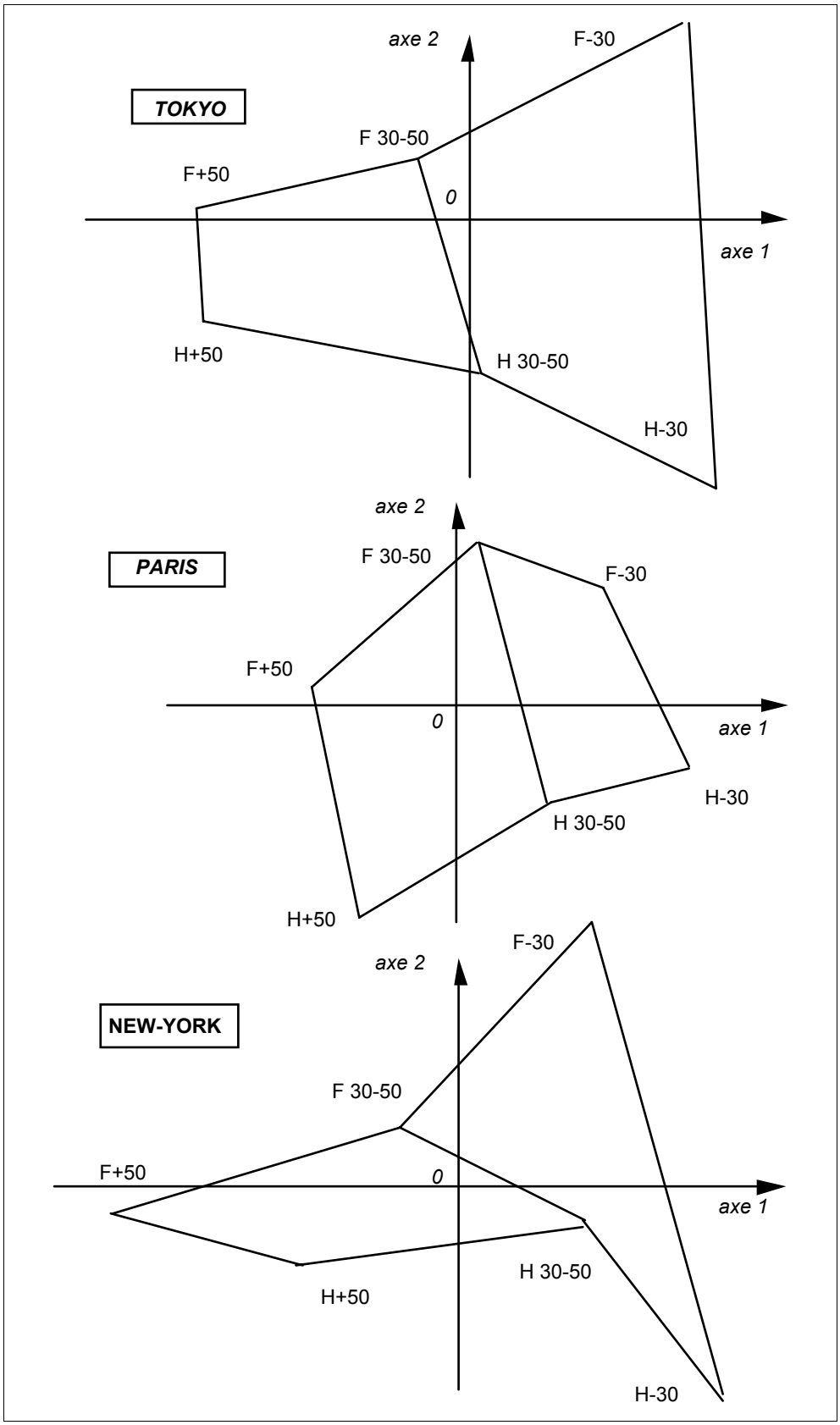


Figure 8.4.

Comparaison des patterns (genre - âge) des plans (1,2) pour les trois villes

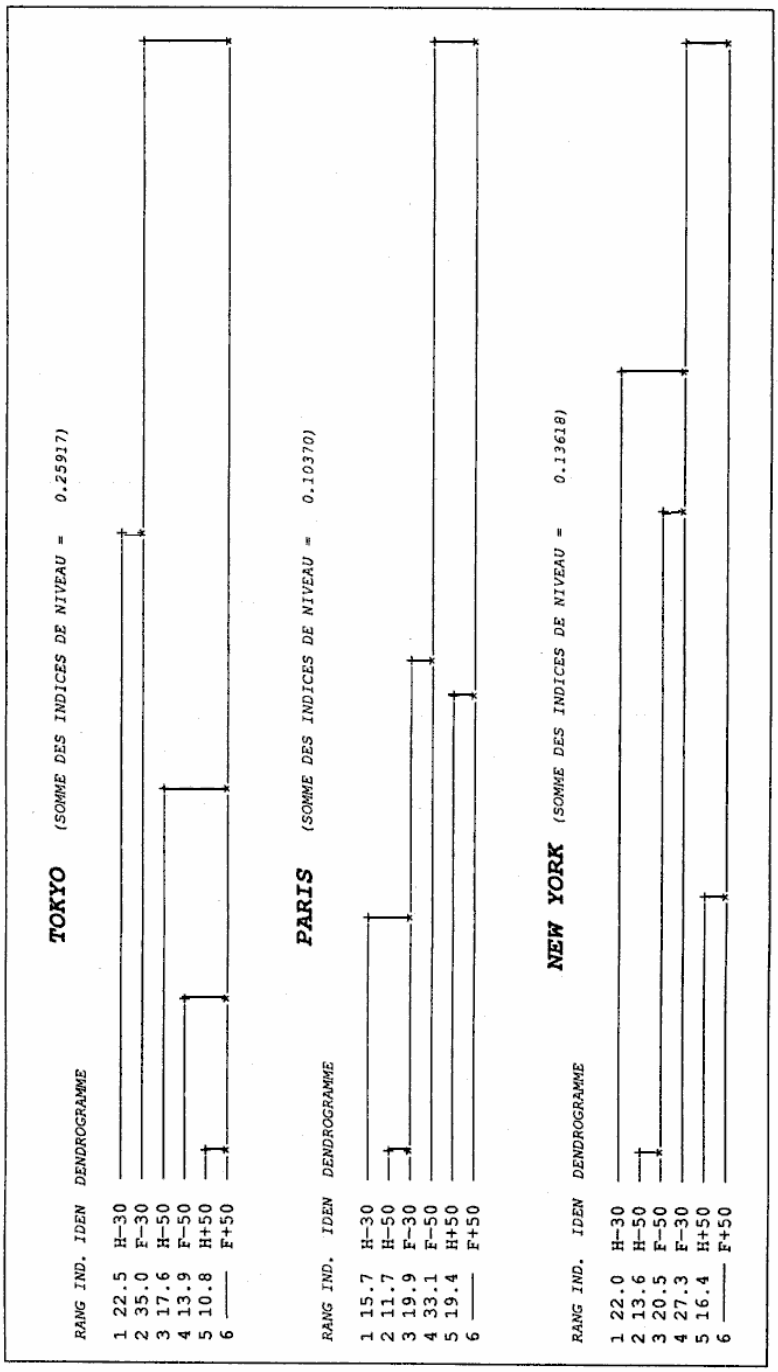


Figure 8.5
Comparaison des trois dendrogrammes (Tokyo, Paris, New York)
(Indices exprimés en pourcentages de la somme des indices)

La Figure 8.4 schématise, en effet, les trois "patterns" observés : notons que dans les trois cas, les âges s'échelonnent sur l'axe 1, et les genres se séparent le long de l'axe 2.

Quelle est donc la réalité des configurations observées ? Quelles sont les limites des interprétations entre différences de configurations ?

Une des façons de répondre à ces questions est de décrire par classification hiérarchique (sur les 5 coordonnées factorielles, en utilisant le critère dit de Ward, cf. chapitre 4) les distances entre points catégories. La Figure 8.5 nous montre ainsi les trois dendrogrammes de Tokyo, Paris et New York. Les indices correspondant à chaque noeud sont exprimés en pourcentages de la somme des indices.

On note par exemple que, pour Tokyo, la *convergence entre les genres à mesure que l'âge augmente* constatée sur l'analyse factorielle (figure 8.3), est bien visible sur la figure 8.5 : les plus âgés s'agrègent pour un niveau bas de l'indice (11% de la somme), et les plus jeunes à un niveau plus élevé (22.5%).

C'est effectivement à Tokyo que les personnes les plus âgées ont les comportements les plus homogènes (niveau 11%, contre 16% à New York et 19% à Paris) Une autre particularité de Tokyo concerne le rattachement tardif des jeunes au reste de la population : la coupure la plus nette est à 30 ans dans cette ville, alors qu'elle est autour de 50 ans dans les deux autres.

A Paris et New York, le rattachement des personnes les plus âgées au reste de la population se fait au plus haut niveau du dendrogramme.

On remarque que les femmes d'une classe d'âge ne se rattachent jamais sur ces dendrogrammes à des hommes d'une classe d'âge inférieure :

- A Tokyo, le point F-50 rejoint les points H+50 et F+50 avant que ne le fasse le point H-50.
- A Paris, le point F-30 s'agrège très bas au point H-50.
- A New York, le point F-30 s'agrège à la classe -50 (H et F) avant de s'agréger au point H-30.

On peut en somme conclure que les préférences alimentaires des femmes, dans les trois pays, les rapprochent plutôt des hommes d'une classe d'âge supérieure (du point de vue du contenu des réponses, à partir des formes caractéristiques et des réponses modales : moindre intérêt pour les viandes, intérêt plus marqué pour les légumes et le poisson). La classification, fondée ici sur des distances dans l'espace à cinq dimensions, a également confirmé que les positions relatives des classes (figure 8.4), notamment les divergences

à l'intérieur des classes de jeunes pour New York et Tokyo, n'étaient pas des artefacts liés à la projection.

8.5.4 Analyse discriminante et matrices de confusion

Les unités choisies relèvent des options a) et e) de la section 8.4.2 : La discrimination se fait dans l'espace des formes graphiques seules après sélection suivant un critère de fréquence, et après régularisation par analyse des correspondances préalable.

Régularisation par les axes principaux

On illustrera l'effet de filtrage des sous-espaces issus de l'analyse des correspondances par la figure 8.6, qui représente les taux de succès (donc de bien classés) d'une série d'analyses discriminantes en fonction du nombre d'axes factoriels (ou principaux) retenus (nombre porté sur l'axe des abscisses de la figure 8.6) et aussi en fonction de l'importance relative des échantillons test.

Cette discrimination concerne le sous-échantillon de New York, les variables de prédiction sont les 83 formes graphiques, et la variable à prédire est l'âge en trois classes.

- Première constatation : Pour les échantillons d'apprentissage (symboles noirs), les taux de succès augmentent continûment avec le nombre d'axes.

- Seconde constatation, pour les échantillons test (symboles blancs) qui mesurent le vrai pouvoir discriminant en situation réelle, on observe une saturation du taux de succès aux alentours de 40 axes. Ainsi, à partir d'une certaine dimension, les axes supplémentaires n'améliorent plus la vraie discrimination, mais améliorent quand même les taux de succès (illusoire, parce qu'optimiste) de l'échantillon d'apprentissage. C'est le phénomène qualifié en Intelligence Artificielle et en réseaux neuronaux d'*apprentissage par coeur* , qui fonctionne à l'intérieur de l'apprentissage, mais qui ne permet pas d'inférences extérieures.

- Troisième constatation : Le taux de succès sur l'échantillon test est meilleur si celui-ci ne constitue qu'un petit prélèvement (toutefois, la différence est petite entre les taux de 30% (symboles carrés) et 10% (symboles losanges). Ce dernier point est intuitif, car l'apprentissage ne peut qu'être meilleur si l'échantillon est important.

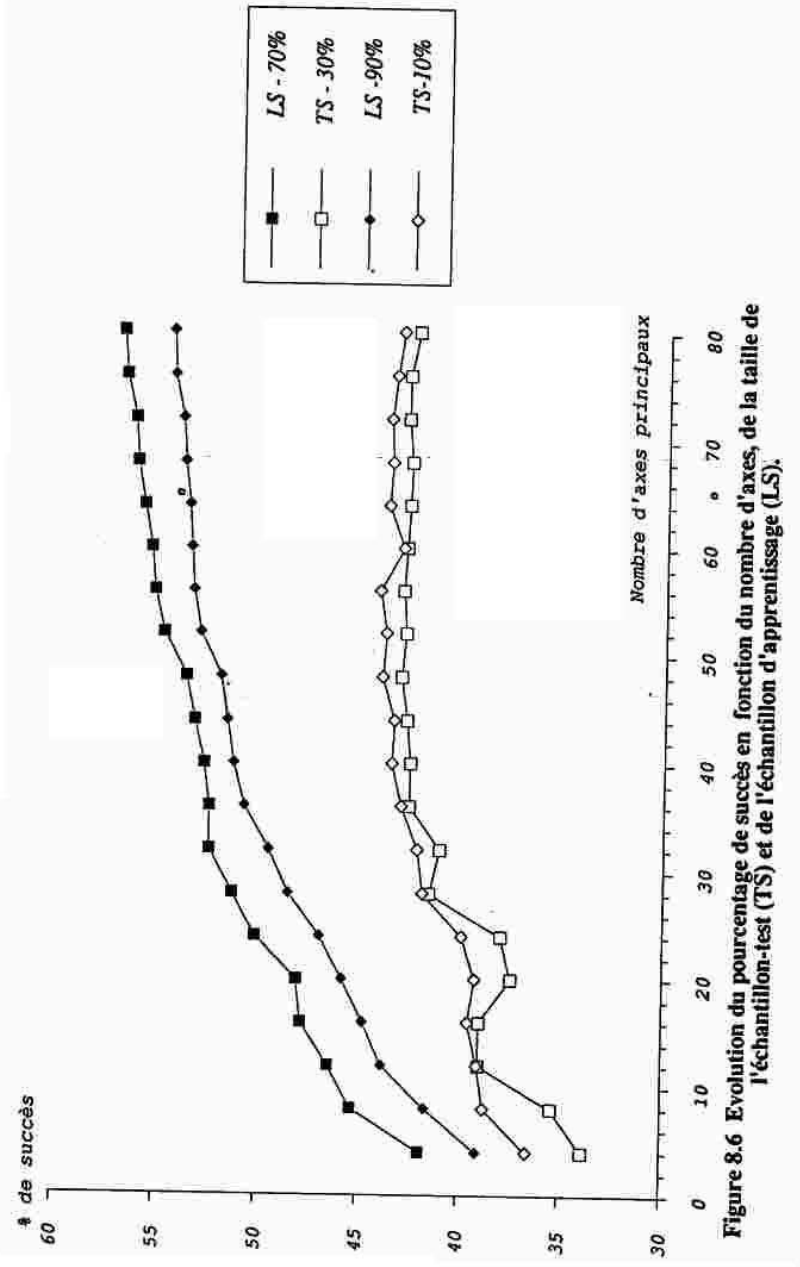


Figure 8.6 Evolution du pourcentage de succès en fonction du nombre d'axes, de la taille de l'échantillon-test (TS) et de l'échantillon d'apprentissage (LS).

En revanche, si l'échantillon d'apprentissage est réduit, les taux de succès sur cet échantillon sont alors trop optimistes, et donc trompeurs (cas des carrés noirs).

Ces trois constatations nous ont conduit à travailler sur un nombre d'axes réduits (36 pour l'exemple de Tokyo), avec un prélèvement d'échantillon- test de 10%.

Tableau 8.7

Exemples de Matrices de confusion (Tokyo)
Nombre d'axes principaux : 36.
(Echantillon-test : 10% de chacun des 30 ech.)

a - Echantillon d'APPRENTISSAGE
Catégories initiales [lignes] et catégories calculées[colonnes]

	H-30	H-50	H+50	F-30	F-50	F+50	Taux de succès
H-30	1469	541	244	506	694	388	38.24 %
H-50	596	1457	620	676	1535	808	25.60 %
H+50	93	361	999	229	1097	833	27.66 %
F-30	377	222	76	1342	835	316	42.36 %
F-50	361	538	486	1146	2610	1023	42.34 %
F+50	268	492	635	391	1195	1841	38.18 %

Taux de succès moyen : 35.60 %

b - Echantillon TEST
Catégories initiales [lignes] et catégories calculées[colonnes]

	H-30	H-50	H+50	F-30	F-50	F+50	Taux de succès
H-30	125	71	20	59	95	48	29.90 %
H-50	61	117	79	71	175	105	19.24 %
H+50	16	51	79	37	115	110	19.36 %
F-30	51	31	6	118	98	38	34.50 %
F-50	54	67	60	143	237	115	35.06 %
F+50	26	63	74	49	125	151	30.94 %

Taux de succès moyen : 28.13 %

Les matrices de confusion

Trente échantillons-test d'environ 100 individus ont été tirés au hasard sans remise dans l'ensemble des 1 008 individus.

Pour chacun de ces 30 tirages, les 6 centres de gravité (points moyens) des catégories *âge-genre* sont calculés dans l'échantillon d'apprentissage (sur les 1 008 - 100 = 908 individus) dans l'espace à 36 dimensions des 36 premiers facteurs de l'analyse des correspondances du tableau **T** (1 008 lignes, 139 colonnes).

Le tableau 8.7 présente la matrice de confusion moyenne relative à l'échantillon de Tokyo. Les individus des échantillons d'apprentissage (tableau 8.7-a), puis ceux de l'échantillon-test (tableau 8.7-b) sont alloués aux centres les plus proches.

La description des matrices de confusion relatives aux trois pays par analyse des correspondances permet de procéder à une comparaison de structures situées dans des espaces différents (Lebart, 1992b).

Le tableau 8.8 nous montre les taux de succès calculés de façons analogues pour les trois pays. On voit ainsi que la discrimination est bien meilleure à Tokyo que dans les autres villes, ce qui n'était pas visible sur les arbres hiérarchiques. En effet, les échelles des indices d'agrégation ne sont pas comparables car au départ ni les vocabulaires utilisés, ni les échantillons n'ont la même taille.

Tableau 8.8

Comparaison des taux de succès pour les trois villes

	Ech. d'apprentissage			Ech. Test		
	TOKYO	*PARIS*	*NEW YORK*	*TOKYO*	*PARIS*	*NEW YORK*
H-30	38.24	20.77	43.45	29.90	14.44	33.00
H-50	25.60	13.16	26.23	19.24	6.89	20.05
H+50	27.66	31.50	17.08	19.36	23.13	10.14
F-30	42.36	29.45	38.00	34.50	24.59	30.48
F-50	42.34	29.55	15.78	35.06	25.28	12.17
F+50	38.18	30.26	42.00	30.94	29.41	37.11
Total	35.60	26.23	28.60	28.13	21.39	22.64

La figure 8.7 donne une visualisation des points lignes des deux matrices de confusion (échantillon d'apprentissage actif en traits pleins, échantillon-test supplémentaire en traits pointillés) relatives à l'échantillon de Tokyo.

La structure correspondant à l'échantillon-test est contractée (donc moins forte) mais reproduit fidèlement celle de l'échantillon d'apprentissage.

C'est une confirmation de la réalité de la structure observée sur la figure 8.3, avec cet élément d'information supplémentaire : la *confusion* semble très importante entre les hommes et les femmes de plus de 50 ans.

Ce que l'on interprète dans les termes suivants : rien, dans leurs réponses à la question ouverte, ne permet de distinguer ces deux catégories, alors que la distinction entre hommes et femmes était aisée pour les deux autres classes d'âge.

Une matrice de confusion sur échantillon-test représente bien le pouvoir prédicteur réel du texte sur les catégories. Etant donné que cette représentation a un caractère intrinsèque, elle peut permettre des comparaisons entre pays différents et entre réponses exprimées dans des langues différentes.

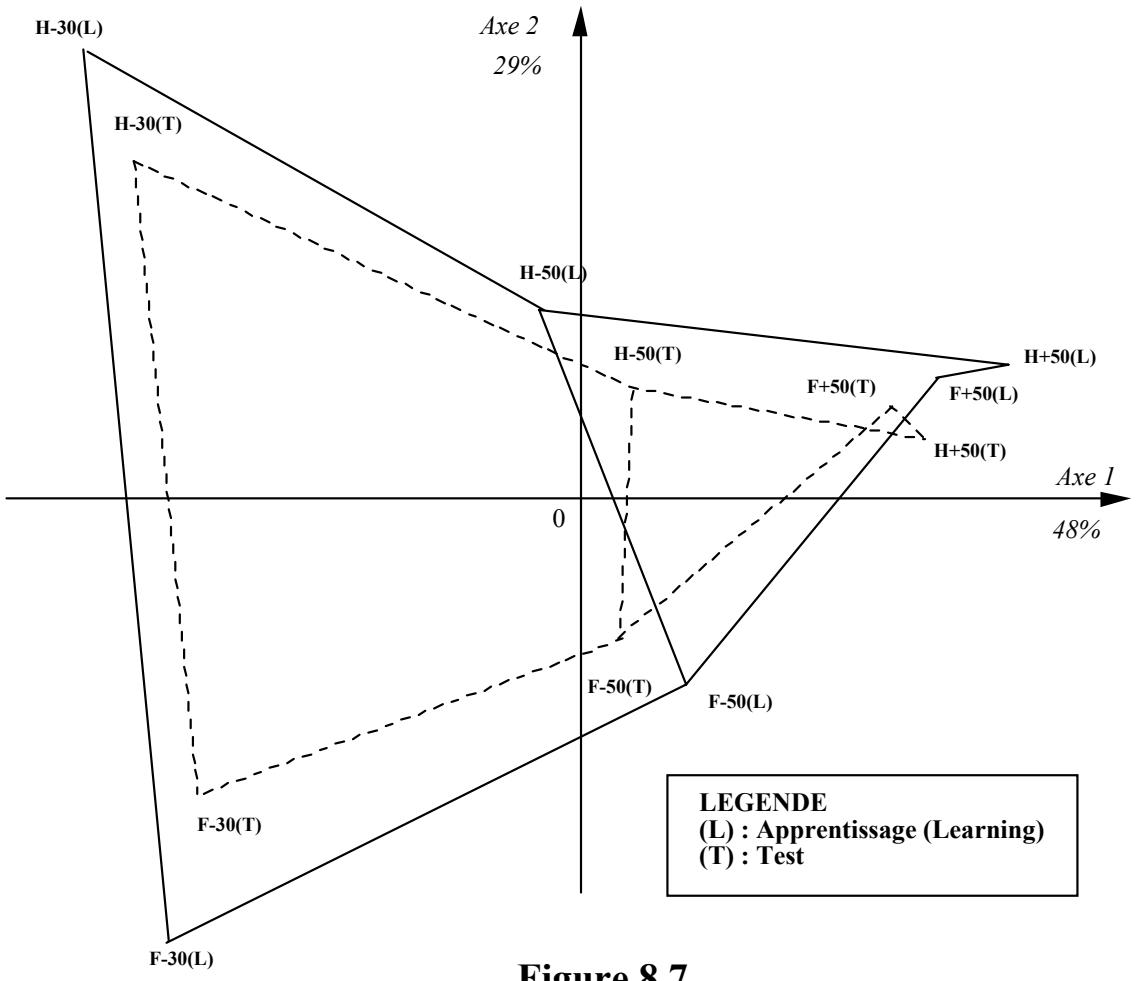


Figure 8.7

Analyse des correspondances des deux matrices de confusion relatives à l'échantillon de Tokyo.

On peut en effet comparer ces matrices de confusion, et donc les pouvoirs prédicteurs de différents textes sur une même variable. C'est dans cet esprit que la figure 8.8, qui n'est complexe qu'en apparence, représente la synthèse des trois matrices de confusion (sur échantillons-tests) relatives aux trois villes, obtenues en soumettant à l'analyse des correspondances leur juxtaposition en ligne. Cette juxtaposition comprend trois matrices analogues de la matrice de confusion du tableau 8.7b qui ne concernait que Tokyo (il s'agit d'échantillons-test, c'est à dire de pouvoir de prédiction réel). Le tableau décrit a donc 18 lignes (six pour chaque ville) et six colonnes.

Lecture de la figure 8.8

La grille en traits gras joignant les points encadrés représente les *catégories affectées* (colonnes de la juxtaposition de matrices) alors que l'on distingue les trois grilles de *catégories réelles* : New York en bas, Paris en pointillé au milieu, et Tokyo en haut (lignes de la juxtaposition de matrices). Remarquons que pour chacune des grilles, les jeunes sont à droite, les plus âgés à gauche, les femmes vers le haut, et les hommes vers le bas.

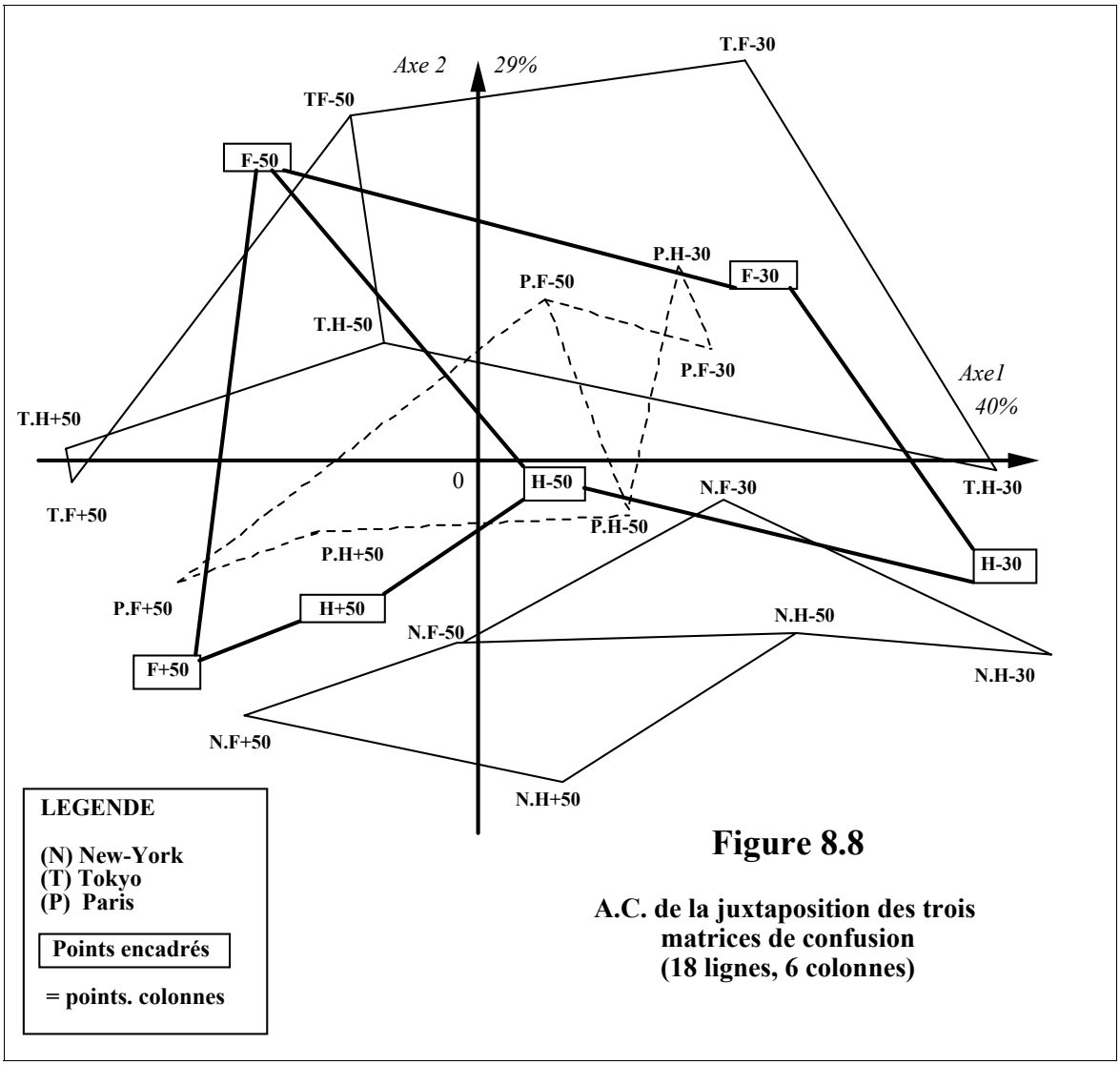


Figure 8.8

A.C. de la juxtaposition des trois matrices de confusion (18 lignes, 6 colonnes)

En regardant la position d'un point-catégorie réel par rapport à la grille en traits gras, on a une idée de la façon dont l'analyse discriminante répartit les individus correspondants. Ainsi, à gauche, les trois points F+50 correspondant aux trois villes (N.F+50, P.F+50, T.F+50) sont plus proches du point encadré F+50 que des autres points encadrés, ce qui veut dire que la catégorie des femmes de plus de 50 ans est bien reconnue à partir des réponses libres sur l'alimentation dans les trois villes (et les trois langues). De

même, à droite, les jeunes hommes de New York et Tokyo sont bien identifiés par leurs discours (points T.H-30, N.H-30), cela est moins vrai pour Paris (P.H-30), pour lequel les jeunes hommes sont beaucoup moins différenciés des jeunes femmes et de leurs aînés par leurs réponses libres. La grille en pointillé est en effet recroquevillée dans sa partie droite.

Le premier axe, lié à l'âge, ne permet pas de distinguer les trois villes (avec l'exception déjà signalée des jeunes parisiens des deux sexes).

En revanche, le second axe, lié au genre, sépare très nettement Tokyo et New York. Ce sont essentiellement les femmes de 30-50 ans (points F-50) et les femmes de moins de 30 ans (points F-30) qui sont responsables de cet état de choses. Les femmes de la catégorie d'âge intermédiaire de Tokyo sont en effet beaucoup mieux identifiées à partir de leurs réponses libres que celles de Paris, et surtout que celles de New York. Notons que ce résultat était déjà visible sur le tableau 8.8 donnant les pourcentages d'individus bien-classés par ville et par catégories.

8.5.5 Conclusions du paragraphe 8.5

La possibilité de comparer (assez facilement, au moins dans une première approche) des comportements à partir de textes dans des langues différentes ouvre évidemment des directions de recherche intéressantes.

Les phases de traduction et de codage de l'ensemble des réponses n'ont pas été nécessaires pour procéder à des descriptions et des inductions confirmées par des procédures de validation puissantes.

Mais une autre caractéristique de cet exemple est que, dans cette première phase exploratoire, l'on n'aurait su que faire des traductions, car beaucoup des variables de bases exprimant des habitudes alimentaires spécifiques n'ont pas d'équivalent d'un pays à l'autre.

Bien entendu, en montrant l'étendue des possibilités d'un traitement aussi formel de l'information de base, on pense mieux préparer les phases ultérieures de travail, et non les remplacer.

Annexe A

Description sommaire de quatre logiciels utilisant la statistique textuelle multidimensionnelle

De nombreux logiciels permettent désormais de réaliser les opérations de segmentation, de comptage et de documentation (index, concordances, sélection de contextes) que l'on a définies au chapitre 2. Dans un proche avenir, on peut prévoir que leurs fonctionnalités communes seront de plus en plus intégrées au sein des logiciels de traitement de texte les plus courants, comme c'est déjà le cas pour le comptage des occurrences et la recherche des contextes d'une même forme graphique ou chaîne de caractères.

On ne tentera pas ici de dresser un inventaire des possibilités et des limites de chacun de ces logiciels ni de comparer leurs performances respectives. On se limitera au contraire à un survol rapide des seuls logiciels qui font appel à l'analyse statistique multidimensionnelle des données textuelles, objet du présent ouvrage.

Les deux premières sections seront consacrées à la description des deux logiciels qui ont servi à produire l'ensemble des analyses dont rendent compte les différents chapitres de ce livre : *SPAD.T* (section A.1), plus particulièrement orienté vers le traitement des réponses à des questions ouvertes fournies par des individus nombreux sur lesquels on possède par ailleurs des renseignements de type socio-économique, et *Lexico1* (section A.2) conçu pour le traitement lexicométrique de textes, moins nombreux, mais comportant chacun plusieurs milliers, voire centaines de milliers d'occurrences.
Les deux sections suivantes décrivent succinctement deux autres logiciels (également consacrés à l'analyse multidimensionnelle des données textuelles) auxquels on s'est référé dans le présent ouvrage. La section A.3 présente le logiciel *ALCESTE*, qui permet de mettre en oeuvre des méthodes voisines de celles que nous avons décrites dans le présent ouvrage, pour un contexte d'applications assez différent.
La section A.4 présente le logiciel *HYPERBASE.* Ce logiciel diffère peu, au plan proprement statistique, de ceux qui sont présentés dans les deux premières sections. Son originalité réside avant tout dans le recours aux méthodes de ce que l'on appelle l'*hypertexte*, qui permettent de *naviguer* aisément dans de grandes masses de textes.
Enfin, la section A.5 donnera les références de quelques autres logiciels, en général moins centrés sur la statistique textuelle.

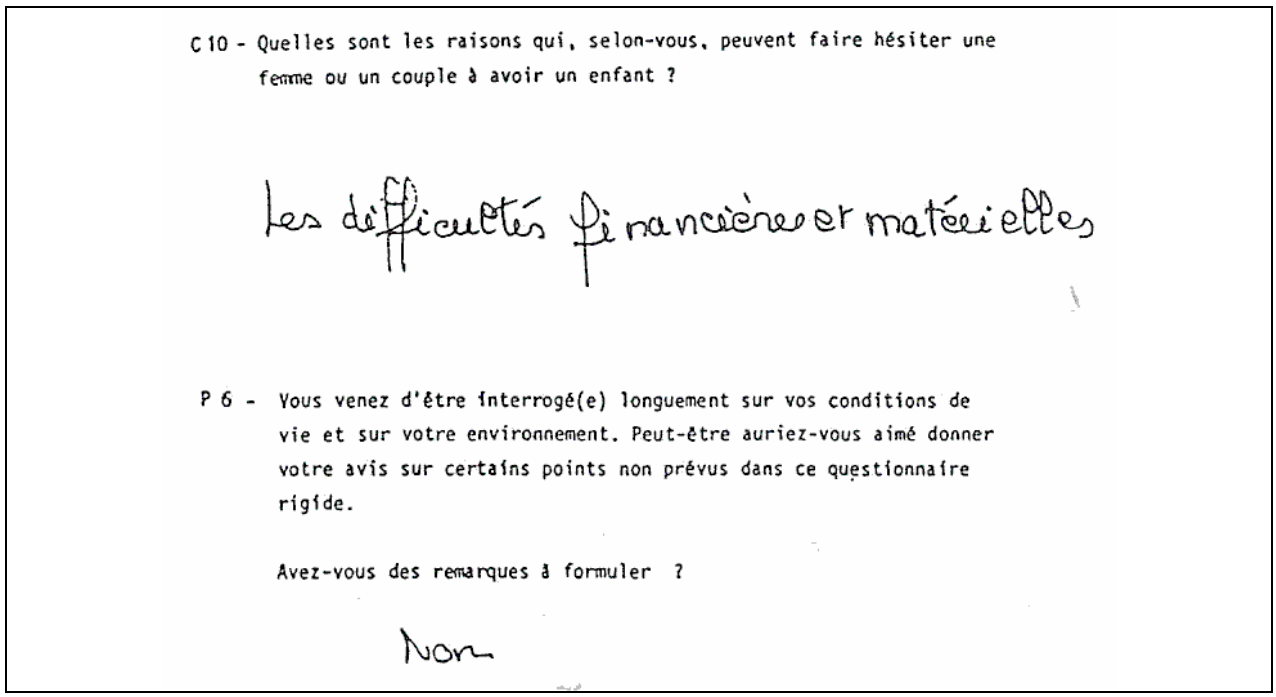
A.1 Le logiciel SPAD.T

On présente ici l'enchaînement des tâches, depuis la saisie de l'information, jusqu'à la production des listages et graphiques par SPAD.T.[1] Bien entendu, il ne s'agit que d'une illustration, et non d'un guide d'utilisation complet du logiciel.

a) Le document de base

La figure A-1 est un fac-similé des libellés de deux questions ouvertes du questionnaire de l'enquête précitée sur les conditions de vie et aspirations des Français, avec les réponses telles qu'elles ont été transcrites par un enquêteur.

Comme le montrent les numéros de ces questions, celles-ci ne sont pas consécutives dans le questionnaire. Elles sont en fait séparées par des questions fermées, qui concernent des thèmes distincts.


C 10 - Quelles sont les raisons qui, selon-vous, peuvent faire hésiter une femme ou un couple à avoir un enfant ?

les difficultés financières et matérielles

P 6 - Vous venez d'être interrogé(e) longuement sur vos conditions de vie et sur votre environnement. Peut-être auriez-vous aimé donner votre avis sur certains points non prévus dans ce questionnaire rigide.

Avez-vous des remarques à formuler ?

Non

Figure A1.1. Fac-similé de réponses à deux questions ouvertes

On reconnaît en la question C10, la question ***Enfants*** qui a illustré plusieurs exemples des chapitres 2 à 6. Les réponses à la question P6 ont, quant à elles, illustré la partition en noyaux factuels des chapitres 5 et 6.

b) Les fichiers de base

La figure A1.2 reproduit le listage du fichier de saisie des réponses à ces quatre questions pour les quatre premiers individus enquêtés. On reconnaît, pour l'individu 1 les réponses de la figure A.1.

[1] SPAD.T réalisé au départ dans un cadre universitaire, fonctionne sur PC et MacIntosh. Il est actuellement maintenu et distribué par le CISIA (1 avenue Herbillon, 94160, Saint-Mandé).

Le format de saisie est donc simple : le séparateur "----" annonce un nouvel individu, alors que le séparateur "++++" indique la fin d'une réponse, "====" indiquant la fin du fichier. Deux séparateurs de réponses consécutifs indiquent une réponse vide. Pour un individu donné, les réponses doivent toujours figurer dans le même ordre.

Dans le cas d'une seule question, il est possible d'introduire un séparateur "****" , dit séparateur de texte, qui permet de séparer des groupes d'individus consécutifs. Ainsi, les individus peuvent être des strophes et les textes des poèmes, les individus peuvent être des phrases et les textes des articles de journaux, etc...

```
----1
les difficultés financières et matérielles
++++
non
----2
l'avenir incertain,les problèmes financiers
++++
je trouve ce questionnaire intéressant
----3
les difficultés financières
++++
c'est un peu long
----4
les raisons matérielles et l'avenir qui les attend
++++
toutes les réponses ne sont pas formulées,on est obligé de choisir entre des
réponses qui ne correspondent pas toujours à ce qu'on pense
====
```

Figure A1.2. Listages du fichier textuel de base

SPAD.T suppose de plus qu'il existe un tableau numérique apparié au précédent décrivant, pour les mêmes individus, disposés dans le même ordre, leurs réponses à des questions fermées codées sous forme de variables nominales (variables dont les modalités s'excluent mutuellement). Ce fichier doit avoir dans la version actuelle, la forme d'un fichier standard SPAD.N, (logiciel de dépouillement d'enquête distribué par le CISIA) ou d'un fichier interfacé avec SPAD.N.

C'est la présence simultanée de ces deux fichiers qui permet de regrouper les réponses à une question ouverte selon les modalités de réponse à une question fermée, de positionner des formes graphiques dans des espaces factoriels calculés à partir de variables nominales, et de façon symétrique, de positionner des caractéristiques des individus dans les espaces calculés à partir des formes graphiques figurant dans leurs réponses libres.

c) Principes du logiciel

Ce logiciel admet donc comme entrée un fichier de données textuelles du type de celui de la figure A1.2 (fichier *texte*). Dans le cas (le plus fréquent) où il existe des variables nominales, il admet également des fichiers du type SPAD.N (fichier *données* et fichier *dictionnaire*).

Le logiciel comporte 15 procédures de base, qui seront désignées chacune par un mot-clé, suivi des valeurs d'un ou plusieurs paramètres. Divers enchaînements de ces procédures permettent de réaliser l'essentiel des traitements décrits précédemment.

Les procédures peuvent être réalisées une par une, car elles ne communiquent entre elles que par des fichiers externes qui peuvent être sauvegardés individuellement.

d) Les principales procédures

ARTEX Archivage du texte. Cette procédure lit le fichier de base (type figure A1.2) et le structure de façon à le rendre facilement accessible pour les procédures suivantes.

SELOX Sélection de la question à traiter dans le fichier archive, avec éventuellement filtre sur les individus.

NUMER Numérisation du texte de la question choisie. Les tableaux 5.1 et 5.2 du chapitre 5 font partie des sorties de cette procédure.

CORTE Cette procédure permet de supprimer des formes (par exemple : des mots-outil), et de déclarer équivalentes des listes de formes (par exemle : relatives à un même lemme)

MOTEX Construction d'un tableau lexical (après NUMER) ou segmental (après SEGME) pour une variable nominale à préciser.

APLUM Analyse des correspondances d'un tableau lexical issu de MOTEX, avec graphiques. Les figure 5-1, 8.3 sont issues des sorties de cette procédure.

MOCAR Calcul et impression des formes caractéristiques pour chacune des modalités de la variable nominale désignée dans MOTEX. Eventuellement, calcul des réponses modales pour le critère de la fréquence lexicale. Le tableau 6.6 est une des sorties de MOCAR.

RECAR Calcul et impression de réponses modales (pour chacune des modalités de la variable nominale sélectionnée dans MOTEX) selon le critère du chi-2.

ASPAR Analyse des correspondances directe d'un tableau de réponses numérisées issu de la procédure NUMER, sans regroupement. Cette procédure utilise des algorithmes particuliers adaptés aus grands tableaux clairsemés. La figure 5.4 est issue d'une des sorties graphiques de la procédure ASPAR.

SEGME Construction du tableau segmental donnant, pour chaque individu, les numéros des segments contenus dans sa réponse. La procédure SEGME doit suivre la procédure NUMER.

POLEX Cette procédure permet de positionner des formes ou des segments comme éléments supplémentaires sur une analyse factorielle concernant le même ensemble d'individus. La figure 3.4, où quelques formes lexicales sont positionnées sur le premier plan factoriel d'une analyse des correspondances multiples, a été construites avec les résultats d'une procédure POLEX.

TALEX Construction d'un juxtaposition de tableaux lexicaux (après NUMER) ou segmentaux (après SEGME) pour une liste de variables nominales à sélectionner par la procédure SELEC.

Les procédures suivantes sont communes aux logiciels SPAD.N et SPAD.T :

ARDIC Archivage d'un dictionnaire décrivant un fichier d'enquête (variables nominales et numériques continues).

ARDON Archivage du fichier de données décrit par ARDIC.

SELEC Sélection de groupes de variables nominales ou continues en vue d'une analyse factorielle, et en vue des étapes POSIT ou TALEX.

POSIT Positionnement de variables nominales illustratives sur des plans factoriels calculés par ailleurs. La figure 5.5 est un exemple de sortie de la procédure POSIT, qui fait suite à une procédure ASPAR.

e) Exemples d'enchaînements

On donnera ci-dessous quelques exemples d'enchaînements de procédures permettant d'effectuer les opérations les plus usuelles. Le fichier de variables nominales est supposé "archivé" (par les procédures ARDIC et ARDON, identiques à celles du logiciel SPAD.N)
Les instructions sont en format libre, et n'ont nul besoin d'être cadrées. Les titres (obligatoires sur la ligne qui suit l'instruction "PROC") sont à l'initiative de l'utilisateur. Les commentaires peuvent être insérés à droite du séparateur " : ". Lorsqu'un paramètre ne figure pas, sa valeur par défaut est utilisée. Dans la version 1994 pour PC, les commandes se font par menus déroulants.

1) Analyse des correspondances d'un tableau lexical : Séquence des instructions de commande

```
 PROC  ARTEX
 Saisie des questions ouvertes     (titre de la procédure)
 ITYP = 2,   NCOL= 68,  NBQT = 4
                  : Fichier type enquête (ITYP=2), enregistré sur 68 colonnes (NCOL),
                  : avec ici 4 questions ouvertes  (NBQT=4)
 PROC  SELOX
 Choix de la question à traiter   (titre de la procédure)
 NUMQ=2,  LSELI= 0
  : On choisit la question n°2 (NUMQ), sans filtre sur les individus (LSELI=0)
 PROC  NUMER
 Numérisation de la question "Enfants"      (titre de la procédure)
 NSEU= 13,   NXMAX= 160,  NXSIG= 60
 FAIBLE                    : Le nombre de formes retenues sera inférieur à 160
 ,   ;   '   -   (   )     : (NXMAX). De plus, celles-ci devront avoir une
 FORT                      : fréquence supérieure à 13 (NSEU), Il y a moins de
 .   !   ?                 : 60 caractères par ligne (NSIG). Suivent des listes de
 FIN                       : séparateurs faibles et forts.
 PROC MOTEX
 Croisement des formes avec la variable 341 ( âge-diplôme)  (titre )
 NVSEL= 341   : NVSEL est le numéro de la variable nominale.
PROC  APLUM
Analyse du tableau lexical et graphiques    (titre de la procédure)
NAXE=6,   LIMPR=1,   NGRAF=1
   : On calcule 6 axes factoriels, puis on liste lescoordonnées des lignes (formes)
```

: (LIMPR=1), on demande enfin un graphique, qui sera ici le plan factoriel (1,2).
STOP

Parmi les listages correspondant à cet enchaînement figurent les tableaux 5.1 et 5.2 (procédure NUMER), le tableau 5.3 (Procédure MOTEX) et les figure 5.1 (Procédure APLUM).

2) Formes caractéristiques et réponses modales

On ne répétera pas les commentaires relatifs aux procédures précédentes.

```
PROC  ARTEX
Saisie des questions ouvertes
ITYP = 2,   NCOL= 68,  NBQT = 4

PROC  SELOX
Choix de la question à traiter
NUMQ=2,  LSELI= 0

PROC  NUMER
Numérisation de la question "enfants"
NSEU= 13,   NXMAX= 160,  NXSIG= 60
FAIBLE
,   ;   '   -   (   )   :
FORT
.   !   ?
FIN

PROC MOTEX
Croisement des formes avec la variable 341 ( âge-diplôme)  (titre )
NVSEL= 341

PROC MOCAR
Formes caractéristiques et réponses modales   (titre de la procédure)
NOMOT = 10,   NOREP = 3
      : Il s'agit ici, pour les réponss modales, du critère de fréquence lexicale.
      : On demande 10 formes caractéristiques (NOMOT) et 3 réponses modales
      : (NOREP) par modalité de la variable nominale choisie dans MOTEX.

PROC RECAR
Réponses modales, Critère du Chi-2
NOREP = 4           : On demande 4 réponses modales par modalité.

STOP
```

3) Analyse directe, avec illustration par des variables nominales

On supposera que les étapes ARTEX, SELOX, NUMER ont déjà été effectuées, et que les fichiers correspondant sont sauvegardés.

```
PROC  ASPAR
Analyse directe du tableau de numérisation issu de NUMER   (titre)
```

```
NAXE= 5,  NGRAF= 2, NPAGE= 2,  NLIGN= 110
      : On demande 5 axes factoriels, 2 graphiques (plans 1,2  et 2,3), chacun
      : sur 2 pages en largeur et 110 lignes en longueur.
PROC  SELEC
Sélection des variables nominales supplémentaires  (titre )
NOPAR                          : Valeurs par défaut pour les paramètres
NOMI  ILL  313  326  17  34    : Numéros des 4 variables nominales illustratives
FIN

PROC  POSIT
Positionnement des nominales sélectionnées plus haut (titre de la procédure)
NAXED= 5,   NGRAF = 2
                    : On demande l'impression des coordonnées sur les
                    : 5 premiers axes factoriels, et 2 graphiques.
STOP
```

La figure 5.4 est un exemple de graphique obtenu à l'issue de la procédure ASPAR. La figure 5.5 est un exemple de graphique (il s'agit du même plan factoriel) obtenu à l'issue de l'étape POSIT.

4) Analyse des correspondances d'un tableau segmental

On supposera encore que les étapes ARTEX, SELOX, NUMER ont déjà été effectuées, et que les fichiers correspondant sont sauvegardés.

```
PROC  SEGME
Recherche des segments répétés   (titre de la procédure)
NXLON= 6,   NXSEG= 600,  NFRE1= 8,   NFRE2 = 12

NSPA= 'NSPB'      : On précise que MOTEX ne doit pas fonctionner
                  : avec le tableau issu de NUMER, mais avec le
                  : tableau de même type issu de SEGME.

PROC MOTEX
Croisement des segments avec la var. 341 ( âge-diplôme)  (titre)
NVSEL= 341        : NVSEL est le numéro de la variable nominale.

PROC  APLUM
Analyse du tableau lexical et graphiques    (titre de la procédure)
NAXE=6,   LIMPR=1,   NGRAF=1
                  : On calcule 6 axes factoriels, puis on liste les
                  : coordonnées des lignes (formes) (LIMPR=1), on
                  : demande enfin un graphique, qui sera ici le plan
                  : factoriel (1,2).

STOP
```

Ces quelques exemples n'épuisent pas les possibilités du logiciel. Ils doivent montrer que le langage de commande permet des enchaînements assez variés.

A.2 Le logiciel Lexico1

Lexico1 est un ensemble de programmes lexicométriques fonctionnant sur micro-ordinateur.[1] L'ensemble est pour l'instant composé de cinq modules distincts :

SEGMENTATION crée une base de données numérisées à partir d'un fichier texte fourni par l'utilisateur. Cette base est constituée d'un dictionnaire des formes rencontrées dans le texte qui leur affecte également un numéro d'ordre, et d'une version numérisée du texte.

DOCUMENTATION permet de lancer plusieurs types de requêtes documentaires dont les résultats seront, selon le désir de l'utilisateur, affichés à l'écran et/ou stockés dans un fichier éditable par la suite.

STAT1 (Statistiques - module 1) Calcule les segments répétés du texte (cf. chapitre 2), construit des tableaux lexicaux à partir des partitions du corpus décidées par l'utilisateur, opère des calculs statistiques portant à la fois sur les formes et les segments répétés du corpus.

AFC réalise l'analyse des correspondances du tableau lexical constitué à partir d'une partition du texte.

CHRON2 calcule, pour une série textuelle chronologique (cf. chapitre 7) les spécificités chronologiques du corpus ainsi que les accroissements spécifiques pour chacune des parties.

A.2.1 Le texte en entrée

Lexico1 accepte en entrée tout texte, saisi sur traitement de texte[2] mais sauvegardé avec les options : "texte seulement avec ruptures de lignes"[3].

Les deux caractères < et > sont des caractères réservés à l'encodage des clefs. Ils ne doivent figurer dans l'enregistrement d'entrée que pour introduire des informations péritextuelles.

ex : <Epg=238> introduit la page 238
<AD=25> assigne la valeur 25 à la clef AD

Le type des clefs (i.e. la zone située entre le signe < et le signe =) peut être quelconque. Le contenu des clefs (i.e. la zone située entre le signe = et le signe >) est obligatoirement numérique dans cette version.
Dans l'exemple ci-dessous, la clef AD permet d'affecter un code âge-diplôme à chacun des répondants d'une enquête socio-économique (le chiffre des dizaines correspond à une catégorie d'âge, celui des unités à une catégorie de diplôme).

[1] Lexico1 est développé par A. Salem au laboratoire *Lexicométrie & textes politiques* de l'E.N.S. de Fontenay-Saint-Cloud. Une version de ce logiciel fonctionne actuellement sur les micro-ordinateurs de type MacIntosh. Elle permet de traiter des corpus comptant jusqu'à 700 000 occurrences environ.
[2] Tels les traitements de texte WORD, MacWrite, QUED, Edit, etc. que l'on trouve actuellement sur le marché.
[3] Sur les traitements de textes anglo-saxons option "text only".

```
<AD=23>   *les difficultes financieres et materielles
<AD=13>   *les problemes materiels,une certaine angoisse vis a vis
          de l'avenir
<AD=23>   *la peur du futur,la souffrance,la mort,le manque
          d'argent
<AD=23>   *l'avenir incertain,les problemes financiers
<AD=23>   *les difficultes financieres
<AD=12>   *les raisons materielles et l'avenir qui les attend
<AD=13>   *des problemes financiers
<AD=31>   *l'avenir difficile qui se prepare,la peur du chomage
<AD=13>   *l'insecurite de l'avenir
<AD=12>   *le manque d'argent
<AD=33>   *la guerre eventuelle
<AD=23>   *la charge financiere que ca represente,la
          responsabilite morale aussi
<AD=13>   *la situation economique,quand le couple ou la femme
          n'est pas psychologiquement pret pour accueillir un
          enfant
```

Figure A2.1. Lexico1 : exemple de texte en entrée

A.2.2 La segmentation du texte

A partir d'un tel fichier texte, en s'appuyant sur la liste des caractères délimiteurs fournie par l'utilisateur, le premier module opère la segmentation automatique du texte, calcule le nombre des occurrences et l'ordre alphabétique pour chacune des formes graphiques contenues dans le corpus.

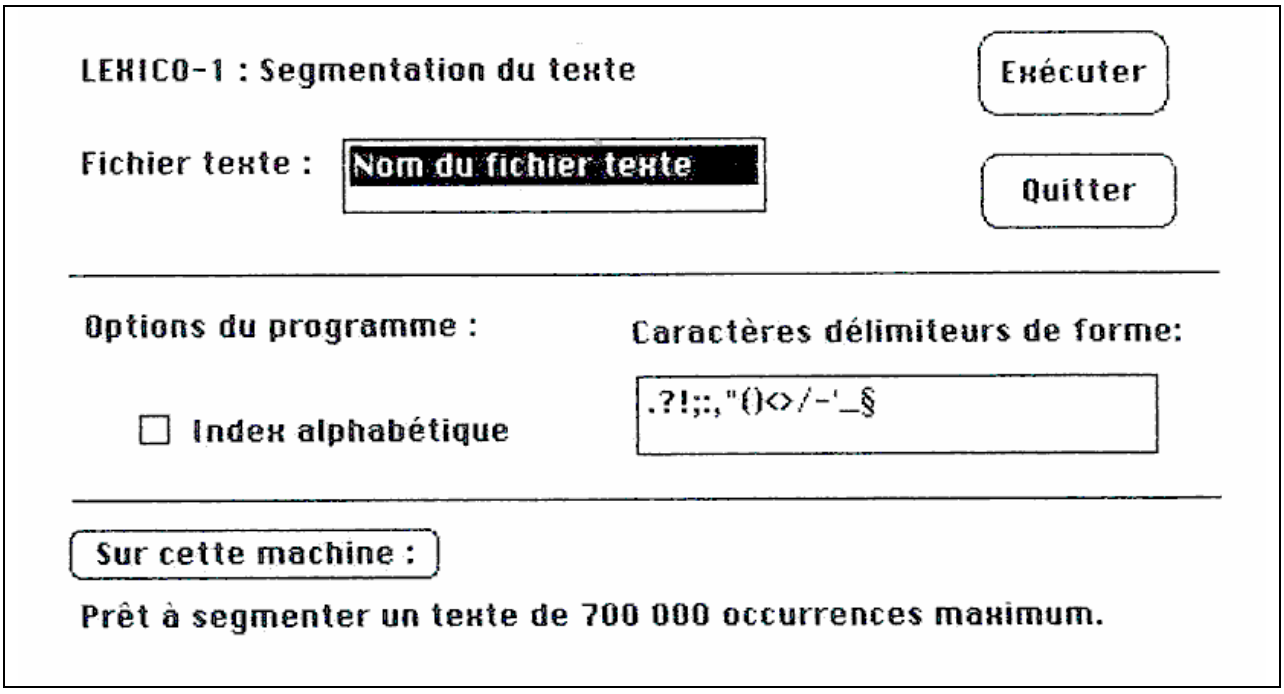


Figure A2.2. Réglage des options de segmentation du texte

Le programme crée ensuite une base de données numérisées qui servira de point de départ pour les travaux documentaires et pour l'étude statistique du texte. Cette base est constituée d'un dictionnaire des formes rencontrées dans le texte lequel contient

également pour chacune d'entre elles un numéro d'ordre lexicométrique (ordre par fréquence décroissante, l'ordre alphabétique départageant les formes en concurrence) et un numéro d'ordre alphabétique, ainsi que d'une version numérisée du texte. En cochant la case "index alpha", on obtient en plus un fichier des formes graphiques classées par ordre lexicographique

A.2.3 La phase documentaire

Le module de documentation permet de retrouver l'ensemble des contextes d'une forme sélectionnée par l'utilisateur. Pour une forme-pôle donnée cet ensemble (un contexte par occurrence de la forme pôle) peut être ordonné au gré de l'utilisateur :

— en fonction de la forme qui précède la forme-pôle (tri avant)
— en fonction de la forme qui suit la forme-pôle (tri après)
— en fonction de l'ordre d'apparition dans le texte.

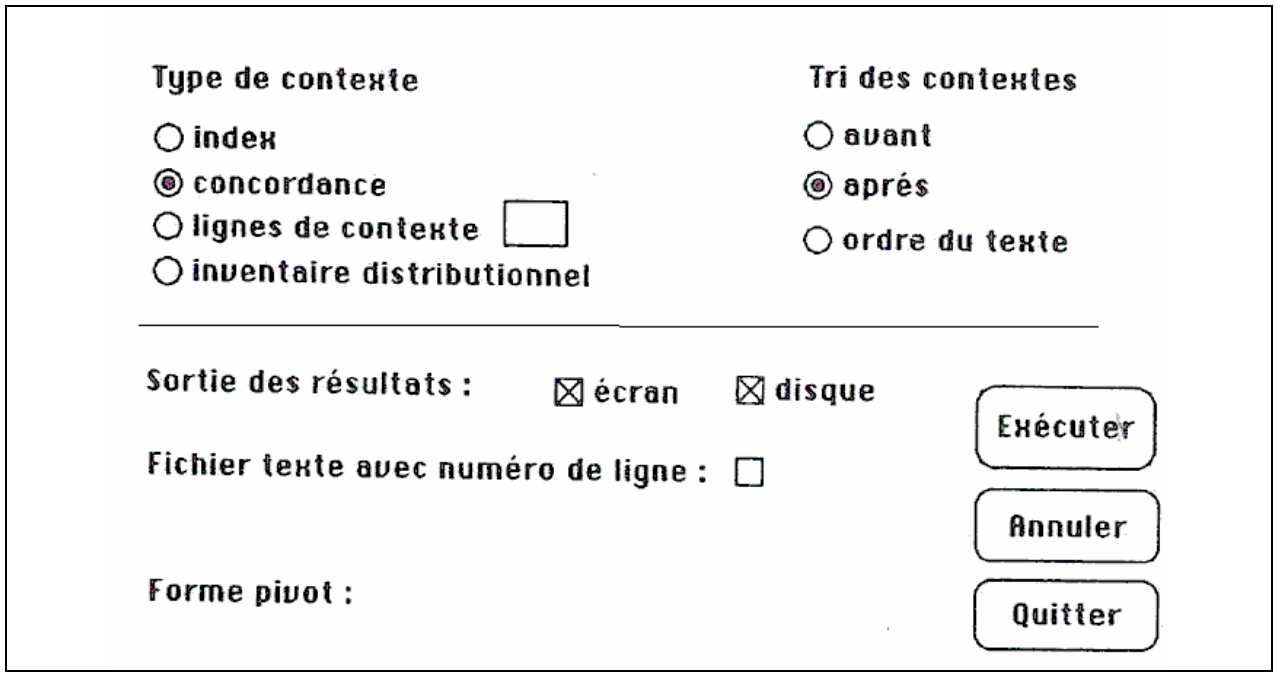


Figure A2.3

Lexico1 : Réglage des options du module de documentation

Les types de contextes disponibles dans cette version sont :

— Index : aucun contexte, mention des lignes ou la forme est attestée.
— concordance : une ligne de contexte centrée sur la forme pivot comportant la mention du numéro de ligne de l'occurrence de la forme pivot.
— contexte: un nombre, définit par l'utilisateur, de lignes de contexte avant et après chaque occurrence de la forme pivot comportant la mention du numéro de ligne de cette occurrence.

Dans ce programme, index, concordances, lignes de contextes renvoient un numéro de ligne qui est le numéro de la ligne dans le fichier d'entrée. Il est donc recommandé d'établir une édition de référence dans laquelle les lignes du texte sont numérotées

A.2.4 Partition du corpus et segments répétés

Ce module permet d'opérer une partition du corpus d'après les différentes valeurs d'une des clefs introduites avant l'étape de segmentation. Les différentes parties du corpus permettent ensuite de construire le tableau lexical qui servira de base aux différentes analyses statistiques. Ce même module calcule également les segments répétés du texte, dont la fréquence dépasse un seuil minimal fixé par l'utilisateur, ainsi que leur ventilation dans les parties du corpus.

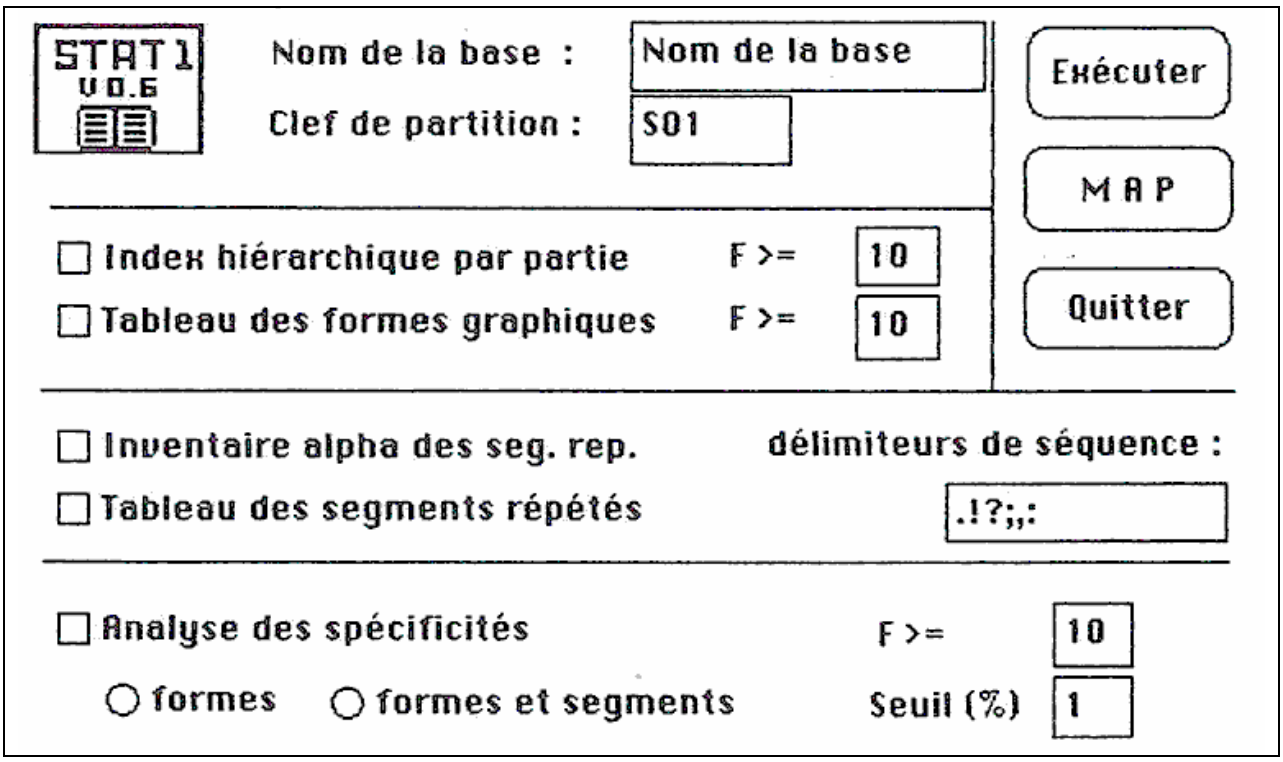


Figure A2.4 . Lexico1 : réglage des options du module *statistiques -1*

On obtient, un *index hiérarchique par partie*, pour chacune des parties du corpus, en cochant la case correspondante et en indiquant une fréquence minimale. Chaque index est classé par ordre de fréquence des formes dans la partie sélectionnée. Chaque index est suivi du rappel des principales caractéristiques lexicométriques de la partie.

En cochant la case "Tableau des formes graphiques", et en indiquant une fréquence minimale on obtient un tableau lexical (formes x parties). On peut également obtenir le tableau qui donne la ventilation des segments dans les mêmes parties en cochant la case correspondante.

On effectue l'Analyse des Spécificités du corpus sur les formes, ou sur les formes et les segments dont la fréquence dépasse un seuil sélectionné, en indiquant une fréquence minimale et un seuil en probabilité.

A.2.4 Analyse des correspondances

Le module AFC effectue l'analyse des correspondances des tableaux (formes x parties) ou (formes + segments x parties)[1].

Tous les paramètres du programme peuvent être modifiés par l'utilisateur en éditant le fichier "afc.param" qui contient les options par défaut. Sans modification explicite de la part de l'utilisateur, ce programme effectue une analyse des correspondances du tableau des formes (à partir de la fréquence minimale sélectionnée dans le module précédent), les segments répétés au-delà de la même fréquence intervenant en qualité d'éléments supplémentaires.

A.2.5 Spécificités chronologiques

A partir d'une série textuelle chronologique divisée en périodes, ce dernier module calcule les spécificités chronologiques du corpus ainsi que les accroissements spécifiques. Les diagnostics de spécificités sont chaque fois triés par probabilité croissante (i. e. des plus remarquables aux moins remarquables) et par période.

A.3 Le logiciel ALCESTE

Le logiciel *ALCESTE*[2] est le résultat d'un approche d'analyse des données textuelles plus spécifiquement orientée vers l'analyse de corpus de textes homogènes (par exemple, un roman, un recueil de poèmes, un corpus d'entretiens, un recueil d'articles sur un même thème, un ensemble de réponses à une question ouverte, etc…). Il est conçu pour être utilisé en liaison avec une analyse de contenu.

L'objectif est d'obtenir un premier classement des "phrases" (i.e. unités de contexte) du corpus étudié en fonction de la répartition des mots dans ces "phrases" (deux phrases se "ressemblent" d'autant plus que leur vocabulaire est semblable, les mots grammaticaux étant exclus) ceci afin d'en dégager les principaux "*mondes lexicaux"*.

L'auteur du logiciel fait l'hypothèse que l'étude statistique de la distribution de ce vocabulaire peut permettre de retrouver la trace des "espaces référentiels" investis par

[1] Ce module recouvre en fait un interface du programme ANCORR du logiciel LADDAD mis à notre disposition par l'association ADDAD.

[2] ALCESTE est développé par M. Reinert (CNRS, Université de Toulouse-Le Mirail ; 5, allée Antonio Machado, 31058 Toulouse-cedex) pour micro-ordinateurs de type MacIntosh. Capacité de la version actuelle : Corpus traité : environ 20 000 lignes de 70 caractères (environ 1 MO). Nombre d'unités de contexte élémentaires (u.c.e.) : 10 000 ; Longueur maximum d'une u.c.e : environ 240 caractères. Nombre maximum d'unités de contexte naturelles : 4 000.

l'énonciateur lors de l'élaboration du discours, trace perceptible sous forme de "mondes lexicaux" (ensemble des mots plus spécifiquement associés à telle classe caractéristique de phrases). Chaque "monde" est ensuite décrit de différentes manières ; notamment à l'aide d'un ensemble de *phrases caractéristiques*, à l'aide de *segments de texte répétés*, à l'aide aussi des lignes du corpus contenant tel ou tel mot caractéristique du monde lexical étudié.
Au niveau informatique, *la version 2.0* du logiciel ALCESTE fonctionne pratiquement sur tout Macintosh muni d'un coprocesseur arithmétique et d'une mémoire centrale égale ou supérieure à 5 MO. Cette version comprend deux modules :

- *Un module de calcul,*
- *Un module de documentation et d'aide à la préparation du plan d'analyse* écrit en HyperTalk sous HyperCard.[1] Ce module se présente sous la forme d'une pile HyperCard d'une centaine de cartes environ.

1) La procédure de mise en oeuvre du logiciel :

a) Crée un dossier d'analyse ne contenant au départ que le seul fichier texte à analyser (en format ascii).
b) Prépare ce dossier à l'aide de la pile HyperCard en le complétant d'un certain nombre de fichiers auxiliaires dont *un plan d'analyse.*
c) Exécute le logiciel en "batch" qui se contente d'appliquer le plan d'analyse au traitement du corpus.
d) Lit les résultats à l'aide d'un éditeur quelconque, la description des principaux fichiers étant présentée dans la pile ou la notice du logiciel.

2) Cette exécution "batch" du logiciel enchaîne 3 étapes, chacune d'elles comprenant plusieurs opérations (programmables) dont voici la liste :

L'étape A permet la définition des "unités de contexte" (sensiblement des phrases), la recherche et la réduction du vocabulaire et le calcul des tableaux de données ; le calcul des couples et segments répétés…

A1 : Lecture et calcul des unités de contexte élémentaires (environ les phrases).
A2 : Recherche et réduction des formes (distinction des mots pleins et des mots outils ; réduction des désinences de conjugaison, des pluriels, etc…).
A3 : Calcul du tableau des données croisant unités de contexte par formes.
A4 : Calcul des couples de formes successives et des segments de texte répétés.

L'étape B effectue une classification des unités de contexte en fonction de la distribution du vocabulaire, classification simple ou double selon que l'on veut tester ou non la stabilité des résultats, en fonction d'une variation de longueur de l'unité de contexte ;

[1] Hypertalk et Hypercard sont des produits Apple.

B1 : Classification simple par la technique de classification descendante hiérarchique mise au point par l'auteur pour le traitement de tableaux clairsemés.

B2 : Classification double : deux classifications successives sur des tableaux ayant en lignes des unités de contexte de longueur différentes afin d'apprécier la stabilité des classes en fonction d'une variation du découpage en unités de contexte.

L'étape C permet plusieurs calculs auxiliaires pour aider à l'interprétation des classes ;

C1 : Choix des classes d'unités de contexte retenues.

C2 : Vocabulaire spécifique de chaque classe.

C3 : Analyse des correspondances sur le tableau de cooccurrences croisant classes par vocabulaire.

C4 : Choix des unités de contexte les plus représentatives de chaque classe.

C5 : Liste des segments répétés par classe.

C6 : Liste des formes d'origine par classe.

C7 : Calcul du concordancier pour les formes les plus spécifiques des classes.

A.4 Le logiciel Hyperbase

Ce logiciel[1] est construit à partir d'un langage à objets[2] et intègre la notion d'Hypertexte. Il en résulte pour l'utilisateur une commodité incomparable pour toute la partie qui concerne la "navigation" entre le texte et les outils documentaires, c'est à dire entre une forme de dictionnaire et ses différents contextes, concordances, ventilation dans les parties d'un corpus etc.
Les commandes sont d'accès facile, en général lancées par la simple sollicitation d'un "bouton". Les aides en ligne sont faciles d'accès et très explicites.

La partie statistique du logiciel, rançon inévitable du langage choisi pour l'écriture du logiciel, est moins développée que la partie documentaire. Elle fournit cependant l'accès aux principales méthodes lexicométriques, spécificités, analyse des correspondances, ainsi qu'une partie permettant de produire des histogrammes à partir des ventilations de formes sélectionnées.

[1] Hyperbase à été conçu et développé par E. Brunet, professeur à l'Université de Nice, pour micro-ordinateurs de type MacIntosh. Il permet de traiter des corpus qui comptent plusieurs millions d'occurrences. Cf. la publication : *Hyperbase*, CUMFID n° 17, URL 9, INaLF (CNRS), Faculté des Lettres, 98 Bd Herriot, 06007, Nice. Ce logiciel est interfacé avec le logiciel ADDAD pour la partie *analyses multidimensionnelles*.

[2] Le logiciel Hypercard est distribué par la firme Apple.

Figure A.4.1 Exemple d'écran *Hyperbase*
Choix des mots dans la version complète du logiciel

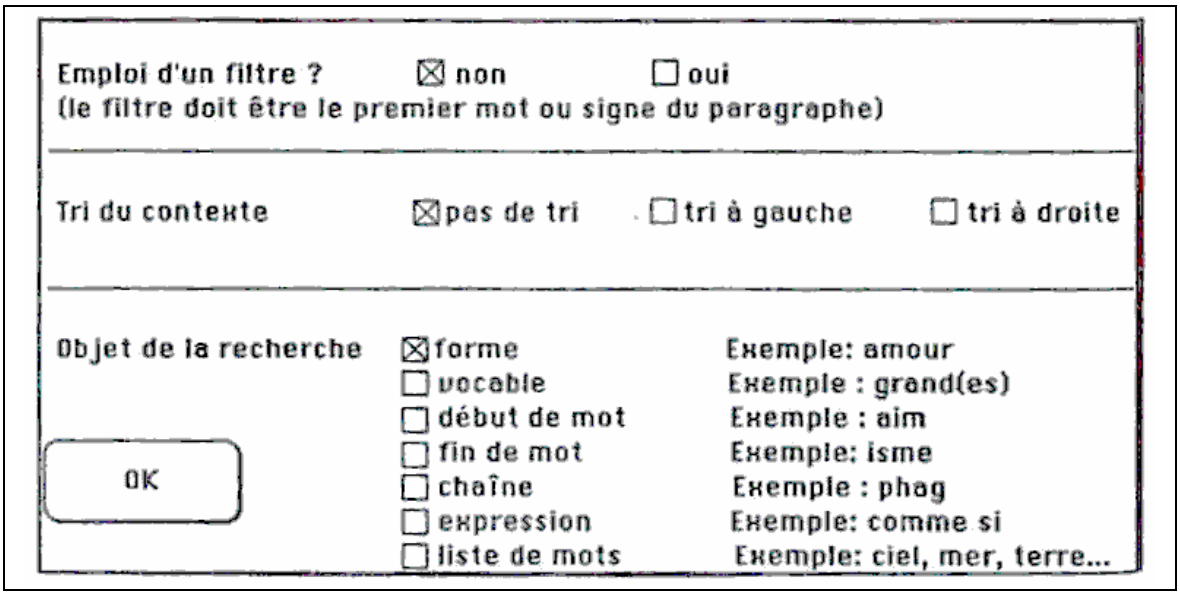


Figure A.4.2. Exemple d'écran *Hyperbase*
Dialogue de la commande *concordance*

Une dernière particularité du logiciel est de fournir, pour tout corpus entré par l'utilisateur, une comparaison statistique avec les données du trésor de la langue française (TLF) pour un corpus comportant plusieurs millions d'occurrences.

A.5 Autres Logiciels

En sus des quatre logiciels qui viennent d'être très brièvement présentés, on mentionnera, sans prétendre à l'exhaustivité, quelques autres produits, en général moins centrés sur la statistique textuelle.

On se limitera aux produits disponibles sur micro-ordinateurs.

SATO[1] qui remplit toutes les fonctions d'indexation du texte que nous avons citées plus haut et permet en outre d'affecter à chacune des occurrences du texte un certain nombre de propriétés (grammaticales, sémantiques ou autres) laissées au choix de l'utilisateur.

SAINT-CHEF[2] logiciel tout particulièrement consacré à la réalisation et à l'édition de concordances sur microordinateur.

PISTES[3] consacré à l'indexation et à l'analyse des spécificités d'un texte découpé en parties.

PHRASEA[4] consacré à la recherche documentaire au sein de vastes corpus de textes.

Le SPHINX[5], logiciel de dépouillement d'enquête doté d'une interface-utilisateur élaborée et comportant des modules de traitements de données textuelles.

LEXIS[6], un des modules de la série de logiciels de dépouillement d'enquêtes développée et distribuée par la société Eole.

---

[1] Sur PC, F Daoust, Centre d'analyse de textes par ordinateur, (ATO), UQAM, Case postale 8888, succursale A, Montréal, Québec, Canada H3C 3P8.
[2] Sur PC, M Sekhraoui, Lexicométrie & textes politiques ENS de Fontenay-St.Cloud.
[3] Sur PC, P. Muller, diffusé par le Centre National de Documentation Pédagogique, Paris.
[4] Sur MacIntosh, C. Poveda et J. Y. Jourdain, B&L Parenthèses, 79 av. Guynemer, 59 700, Marc en Bareuil
[5] Sur PC ou MacIntosh, J. Moscarola, Le SPHINX Développement, 13 Chemin des Amarantes, 74600, Seynod
[6] Sur PC, EOLE, 6 rue du Quatre Septembre, 92130, Issy-les-Moulineaux.

Annexe B

Esquisse des algorithmes et structures de données pour la statistique textuelle

Les logiciels décrits ou évoqués dans l'annexe A permettent la segmentation automatique d'un texte, l'indexation des unités textuelles, la recherche des contextes, etc. Les similitudes que l'on note entre ces différents logiciels montrent que leurs auteurs ont trouvé des solutions souvent très proches aux problèmes rencontrés ; les différences qui existent entre ces logiciels témoignent par ailleurs de leurs objectifs distincts.

Il nous a semblé utile de décrire sommairement dans cette annexe les principaux algorithmes qui sont à la base du logiciel *Lexico1*. Pour ce logiciel, les méthodes relatives à la segmentation automatique et à l'indexation des occurrences d'une forme sont fondées sur une structure de données particulièrement simple.[1] La numérisation du texte proposée permet, dans la plupart des cas, de stocker à la fois en mémoire vive le texte numérisé et le dictionnaire des formes. Cette structure s'est révélée très efficace pour la réalisation des principales tâches au sein d'un logiciel dont l'ambition se limite au domaine de la recherche lexicométrique.

Classification des items textuels

Dans ce qui suit, on appellera *item* toutes les occurrences des unités que l'on peut rencontrer lors du dépouillement d'un texte (occurrences de formes graphiques, occurrences de ponctuations diverses, etc.), on appellera ces unités elles-mêmes des *articles*.

L'algorithme de segmentation automatique repose sur une classification des items (et articles) du fichier-texte. Cette classification implique qu'au moment de soumettre le texte au processus de segmentation un certain nombre d'ambiguïtés aient été levées et que l'on puisse considérer comme réalisée, par rapport au processus de segmentation, la condition :

un signe de l'enregistrement = un statut

Les items que l'on s'attend à rencontrer appartiennent à deux grandes catégories : les *occurrences de forme* et les *jalons textuels*. Les items de la première catégorie sont les unités dont le décompte motive l'entreprise de la statistique textuelle telle qu'elle a été définie aux deux premiers chapitres.

[1] Cette structure a pu être améliorée à la suite de discussions avec des spécialistes du domaine de l'informatique textuelle et avec J. Dendien en particulier.

L'ensemble des jalons textuels se subdivise en deux catégories les *ponctuations* et les *clés*.[1]

Les clés

Les clés permettent d'introduire dans le texte des informations péritextuelles de toutes sortes. Elles sont du type :

<Type1=va11>

Les clés sont introduites entre deux caractères réservés < et >. Le signe "=" sépare le type et le contenu de la clé.

Type1 — indique le *type* de la clé.

val1 — qui peut être précédé par des caractères "blancs" ne faisant pas partie de ce contenu, indique un *contenu* de clé, c'est-à-dire une valeur que l'on assigne à cette clé.

Les ponctuations

La classe "ponctuation" rassemble une classe de jalons textuels qui dépasse largement l'ensemble des signes que l'on regroupe sous cette appellation dans le langage courant.[2]

La liste des signes qui seront considérés comme signes délimiteurs de forme par le programme de segmentation est déterminée par l'utilisateur.[3]

A ces signes délimiteurs de formes s'ajoutent obligatoirement[4] :

— les caractères < et > qui servent à introduire les clés
— le caractère "blanc"
— le caractère "RC" ou retour chariot qui sert à découper le flot d'entrée en lignes.

[1] Le système de codage des clés présenté ici s'inspire très largement de celui présenté dans Lafon et al. (1985).

[2] Bien que ce signe ait une fonction complètement distincte au plan de l'encodage du texte, de celle des signes usuels de ponctuation, le caractère "retour chariot" peut être considéré du point de vue de la segmentation automatique comme un caractère de type "ponctuation".

[3] Le programme Segmentation propose à l'utilisateur une liste contenant les signes de ponctuations les plus courants [. ? ! ; : , " ( ) / ' _ §]. Cette liste est entièrement modifiable par l'utilisateur. Le caractère § peut être utilisé pour marquer le début de chacun des paragraphes du texte de référence.

[4] Ces signes délimiteurs de forme sont rajoutés automatiquement par le programme de segmentation à la liste des caractères délimiteurs choisis par l'utilistateur.

Avec cette convention on peut regrouper en trois classes l'ensemble des jalons textuels que l'on trouve dans un texte :

ponctuation	—	caractère délimiteur de forme (à l'exclusion des signes réservés < et >)
type de clé	—	suite de caractères non-délimiteurs, précédée par le signe < et terminée par le signe "="
contenu de clé	—	suite de caractères non-délimiteurs terminée par le signe >

La segmentation automatique du texte

Le but de l'étape de segmentation automatique du texte est de permettre la construction, à partir d'un fichier texte que l'on appellera *text1*, d'une base textuelle numérisée qui servira de point de départ à l'ensemble des algorithmes de documentation et d'analyse statistique. Cette base est constituée par deux fichiers complémentaires :

text1.dicnum ou *dictionnaire*
et *text1*.textnum ou *fichier texte numérisé*

Le dictionnaire

Le fichier dicnum contient un enregistrement pour chacun des articles du texte à l'exception des articles du type "ponctuation". Les enregistrements correspondant aux articles sont classés en fonction de leur type dans l'ordre suivant :

formes	—	interclassées par ordre lexicométrique[1] à l'intérieur de cette catégorie (i. e. par ordre de fréquence décroissante, l'ordre lexicographique départageant les formes de même fréquence).
types de clés	—	classés par ordre lexicographique
contenus de clés	—	classés dans le même ordre

Chacun de ces enregistrements contient les renseignements suivants :

lexicog.	—	ordre lexicographique de l'article dans la liste ci-dessus (l'ordre lexicographique pour les formes).
fréq.	—	la fréquence de l'article
fgraph.	—	la forme graphique de l'article : forme dans le cas d'une occurrence textuelle, suite de caractères donnant le type ou le contenu d'une clé dans les autres cas.

Le nombre des formes différentes est noté *nbform,* celui des articles *nbarticles*.

[1] Cf. les définitions du chapitre 2.

Le texte numérisé

Le fichier dicnum se présente sous la forme d'une suite de **nbitem** (= nombres des items du texte). Chacun de ces nombres correspond à un des items du texte.

B.1 Tâche n°1 : La numérisation du texte

En fonction de la place mémoire disponible[1] le programme commence par fixer la taille maximale du problème que l'on pourra traiter dans l'environnement considéré.

Les deux principaux paramètres du problème sont *ItemMax* et *ArtMax* respectivement égaux au nombre maximal des items que l'on se propose de traiter et au nombre maximal d'articles que l'on s'attend à trouver dans un texte ItemMax items.

On réserve ensuite les tableaux à une dimension **T**(*ItemMax*), **V**(*ArtMax*) ainsi qu'une série de tableaux de travail. L'un de ces tableaux est muni d'une structure d'arbre binaire dont le noeud élémentaire est structuré de la manière suivante :

structure ***tdic*** *{*

mot :	pointeur sur une zone texte qui contient la forme graphique de l'article;
freq	variable entière contenant le nombre des occurrences de l'article rencontrées depuis le début du processus;
lexicog	rang lexicographique de l'article (qui sera calculé après la fin de la lecture du texte);
lexicom	rang lexicometrique de l'article (idem);
numorg	numéro de l'article dans la liste des articles classés par ordre d'apparition dans le texte;
suivant	pointeur qui pointe sur le noeud suivant;

}

Le tableau **Ftex** est un tableau de caractères dans lequel sont mis bout à bout les suites de caractères qui composent tant les forme graphiques que les types et contenus des clés.

L'adresse du premier de ces caractères est stockée dans la variable *mot* de la structure **tdic**. Le dernier caractère de la forme est suivi d'un caractère spécifique qui permet de le repérer en tant que tel.

La deuxième phase de l'algorithme voit une lecture du fichier texte et une première numérisation item par item. Les items qui se présentent dans le flot d'entrée sont isolés et analysés par la procédure **LireItem** qui calcule pour chacun d'eux :

[1] A titre indicatif, cette structuration des données permet de traiter environ 700 000 occurrences sur une machine disposant de 4MO de mémoire vive.

— le numéro d'item dans le texte depuis le début du texte
— le statut de l'item par rapport au 4 catégories d'articles (occurrence, ponctuation, type de clé, contenu de clé)
— la forme graphique de l'item

Chaque item est ensuite présenté à la racine de l'arbre binaire (ou d'un B-arbre) pour être comparé aux articles déjà stockés dans l'arbre. Les articles sont comparés sur une base lexicographique (aa < ab).

Ce processus permet de calculer un code numérique *CodNum* pour chacun des items entrés. Si l'item considéré correspond à l'occurrence d'un article qui est déjà apparu lors de l'exploitation en cours, sa présentation à l'arbre de stockage permet de retrouver et d'incrémenter la variable qui comptabilise le nombre de ses occurrences. Dans le cas contraire, il s'agit d'un nouvel article dont le numéro de stockage est calculé par incrémentation de la variable qui comptabilise le nombre des objets stockés dans l'arbre.

A la fin de la lecture du fichier texte, l'arbre contient tous les articles présents dans le texte. Pour chaque article on connaît en outre le nombre de ses occurrences dans le texte. Les articles sont numérotés d'après l'ordre de leur apparition dans le texte. On en profite pour affecter ce numéro d'ordre à la variable *numorg*.

Les tris

Le calcul des rangs lexicographiques et lexicométriques de chacun des articles est réalisé à partir du fichier des articles.[1] On commence par remplir *List (*)* un tableau de pointeurs de longueur *nbarticles* dont chacun pointe sur un des noeuds de l'arbre de stockage des articles.

Ce tableau de pointeurs est trié une première fois d'après l'ordre lexicographique de chacun des articles.[2] Les articles correspondant aux formes graphiques se trouvent placés en tête de cette liste du fait de l'adjonction de caractères de poids très élevé devant les chaînes de caractères correspondant aux autres catégories. Les articles du type "type de clé" sont suivis eux-même par les articles du type "contenu de clé". Après cette opération, on peut affecter à la variable *lexicog* une valeur égale au rang de l'article après tri de l'ensemble selon l'ordre lexicographique.

Un second tri par ordre décroissant des pointeurs du tableau *List (*)* d'après la valeur de la variable *freq* de l'article sur lequel ils pointent, les formes de même fréquence étant départagées d'après les valeurs ascendantes de la variable *lexicog* que nous venons de calculer, permet de classer les articles correspondant aux formes graphiques d'après l'ordre lexicométrique. Cette opération terminée, on peut assigner à la variable *lexicog* le numéro d'ordre lexicométrique.

[1] Cette manière de procéder constitue un progrès important par rapport aux algorithmes de segmentation qui effectuent les tris sur le fichier des items beaucoup plus volumineux que celui des articles.
[2] Dans la pratique, ce tri est notablement accéléré par une technique de "tri par morceaux". Les articles sont d'abord regroupés en sous-groupes en fonction du premier caractère. On effectue ensuite un tri à l'intérieur de chacun des sous-groupes.

Il est possible alors de passer à l'écriture sur support externe du dictionnaire. Cette écriture se fait en suivant l'ordre lexicométrique dans lequel sont actuellement triés les pointeurs contenus dans le tableau list. Pour chaque article on écrit successivement :

freq	le nombre des occurrences de l'article dans le texte;
lexicog	rang lexicographique de l'article;
mot	suite de caractères composant l'article terminée par un retour chariot.

Un dernier tri de la liste de pointeurs contenue dans le tableau list, effectué d'après les valeurs de la variable *numorg,* replace cette liste dans l'ordre initial.

La numérisation finale

La dernière phase de l'algorithme de segmentation permet de stocker sur support externe le texte numérisé d'après l'ordre lexicométrique des articles du texte. Pour effectuer cette opération il suffit de reprendre le tableau T dans lequel les articles sont numérisés d'après l'ordre de leur apparition dans le texte et de substituer le numéro lexicométrique de chaque article à ce numéro. On trouve le numéro lexicométrique correspondant à l'article dont le numéro d'apparition est égal à i dans la variable *lexicom* du noeud de l'arbre de stockage pointé par l'élément i.

B.2 Tâche n°2 : La recherche de contextes

On regroupe sous cette appellation la recherche de différents sites textuels pour un ensemble d'occurrences correspondant à une *forme-pôle* donnée ou à une liste de ces formes. On construit une telle liste en introduisant l'une des sélections suivantes :

- une forme graphique
- une liste de formes graphiques
- un groupe de caractères constituant le début d'une ou de plusieurs formes présentes dans le texte.
- un groupe de caractères constituant la fin d'une ou de plusieurs formes présentes dans le texte.
- un groupe de caractères présent dans le corps d'une ou de plusieurs formes présentes dans le texte.

Les unités de contexte retenues pour l'ensemble des formes-pôle sélectionnées peuvent être :

- une ligne de contexte

— un nombre fixe de lignes de contexte pour chacune des occurrences de la liste des formes-pôle.
— un inventaire distributionnel des segments répétés après la forme-pôle.

Les contextes peuvent-être triés selon deux types d'arguments dont chacun peut être choisi comme argument majeur du tri ou comme argument mineur :

— en fonction des formes graphiques
— en fonction des valeurs d'une clé sélectionnée dans le texte

Pour une forme graphique donnée, les contextes peuvent être également triés, par rapport à la forme-pôle :

— en fonction de l'ordre lexicographique des formes qui précèdent chacune des occurrences de cette forme.
— en fonction de l'ordre lexicographique des formes qui suivent chacune des occurrences de cette forme.
— en fonction de l'ordre d'apparition des occurrences de cette forme dans le texte.

B.3 Tâche n°3 : Le calcul de la gamme des fréquences.

Le calcul de la gamme des fréquences peut être réalisé à partir du seul dictionnaire dans lequel les formes qui ont une même fréquence sont classées côte à côte. En commençant par la fréquence la plus élevée (ou au contraire par la fréquence 1) on calcule par simple cumul les effectifs V_i qui correspondent à chacune des fréquences pour i variant de 1 à fmax.

B.4 Tâche n°4 : La construction des Tableaux Lexicaux.

Le tableau lexical est déterminé par deux paramètres fixés par l'utilisateur :

— *fmin* : le seuil minimal retenu pour la sélection des formes qui correspondront aux lignes du tableau lexical (on ne retiendra que les formes dont la fréquence est égale ou supérieure à ce seuil).

— *Sxx* : un type correspondant à une clé (le tableau lexical comptera une colonne pour chacune des valeurs différentes prises par le contenu de ce type de clé dans le fichier texte).

Une fois ces valeurs fixées, on peut calculer *x1*, le nombre des lignes du tableau lexical qui est égal au nombre des formes dont la fréquence est au moins égale à *fmin* dans le corpus. Remarquons que, par suite de la définition de l'ordre lexicométrique que nous

avons adoptée, les numéros lexicométriques de toutes les formes dont la fréquence est inférieure à ce seuil sont supérieurs au numéro lexicométrique de la dernière forme de fréquence est égale au seuil retenu. Appelons ce numéro *fder*.

Une exploration des contenus existants dans le texte pour la clé sélectionnée permet de fixer *x2,* le nombre des colonnes du tableau égal au nombre des valeurs différentes prise par la variable "contenu de la clé sélectionnée". Les différentes valeurs prises par cette variable sont triées par ordre lexicographique ascendant et numérotées dans cet ordre. Cette numérotation permet d'affecter à chaque code un numéro de partie.

On peut alors réserver une zone mémoire de taille (*x1* x *x2*) qui contiendra le tableau que l'on se propose de calculer. Le calcul de ce tableau s'effectue en une seule lecture du fichier texte numérisé. En parcourant ce tableau à partir de la première de ses cases on commence par trouver la première occurrence de la clé sélectionnée et le premier contenu de cette clé. Ce contenu nous renvoie à un numéro de partie d'après la table établie précédemment. Ce numéro s'appellera le "numéro de partie courant".

La poursuite de ce traitement pour chacune des cases du tableau "tnum" amène à trois situations différentes :

— le code tnum[k] est :

  soit le code d'une forme dont la fréquence est inférieure au seuil retenu (i.e. un code supérieur à *fder*) ;

  soit le code d'une ponctuation ; soit le code d'un type ou d'un contenu de clé correspondant à une autre clé que la clé sélectionnée et dans ce cas on ignore la case tnum[k] pour passer au traitement du code contenu dans la case tnum[k+1].

— le code tnum[k] est le code de la clé sélectionnée. On lit alors dans la case tnum[k+1] la valeur du contenu de la clé qui permet de mettre à jour le code de partie courant (puisque l'on passe au traitement des occurrences situées sans une autre partie du texte).

— le code tnum[k] est le code d'une occurrence de la forme *forme* dont la fréquence est supérieure ou égale au seuil retenu (i.e. un code inférieur à *fder*). Si ce code est égal à i, on ajoute une unité à la case [i, numéro de partie courant] du tableau lexical que l'on est en train de calculer.

0	de	31	44	52	112	34	72	136	36	48
1	l	35	36	57	82	29	51	82	27	30
2	la	22	28	42	82	27	43	118	25	24
3	les	24	24	32	48	21	33	65	13	26
4	le	36	20	33	61	13	28	71	8	14
5	d	18	20	25	47	14	26	55	7	16
6	pas	15	11	21	43	10	18	72	9	11
7	avenir	21	22	31	34	12	30	28	12	8
8	chômage	21	18	13	35	5	9	58	6	10

Tableau B.1
Exemple de tableau lexical

Lorsque ce processus est achevé on a calculé le tableau lexical des formes de fréquence supérieure ou égale au seuil fixé. On peut éditer ce tableau en faisant précéder la ventilation de chacune des formes sélectionnées par son numéro lexicométrique et son libellé, que l'on trouvera dans le dictionnaire, comme dans le tableau B.1 ci-dessus.

B.5 Tâche n°5 : Tableaux des segments répétés

Les calculs portant sur la ventilation des segments répétés dans les parties d'un corpus de textes permettent de compléter les analyses effectuées à partir des tableaux de formes graphiques (tableaux lexicaux).[1]

Cependant les segments répétés d'un texte sont avant tout caractérisés par leur énorme redondance.

Pour réduire le volume des segments répétés, on écarte de la liste des segments ceux d'entre eux qui sont des segments contraints (i.e. qui sont toujours précédés ou suivis par une même forme et qui entrent donc dans la composition de segments plus longs).

Pour certaines applications statistiques, il est en outre possible de ne considérer que les segments dont la fréquence dépasse un certain seuil de répétition (on abaissera ce seuil à 2 occurrences si l'on désire obtenir la liste de tous les segments répétés dans le texte).

On peut distinguer pour la réalisation de la tâche n°5 deux sous-tâches, qui interviennent alternativement lors de la construction du TSR d'un corpus.

Ces sous-tâches sont :

5-1 — le repérage des segments non contraints dans le corpus qui commencent par une forme graphique donnée et dont la fréquence dépasse un seuil fixé.

5-2 — le calcul de la ventilation dans les parties du corpus de l'ensemble des occurrences d'un segment ainsi repéré.

Le repérage de l'ensemble des segments non-contraints débutant par une forme donnée, pour chacune des formes graphiques dont la fréquence est égale ou supérieure à un seuil fixé réalise l'ensemble de la tâche n°5.

B.5.1 Sous-tâche 5.1 : Repérage des segments répétés non contraints

Pour une forme graphique donnée *Form1* et un seuil de fréquence *Sfr* fixé, le but de la sous tâche 5.1 consiste donc dans le repérage de l'ensemble des segments non-contraints débutant par *Form1* dont la fréquence est au moins égale à *Sfr* occurrences.

[1] Cf. chapitres 2 et 5.

Ainsi définie, cette sous tâche devient très simple à réaliser. On commence par définir une liste de pointeurs list1[k] qui pointent chacun sur une des occurrences de la forme *Form1*.[1]

Dans un deuxième temps ces pointeurs sont triés en fonction de l'ordre lexicographique des formes qui suivent la forme *Form1*.[2]

A la fin de ce tri les segments répétés éventuels qui commencent par la forme *Form1* se trouvent placés les uns à la suite des autres dans l'ordre lexicométrique des segments répétés (i.e. les segments sont classés en fonction de l'ordre lexicographique de la première forme, départagés en cas d'égalité par l'ordre lexicographique de la forme qui suit et ainsi de suite) comme c'est le cas dans l'exemple ci-dessous.[3]

A B
A B
A B C
A B C A C
A B C A C D
A B C A C D
A C B A

Dans le processus de comparaison des segments en vue de compter les occurrences de segments répétés, on procède par comparaison des formes graphiques situées à chacune des positions du segment.

Si le tri effectué nous assure que les segments répétés identiques se trouvent bien côte à côte, il reste à repérer, en parcourant la liste des segments commençant par la forme *Form1*, les segments qui marquent la fin d'un groupe d'occurrences relatives à un même segment de longueur *L1*.

Remarquons qu'à l'intérieur des occurrences des segments commençant par *L* formes identiques, les occurrences d'un segment répété comportant *L+1* formes identiques constituent un ensemble de segments consécutifs après le tri lexicographique des segments.

Comparaison de segments

Les cases des tableaux SEGn[k] sont remplies par des valeurs relatives à des occurrences de la forme *Form1*. Pour le repérage des segments répétés dont la fréquence est supérieure ou égale à *Sfr,* on ne s'intéresse qu'aux séquences ne comprenant aucune forme de fréquence inférieure à ce seuil ne chevauchant pas en outre de délimiteurs de séquences.

[1] La taille de cette liste est égale, au maximum, à *fmax* la fréquence de la forme la plus fréquente.
[2] Remarquons que cet ordre correspond exactement à celui dans lequel apparaîssent les contextes-lignes lors de la réalisation d'une concordance triée sur le contexte qui suit.
[3] Comme dans les exemples du chapitre 2 relatifs au corpus **P**, les lettres capitales symbolisent ici chacune une forme graphique.

Les séquences à comparer sont donc constituées par des suites de codes correspondant au numéro lexicométrique de formes de fréquence supérieure au seuil retenu. On sélectionne un code particulier noté *VIDE* pour remplir les autres cases du tableau SEGn[k].

Par rapport à une occurrence de forme graphique, on appellera dans ce qui suit **forme suivante**, soit la forme *VIDE,* soit la première forme graphique située après cette forme, en négligeant les jalons textuels et signes de ponctuation non-délimiteurs de séquence situés entre les deux formes. Les mêmes conventions s'appliquent pour la définition de la **forme précédente**.

Les séquences débutant par la forme *Form1* qui vont être soumises à la comparaison sont rangées dans les tableaux SEGn[k] de la manière suivante: on range dans la case SEGn[0] le numéro lexicométrique de la forme précédant *Form1*, en respectant les conventions ci-dessus.

La case SEGn[1] contient le numéro lexicométrique de la forme *Form1* qui sert de pivot à la recherche des segments.

Les cases suivantes SEGn[2], SEGn[3],..., SEGn[longmax] contiennent soit les numéros lexicométriques des formes qui suivent soit le code *VIDE* (à partir d'une certaine position).

Recherche de l'accident

La comparaison entre deux segments successifs permet de déterminer la longueur du sous-segment répété le plus long qu'ils ont en commun, c'est-à-dire le nombre des premières occurrences qui sont identiques pour les deux segments.

Nous appellerons *accident* la position pour laquelle une occurrence du second segment ne correspond pas, pour la première fois à l'occurrence du segment précédent. Ainsi, par exemple, dans la comparaison des deux segments

```
    1 2 3 4 5 6 7
    A B C D E F G
et  A B C D E G H
```

nous dirons que l'accident se situe en position numéro 6.

Recherche des segments répétés

Le tableau SEGc[n], qui contient le "segment courant", est initialisé par les valeurs du segment placé en tête par le tri lexicographique des segments. On gère parallèlement un tableau FreqSeg[n] qui contient pour n variant de 2 à *longmax* la fréquence courante des segments de longueur 2 à *longmax*.

On considère ensuite le segment placé immédiatement après par le tri lexicographique. La détermination d'un accident en position *x* par rapport au segment précédent indique que l'on en a terminé avec les occurrences du sous-segment précédent (dont la longueur est égale à *x*). Si la fréquence des segments répétés ainsi répertoriés dépasse le seuil *Sfr,* nous avons repéré un segment répété dont il nous reste à calculer la ventilation dans les parties du corpus.

B.5.2 Sous-tâche 5.2 : Calcul de la ventilation des segments répétés

Le calcul de la ventilation des occurrences du segment répété s'opère à partir du numéro de chacune des occurrences de la forme *Form1*. Ce problème suppose que l'on ait un moyen de déterminer le numéro de la partie à laquelle appartient une occurrence du texte. Plusieurs possibilités s'offrent pour réaliser cette tâche dont la plus simple consiste à construire un tableau récapitulatif indiquant les cases du tableau *tnum* qui correspondent à des changements de partie.

Glossaire pour la statistique textuelle

NB : Les astérisques renvoient à une entrée de ce même glossaire. Les abréviations qui suivent entre parenthèses précisent le domaine auquel s'applique plus particulièrement la définition.

Abréviations :

ac	Analyse factorielle des correspondances
acm	Analyse des correspondances multiples
cla	Classification
sp	Méthode des Spécificités
sr	Analyse des segments répétés
ling	Linguistique
stat	Statistique
sa	Segmentation automatique

accroissement spécifique - (sp) spécificité* calculée pour une partie d'un corpus par rapport à une partie antérieure

algorithme - ensemble des règles opératoires propres à un calcul.

analyse factorielle (stat) - famille de méthodes statistiques d'analyse multidimensionnelle, s'appliquant à des tableaux de nombres, qui visent à extraire des "facteurs" résumant approximativement par quelques séries de nombres l'ensemble des informations contenues dans le tableau de départ.

analyse des correspondances (stat)- méthode d'analyse factorielle s'appliquant à l'étude de tableaux à double entrée composés de nombres positifs. L'AC est caractérisée par l'emploi d'une distance (ou métrique) particulière dite distance du chi-2 (ou χ^2).

analyse des correspondances multiples (stat) **-** méthode d'analyse des correspondances s'appliquant à l'étude de tableaux disjonctifs complets. Tableaux binaires dont les lignes sont des individus ou observations et les colonnes la juxtaposition des modalités de réponse à des questions (les modalités de réponse à une question s'excluant mutuellement).

caractère (sa) - signe typographique utilisé pour l'encodage du texte sur un support lisible par l'ordinateur.

caractères délimiteurs / **non-délimiteurs** (sa) - distinction opérée sur l'ensemble des caractères, qui entrent dans la composition du texte permettant aux procédures informatisées de segmenter le texte en occurrences* (suite de caractères non-délimiteurs bornée à ses extrémités par des caractères délimiteurs).

On distingue parmi les caractères délimiteurs:

- les caractères délimiteurs d'occurrence (encore appelés "**délimiteurs de forme**") qui sont en général : le blanc, les signes de ponctuation usuels, les signes de préanalyse éventuellement contenus dans le texte.

- les caractères **délimiteurs de séquence** : sous-ensemble des délimiteurs d'occurrence correspondant, en général, aux ponctuations faibles et fortes contenues dans la police des caractères.

- les caractères **séparateurs de phrase** : (sous-ensemble des délimiteurs de séquence) qui correspondent, en général, aux seules ponctuations fortes.

classification (stat) - technique statistique permettant de regrouper des individus ou observations entre lesquels a été définie une distance.

classification hiérarchique (cla) - technique particulière de classification produisant par agglomération progressive des classes ayant la propriété d'être, pour deux quelconques d'entre-elles, soit disjointes, soit incluses.

concordance (sa) - l'ensemble de lignes de contexte se rapportant à une même forme-pôle.

contribution absolue (ou contribution) - (ac) contribution apportée par un élément au facteur . Pour un facteur donné, la somme des contributions sur les éléments de chacun des ensembles mis en correspondance est égale à 100.

contribution relative (ou cosinus carré) - (ac) contribution apportée par le facteur à un élément. Pour un élément donné, la somme des contributions relatives sur l'ensemble des facteurs est égale à 1.

cooccurrence (sa) - (une c.) - présence simultanée, mais non forcément contiguë, dans un fragment de texte (séquence, phrase, paragraphe, voisinage d'une occurrence, partie du corpus etc.) des occurrences de deux formes données.

corpus (ling) - ensemble limité des éléments (énoncés) sur lesquels se base l'étude d'un phénomène linguistique.

(lexicométrie) ensemble de textes réunis à des fins de comparaison; servant de base à une étude quantitative.

délimiteurs de séquence - (sa) sous-ensemble des caractères délimiteurs* de forme* correspondant aux ponctuations faibles et fortes (en général - le point, le point d'interrogation, le point d'exclamation, la virgule, le point-virgule, les deux points, les guillemets, les tirets et les parenthèses).

dendrogramme - (cla) représentation graphique d'un arbre de classification hiérarchique, mettant en évidence l'inclusion progressive des classes.

discours/langue - La langue est un ensemble virtuel qui ne peut être appréhendé que dans son actualisation orale ou écrite; "discours" est un terme commode qui recouvre les deux domaines de cette actualisation.

distance du chi-2 - distance entre profils* de fréquence utilisée en analyse des correspondances* et dans certains algorithmes* de classification*.

éditions de contextes (sa) - éditions de type concordanciel dans lesquelles les occurrences d'une forme sont accompagnées d'un fragment de contexte pouvant contenir plusieurs lignes de texte autour de la forme-pôle. La longueur de ce

contexte est définie en nombre d'occurrences avant et après chaque occurrence de la forme-pôle.

éléments d'un segment (sr) - chacune des formes correspondant aux occurrences qui entrent dans sa composition. ex : A, B, C sont respectivement les premier, deuxième et troisième éléments du segment ABC.

éléments actifs- (ac ou acm) ensemble des éléments servant de base au calcul des axes factoriels, des valeurs propres relatives à ces axes et des coordonnées factorielles.

éléments supplémentaires (ou illustratifs)- (ac ou acm) ensemble des éléments ne participant pas aux calculs des axes factoriels, pour lesquels on calcule des coordonnées factorielles qui auraient été affectées à une forme ayant la même répartition dans le corpus mais participant à l'analyse avec un poids négligeable.

énoncé/énonciation - (ling) à l'intérieur du texte un ensemble de traces qui manifestent l'acte par lequel un auteur a produit ce texte.

expansion contrainte -(sr) terme dont les occurrences constituent chaque fois les expansions (du même côté) d'un même terme ayant plusieurs occurrences dans le corpus.

expansion d'un segment - (sr) segment situé immédiatement avant (expansion gauche) ou après (expansion droite) d'un segment donné, non séparé de ce segment par un délimiteur de séquence.

expansion récurrente d'un terme - terme dont les occurrences constituent plusieurs fois l'expansion des occurrences d'un terme donné.

facteur- (ac ou acm) variables artificielles construites par les techniques d'analyse factorielle permettant de résumer (de décrire brièvement) les variables actives initiales.

forme- (sa) ou "**forme graphique**" archétype correspondant aux occurrences* identiques dans un corpus de textes, c'est-à-dire aux occurrences composées strictement des mêmes caractères non-délimiteurs d'occurrence.

forme banale - (sp) pour une partie du corpus donnée, forme ne présentant aucune spécificité (ni positive ni négative) dans cette partie .

forme caractéristique - (d'une partie) synonyme de spécificité positive*.

forme commune - forme attestée dans chacune des parties du corpus.

forme originale- (pour une partie du corpus) forme trouvant toutes ses occurrences dans cette seule partie.

fréquence (sa) - (d'une unité textuelle) le nombre de ses occurrences dans le corpus.

fréquence d'un segment (sr) - (ou d'une polyforme) le nombre des occurrences de ce segment, dans l'ensemble du corpus.

fréquence maximale (sa) - fréquence de la forme la plus fréquente du corpus (en français, le plus souvent, la préposition "de").

fréquence relative (sa) - la fréquence d'une unité textuelle dans le corpus ou dans l'une de ses parties, rapportée à la taille du corpus (resp. de cette partie).

gamme des fréquences (sa) **-** suite notée V_k, des effectifs correspondant aux formes de fréquence k, lorsque k varie de 1 à la fréquence maximale.

hapax - gr. hapax (legomenon), "chose dite une seule fois".

(sa) forme dont la fréquence est égale à un dans le corpus (hapax du corpus) ou dans une de ses parties (hapax de la partie).

identification - (stat, ling, sa) reconnaissance d'un seul et même élément à travers ses multiples emplois dans des contextes et dans des situations différentes.

index - (sa) liste imprimée constituée à partir d'une réorganisation des formes et des occurrences d'un texte, ayant pour base la forme graphique et permettant de regouper les références* relatives à l'ensemble des occurrences d'une même forme.

index alphabétique (sa) - index* dans lequel les formes-pôles* sont classées selon l'ordre lexicographique* (celui des dictionnaires).

index hiérarchique (sa) - index* dans lequel les formes-pôles* sont classées selon l'ordre lexicométrique*.

index par parties - ensemble d'index (hiérarchiques ou alphabétiques) réalisés séparément pour chaque partie d'un corpus.

item de réponse - (ou modalité de réponse) élément de réponse préétabli dans une question fermée.

lemmatisation - regroupement sous une forme canonique (en général à partir d'un dictionnaire) des occurrences du texte. En français, ce regroupement se pratique en général de la manière suivante :

- les formes verbales à l'infinitif,
- les substantifs au singulier,
- les adjectifs au masculin singulier,
- les formes élidées à la forme sans élision.

lexical - (ling) qui concerne le lexique* ou le vocabulaire*.

lexicométrie ensemble de méthodes permettant d'opérer des réorganisations formelles de la séquence textuelle et des analyses statistiques portant sur le vocabulaire* d'un corpus de textes.

lexique - (ling) ensemble virtuel des mots d'une langue.

longueur (sa) - (d'un corpus, d'une partie de ce corpus, d'un fragment de texte, d'une tranche, d'un segment, etc.) le nombre des occurrences contenues dans ce corpus (resp. : partie, fragment, etc.). Synonyme de taille.

On note: T la longueur du corpus; t_j celle de la partie (ou tranche) numéro j du corpus.

longueur d'un segment (sr) - le nombre des occurrences entrant dans la composition de ce segment.

noyaux factuels - (cla) classes d'une partition* d'un ensemble d'observations synthétisant une batterie de descripteurs objectifs (sexe, âge, profession, etc.).

occurrence (sa) **-** suite de caractères non-délimiteurs bornée à ses extrémités par deux caractères délimiteurs* de forme.

ordre lexicographique -

_ pour les formes graphiques :

l'ordre selon lequel les formes sont classées dans un dictionnaire.

NB : Les lettres comportant des signes diacrisés sont classées au même niveau que les mêmes caractères non diacrisés, le signe diacritique n'intervenant que dans les cas d'homographie complète. Dans les dictionnaires, on trouve par exemple, rangées dans cet ordre, les formes : *mais, maïs, maison, maître* .

_ pour les polyformes:

ordre résultant d'un tri des polyformes par ordre lexicographique sur la première composante, les polyformes commençant par une même forme graphique sont départagées par un tri lexicographique sur la seconde, etc.

ordre lexicométrique (sa) **-**

_ pour les formes graphiques :

ordre résultant d'un tri des formes du corpus par ordre de fréquences décroissantes; les formes de même fréquence sont classées par ordre lexicographique.

_ pour les polyformes:

ordre résultant d'un tri par ordre de longueur décroissante des segments, les segments de même longueur sont départagés par leur fréquence, les segments ayant même longueur et même fréquence par l'ordre lexicographique.

paradigme- (ling) ensemble des termes qui peuvent figurer en un point de la chaîne parlée.

paradigmatique- (sa) qui concerne le regroupement en série des unités textuelles, indépendamment de leur ordre de succession dans la chaîne écrite.

partie - (d'un corpus de textes) fragment de texte correspondant aux divisions naturelles de ce corpus ou à un regroupement de ces dernières.

partition - (d'un corpus de textes) division d'un corpus en *parties* constituées par des fragments de texte consécutifs, n'ayant pas d'intersection commune et dont la réunion est égale au corpus.

(d'un ensemble, d'un échantillon) division d'un ensemble d'individus ou d'observations en *classes* disjointes dont la réunion est égale à l'ensemble tout entier.

partition longitudinale - (sa) partition d'un corpus en fonction d'une variable qui définit un ordre sur l'ensemble des parties

périodisation (sa) - regroupement des parties naturelles du corpus respectant l'ordre chronologique d'écriture, d'édition ou de parution des textes réunis dans le corpus.

phrase - (sa) fragment de texte compris entre deux séparateurs* de phrase.

places - (sa) pour un texte comptant T occurrences, suite de T objets correspondant chacun à une des occurrences du texte, éventuellement séparés par des délimiteurs* de séquence correspondant aux ponctuations du texte de départ.

polyforme (sr) - archétype des occurrences d'un segment; suite de formes non séparées par un séparateur de séquence, qui n'est pas obligatoirement attestée dans le corpus.

ponctuation - Système de signes servant à indiquer les divisions d'un texte et à noter certains rapports syntaxiques et/ou conditions d'énonciation.

(*sa*) caractère (ou suite de caractères) correspondant à un signe de ponctuation.

post-codage - opération manuelle qui consiste à repérer les principales catégories de réponses libres sur un sous-échantillon de réponses, puis à fermer la question ouverte correspondante, en affectant toutes les réponses à ces catégories.

pourcentages d'inertie - (ac ou acm) quantités proportionnelles aux valeurs propres* dont la somme est égale à 100. Notées τ_α.

profil - (stat et ac) (d'une ligne ou d'une colonne d'un tableau à double entrée) vecteur constitué par le rapport des effectifs contenus sur cette ligne (resp. colonne) à la somme des effectifs que contient la ligne (resp. la colonne).

question fermée - question dont les seules réponses possibles sont proposées explicitement à la personne interrogée.

question ouverte - question posée sans grille de réponse préétablie, dont la réponse peut être numérique (ex: *Quelles doivent être, selon vous, les ressources minimum d'une famille ayant trois enfants de moins de 16 ans?*), ou textuelle (ex: *pouvez-vous justifier votre choix?*).

références (sa) - système de coordonnées numériques permettant de repérer dans le texte d'origine chacune des occurrences issues de la segmentation (ex : le tome, la page, la ligne, la position de l'occurrence dans la ligne) ou de situer rapidement cette occurrence parmi des catégories prédéfinies (auteur, année de parution, citation, mise en valeur, etc.).

répartition (sa) - (des occurrences d'une forme dans les parties du corpus) nombre des parties du corpus dans lesquelles cette forme est attestée.

réponse modale - (d'une classe d'individus, d'une partie* de corpus*) réponse sélectionnée en fonction de son caractère représentatif d'une classe ou d'une partie en général à partir des formes* caractéristiques qu'elle contient.

segment - (sr) toute suite d'occurrences consécutives dans le corpus et non séparées par un séparateur* de séquence est un segment du texte.

segment répété (sr) - (ou polyforme répétée) suite de forme dont la fréquence est supérieure ou égale à 2 dans le corpus.

segmentaire - (sr) ensemble des termes* attestés dans le corpus.

segmentation - opération qui consiste à délimiter des unités minimales* dans un texte.

segmentation automatique - ensemble d'opérations réalisées au moyen de procédures informatisées qui aboutissent à découper, selon des règles prédéfinies, un texte

stocké sur un support lisible par un ordinateur en unités distinctes que l'on appelle des unités minimales*.

séparateurs de phrases - (sa) sous-ensemble des caractères délimiteurs* de séquence* correspondant aux seules ponctuations fortes (en général : le point, le point d'interrogation, le point d'exclamation).

séquence - (sa) suite d'occurrences du texte non séparées par un délimiteur* de séquence.

seuil - (stat) quantité arbitrairement fixée au début d'une expérience visant à sélectionner parmi un grand nombre de résultats, ceux pour lesquels les valeurs d'un indice numérique dépassent ce seuil (de fréquence, en probabilité, etc.).

seuil d'absence spécifique - (sp) pour un seuil fixé, pour une partie du corpus fixée, fréquence A(j) pour laquelle toute forme de fréquence supérieure à A(j) dans le corpus et absente dans la partie j est spécifique (négative) pour la partie j.

seuil de présence spécifique - (sp) pour un seuil fixé, pour une partie du corpus fixée, fréquence B(j) pour laquelle toute forme de fréquence inférieure à A(j) dans le corpus et présente dans la partie j est spécifique (positive) pour la partie j.

sous-fréquence (sa) - (d'une unité textuelle dans une partie, tranche, etc.) nombre des occurrences de cette unité dans la seule partie (resp. tranche, etc.) du corpus.

sous-segments (sr) - pour un segment donné, tous les segments de longueur inférieure et compris dans ce segment sont des sous-segments. ex : AB et BC sont deux sous-segments du segment ABC.

spécificité chronologique - (sp) spécificité* portant sur un groupe connexe de parties d'un corpus muni d'une partition longitudinale*.

spécificité positive - (sp) pour un seuil de spécificité fixé, une forme i et une partie j données, la forme i est dite spécifique positive de la partie j (ou forme caractéristique* de cette partie) si sa sous-fréquence est "anormalement élevée" dans cette partie. De façon plus précise, si la somme des probabilités calculées à partir du modèle hypergéométrique pour les valeurs égales ou supérieures à la sous-fréquence constatée est inférieure au seuil fixé au départ.

spécificité négative - (sp) pour un seuil de spécificité fixé, une forme i et une partie j données, la forme i est dite spécifique négative de la partie j si sa sous-fréquence est anormalement faible dans cette partie. De façon plus précise, si la somme des probabilités calculées à partir du modèle hypergéométrique pour les valeurs égales ou inférieures à la sous-fréquence constatée est inférieure au seuil fixé au départ.

stock distributionnel du vocabulaire - (d'un fragment de texte) le vocabulaire* de ce fragment assorti de comptages de fréquence pour chacune des formes entrant dans sa composition.

syntagmatique- (sa) qui concerne le regroupement des unités textuelles, selon leur ordre de succession dans la chaîne écrite.

syntagme- (ling) groupe de mots en séquence formant une unité à l'intérieur de la phrase.

tableau de contingence (stat) - synonyme de tableau de fréquences ou de tableau croisé: tableau dont les lignes et les colonnes représentent respectivement les modalités

de deux questions (ou deux variables nominales) , et dont le terme général représente le nombre d'individus correspondant à chaque couple de modalités.

tableau lexical entier (TLE) - tableau à double entrée dont les lignes sont constituées par les ventilations* des différentes formes dans les parties du corpus. Le terme générique *k(i,j)* du TLE est égal au nombre de fois que la forme *i* est attestée dans la partie *j* du corpus. Les lignes du TLE sont triées selon l'ordre lexicométrique* des formes correspondantes.

tableau des segments répétés (TSR) - tableau à double entrée dont les lignes sont constituées par les ventilations* des segments répétés dans les parties du corpus. Les lignes du TSR sont triées selon l'ordre lexicométrique* des segments. (i.e. longueur décroissante, fréquence décroissante, ordre lexicographique).

tableau lexical- tableau à double entrée résultant du TLE par suppression de certaines lignes (par exemple celles qui correspondent à des formes dont la fréquence est inférieure à un seuil donné).

taille- (sa) (d'un corpus) sa longueur* mesurée en occurrences (de formes simples).

terme - (sr) nom générique s'appliquant à la fois aux formes* et aux polyformes*. Dans le premier cas on parlera de termes de longueur 1. Les polyformes sont des termes de longueur 2,3, etc.

termes contraints / termes libres - Un terme S1 est contraint dans un autre terme S2 de longueur supérieure si toutes ses occurrences* sont des sous-segments* de segments correspondant à des occurrences du segment S2. Si au contraire un terme possède plusieurs expansions distinctes, qui ne sont pas forcément récurrentes, c'est un terme libre.

unités minimales (pour un type de segmentation) - unités que l'on ne décompose pas en unités plus petites pouvant entrer dans leur composition (ex : dans la segmentation en formes graphiques les formes ne sont pas décomposées en fonction des caractères qui les composent).

valeur modale - (stat) valeur pour laquelle une distribution atteint son maximum.

valeurs propres - (ac ou acm) quantités permettant de juger de l'importance des facteurs successifs de la décomposition factorielle. La valeur propre notée λ_α. mesure la dispersion des éléments sur l' axe.α.

valeurs-tests - (ac ou acm) quantités permettant d'apprécier la signification de la position d'un élément supplémentaire* (ou illustratif) sur une axe factoriel. Brièvement, si une valeur test dépasse 2 en valeur absolue, il y a 95 chances sur 100 que la position de l'élément correspondant ne puisse être due au hasard.

variables actives - variables utilisées pour dresser une typologie, soit par analyse factorielle, soit par classification. Les typologies dépendent du choix et des poids des variables actives, qui doivent de ce fait constituer un ensemble homogène.

variables supplémentaires (ou illustratives) - variables utilisées a posteriori pour illustrer des plans factoriels ou des classes. Une variable supplémentaire peut-être considérée comme une variable active munie d'un poids nul.

variables de type T - variable dont la fréquence est à peu près proportionnelle à l'allongement du texte. (ex : la fréquence maximale)

variables de type V- variable dont l'accroissement a tendance à diminuer avec l'allongement du texte (ex : le nombre des formes, le nombre des hapax).

ventilation (sa) - (des occurrences d'une unité dans les parties du corpus) La suite des n nombres (n = nombre de parties du corpus) constituée par la succession des sous-fréquences* de cette unité dans chacune des parties, prises dans l'ordre des parties.

vocabulaire (sa) - ensemble des formes* attestées dans un corpus de textes.

vocabulaire commun - (sa) l'ensemble des formes attestées dans chacune des parties du corpus.

vocabulaire de base - (sp) ensemble des formes du corpus ne présentant, pour un seuil fixé, aucune spécificité (négative ou positive) dans aucune des parties , (i.e. l'ensemble des formes qui sont "banales" pour chacune des parties du corpus).

vocabulaire original- (sa) (pour une partie du corpus) l'ensemble des formes* originales* pour cette partie.

voisinage d'une occurrence - (sa) pour une occurrence donnée du texte, tout segment (suite d'occurrences consécutives, non séparées par un délimiteur de séquence) contenant cette occurrence.

voisinages d'une forme - (sa) ensemble constitué par les voisinages de chacune des occurrences correspondant à la forme donnée.

Références bibliographiques

Abi Farah A. (1988) - Reconnaissance de l'auteur d'un texte d'après les caractères utilisés, *Les Cahiers de l'Analyse des Données*, XIII, n°1, p 95-96.

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

Adams L. L.(1975) - *A statistical analysis of the book of Isaiah in relation to the Isaiah problem,* Brigham Young University, Provo.

Aitchison J., Aitken C. G. G. (1976) - Multivariate binary discrimination by the kernel method, *Biometrika,* 63, p 413-420.

Akuto H. (Ed.) (1992) - *International Comparison of Dietary Cultures*, Nihon Keizai Shimbun, Tokyo.

Akuto H., Lebart L. (1992) - Le repas idéal. Analyse de réponses libres en anglais, français, japonais, *Les Cahiers de l'Analyse des Données*, vol XVII, n°3, Dunod, Paris, p 327-352.

Aluja Banet T., Lebart L. (1984) - Local and partial principal component analysis and correspondence analysis, *COMPSTAT, Proceedings in Computational Statistics*, Physica Verlag, Vienna, p 113-118.

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

Babeau A., Lebart L. (1984) - Les conditions de vie et les aspirations des Français, *Futuribles,* Avril, p 37-51.

Bardin L. (1989) - *L'analyse de contenu,* PUF, Paris.

Bartell B.T., Cottrell G.W., Belew R.K. (1992) - Latent semantic indexing is an optimal special case of multidimensional scaling, *Proceedings of the 15th Int. ACM-SIGIR Conf. on Res. and Dev. in Information Retrieval*, Belkin N and al. Ed., p 161-167, ACM Press, New York.

Barthélémy J.P., Luong X. (1987) - Sur la typologie d'un arbre phylogénétique : Aspects théoriques, algorithmes et applications à l'analyse de données textuelles, *Math. Sci. Hum.* n°100, p 57-80.

Baudelot C. (1988) - Confiance dans l'avenir et vie réussie (Réponse à un questionnaire), *Recherches économiques : études en hommage à Edmond Malinvaud*, Economica et EHESS, Paris.

Bécue M. (1988) - Characteristic repeated segments and chains in textual data analysis, *COMPSTAT, 8th Symposium on Computational Statistics*, Physica Verlag, Vienna.

Bécue M. (1989) - *Un sistema informatico para el analisis de datos textuales*, Tesis. Facult. d'Informatica, Univ. Politecnica de Catalunya, Barcelona.

Becue M., Peiro R. (1993) - Les quasi-segments pour une classification automatique des réponses ouvertes, in *Actes des 2ndes Journées Internationales d'analyse des données textuelles*, (Montpellier), ENST, Paris, p 310-325.

Behrakis T., Nicolaidis E. (1990) - Typologie des prologues des livres grecs de sciences édités de 1730 à 1820 : Humanisme et esprit des Lumières, *Les Cahiers de l'Analyse des Données*, p 9-20.

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

Benveniste E. (1966) - *Problèmes de linguistique générale*, Gallimard, Paris.

Benzécri J.-P.(1964) - *Cours de linguistique mathématique*, Document mimeographié, Faculté des Sciences de Rennes (cf. aussi Benzécri et al., 1981a).

Benzecri J.-P.(1977) - Analyse discriminante et analyse factorielle, *Les Cahiers de l'Analyse des Données*, II, n °4, p 369-406.

Benzécri J.-P. & coll. (1973) - *La taxinomie*, Vol. I ; *L'analyse des correspondances,* Vol. II, Dunod, Paris.

Benzécri J.-P. (1982) - *Histoire et préhistoire de l'analyse des données,* Dunod, Paris.

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

Benzécri J.-P. (1991a) - Typologies de textes grecs d'après les occurrences des formes des mots-outil, *Les Cahiers de l'Analyse des Données*, XVI, n°1, p 61-86.

Benzécri J.-P. (1991b) - Typologies de textes latins d'après les occurrences des formes des mots-outil, *Les Cahiers de l'Analyse des Données*, XVI, n°4, p 439-465.

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

Benzécri J.-P. et F. (1992b) - Typologie de textes espagnols de la littérature du Siècle d'Or d'après les occurrences des formes des mots outil, *Les Cahiers de l'Analyse des Données*, vol XVII, n°4, Dunod, Paris, p 425-464.

Benzécri J.-P. (1992c) - *Correspondence Analysis Handbook*, (Transl : T.K. Gopalan) Marcel Dekker, New York.

Bergounioux A., Launay M.-F., Mouriaux R., Sueur J-P., Tournier M. (1982) - *La parole syndicale*, P.U.F., Paris.

Bernet C. (1983) - *Le vocabulaire des tragédies de Jean Racine*, *Analyse statistique*, Slatkine-Champion, Genève 1983.

Blosseville M.J., Hébrail G., Monteil M.G., Pénot N. (1992) - Automatic document classification : natural language processing, statistical analysis and expert system techniques used together, *Proceeding of the ACM-SIGIR*, Copenhagen, p 51-58.

Bochi S., Celeux G., Mkhadri A. (1993) - Le modèle d'indépendance conditionnelle : le programme DISIND, *La revue de MODULAD (INRIA),* Juin, p 1-5.

Boeswillwald E. (1992) - L'expérience du CESP en matière de qualité des mesures d'audience, in : *La qualité de l'information dans les enquêtes*, ASU, Dunod, Paris.

Bolasco S. (1992) - Sur différentes stratégie dans une analyse des formes textuelles : Une expérimentation à partir de données d'enquête, *Jornades Internacionals d'Analisi de Dades Textuals*, UPC, Barcelona, p 69-88.

Bonnafous S. (1991) - *L'immigration prise aux mots. Les immigrés dans la presse au tournant des années quatre-vingt,* Kimé, Paris.

Bonnet A. (1984) - *L'intelligence artificielle, promesses et réalités*, InterEditions, Paris.

Boscher F. (1984) - *Le système d'enquêtes sur les conditions de vie et aspirations des Français*, Phase 5, Thème Transport, Rapport Credoc.

Bouchaffra D., Rouault J. (1992) - Solving the morphological ambiguities using a nonstationnary hidden Markov model, *AAAI Fall Symposium on Probabilistic Approach to Natural Language processing*, Cambridge, Massachusetts.

Bouroche J.M., Curvalle B. (1974) - La recherche documentaire par voisinage, R.A.I.R.O., V-1, p 65-96.

Bradburn N., Sudman S., and Associates (1979) - *Improving Interview Method and Questionnaire Design,* Jossey Bass. San Francisco.

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

Brunet E. (1981) - *Le vocabulaire français de 1789 à nos jours, d'après les données du Trésor de la langue française*, Slatkine-Champion, Genève-Paris.

Burt C. (1950) - The factorial analysis of qualitative data, *British J. of Stat. Psychol.,* vol 3, n°3, p 166-185.

Burtschy B., Lebart L. (1991) - Contiguity analysis and projection pursuit, in : *Applied Stochastic Models and Data Analysis,* R. Gutierrez and M.J.M. Valderrama, Eds, World scientific, Singapore, p 117-128.

Callon M., Courtial J.-P., Turner W. (1991) - La méthode Leximappe, in : *Gestion de la recherche, nouveaux problèmes, nouveaux outils*. Vinck D. (Ed.), De Boeck-Wesmael, Bruxelles.

Carré R., Dégremont J.F., Gross M., Pierrel J.M., Sabah G. (1991) - *Langage humain et machine*, Presses du CNRS, Paris.

Celeux G., Hébrail G., Mkhadri A., Suchard M. (1991) - Reduction of a large scale and ill-conditioned statistical problem on textual data, in : *Applied Stochastic Models and Data Analysis, Proceedings of the 5th Symposium in ASMDA*, Gutierrez R. and Valderrama M.J. Eds,World Scientific, p 129-137.

Chartron G. (1988) - *Analyse des corpus de données textuelles, sondage de flux d'Information*, Thèse. Univ. Paris 7.

Church K.W., Hanks P. (1990) - Word association norms, mutual information and lexicography, *Computational Linguistics,* Vol 16, p 22-29.

Cibois P. (1992) - Eclairer le vocabulaire des questions ouvertes par les questions fermées : le tableau lexical des questions, *Bull. de Method. Sociol.*, 26, p 24-54.

Cohen M. (1950) - Sur la statistique linguistique, *Conférence de l'institut de linguistique de l'université de Paris*, Vol IX, Année 1949, Klincksieck, Paris.

Cooper W., Gey F.C., Dabney D.P. (1992) - Probabilistic retrieval based on staged logistic regression, *Proceedings of the 15th Int. ACM-SIGIR Conf. on Res. and Dev. in Information Retrieval*, Belkin N and al., Eds, ACM Press, New York, p 198-209.

Coulon D., Kayser D. (1986) - Informatique et langage naturel : Présentation générale des méthodes d'interprétation des textes écrits, *Techniques et sciences informatiques*, Vol.5, n°2, p 103-128.

Cutting D. R., Karger D.R., Pedersen J.O., Tukey J.W. (1992) - Scatter / Gather : A cluster based approach to browsing large document collections. *Proceedings of the 15th Int. ACM-SIGIR*

Conf. on Res. and Dev., in Information Retrieval, Belkin N. and al., Ed, ACM Press, New York, p 318-329.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshman R. (1990) - Indexing by latent semantic analysis, *J. of the Amer. Soc. for Information Science,* 41 (6), p 391-407.

Demonet M., Geffroy A., Gouaze J., Lafon P., Mouillaud M., Tournier M. (1975) - *Des tracts en Mai 68. Mesures de vocabulaire et de contenu*, Armand Colin et Presses de la Fondation Nat. des Sc. Pol., Paris.

Dendien J. (1986) - La Base de données de l'Institut National de la Langue Française, *Actes du colloque international CNRS,* Nice, juin 1985, 2 vol., Slatkine-Champion Genève, Paris.

Desval H., Dou H. (1992) - *La veille technologique. L'information scientifique, technique et industrielle*, Dunod, Paris.

Diday E. (1971) - La méthode des nuées dynamiques, *Revue Stat. Appl.*,vol.19, n°2, p 19-34.

Diday E. (1992) - From data to knowledge : Probabilist objects for a symbolic data analysis, in : *Computational Statistics*, Dodge Y., Whittaker J. (Eds), Physica Verlag, Heidelberg, p 193-214.

Efron B. (1982) - *The Jacknife, the Bootstrap and other Resampling Plans*, SIAM, Philadelphia.

Efron B., Thisted R. (1976) - Estimating the number of unseen species : how many words did Shakespeare know ?, *Biometrika*, 63, p 435-437.

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

Escofier B. (1978) - Analyse factorielle et distances répondant au principe d'équivalence distributionnelle, *Revue de Statist. Appl.*, vol. 26, n°4, p 29-37.

Escofier B. [Cordier] (1965) - *Analyse des correspondances*, Thèse, Faculté des Sciences de Rennes.

Escoufier Y. (1985) - L'Analyse des correspondances, ses propriétés, ses extensions. *Bull. of the Int. Stat. Inst.*, 4, p 28-2.

Estoup J. B. (1916) - *Gammes sténographiques,* 4ème Edition, Paris. (cf. aussi Mandelbrot, 1961).

Fiala P., Habert B., Lafon P., Pineira C. (1987) - Des mots aux syntagmes. Figements et variations dans la résolution générale du congrès de la CGT de 1978, *Mots*, N° 14, p 47-87.

Fisher R. A. (1936) - The use of multiple measurements in taxonomic problems, *Annals of Eugen*ics, 7, p 179-188.

Fowler R.H., Fowler W.A.L., Wilson B.A. (1991) - Integrating query, thesaurus, and documents through a common visual representation, *Proceedings of the 14th Int. ACM Conf. on Res. and Dev. in Information Retrieval*, Bookstein A. and al., Ed, ACM Press, New York, p 142-151.

Friedman J.H. (1989) - Regularized discriminant analysis, *Journal of the American Statistical Association*, 84, p 165-175.

Froeschl K.A. (1992) - Semantic metadata : query processing and data aggregation, in : *Computational Statistics*, Dodge Y., Whittaker J. (Eds), Physica Verlag, Heidelberg, p 357-362.

Fuchs W. (1952) - On the mathematical analysis of style, *Biometrika,* 39, p 122-129.

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

Furnas G. W., Deerwester S., Dumais S.T., Landauer T.K., Harshman R. A., Streeter L.A., Lochbaum K.E. (1988) - Information retrieval using a singular value decomposition model of latent semantic structure, *Proceedings of the 14th Int. ACM Conf. on Res. and Dev. in Information Retrieval*, p 465-480.

Geary R.C. (1954) - The contiguity ratio and statistical mapping, *The Incorporated Statistician,* Volume 5, N° 3, p 115-145.

Geffroy A., Lafon P., Tournier M. (1974) - L'indexation minimale, Plaidoyer pour une non-lemmatisation, Colloque sur l'analyse des corpus linguistiques : *"Problèmes et méthodes de l'indexation minimale",* Strasbourg 21-23 mai 1973.

Geisser S. (1975) - The predictive sample reuse method with applications, *J. Amer. Statist. Assoc.* 70, p 320-328.

Gifi A. (1990) - *Non Linear Multivariate Analysis*, Wiley, Chichester.

Gobin C., Deroubaix J. C. (1987) - Du progrès, de la réforme de l'Etat, de l'austérité. Déclarations gouvernementales en Belgique, *Mots,* n°15, p 137-170.

Goldstein M., Dillon W. R. (1978) - *Discrete Discriminant Analysis,* Wiley, Chichester.

Good I. J., Toulmin G.H. (1956) - The number of new species and the increase in population coverage when a sample is increased, *Biometrika*, 43, p 45-63.

Greenacre M.(1984) - *Theory and Applications of Correspondence Analysis*, Academic Press, London.

Greenacre M. (1993) - *Correspondence Analysis in Practice*, Academic Press, London.

Gruaz C. (1987) - *Le mot français, cet inconnu. Précis de morphographémologie,* Publ. de l'Univ. de Rouen, n°138, Rouen.

Guilbaud G.-Th. (1980) - Zipf et les fréquences, *Mots N° 1*, p 97-126.

Guilhaumou J. (1986) - L'historien du discours et la lexicométrie. Etude d'une série chronologique : Le père Duchesne de Hébert, juillet 1793- mars 1794, *Histoire & Mesure ,* Vol. I, n° 3-4.

Guiraud P. (1954) - *Les caractères statistiques du vocabulaire*, P.U.F., Paris.

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

Guttman L. (1941) - The quantification of a class of attributes: a theory and method of a scale construction, in *The prediction of personal adjustment* (P. Horst, ed.), SSCR New York, p 251 -264.

Habbema. D. F., Hermans J., van Den Broek K. (1974) - A stepwise discriminant analysis program using density estimation, in *COMPSTAT 1974*, Bruckman G.(ed), Physica Verlag, Vienna.

Habert B., Tournier M. (1987) - La tradition chrétienne du syndicalisme français aux prises avec le temps. Evolution comparée des résolutions confédérales (1945 - 1985), *Mots*, n°14.

Hand D. J. (1981 and 1986) - *Discrimination and Classification*, Wiley, New-York.

Hand D. J. (1992) - Microdata, macrodata, and metadata, *Computational Statistics*, Dodge Y., Whittaker J. (Eds), Physica Verlag, Heidelberg, p 325-340.

Haton J. P. (1985) - Intelligence artificielle en compréhension automatique de la parole : état des recherches et comparaison avec la vision par ordinateur, *Techniques et sciences informatiques*, 4 (3), p 265-287.

Hayashi C. (1956) - Theory and examples of quantification (II), *Proc. of the Institute of Stat. Math.* 4 (2) p 19-30.

Hayashi C. (1987) - Statistical Study of Japanese National Character, *J. Japan Statistical Soc.*, Special Issue, p 71-95.

Hayashi C., Suzuki T., Sasaki M. (1992) - *Data Analysis for Social Comparative research : International Perspective*. North-Holland, Amsterdam.

Hébrail G., Marsais J. (1993) - Experiment in textual data analysis, *EDF, Centre de Recherche*, Clamart, France.

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

Herdan G. (1964) - *Quantitative Linguistics*, Londres, Butterworths.

Hirschfeld H.D. (1935) - A Connection between correlation and contingency, *Proc. Camb. Phil. Soc.* 31, p 520-524.

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

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

Jambu M. (1978) - *Classification automatique pour l'analyse des données, Tome 1 : méthodes et algorithmes*, Dunod, Paris.

Juan S. (1986) - L'ouvert et le fermé dans la pratique du questionnaire. *R. Fr. Socio.* XXVII, Ed. du CNRS, Paris, p 301-306.

Kasher A. (1972) - The book of Isaiah : Caracterization of Authors by Morphological Data Processing, *Revue (R.E.L.O) LASLA N°3,* Liège, p 1- 62.

Labbé D. (1990) - *Le vocabulaire de François Mitterand,* Presses de la Fond. Nat. des Sciences Politiques, Paris.

Labbé D. (1983) - *François Mitterrand - Essai sur le discours*, La pensée sauvage, Grenoble.

Labbé D. (1990) - *Normes de dépouillement et procédures d'analyse des textes politiques*, CERAT, Grenoble.

Labbé D., Thoiron P., Serant D. (Ed.) (1988) - *Etudes sur la richesse et la structure lexicales*, Slatkine-Champion, Paris-Genève.

Lachenbruch P.A., Mickey M.R. (1968) - Estimation of error rate in discriminant analysis, *Technometrics*, 10, p 1-11.

Lafon P. (1980) - Sur la variabilité de la fréquence des formes dans un corpus, *Mots N°1* , p 127-165.

Lafon P. (1981) - Analyse lexicométrique et recherche des cooccurrences, *Mots N°3* , p 95-148.

Lafon P. (1981) - *Dépouillements et statistiques en lexicométrie*, Slatkine-Champion, 1984, Paris.

Lafon P., Salem A. (1983) - L'Inventaire des segments répétés d'un texte, *Mots N°6*, p 161-177.

Lafon P., Salem A., Tournier M. (1985) - Lexicométrie et associations syntagmatiques (Analyse des segments répétés et des cooccurrences appliquée à un corpus de textes syndicaux). *Colloque de l'ALLC*, Metz -1983, Slatkine-Champion, Genève, Paris, p 59-72.

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

Lebart L. (1969) - L'Analyse statistique de la contiguïté, *Publications de l'ISUP*, XVIII- p 81 - 112.

Lebart L., Houzel van Effenterre Y. (1980) - Le système d'enquête sur les aspirations des français, une brève présentation, *Consommation* n°1, 1980, Dunod, p 3-25.

Lebart L. (1987) - Conditions de vie et aspirations des français, évolution et structure des opinions de 1978 à 1986, *Futuribles ,* sept 1987, p 25-56.

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

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

Lebart L. (1992a) - Discrimination through the regularized nearest cluster method, *in : Computational Statistics*, (Y. Dodge, J. Whittaker, Eds) Physica Verlag, Heidelberg, p 103-118.

Lebart L. (1992b) - Assessing and comparing patterns in multivariate analysis, *Second Japanese French Seminar on Data Science*, in Data Science and application, Hayashi et al. Eds, HBJ, Tokyo, Japan..

Lebart L., Memmi D. (1984) - Analisi dei Dati Testuali : Applicazione al Discorso Politico. Atti delle XXXII Riunione Scient. (Soc. Ital. Statis.), p 27-41.

Lebart L., Morineau A., Bécue M. (1989) - *SPAD.T, Système Portable pour l'Analyse des Données Textuelles*, *Manuel de Référence,* CISIA , Paris.

Lebart L., Morineau A., Warwick K. (1984) *Multivariate Descriptive Statistical Analysis,* Wiley, New York.

Lebart L., Salem A. (1988) - *Analyse statistique des données textuelles*, Dunod, Paris.

Lebart L., Salem A., Berry E. (1991) - Recent development in the statistical processing of textual data, *Applied Stoch. Model and Data Analysis,* 7, p 47-62.

Lelu A., Rozenblatt D. (1986) - Représentation et parcours d'un espace documentaire. Analyse des données, réseaux neuronaux et banques d'images, *Les Cahiers de l'Analyse des Données,* Vol XI - n° 4, p 453-470.

Lelu A. (1991) - From data analysis to neural networks : new prospects for efficient browsing through databases, *Journal of Information Science,* vol. 17, p 1-12.

Lewis D. D., Croft W. B. (1990) - Term clustering of syntactic phrases, *Proceedings of the 13th Int. ACM Conf. on Res. and Dev. in Information Retrieval*, Vidick J.L., Ed, A.C.M.Press, New York, p 385-395.

Lewis D.D. (1992) - An evaluation of phrasal and clustered representation on a text categorization task, *Proceedings of the 15th Int. ACM-SIGIR Conf. on Res. and Dev. in Information Retrieval*, Belkin N. and al., Eds, ACM Press, New York, p 37-50.

Luong X. (Ed.) (1989) - *Analyse arborée des données textuelles, INALF, CUMFID/CNRS*, Nice.

Mahalanobis P.C. (1936) - On the generalized distance in statistics, *Proc. Nat. Inst. Sci. India*, 12, p 49-55.

Mandelbrot B. (1961) - On the theory of words frequency and on related Markovian models of discourses, *Structure of Language and its Mathematical Aspects*, Am. Math. Soc., Providence, R.I., p 190-219.

Mandelbrot B. (1968) - Les constantes chiffrées du discours, *Le Langage*, Encyclopédie de la Pléiade, vol XXV, Gallimard, Paris.

Margulis E.L. (1992) - N-Poisson document modelling, *Proceedings of the 15th Int. ACM-SIGIR Conf. on Res. and Dev. in Information Retrieval*, Belkin N and al., Ed, ACM Press, New York, p 177-189.

McKevitt P., Partridge D., Wilks Y. (1992) - Approaches to natural language discourse processing, *Artificial Intelligence Review,* 6, p 333-364.

McKinnon A., Webster R. (1971) - A method of "author" identification, *The Comput. in Liter. and Linguist. Res*., R.A. Wisbey, Ed., Cambridge Univ. Press.

McLachlan G.J. (1992) - *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New-York.

Menard N. (1983) - *Mesure de la richesse lexicale, théorie et vérifications expérimentales*, Slatkine-Champion, Paris.

Michelet B. (1988) - *La logique d'association,* Thèse, Université Paris 7.

Moran P.A.P. (1954) Notes on continuous stochastic phenomena, *Biometrika,* 37, p 17-23.

Morton A.Q. (1963) - The authorship of the Pauline corpus, *The New Testament and Contemporary Perspective*, W. Barclay, ed., Oxford.

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

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

Muller C. (1964) - *Essai de statistique lexicale : L'illusion comique de P. Corneille,* Klincksieck, Paris.

Muller C. (1968) - *Initiation à la statistique linguistique*, Larousse, Paris.

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

Muller C.(1967) - *Etude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille*, Paris, Larousse.

Nishisato S.(1980) - *Analysis of Categorical Data. Dual Scaling and its Application,* Univ. of Toronto Press.

Palermo D., Jenkins J. (1964) - *Word Association Norm*, University of Minnesota Press, Minneapolis.

Pêcheux M. (1969) - *Analyse automatique du discours*, Dunod, Paris.

Peschanski D. (1988) - *Et pourtant, ils tournent. Vocabulaire et stratégie du PCF (1934 - 1936)*, Klincksieck, Paris.

Petruszewycz M. (1973) - L'histoire de la loi d'Estoup-Zipf, *Math. Sciences Hum.*, n°44.

Pitrat J. (1985) - *Textes, ordinateurs, et compréhension*, Eyrolles, Paris.

Plante P.(1985) - La structure des données et des algorithmes en Dérédec, *Revue Québecoise de Linguistique,* 14:2.

Radday Y. T.(1974) - *"And* " in Isaiah, *Revue (R.E.L.O) LASLA N°2* , Liége, p 25-41.

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

Reinert M. (1983) - Une méthode de classification descendante hiérarchique : Application à l'analyse lexicale par contexte, *Les Cahiers de l'Analyse des Données*, 3, Dunod, p 187-198.

Reinert M. (1986) - Un Logiciel d'analyse lexicale, *Les Cahiers de l'Analyse des Données*, 4, Dunod, p 471-484.

Reinert M. (1990) - Alceste, Une méthodologie d'analyse des données textuelles et une Application : Aurélia de Gérard de Nerval, *Bull. de Méthod. Sociol.* n°26, p 24-54.

Romeu L. (1992) - *Approche du discours éditorial de Ya et Arriba (1939 - 1945)*, Thèse Paris 3.

Rugg D. (1941) - Experiments in wording questions, *Public Opinion Quarterly*, 5, p 91-92.

Salem A. (1979) - *Contribution à une méthodologie de la validation en analyse des données textuelles*, Thèse, Université Paris 6.

Salem A. (1982) - Analyse factorielle et lexicométrie. Synthèse de quelques expériences, *Mots N°4* , p 147-168.

Salem A. (1984) - La typologie des segments répétés dans un corpus, fondée sur l'analyse d'un tableau croisant mots et textes, *Les Cahiers de l'Analyse des Données,* Vol IX, n° 4, p 489-500.

Salem A. (1986) - Segments répétés et analyse statistique des données textuelles, Etude quantitative à propos du père Duchesne de Hébert, *Histoire & Mesure, Vol. I- n° 2*, Paris, Ed. du CNRS.

Salem A. (1987) - *Pratique des segments répétés, Essai de statistique textuelle*, Klincksieck, Paris.

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

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

Salton G., Mc Gill M.J. (1983) - *Introduction to Modern Information Retrieval,* International Student Edition.

Sasaki M., Suzuki T. (1989) - New directions in the study of general social attitudes : trends and cross-national perspectives, *Behaviormetrika*, 26, p 9-30.

Saussure F. de (1915) - *Cours de linguistique générale*, Payot, Genève (rééd. 1986).

Schank, R. C., (Ed.) - *Conceptual Information Processing*, North-Holland, Amsterdam, 1975.

Schuman H. (1966) - The random probe : a technique for evaluating the validity of closed questions, *Amer. Socio. Rev.* n°21, p 218-222.

Schuman H., Presser F. (1981) - *Question and Answers in Attitude Surveys*, Academic Press, New York.

Sekhraoui M. (1981) - *La saisie des textes et le traitement des mots : Problèmes posés, essai de solution,* Mémoire, Ecole des hautes études en sciences sociales, Paris.

Shannon C.E. (1948) - A mathematical theory of communication, *B.S.T.J.*, n° 27.

Simpson E.H. (1949) - Measurement of diversity, *Nature,* 163, p 688.

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

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

Spevack M. (1968) - *A Complete and Systematic Concordance of the Work of Shakespeare*, George Olms, Hildesheim.

Stone C.J. (1974) - Cross-validatory choice and assessment of statistical predictions. *J.R. Statist. Soc., B 36, p 111-147.*

Sudman S., Bradburn N. (1974) - *Response Effects in Survey,* Aldine, Chicago.

Suzuki T. (1989) - Cultural link analysis : its application to social attitudes - A study among five nations, *Bull. of the Int. Stat. Inst.*, 47° Session, vol.1, p 363-379.

Tabard (1974) - *Besoins et aspirations des familles et des jeunes*, Coll. Etudes CAF, n° 16, CNAF, Paris.

Tabard N. (1975) - Refus et approbations systématiques dans les enquêtes par sondage, *Consommation,* n°4, Dunod, p 59-76.

Tenenhaus M., Young F. W. (1985) - An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data, *Psychometrika*, vol 50, p 91-119.

Thisted R., Efron B. (1987) - Did Shakespeare write a newly discovered poem?, *Biometrika*, 74, p 445-455.

Tournier M. (1985a) - Sur quoi pouvons-nous compter ? Hommage à Hélène Nais, *Verbum*.

Tournier M. (1985b) - Texte propagandiste et cooccurrences. Hypothèses et méthodes pour l'étude de la sloganisation, *Mots N°11*, p 155-187.

Tournier M. (1980) - D'ou viennent les fréquences de vocabulaire?, *Mots N°1*, p 189-212.

Van Rijckevorsel J. (1987) - *The application of fuzzy coding and horseshoes in multiple correspondances analysis,* DSWO Press, Leyde.

Van Rijsbergen C.J. (1980) - *Information Retrieval*, 2nd Ed., Butterworths, London.

Von Neumann J. (1941) - Distribution of the ratio of the mean square successive difference to the variance, *Annals of Math. Stat.*, vol.12.

Warnesson I., Parisot P., Bedecarrax C., Huot C. (1993) - Traitements linguistiques et analyse des données pour une exploitation systématique des banques de données, *Revue Française de bibliométrie*, i 21.

Weber R.P. (1985) - *Basic Content Analysis*, Sage, Beverly Hills.

Weil G.E., Salem A., Serfaty M. (1976) - Le livre d'Isaïe et l'analyse critique des sources textuelles, *Revue (R.E.L.O) LASLA , N°2* , Liège.

Wilks Y., Fass D., Guo C., McDonald J. E., Plate T., Slator B. M. (1991) - Providing machine tractable dictionnary tools, *Machine translation,* p 99-154.

Wold S. (1976) - Pattern recognition by means of disjoint principal component models, *Pattern Recognition,* 8, p 127-139.

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

Zipf G. K. (1935) - *The Psychobiology of Language, an Introduction to Dynamic Philology*, Boston, Houghton-Mifflin.

Bibliographie complémentaire

On trouvera ci-dessous une liste (non-exhaustive) d'***ouvrages généraux*** ou de ***thèses,*** pour la plupart en langue française, qui complètent les références précédente du point de vue des domaines de recherche abordés et des domaines connexes (Statistique et analyse des données, linguistique et analyse de discours, méthodes d'enquêtes, applications).

Andreewsky A., Fluhr C. (1973) - *Apprentissage, analyse automatique du langage, application à la documentation,* Documents de linguistique quantitative, n° 21, Dunod, Paris.

Antoine J. (1982) - *Le Sondage, outil du marketing*, Dunod, Paris.

Balpe J. P. (1990) - *Hyperdocuments, hypertextes, hypermedias,* Eyrolles, Paris.

Bar-Hillel Y.(1964) - *Language and Information*, Addison-Wesley,

Baudelot C., Establet R. (1989) - *Le niveau monte,* Seuil, Paris.

Bourdieu P. (1982) - *Ce que parler veut dire*, *l'économie des echanges linguistiques,* Fayard, Paris.

Bouroche J.M., Saporta G. (1980) - *L'analyse des données*, coll."Que sais-je", n°1854, PUF, Paris .

Bourques G., Duchastel J. (1988) - *Restons traditionnels et progressifs, pour une nouvelle analyse du discours politique*, Boréal, Montréal.

Brian E. (1984) - *Analyse des données lexicométriques,* Rapport Credoc / D.G.T.

Caillez F., Pagès J.P. (1976) - *Introduction à l'analyse des données*, S.M.A.S.H., Paris.

Cibois P. (1984) - *Analyse des données en sociologie*, P.U.F. Paris.

Cliff A.D. and Ord J.K. (1981) - *Spatial Processes : Models and Applications*, Pion, London.

Daoust F. (1990) - *SATO: Système de base d'analyse de textes par ordinateurs*, ATO-UQAM, Montréal.

Deroo M., Dussaix A. M. (1980) - *Pratique et analyse des enquêtes par sondage*, P.U.F., Paris.

Dodge Y. (1993) - *Statistique. Dictionnaire encyclopédique*, Dunod, Paris.

Ducrot O., Todorov T. (1972) - *Dictionnaire encyclopédique des sciences du langage,* Ed. du Seuil, Paris.

Eissfeldt O. (1965) - *The Old Testament. An introduction*, transl. by P.R. Ackroyd, Oxford.

Fénelon J.P. (1981) - *Qu'est-ce-que l'analyse des données?*, Lefonen, Paris.

Fuchs C. (1993) - *Linguistique et traitements automatiques des langues*, Hachette, Paris.

Gardin J-C. et coll. (1981) - *La logique du plausible : essai d'epistémologie pratique*, Ed. de la Maison des sciences de l'homme, Paris.

Gardin J-C. (1971) - *Les analyses de discours*, Delachaux & Niestlé, Neuchâtel.

Georgel A. (1992) - Classification statistique et réseaux de neurones formels pour la représentation des banques de données documentaires,Thèse - Université Paris-7.

Ghiglione R. (1986) - *L'homme communicant,* Armand Colin, Paris.

Ghiglione R., Matalon B. (1991) - *Les enquêtes sociologiques,* Armand Colin, Paris.

Grangé D., Lebart L. (Ed.) (1993) - *Traitements statistiques des enquêtes,* Dunod, Paris.

Groves R.M. (1989) - *Survey Errors and Survey Costs*, J. Wiley, New-York.

Lafon P., Lefevre J., Salem A., Tournier M. (1985) - *Le Machinal*, *Principes d'enregistrement informatique des textes*, Publications de l'INaLF, Klincksieck, Paris.

Lebart L., Morineau A., Fénelon J.P. (1979) - *Traitement des données statistiques*. Dunod, Paris.

Lebart L., Morineau A., Tabard N. (1977) - *Technique de la description statistique,* Dunod, Paris.

Lelu A. (1993) - *Modèles neuronaux pour l'analyse documentaire et textuelle*, Thèse, Univ. Paris 6.

Maingueneau D. (1976) - *Initiation aux méthodes de l'analyse du discours*, Hachette, Paris.

Marchand P. (1993) - *Engagement politique et rationalisation - Analyse psychosociale du discours militant,* Thèse de l'Université de Toulouse II.

Meynaud H., Duclos D. (1985) - *Les Sondages d'opinions,* La Découverte, Paris.

Muller P., (1991) - *Vocabulaire et rhétorique dans Etudes socialistes de Jean Jaurès*, Thèse, Paris 3.

Prost A. (1974) - *Vocabulaire des proclamations électorales de 1881,1885 et 1889*, PUF, Paris.

Rey A. (1970) - *La Lexicologie*, Klincksieck, Paris.

Robin R. (1973) - *Histoire et Linguistique,* Armand Colin, Paris.

Rouanet H., Le Roux B. (1993) - *Analyse des données multidimensionnelles,* Dunod, Paris.

Rouault J. (1986) - *Linguistique automatique : Applications documentaires,* Berne P. Lang.

Roux M. (1985) - *Algorithmes de classification*, Masson, Paris.

Saporta G. (1990) - *Probabilité, analyse des données et statistique*, Technip, Paris.

Searle, J.R. (1979) - *Sens et expression : étude de la théorie des actes de langages.* Editions de Minuit, Paris.

Tournier M. (1975) - *Un Vocabulaire ouvrier en 1848, essai de lexicométrie*, Thèse d'état, multigr., ENS de Saint-Cloud.

Volle M. (1980) - *Analyse des données,* Economica, Paris.

Wagner R.-L. (1970) - *Les vocabulaires français*, tomes 1 et 2, Didier, Paris.

Wisbey R. A. (Ed) (1971) - *The computer in Literary and Linguistic Research*, Cambridge Univ. Press, Cambridge.

Index des auteurs