Retour aux origines de la statistique textuelle : Benzécri et l'école française d'analyse des données

Valérie Beaudouin

Retour aux origines de la statistique textuelle : Benzécri et l'école française d'analyse des données

Valérie Beaudouin

I3 (Institut Interdisciplinaire de l'Innovation), UMR 9217 CNRS – Telecom Paris Tech
46 rue Barrault, 75013 Paris, France
valerie.beaudouin@telecom-paristech.fr

Abstract

In this article, we have attempted to trace the history of the statistical analysis of textual data, focusing on the influence of Benzécri's work and school, and to make explicit their theoretical positions, clearly opposed to AI and to Chomskyan linguistics. After a presentation of the intellectual project, as an inductive approach to language based on the exploration of corpora, we present the principles of correspondence analysis, which is the main method developed in the Data Analysis School, used for corpus analysis but also for many other types of datasets. Then, we will focus on textual data analysis. Based on the fact that software programmes have played a major role in the use of these statistical techniques, we shall examine a selection of these, display their specificities and their underlying theoretical bases.

Résumé

Cet article revient sur une des deux branches à l'origine de la statistique textuelle, l'école d'analyse des données « à la française », dont Jean-Paul Benzécri peut être considéré comme l'initiateur. Après avoir explicité les orientations théoriques de l'analyse des données, et le rôle joué par une approche inductive du langage, nous présentons rapidement les principes de l'analyse des correspondances. Ensuite, nous explorerons l'application de l'analyse des données aux corpus de textes en montrant le rôle joué par les logiciels dans la diffusion de cette approche.

Mots-clés : analyse des données, analyse de correspondance, statistitique textuelle, sociologie des sciences.

Introduction

De nombreux travaux en « humanités digitales » utilisent des outils de statistique textuelle. Le « rebranding » des Humanities Computing en Digital Humanities opéré au milieu des années 2000, la modernisation des interfaces des logiciels font parfois perdre de vue que les méthodes utilisées sont anciennes et qu'elles ont un ancrage épistémologique qui leur est propre.

Dès la fin des années 60, les potentialités offertes par l'outil informatique ont favorisé l'émergence d'un champ de recherche très actif autour de l'analyse « automatique » des textes s'appuyant sur la numérisation des corpus (à commencer par la base Frantext constituée par l'Inalf pour le Trésor de la Langue Française), le développement d'algorithmes de calculs statistiques et la puissance de calcul[1].

À côté des travaux en intelligence artificielle, qui exploraient l'idée d'une machine capable de dialoguer en langage naturel, de comprendre les textes et donc de les générer, ou même de les

[1] Je remercie Michèle Audin, Saadi Lahlou, Ludovic Lebart, Olivier Martin et Jacques Roubaud pour leurs apports et éclairages sur les travaux de Benzécri.

traduire, à base de règles ou d'heuristiques, des approches inductives se sont développées pour explorer les textes, avec des ambitions théoriques moindres mais sans doute avec plus d'efficacité. La visée de ces approches était de révéler des phénomènes, des régularités dans les corpus de textes et d'en inférer des lois.

Un champ de recherche s'est constitué avec ses laboratoires, ses revues, ses livres de référence, ses colloques, ses controverses, ses courants… Il a la particularité de regrouper des chercheurs provenant de différentes disciplines (littérature, linguistique, politique, sociologie…). La pluridisciplinarité et la diversité des objets de recherche sur lesquelles sont mobilisées ces méthodes en est un des traits caractéristiques, qui provient du langage comme instrument commun.

La diffusion de ces méthodes dans les sciences sociales est liée à l'engagement de chercheurs qui ont consacré une partie importante de leur activité à développer et diffuser des logiciels mettant en œuvre ces méthodes.

Deux courants majeurs peuvent être distingués : celui de la statistique lexicale porté par Charles Muller et celui de l'analyse des données dans lequel Jean-Paul Benzécri a joué un rôle clef, sachant que les croisements ont été nombreux entre ces deux écoles.

Dans ce papier, nous revenons sur le rôle qu'a joué l'école française d'analyse des données dans le développement de l'analyse statistique des données textuelles, en nous concentrant sur l'influence du travail de Benzécri et de son école, et en rendant explicite les positions théoriques, clairement opposées aux postulats de l'Intelligence Artificielle ou de la linguistique chomskyenne.

Après une présentation du projet intellectuel défendant une approche inductive du langage fondée sur l'exploration de corpus, nous reviendrons sur les principes de l'analyse des correspondances, qui a été la méthode fondatrice de l'école de l'analyse des données, utilisée pour les corpus de textes mais aussi pour tout autre type de données. Ensuite, nous en viendrons à l'analyse des données textuelles (réponses à des questions ouvertes, corpus d'articles de presse, de textes littéraires, etc.) Partant du constat que les logiciels ont joué un rôle majeur dans l'usage de ces méthodes statistiques, nous examinerons une sélection d'entre eux, en soulignant leurs spécificités et leurs fondements épistémologiques.

Ce faisant, nous sommes confrontés à la question de la dénomination de ce champ qui a beaucoup évolué. Pour la lisibilité du propos, nous reprendrons comme terme générique analyse des données textuelles (ADT), le terme utilisé dans le colloque emblématique de cette communauté les JADT (Journées Internationales d'Analyse des Données Textuelles – Textual Data Statistical Analysis), même si le terme le plus utilisé aujourd'hui est text-mining.

Les origines de l'analyse des données textuelles

À partir du milieu des années 1960, Jean-Paul Benzécri, ses collègues et leurs étudiants ont introduit et développé une série de méthodes statistiques, plus tard désignées sous le label "Analyse des Données", qui peuvent être considérées comme des précurseurs des méthodes de fouille de données. Les méthodes peuvent être appliquées à tous les types de données qu'elles soient numériques ou textuelles.
Jean-Paul Benzécri, né en 1932, ancien élève de l'École Normale Supérieure, a obtenu son doctorat en mathématiques (topologie) en 1955 à l'Université de Princeton sous la direction du mathématicien Henri Cartan. Il a commencé sa carrière à l'Université de Rennes en tant que professeur adjoint en 1960. En 1965, il est promu professeur à l'ISUP, l'Institut de statistique de l'Université de Paris, où il a passé le reste de sa carrière (Armatte, 2008). Son

intérêt pour la linguistique s'exprime dans ses enseignements à Rennes, dont l'un s'intitule *Linguistique Mathématique. Il* y déploie un formalisme à base de morphismes pour décrire la formation syntaxique des phrases à partir d'ensemble de mots, en mobilisant des structures d'imbrication. La dernière section du polycopié de son cours (Benzécri, 1964), intitulée « Langage de programmation », récapitule les notations « en vue du traitement sur machine des problèmes de linguistique ».. L'horizon est bien l'utilisation de ce langage formel par une machine.. Ses premiers travaux à Paris témoignent d'un abandon de la mathématique formaliste au profit de la statistique multidimensionnelle. Comment s'est opéré chez Benzécri le passage aux méthodes statistiques d'analyse des données reste un mystère. Sans doute la multiplicité des champs d'application possibles (linguistique, psychologie…) de la statistique multidimensionnelle l'a-t-elle emporté sur le plaisir du formalisme mathématique.
Toujours est-il que Benzécri est considéré comme le père de l'École française d'analyse de données, domaine dans lequel se sont déployées ses recherches dès la fin des années 60.

En un mot, le principe de l'analyse des correspondances, expression centrale dans les publications de Benzécri, consiste à mettre les données dans des tableaux rectangulaires, sous forme de matrice, afin d'être en mesure d'appliquer des méthodes de calcul statistique. L'analyse des correspondances, initialement adaptée aux tables de contingence (ou tableaux croisés qui représentent la fréquence de distribution de deux variables qualitatives) a été étendue à d'autres types de matrices, sous forme de tableaux disjonctifs (analyse des correspondances multiples) et peut être utilisée sur toute sorte de matrice avec des nombres positifs. L'idée est d'identifier le modèle de relation entre les deux ensembles d'éléments mis dans le tableau. Dans le cas d'un corpus de texte, les tableaux contiennent des textes en ligne et des mots en colonne ; à l'intersection d'une ligne et d'une colonne se trouve un indicateur de la présence ou de la fréquence du mot dans le texte.

Les algorithmes d'analyse de données permettent de synthétiser l'information contenue dans les matrices. L'analyse factorielle tente de réorganiser les matrices de sorte que les premières dimensions contiennent la quantité maximale d'information ; les méthodes de classification permettent l'identification de sous-groupes homogènes de textes et de mots. L'école de l'analyse des données combine souvent les méthodes d'analyse factorielle et de classification.

Les origines de l'analyse des données

Dans *Histoire et préhistoire de l'analyse des données*, écrit en 1975 et publié en 1982, Benzécri retrace les origines de l'analyse des données, explique l'analyse des correspondances et tente de la mettre en perspective par rapport aux travaux contemporains (Benzécri, 1982). Après un chapitre sur la « science du hasard », il distingue les trois étapes qu'il considère comme décisives dans les progrès de la statistique multidimensionnelle (ou analyse des données multivariées) et qui ouvrent la voie à l'analyse des correspondances : la biométrie de Quételet à Pearson, les travaux de Sir Ronald Fisher et enfin la psychométrie (de Spearman à Guttman). Ce faisant, il dessine une histoire personnelle des origines de l'analyse des correspondances (Armatte, 2008) à laquelle il consacre la dernière partie du livre. Bien qu'il souligne l'originalité et l'homogénéité introduite par sa méthode, il présente également les travaux contemporains relevant du même domaine.

Les origines de l'analyse des données remontent au début du siècle. Les psychologues ont été les pionniers dans l'exploration des données multidimensionnelles et de l'analyse factorielle, et ont une influence décisive sur les travaux de Benzécri, comme le montrent les travaux d'Olivier Martin (Martin, 1997). Spearman, psychologue britannique, en analysant les liens entre les résultats scolaires des élèves et leurs aptitudes mentales (Spearman, 1904), a estimé

qu'il avait montré l'existence d'un *facteur* général d'aptitude ou d'intelligence, qui sera plus tard désigné par la lettre G. Ultérieurement, non seulement un mais plusieurs facteurs seront recherchés à partir de données de plus en plus nombreuses. Résumer la complexité des données par quelques facteurs constitue l'originalité de l'approche.

L'analyse des correspondances, branche de l'analyse factorielle dont Benzécri revendique la paternité, s'appuie selon lui sur les avancées de Fisher au cours des années 1940 (Fisher, 1940). Pour Benzécri, en explorant l'analyse discriminante, Fisher a développé l'équation de base de l'analyse des correspondances. Puis, en 1961, Kendall et Stuart en élaborant les méthodes canoniques pour l'analyse des tableaux de contingence (Kendall et Stuart, 1961) ont pu calculer les paramètres utilisés pour tester l'hypothèse d'indépendance entre les lignes et les colonnes. Benzécri explique qu'il a utilisé le nom de l'analyse de correspondance pour la première fois en 1962 et qu'il a présenté la méthode en 1963 au Collège de France (Benzécri, 1982, p. 101). L'analyse des correspondances est un terme générique utilisé pour rendre compte de tout un ensemble de méthodes de traitement de tableaux de données.

Benzécri connaissait les travaux des psychométriciens et était en contact avec Shepard chez Bell Labs qui venait d'introduire l'analyse des proximités (ou MultiDimensional Scaling MDS) : il présente ses travaux à l'été 1965 aux Bell Labs, mais il n'y aura pas de suite à ces échanges avec les États-Unis (Rouanet, 2008).

Contribution principale de Benzécri

L'analyse des correspondances est souvent présentée dans les travaux anglo-saxons comme une adaptation aux données qualitatives de l'Analyse en Composantes Principales (Greenacre et Blasius, 2006 ; Hill, 1974 ; Murtagh, 2005) ou comme étant très proche de l'analyse des proximités ou MultiDimensional Scaling (Hill, 1974). Comment peut-on caractériser l'originalité de la contribution de Benzécri à l'analyse multidimensionnelle ?

Son apport principal est d'avoir mis en évidence toutes les propriétés algébriques de la méthode et montré son intérêt : tester l'indépendance des lignes et des colonnes, mais surtout décrire comment les données s'éloignent de cette hypothèse en représentant par des "proximités" les associations existant entre les lignes et les colonnes (Diday et Lebart, 1977). Le plan factoriel, qui visualise les proximités entre les individus et les variables (sur deux axes factoriels), devient un des principaux supports pour la construction de l'interprétation par le chercheur. L'accent mis sur les méthodes de visualisation est une clé pour comprendre le succès de l'analyse des données « à la française ». Un nuage de données dans un espace multidimensionnel se voit projeté dans un « espace » à deux dimensions accessible et interprétable par le chercheur, même s'il ne connait pas les subtilités de la méthode. La projection ouvre à la construction du sens. Cette approche diffère des approches hypothético-déductives (largement répandues dans la littérature anglo-saxonne) beaucoup plus austères en termes de présentation des résultats.

Benzécri ne s'intéressait pas uniquement aux algorithmes ; l'analyse des données représentait pour lui un *cadre général d'analyse* et ce point constitue sa seconde contribution majeure. Ce cadre porte sur l'ensemble des étapes de l'analyse des données. En amont, pour la préparation des données, des méthodes sont proposées pour transformer n'importe quel type de données en une table avec des nombres positifs analysable par la méthode de l'analyse des correspondances. On a donc un algorithme statistique principal qui s'applique à tout tableau moyennant quelques transformations. En aval, ce cadre offre un ensemble d'aides à l'interprétation : le calcul des contributions permet de mesurer la qualité de la représentation

d'un point sur le plan, la projection des variables supplémentaires donne des éléments d'interprétation complémentaires. L'association de l'analyse des correspondances avec les méthodes de classification (en particulier la classification ascendante hiérarchique) permet d'approfondir la compréhension des données et facilite l'interprétation. « Ainsi une méthode unique dont le formulaire reste simple est parvenue à incorporer des idées et des problèmes nombreux apparus d'abord séparément, certains depuis plusieurs décennies. », écrit (Benzécri, 1982, p. 102). Comme l'explique Ludovic Lebart (communication personnelle), l'idée était de proposer « des panoplies bien étalonnées (comme le duo analyse des correspondances - classification) pour appréhender la complexité, sans chercher à créer systématiquement une nouvelle méthode pour chaque nouveau problème ».
Plutôt qu'une profusion de méthodes, difficiles à appréhender pour des non statisticiens, l'analyse des données offre un cadre unifié d'analyse avec le couple analyse des correspondances et classification. Celui-ci est clairement pensé pour les utilisateurs et les praticiens, en accordant une place importante à la présentation des résultats.

Benzécri a conçu et diffusé un cadre global pour l'analyse de "grandes tables", mais il est avant tout guidé par une ambition théorique et philosophique, qui nous intéresse directement ici.

Philosophie de Benzécri

En tant que mathématicien tourné vers la linguistique, Benzécri s'est intéressé aux méthodes d'analyse des données non pas comme outils pour la psychologie (discipline qui a été à l'origine des développements les plus nombreux), mais bien comme outil de recherche en linguistique : « C'est principalement en vue de l'étude des langues que nous nous sommes engagés dans l'analyse factorielle des correspondances » (Benzécri, 1981). Il avait pour ambition théorique d'ouvrir les portes à une nouvelle linguistique à une époque qui était dominée par la linguistique générative. Il s'oppose à la thèse idéaliste de Chomsky qui, dans les années 60, considère que seule une modélisation abstraite permet de mettre à jour les structures linguistiques. Contre cette thèse, Benzécri propose une méthode inductive d'analyse des données linguistiques « avec à l'horizon l'ambitieux étagement des recherches successives ne laissant rien dans l'ombre des formes, du sens et du style » (Benzécri, 1981, p. X). En ce sens, il est assez proche des objectifs de distributionalistes comme Bloomfield et Harris qui escomptaient construire les lois de la grammaire à partir de corpus d'énoncés (Bloomfield, 1973 ; Harris, 1954).

De son point de vue, les méthodes qu'il développait étaient plus efficaces pour la compréhension en profondeur de la langue que les travaux en linguistique statistique menés par Guiraud et Muller (Guiraud, 1954; Muller, 1977), qu'il trouvait intéressants mais trop centrées sur le vocabulaire (Benzécri, 1981, p. 3).

> Nous proposons une méthode portant sur les problèmes fondamentaux qui intéressent un linguiste. Et cette méthode (…) effectuera une abstraction quantitative, en ce sens que partant de tableaux de données les plus divers, elle construira, par le calcul, des quantités qui pourraient mesurer des entités nouvelles, situées à un niveau d'abstraction supérieur à celui des faits recensés d'abord (Benzécri, 1981, p. 4).

En identifiant les facteurs, il ne fait pas de doute qu'une opération d'abstraction a été réalisée. L'ordinateur ne donne pas de nom, ni de sens à ces entités qu'il a abstraites ; c'est au spécialiste d'apporter son interprétation. L'ambition épistémologique de Benzécri est de redonner de la valeur à la démarche inductive et de s'opposer par là même à l'idéalisme :

> Car nous condamnons que, de principes reçus à la légère, l'idéalisme prétende par une dialectique, fut-elle sous l'emprise des mathématiques, tirer des conclusions sûres ;

> puis à cette déduction a priori nous opposons l'induction qui, a posteriori, des faits observés veut monter à ce qui les ordonne (Benzécri, 1968, p. 11).

Il critique les théories idéalistes qui posent l'existence d'un modèle et en vérifient approximativement la pertinence avec l'observation. Si cette démarche est encore vaguement acceptable pour la physique, elle ne l'est pas du tout pour l'économie. En effet, il n'existe pas de situation expérimentale suffisamment schématique pour satisfaire l'économiste. Il doute qu'il soit possible de réduire un objet complexe en une combinaison d'objets élémentaires, « car l'ordre du composé vaut plus que les propriétés élémentaires des composants » (Benzécri, 1968, p. 16).

L'objectif qu'il lui paraît possible d'atteindre avec l'analyse des données est de pouvoir dégager « de la gangue des données le pur diamant de la véridique nature » (on notera au passage les relents d'idéalisme). Le passage des données aux entités abstraites, de l'ombre à la lumière, est rendu possible à ses yeux grâce à l'analyse des données et au « novius organum » qu'est l'ordinateur : « Les moyens de calculs nouveaux permettent de confronter des descriptions complexes d'un grand nombre d'individus, pour aboutir à placer ceux-ci sur des cartes planes ou spatiales, image fidèle et accessible à l'intuition de la nébuleuse des données initiales » (Benzécri, 1968, p. 21). Auxiliaire de la synthèse, l'ordinateur est un outil mental : après l'organum d'Aristote et le *Novum Organum* de Bacon, n'est-il pas le *Novius Organum*, « l'outil le plus nouveau » ? (Benzécri, 1968, p. 24).

> Somme toute, on voit comment l'analyse affranchit des idées a priori. Des données aux résultats, l'ordinateur, insensible aux espérances comme aux préjugés du chercheur, procède sur la base ample et solide de faits définis et acceptés d'abord dans leur ensemble, puis dénombrés et ordonnés suivant un programme qui, parce qu'il ne sait pas comprendre, ne sait pas non plus mentir. (Benzécri, 1968, p. 24)

> Enfin, parmi toutes les idées a priori, souvent contradictoires, que chaque problème suscite en si grand nombre, un choix opportun s'opère : bien plus, l'idée qui a posteriori, après examen statistique des données, semble avoir été a priori fort naturelle ne se serait pas toujours présentée d'elle-même à l'esprit. (Benzécri, 1968, p. 24)

Les instruments de la diffusion

A Paris, autour de Benzécri s'était constitué un réseau de chercheurs dans le domaine de l'analyse des données, qui ont contribué à de nombreuses publications collectives. Les principales publications se composent de traités, de manuels et d'une histoire de l'analyse des données.

Le traité d'analyse des données est constitué de deux volumes : le premier (Benzécri et coll., 1973) est dédié à la taxinomie et examine toutes les méthodes de classification, le second (Benzécri et coll., 1973b) porte sur l'analyse des correspondances.

Histoire et préhistoire de l'analyse des données, publié en 1982 (Benzécri, 1982), après une publication progressive des chapitres dans les *Cahiers de l'analyse des données* à partir de 1975, propose une généalogie de l'analyse des correspondances et situe l'originalité de l'approche par rapport à d'autres travaux. Pour Benzécri, ce livre est une introduction à la série des manuels *Pratique de l'analyse des données* publiés au début des années 1980 : le premier volume est consacré à l'analyse des correspondances (Benzécri et coll., 1980), avec dans l'édition de 1984, l'ajout d'un chapitre sur la classification. Le second est plus théorique

et le troisième est consacré à la linguistique: *Pratique de l'analyse des données. 3 Linguistique et lexicologie* (Benzécri et coll., 1981).

Chacun de ces volumes impliquait un grand nombre de collaborateurs, 30 par exemple pour *Linguistique et lexicologie*.

Les Cahiers de l'analyse des données fondés sur une idée de Michel Jambu (Armatte, 2008), se présentent comme le principal débouché pour les articles du domaine, et accueille évidemment les travaux portant sur l'analyse des textes. Cette revue sera publiée de 1976 à 1997.

Un élément qui distingue le travail de Benzécri et de ses collaborateurs tient à l'organisation de ces ouvrages collectifs qui proposent tous : des articles théoriques, des exemples d'applications provenant de domaines très diversifiés (sciences naturelles et humaines) et des programmes informatiques qui pourront être réutilisés par les lecteurs. Cette structure est un élément qui explique la diffusion importante des méthodes. Les procédures statistiques étaient explicites et le code partagé (une approche *open source* avant l'heure). À la fin des années 1980, plusieurs procédures d'analyse de correspondance ont été incluses dans les principaux progiciels statistiques de l'époque, notamment SPSS, BMDP et SAS (Greenacre & Blasius, 2006). On en retrouve aujourd'hui dans "R", le package open source pour le calcul statistique (Husson, Lê et Pagès, 2009).

A l'ISUP, Benzécri et ses collègues encadraient un flux important d'étudiants, estimé à 180 étudiants en master par an et 40 doctorats (Armatte, 2008) qui ont également contribué à la diffusion de méthodes.

Influence

La contribution de Benzécri (un cadre unifié pour l'analyse des données orienté vers les utilisateurs) a grandement aidé à la diffusion de l'analyse des correspondances en France dans les sciences de la nature comme dans les sciences sociales, où elles continuent régulièrement d'être utilisées –d'où le nom d'analyse des données « à la française ». En sociologie, Pierre Bourdieu a joué un rôle important de diffusion à mesure que son influence dans les sciences sociales augmentait. La théorie de Bourdieu a été profondément influencée par l'analyse des correspondances. L'espace social analysé comme un champ de tensions et d'oppositions selon deux dimensions liée l'une à l'intensité du capital, l'autre à sa décomposition en capital culturel et économique n'est pas sans affinité avec un plan factoriel à deux dimensions. La *Distinction* théorise à partir des résultats de l'analyse des correspondances (Bourdieu, 1979). Rouanet explique que « pour Bourdieu, l'analyse des correspondances multiples fournit une représentation des deux faces complémentaires de l'espace social, à savoir l'espace de catégories - selon les termes de Bourdieu, l'espace des propriétés - et l'espace des individus. Représenter les deux espaces est devenu une tradition dans la sociologie de Bourdieu (Rouanet, 2006).

L'analyse des données souffre cependant d'une reconnaissance (très) limitée dans les publications anglophones qui favorisent les approches hypothético-déductives. La dimension purement exploratoire, visant à faire ressortir les formes et les modèles à partir de données, n'a pas la même légitimité que d'autres approches ; trop de description, pas assez d'explication, et surtout pas assez d'hypothèses a priori. Il est cependant bien connu que les procédés hypothético-déductifs sont fragiles, en raison de l'ordre de causalité pré-établi au moment où une hypothèse est posée. Même à l'heure actuelle, les travaux relevant de cette école restent rares dans la littérature anglo-saxonne et réussir à publier un article utilisant de

la statistique textuelle dans une revue américaine relève de la performance (Beaudouin et Pasquier, 2016 ; Schonhardt-Bailey, Yager et Lahlou, 2012) : convaincre les relecteurs de l'intérêt de la démarche et faire comprendre comment fonctionne l'outil ne va pas de soi. Revenons à présent à l'analyse des correspondances, qui peut être considérée comme le cœur de l'innovation de Benzécri.

L'analyse des correspondances

La présentation que nous donnons de l'analyse des correspondances est fondée sur le chapitre dédié dans *Histoire et préhistoire de l'analyse des données* (Benzécri, 1982, p. 101-131), sur le chapitre introductif au volume *Linguistique et Lexicologie* (Benzécri, 1981, p. 73-135) et sur le manuel publié en anglais (Benzécri, 1992).

L'analyse des correspondances est une méthode qui donne une représentation géométrique des associations entre deux ensembles d'éléments mis en correspondance dans un tableau. Les tests statistiques sont généralement utilisés pour rejeter l'hypothèse d'indépendance des variables ou des attributs. L'approche de Benzécri vise à représenter, d'une manière géométrique, à quel point l'indépendance des observations et des variables n'est pas vérifiée. Pour Benzécri, l'indépendance entre les lignes et les colonnes manque d'intérêt scientifique ; ce qui est intéressant est précisément la mesure de l'écart par rapport à l'indépendance.

D'une table de correspondance aux profils

L'analyse des correspondances nécessite tout d'abord la transformation de données brutes, ici un corpus, en un tableau de contingence. Ce tableau croise deux ensembles d'éléments, un ensemble I (individus ou observations) et un ensemble J (variables ou attributs). Au croisement d'une ligne et d'une colonne, on obtient le nombre d'occurrences de l'attribut j dans l'observation i, soit k (i, j). Deux exemples permettront de clarifier. Supposons que nous partions d'un corpus de pièces de théâtre. Nous pouvons construire une table où I est l'ensemble des pièces de théâtre et J le vocabulaire présent dans les pièces. Dans ce cas, k (i, j) représentera le nombre d'occurrences du mot j dans la pièce i. Dans le tableau, il y a autant de lignes, m, que d'éléments de l'ensemble I (pièces), et autant de colonnes, n, qu'il y a de mots dans l'ensemble J. Les lignes sont les individus et les colonnes sont les variables. Prenons un autre exemple de (Benzécri, 1982, p. 103). Afin d'analyser la distribution des noms et des verbes dans un corpus, nous pouvons construire une table où les lignes correspondent aux noms et les colonnes aux verbes : à l'intersection d'une ligne et d'une colonne, est indiqué le nombre de phrases où le nom est le sujet du verbe.

Pour comparer la distribution des deux ensembles d'éléments, les profils de ligne et de colonnes sont calculés : $f^i_j = k(i,j)/ki.$ (où $ki. = \sum_{j=1}^{n} k(i,j)$, i.e. la somme des fréquences de la ligne i). Le profil de i sera f^i_J, un vecteur formé par la séquence des f^i_j ($f^i_J = \{f^i_j \mid j \in J\}$)

Symétriquement, le profil d'un élément j sera $f^j_I = \{f^j_i \mid i \in I\}$.

Représenter la distance entre les profils

Comment comparer les profils des différents éléments (lignes ou colonnes de la table) ? Nous avons besoin d'un espace et d'une distance. L'analyse des correspondances utilise un espace

euclidien et la distance du chi-deux, ce qui constitue une caractéristique distinctive de l'analyse des correspondances. La distance entre i et i' est définie comme suit :

$$d^2(i, i') = \sum \{(f_{ij} - f_{i'j})^2 / f_j \mid j \in J\}$$

Chaque élément i (respectivement j) de l'ensemble I est représenté par son profil et se voit attribuer une masse proportionnelle au total de la ligne. L'ensemble des profils fiJ constitue un nuage N(I) dans un espace multidimensionnel. Respectivement, un nuage N(J) est défini pour les profils fjI.

L'idée principale est de réduire la complexité du nuage et de trouver une façon de représenter la plus grande part de l'information dans un espace de dimension inférieure. Pour cela, le centre de gravité du nuage est calculé et la dispersion du nuage autour de son centre est mesurée (inertie). Ensuite, les axes factoriels, ou axes principaux de dispersion, sont construits. Les points sont projetés sur ces axes, et leurs coordonnées sur ces axes sont appelées facteurs. Dans le plan défini par les deux premiers axes, nous pouvons avoir la meilleure projection du nuage (qui minimise la perte d'informations).
Un trait distinctif de l'analyse de correspondance est la symétrie parfaite des rôles assignés aux deux ensembles I et J. Cela permet la représentation simultanée des deux nuages sur les mêmes axes.

L'objectif principal est de visualiser la distance entre les observations ou entre les attributs, à savoir l'écart par rapport à une distribution aléatoire. L'algorithme produit un ensemble d'« aides à l'interprétation » qui permettent au chercheur d'interpréter correctement les résultats.
Souvent, l'analyse des correspondances est combinée avec la classification hiérarchique en utilisant les coordonnées des éléments sur les axes factoriels.

Instruments au service des sciences humaines et sociales

Les innovations sont rarement le fait d'individus isolés. Elles émergent et se diffusent au travers de réseaux, de collectifs, d'institutions dans lesquels les individus se rencontrent, échangent, où les innovations circulent, sont discutées, améliorées, critiquées. La diffusion des méthodes de statistique textuelle ne fait pas exception à la règle.

Des laboratoires, des revues, des conférences se sont progressivement mis en place qui ont stimulé les échanges, mais aussi les débats. Dans ce domaine spécifique de la recherche, les outils informatiques sont devenus des acteurs majeurs dans la transmission des méthodes et dans l'organisation de ce réseau : ils portent la marque de la perspective théorique dans laquelle s'inscrivent leurs auteurs, ils incarnent des prises de positions, ils s'adaptent en fonction des échanges avec les parties prenantes (les utilisateurs principalement) (Akrich, Latour et Callon, 1988). En bref l'adoption des théories passe par les logiciels qui se reconfigurent en fonction des usages. Ces logiciels cristallisent d'une part les débats théoriques au sein de la communauté et d'autre part posent la question des enjeux économiques, ou plus modestement commerciaux, liés à ces méthodes.

En effet, la diffusion de ces méthodes a été soutenue par des enjeux économiques : dans le secteur des études et du marketing, la possibilité de faire du quantitatif sur des données qualitatives, autrement dit la possibilité d'introduire de la mesure dans l'analyse du discours a représenté une opportunité intéressante. Si ces outils permettaient à des cabinets d'accéder à

des financements, comment fallait-il penser ces outils, comme des outils de recherche ou des outils commerciaux ?

Après avoir passé rapidement en revue les lieux institutionnels qui ont contribué à l'animation de la vie scientifique de cette spécialité, nous nous focaliserons sur quelques logiciels emblématiques de la statistique textuelle, en montrant comment chaque outil porte la marque du milieu dans lequel il s'est développé (la discipline, le type de corpus et de questions que se posent les chercheurs) et comment ce milieu interagit avec les objectifs propres aux chercheurs.

Lieux

Après l'université de Rennes, l'ISUP à Paris devint un des lieux principaux d'élaboration et de diffusion de l'analyse des données. D'éminents statisticiens et chercheurs du domaine ont participé au séminaire de Benzécri. Le champ de recherche est beaucoup plus large que l'analyse des données textuelles, mais le public comprenait des chercheurs clés tels que Ludovic Lebart, déjà docteur, qui ont accordé une attention particulière au traitement des textes.

Le Crédoc (Centre de recherche pour l'observation des conditions de vie) a longtemps été un centre actif dans le domaine. Ludovic Lebart, chercheur CNRS, y a travaillé de nombreuses années (1971-1988) en mettant au point et en dirigeant l'enquête Aspiration et Conditions de vie des Français. Avec Alain Morineau, il est à l'origine du développement du logiciel Spad (système portable d'analyse de données) (Lebart et Morineau, 1982) et de son extension dédiée aux textes Spad.T (Lebart et al., 1989) qui s'appuie sur les travaux de recherche d'Éric Brian (Brian, 1986). Les logiciels seront diffusés librement jusqu'à 1987 par l'association Cesia, ce qui favorise la diffusion de ces outils pionniers de la fouille de données et de textes. Spad a été conçu pour l'analyse des enquêtes quantitatives et Spad T pour l'analyse des réponses aux questions ouvertes. L'implémentation des algorithmes est guidée par ce contexte d'usage spécifique : l'enquête quantitative. Un centre de calcul partagé avec le Cepremap, centre de recherche en économie, et relié au Circé (*Centre* Inter Régional de *Calcul* Électronique d' Orsay) permet de mettre au point et tester les outils sur des données et devient le point de rencontre d'un collectif de statisticiens comme Jean-Pierre Fénelon (Fénelon, 1981) ou Nicole Tabard (pionnière dans les systèmes d'information géographique) (Lebart, Morineau et Tabard, 1977).

Quelques années plus tard, dans le département « Prospective de la consommation », Saadi Lahlou prend la relève en développant un axe de recherche sur l'analyse lexicale pour l'analyse des représentations sociales (Beaudouin et Lahlou, 1993 ; Lahlou, 1992 ; Yvon, 1990). Il contribue à la diffusion de ces méthodes dans le champ de la psychologie sociale. Au Crédoc, on utilise Spad, mais aussi Alceste, développé par Max Reinert (Reinert, 1987, 1990), qui permet d'analyser des corpus de textes autres que les questions ouvertes. La statistique lexicale devient un outil pour l'étude des représentations sociales (Lahlou, 1998) et conduit à une réflexion sur les processus d'interprétation (Lahlou, 1995). Une collaboration établie avec Reinert permet de développer les outils sur plateforme Unix et de traiter de plus grands volumes de textes. L'adaptation d'Alceste à des environnements Mac, Windows et Unix favorise la diffusion de l'outil en sciences sociales en France et la traduction des dictionnaires permet son utilisation pour l'analyse d'autres langues que le français (anglais, portugais…).

Le laboratoire « Lexicologie et textes politiques » a été créé en 1967 à l'école Normale supérieure de St-Cloud. Il connaitra différents rattachements au fil du temps et une partie de

l'activité se retrouve aujourd'hui dans le laboratoire Icare de l'ENS de Lyon tandis qu'une autre s'est retrouvée à Paris III. L'analyse du discours politique constitue la colonne vertébrale de l'unité avec une branche de réflexion méthodologique qui explore la place que peuvent occuper les machines en lexicométrie pour l'analyse des textes. Pierre Lafon (Lafon, 1984) et André Salem (Salem, 1987) s'occupent plus spécifiquement de la mise en place d'outils d'analyse statistique : « ces deux linguistes-mathématiciens […] étaient conseillés dans leurs méthodes par les maitres de l'« analyse des données » (Jean-Paul Benzécri) et du calcul des probabilités (Georges-Théodule Guilbaud) » (Tournier, 2010). C'est dans ce laboratoire que nait la réflexion sur la linguistique de corpus (Habert, Nazarenko et Salem, 1997) et plus précisément sur les systèmes d'annotation et d'enrichissement des textes. Longtemps dirigé par Michel Tournier, ce laboratoire est aussi à l'origine de la revue Mots (Mots, Ordinateur, Textes, Société) qui paraît pour la première fois en octobre 1980 et continue d'être publiée sous le nom *Mots. Les langages du politique*. Le logiciel Lexico d'André Salem est un des outils nés dans ce contexte. Il se distingue des autres logiciels sur deux points principaux : l'identification des segments répétés, séquences de mots qui sont une approximation de la syntaxe, (Salem, 1987) et une méthode de mesure de l'évolution chronologique du vocabulaire (Salem, 1995). L'analyse de correspondance permet de mesurer les distances entre sous-parties du corpus et de visualiser, si cela est pertinent, l'évolution chronologique du vocabulaire. Un attachement aux discours politiques et syndicaux marque la spécificité de ce laboratoire.

À l'université de Nice, un autre laboratoire créé en 1980 se met en place qui accorde une place importante à la machine. Etienne Brunet, littéraire passionné d'informatique, dès la fin des années 60, constitue un pôle de recherche actif à l'université autour de ce laboratoire qui deviendra *Bases, Corpus, Langage*. Brunet conçoit Hyperbase un outil particulièrement bien adapté à l'analyse de très grands volumes de textes littéraires (Brunet, 1988), mais aussi aux textes politiques (Mayaffre, 2000), ce qui crée des ponts avec le laboratoire de St Cloud. Cet outil connaît une diffusion importante dans la communauté des chercheurs en sciences humaines. Hyperbase inclut une procédure d'analyse factorielle provenant des programmes développés par Fénelon et ses collègues. Celle-ci donne une visualisation des distances entre les mots et sous parties d'un corpus. Par exemple, la Figure 1 présente le résultat d'une analyse sur une matrice croisant les œuvres de Rabelais avec l'emploi des pronoms personnels.

Proche de l'Inalf, ce laboratoire explore d'importants corpus issus de la base Frantext. Depuis 2001, il s'est doté de sa propre revue Corpus, dont la responsable éditoriale est Sylvie Mellet. Deux volumes regroupent les principaux articles publiés par Etienne Brunet (Brunet, 2009, 2011)

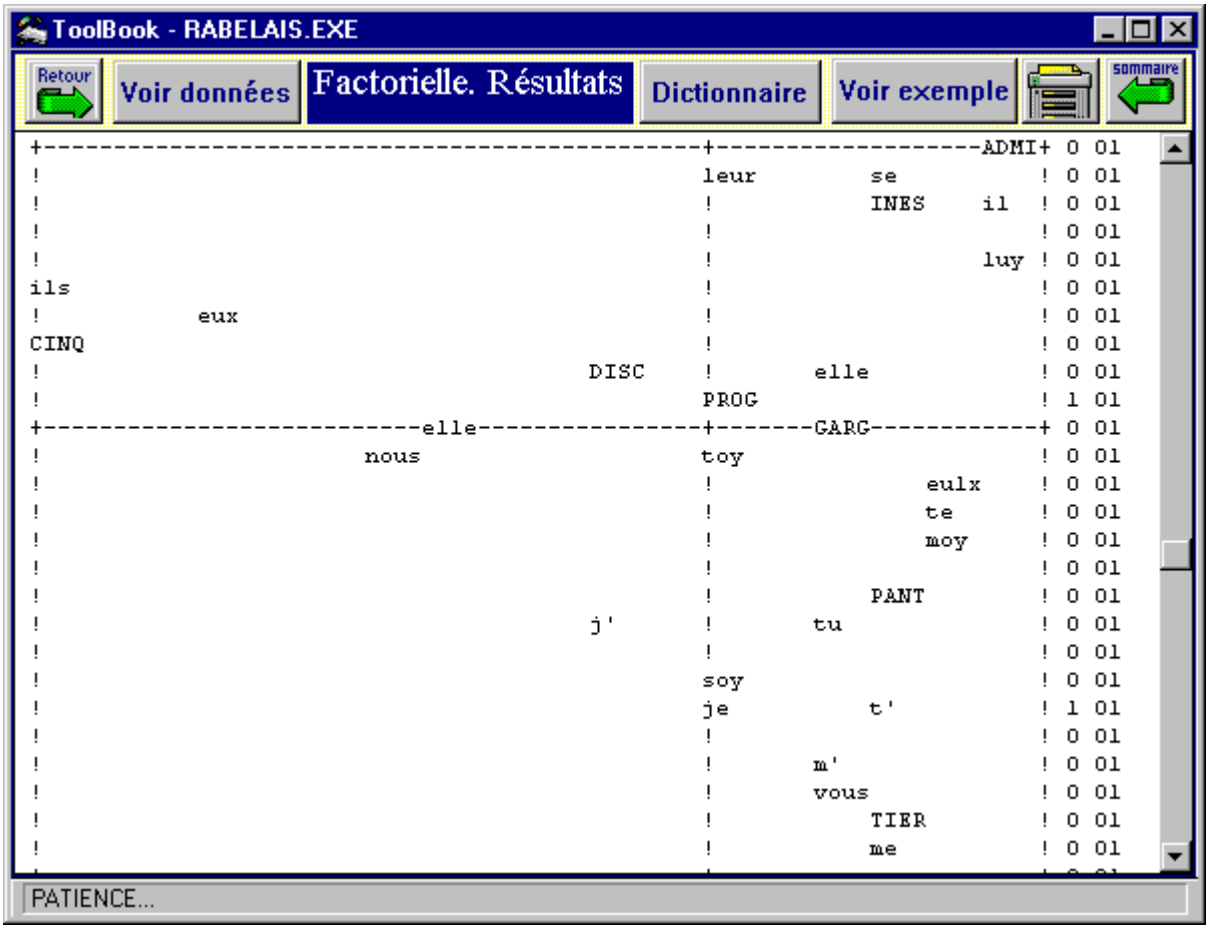

Figure 1. Analyse factorielle dans Hyperbase (http://ancilla.unice.fr/~brunet/PUB/hyperwin/analyse.html)

D'autres lieux ont également joué un rôle important : le centre scientifique d'IBM animé par Marchotorchino, l'équipe autour de Dominique Labbé à Grenoble et d'autres lieux à l'étranger autour de Sergio Bolasco à la Sapienza à Rome…

Les journées internationales d'analyse des données textuelles, organisées tous les deux ans depuis 1991, constituent un point de rassemblement mais aussi d'élargissement de la communauté des chercheurs du domaine. À dominante francophone, elle accueille chercheurs italiens et espagnols issus du même courant. La publication systématique des actes et leur mise en ligne dans la revue en ligne Lexicometrica d'André Salem et Serge Fleury (http://lexicometrica.univ-paris3.fr/jadt/) de Paris III permettent la constitution d'un corpus consistant d'expériences.

Le livre de Lebart et Salem, *Analyse statistique des données textuelles*, publié chez Dunod en 1988 et réédité en 1994 sous le titre, puis traduit en anglais, s'est imposé comme le manuel de référence dans le domaine (Lebart, Salem et Berry, 1998 ; Lebart et Salem, 1988, 1994).

Logiciels

Les publications ont joué un rôle décisif dans la diffusion des méthodes d'analyse des textes, expliquant les algorithmes, montrant les usages possibles sur des corpus, multipliant les exemples d'application. Mais la diffusion des usages s'est faite principalement à travers les outils qui ont été les vecteurs majeurs de l'appropriation de méthodes parfois regardées avec méfiance par le monde des sciences humaines et sociales. Dans chaque cas, nous soulignerons

les spécificités des logiciels : préparation des corpus (sélection des textes et des variables), algorithmes de traitement et interprétation. Nous nous limiterons à deux logiciels, Spad T et Alceste qui ont été les plus innovants dans la tradition de l'école d'analyse des données de Benzécri.

Spad T

Comme nous l'avons vu, Spad T est l'extension de Spad (Système portable pour l'analyse des données) qui permet l'analyse des réponses à des questions ouvertes dans les enquêtes. L'unité d'analyse (chaque ligne du tableau) correspond à l'individu enquêté caractérisé par ses réponses aux questions fermées et ouvertes. Mais elle peut aussi correspondre au regroupement des individus selon des variables comme l'âge, le niveau de diplôme, tous les individus ayant la même modalité de variable constituent *un* texte (une ligne dans le tableau). Par exemple, la Figure 2 représente l'analyse de correspondance d'un tableau croisant les mots employés dans les réponses à une question ouverte[2] avec les individus regroupés selon leur niveau d'éducation.

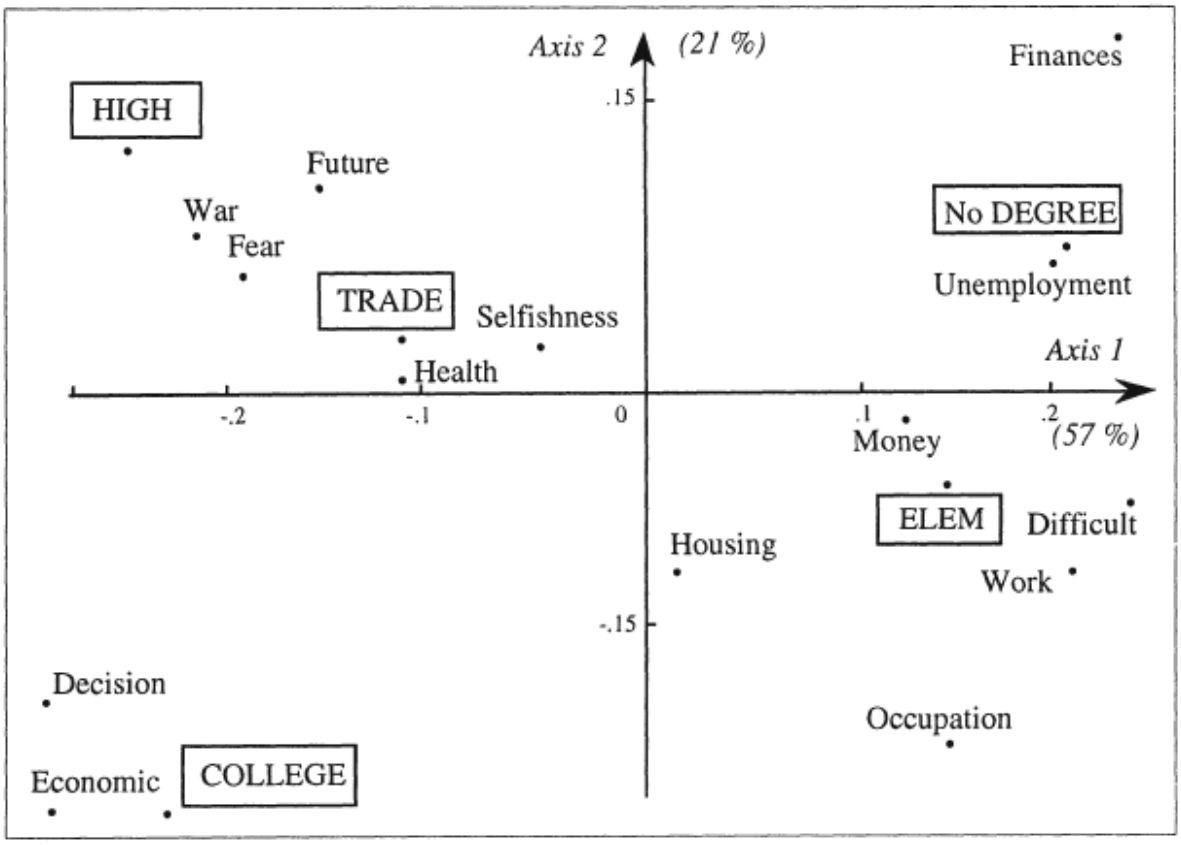

Figure 2. Proximités entre mots et niveaux d'éducation (Lebart et al., 1998, p. 52)

Pour les mots entrant dans les tableaux (i.e. constituant les colonnes du tableau), SpadN procède de la façon suivante : il conserve les formes graphiques, les mots tels qu'ils apparaissent dans le texte, et ne procède à aucune forme de lemmatisation (ramener les formes graphiques à leur racine, à leur entrée dans le dictionnaire) ; il élimine avec un seuil de fréquences les mots rares et les mots très courts (moins de 3 lettres par exemple) ce qui est

[2] La question était : « *Quelles* sont les *raisons* qui, selon vous, *peuvent faire* hésiter une *femme* ou un *couple* à avoir un enfant? » (Lebart, Salem et Berry, 1998).

une façon d'éliminer les mots grammaticaux (déterminants, pronoms…). Comme les réponses ont déjà été réduites tout au long de la chaîne qui va de l'enquête au traitement (l'enquêteur qui prend note de la réponse a tendance à ne garder que l'essentiel, l'opérateur de saisie peut aussi simplifier) et qu'on travaille sur des corpus très redondants, ce « nettoyage » à la hache a peu d'impact sur les résultats.

Spad T propose toute une palette de procédures d'analyse des données. La démarche la plus classique est d'effectuer une analyse factorielle des correspondances sur le tableau croisant réponses en ligne et mots employés en colonne, puis une classification ascendante hiérarchique sur la base des cordonnées factorielles. Le principe consiste à rapprocher deux à deux les réponses qui se ressemblent le plus par le vocabulaire utilisé et à remonter progressivement jusqu'à un regroupement de toutes les réponses. Un calcul statistique permet de déterminer le nombre optimal de classes.

Pour aider à l'interprétation, on peut obtenir pour chaque classe le vocabulaire spécifique (les mots significativement plus présents dans la classe que dans les autres) et les réponses les plus caractéristiques. Comme Spad T est articulé avec Spad, on peut aussi avoir les modalités des autres variables de l'enquête sur- ou sous-représentées dans la classe grâce à une procédure très pratique, le « tamis ». Ceci permet de caractériser une classe et d'aider à construire une interprétation dans l'esprit très exploratoire de l'analyse des données.

En bref, Spad T[3] est particulièrement bien adapté pour un contexte d'usage spécifique (l'enquête quantitative) et des types de corpus bien cadrés (les réponses à des questions ouvertes). Les algorithmes d'analyse des données et d'aides à l'interprétation y sont très robustes et le contexte d'usage fait que les procédures frustes pour réduire le vocabulaire ne posent pas problème. L'originalité de l'approche tient à la possibilité d'inclure des métadonnées (i.e. des informations sur les individus ayant produit les textes) et ensuite de caractériser les textes par les caractéristiques de ceux qui les ont produits.

Rappelons qu'un des débats qui a animé la communauté portait sur la question de la lemmatisation : certains justifiant le fait de travailler sur des formes graphiques « brutes » (Lafon, 1984), d'autres considérant que la lemmatisation (réduction des variations et levée des ambiguïtés) était un préalable indispensable à tout traitement comme le montre le plaidoyer de Muller en introduction au livre de Lafon. Les partisans de la lemmatisation considéraient que c'était une étape indispensable pour éviter l'ambiguïté des formes (distinguer les homonymes) tandis que les opposants regrettaient la perte d'information : les indications de genre, de pluriel, de temps et de personne étant significatives. Ce débat a produit des discussions animées pendant les JADT jusqu'à ce que les outils donnent la possibilité de traiter au choix les données brutes ou lemmatisées.

Alceste

La méthodologie ALCESTE (Analyse des Lexèmes Cooccurrents dans les Énoncés Simples d'un Texte) a été mise au point par Max Reinert (1983, 1993) et s'inspire du courant de l'analyse des données de Benzécri, dont Reinert fut l'élève. Les préoccupations de Reinert constituent cependant une orientation particulière. Il considère un corpus comme une suite d'énoncés élémentaires produits par un sujet-énonciateur. Ainsi le texte est modélisé dans un tableau qui contient en ligne les énoncés, qui portent la marque du sujet énonciateur et en

[3] Ludovic Lebart a mis en ligne le logiciel DTM-VIC (http://www.dtmvic.com/), qui a les mêmes propriétés que Spad pour l'analyse des données numériques et textuelles.

colonne les mots ou lexèmes, qui renvoient à des objets du monde (sans aucunement préjuger de la " réalité " de ces objets). L'objectif est de faire émerger des « mondes lexicaux ».

> Un monde lexical, est donc à la fois la trace d'un lieu référentiel et l'indice d'une forme de cohérence liée à l'activité spécifique du sujet-énonciateur que l'on appellera une logique locale (Reinert, 1993, p.9).

Grâce aux procédures statistiques, qui rapprochent des énoncés employant le même type de lexique, la méthode permet d'identifier différents mondes lexicaux, qui pourront être interprétés comme des " visions du monde ". Par exemple, dans son étude sur Aurélia de Nerval, Reinert identifie trois types de mondes en classant les énoncés : le monde imaginaire, le monde réel et le monde symbolique qui portent chacun la marque d'un certain rapport au monde du narrateur (Reinert, 1990).

Décrivons Alceste brièvement. En entrée du logiciel, on a un texte ou un ensemble de textes, décrits par certaines variables illustratives, qui portent sur la situation de communication (caractéristiques du locuteur, de la situation énonciative…). En sortie, on obtient une typologie des énoncés qui constituent le corpus. Un énoncé est défini comme étant un point de vue du sujet sur le monde. Le processus de regroupement est basé sur la similitude / dissimilitude des mots à l'intérieur des énoncés. Chaque groupe d'énoncés est interprété comme un monde lexical qui reflète une vision du monde.

Cette orientation théorique a des conséquences sur tout le déroulement de l'analyse. Commençons par les unités textuelles. Reinert cherche à identifier la notion d'énoncé : un point de vue sur le monde qui porte la trace d'un sujet. Mais comment définir de façon automatique la notion d'énoncé alors qu'elle ne coïncide pas forcément avec la notion de phrase et qu'aucun signe de ponctuation ne permet clairement de l'identifier ? Comme il n'y a pas de solution satisfaisante à ce problème, Reinert propose une heuristique : deux découpages possibles du corpus en unités textuelles en faisant varier la longueur des unités. Ainsi un tableau contiendra en ligne les unités textuelles du premier découpage, un second celui du second. La comparaison des analyses obtenues avec les deux découpages en gardant ce qui est invariant permettra de contourner le problème initial, puisque le résultat sera alors indépendant du découpage (cf. infra).

Quels éléments du vocabulaire sont conservés dans les colonnes des tableaux ? Comme pour Spad T un seuil de fréquence permet d'éliminer les mots rares (qui ne jouent aucun rôle quand on travaille sur des cooccurrences). Une procédure de lemmatisation réduit les mots à leur racine et permet surtout d'identifier les parties du discours (noms, verbes, pronoms…). Étant donnée la perspective adoptée par Reinert, ne seront conservés dans l'analyse que les mots « pleins », qui renvoient à un référent et pas les mots grammaticaux, qui constituent le ciment du texte.

Sur ces matrices qui croisent segments de textes et mots lemmatisés, Alceste effectue une classification descendante hiérarchique, un algorithme original mis au point en 1983 (Reinert, 1983) et particulièrement bien adapté aux matrices hypercreuses (plus de 90% de « 0 »). L'idée consiste à prendre l'ensemble des segments de textes et à les répartir dans deux groupes, de façon à ce que chaque groupe soit le plus homogène possible en termes de vocabulaire utilisé et le plus éloigné possible de l'autre groupe. Techniquement, la classification s'appuie sur une analyse factorielle : le premier axe factoriel est calculé, on fait glisser un hyperplan sur l'axe qui coupe le nuage en deux de manière à ce que l'inertie interclasse soit maximale et l'inertie intraclasse minimale. On réitère la procédure sur le plus

grand groupe restant jusqu'à atteindre le nombre de classes demandées. Le processus de classification est itératif et conduit à une typologie.

C'est là qu'intervient à nouveau l'heuristique proposée par Reinert : sur chacun des tableaux constitués, le logiciel fera une classification descendante hiérarchique, puis, il comparera les deux analyses réalisées et conservera les classes de la typologie qui sont les plus stables dans les deux analyses. Cela permet en plus d'avoir une procédure pour optimiser le nombre de classes terminales. Par exemple, la figure 3 montre le résultat de la double classification sur Aurélia (Reinert, 1990). Finalement trois classes sont conservées : 8 <->9, 10 <->11 and 11<->10.

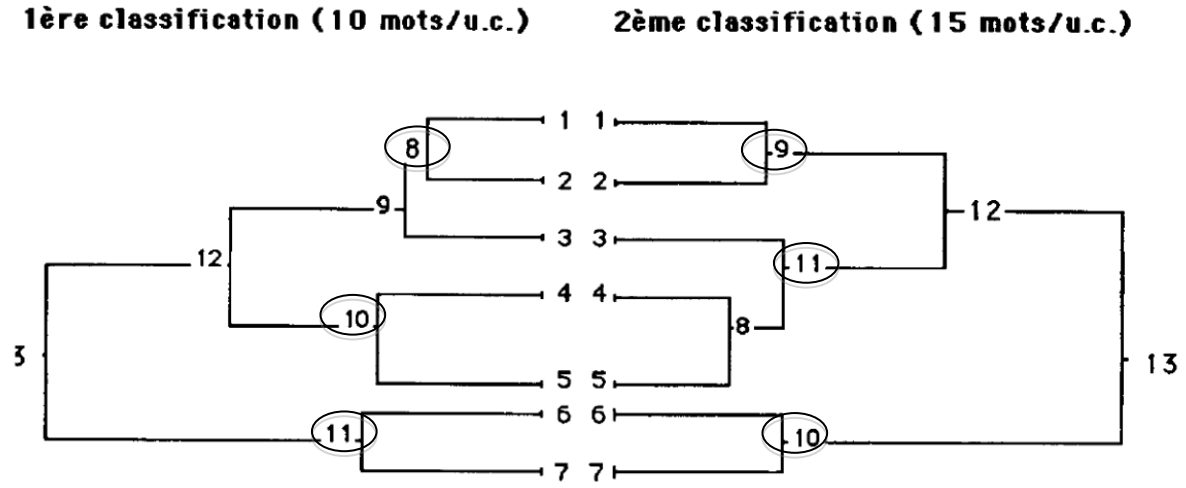

Figure 3. Double classification descendante hiérarchique avec Alceste (Reinert, 1990)

Chaque classe de la typologie est caractérisée par une liste de mots qui constitue le vocabulaire spécifique de la classe par rapport à l'ensemble du corpus, par les segments de textes les plus caractéristiques de la classe, et par les modalités des variables illustratives les mieux représentées. L'ensemble peut être visualisé sur un plan d'analyse factorielle. Ces aides à l'interprétation permettront de caractériser le champ lexico-sémantique propre à chaque classe et de voir quels sont les facteurs de production externes qui expliquent le mieux ses spécificités.

Conclusion et perspectives

Jean-Paul Benzécri et ses collègues ont développé un cadre général pour l'analyse des données qui inclut la phase de préparation des données, la phase d'analyse avec les algorithmes de traitements des matrices (analyse des correspondances et méthodes de classification) et la phase d'interprétation avec l'ensemble des aides à l'interprétation. Cette approche inductive a été pensée pour l'étude des phénomènes linguistiques, mais elle s'applique aussi bien aux données quantitatives qu'aux données textuelles.

Ces méthodes ont été largement utilisées dans un grand nombre de disciplines. L'efficacité de ces approches pour l'exploration des données et pour la construction d'hypothèses de

recherche a été largement prouvée par le très grand volume de publications mobilisant l'analyse des données.

En linguistique, l'analyse des données textuelles a ouvert la voie à une étude systématique de la langue basée sur des corpus, la linguistique de corpus, avec l'hypothèse que les textes recueillis sur le terrain, dans des contextes naturels, sont le meilleur moyen pour faire émerger les règles de la langue.

Bien que la recherche en statistique et en informatique ait beaucoup évolué, notamment avec les techniques d'apprentissage automatique, il est intéressant de noter que ces techniques « anciennes » sont encore utilisées par les chercheurs en sciences sociales. Cela peut être dû à un simple retard dans la maitrise de ces techniques de « machine learning » par les sciences sociales, mais peut-être simplement à leur efficacité comme outil d'exploration de corpus. C'est que ces techniques ont l'avantage sur le machine learning et les approches hypothético-déductives de permettre de construire des hypothèses explicites à partir des corpus (grounded theory), une puissance heuristique pour le moment inégalée.

Les outils d'analyse de données textuelles ont dû être adaptés aux grands corpus. Alors qu'un corpus contenant 2.000 réponses pouvait être considéré comme volumineux dans les années 1980, les corpus actuels peuvent être constitués de millions de textes et de mots. Les langages de programmation utilisés dans les années 70-90 sont parfois devenus obsolètes, comme le Fortran, et les logiciels étaient souvent limités dans leur taille. Leur adaptation à des corpus beaucoup plus volumineux a parfois demandé une réécriture complète du code. Par exemple, le logiciel Alceste de Max Reinert a été entièrement reprogrammé par Pierre Ratinaud, et rebaptisé Iramuteq (http://www.iramuteq.org/), avec une interface plus moderne, des fonctionnalités nouvelles et la capacité de traiter des volumes beaucoup plus importants. Cette réécriture n'est pas sans soulever des problèmes de propriété intellectuelle, dans la mesure où les algorithmes de classification utilisés sont pratiquement identiques : la mention « méthode Reinert » souligne la filiation. De la même manière, TXM développé dans le cadre du projet Textométrie (http://textometrie.ens-lyon.fr/), réutilise et modernise d'anciens algorithmes, tout en proposant un enrichissement des données lexicales avec des traits morphosyntaxiques, phonétiques ou autres. Dans de tels cas, il n'y a pas eu de changements fondamentaux apportés aux algorithmes d'analyse de données eux-mêmes ce qui est une preuve de leur efficacité.

Les méthodes décrites ci-dessus s'appuient principalement sur l'analyse de la répartition des fréquences et des cooccurrences de mots dans les textes. L'unité principale d'analyse est le mot dans son contexte textuel. Mais, depuis toujours, la transformation d'un texte en un "sac de mots" a paru réductrice et l'introduction de traits descriptifs plus fins a paru nécessaire. Benzécri et ses collègues (Benzécri 1981) avaient d'ailleurs imaginé l'introduction d'annotations alors que les technologies n'étaient pas encore au point. Les méthodes se sont progressivement améliorées grâce aux outils de traitement du langage naturel : analyse morphosyntaxique, sémantique voire prosodique. À un texte on peut associer une série de traits descriptifs concernant différents niveaux linguistiques. Dans cette perspective influencée par (Biber, 1989) qui vise à construire de manière inductive des typologies de textes à partir de traits de description, se structure le champ de la linguistique de corpus (Habert, Nazarenko et Salem, 1997). Prenons quelques exemples de son application : le projet TypTex (Habert et al., 2000), la caractérisation des genres selon des traits morphosyntaxiques (Malrieu et Rastier, 2001 ; Rastier, 2011) ou l'articulation des traits phonétiques, morphosyntaxiques, prosodiques et sémantiques pour caractériser le vers (Beaudouin, 2002). En bref, les approches intègrent les avancées en traitement automatique des langues, en ne se

limitant plus aux mots mais en intégrant les différents niveaux d'analyse linguistique (phonétique, syntaxique, sémantique…). L'analyse de correspondance et la classification sont désormais appliquées à des tableaux beaucoup plus riches (plus de textes et beaucoup plus de variables descriptives).

La nouvelle frontière pour l'analyse de données textuelles est l'analyse des documents issus du web. Le texte a été le premier médium à entrer dans le monde numérique avant l'image, le son ou la vidéo. Il est donc tout à fait naturel que les méthodes statistiques aient été appliquées à des textes avant de porter sur d'autres contenus. Aujourd'hui, la numérisation a atteint l'ensemble des productions culturelles et de plus en plus de contenus naissent dans le numérique. Cela ouvre de nouvelles questions de recherche. Il n'est plus possible de réduire le Web à du simple texte, il est donc nécessaire d'enrichir les méthodes actuelles avec des éléments descriptifs qui relèvent des particularités du Web (multimédia, hypertextuel, imbriqué à la réception, dynamique) et de développer des approches qui combinent différentes méthodes, la statistique textuelle n'étant qu'une technique parmi d'autres.

Bibliographie

Akrich M., Latour B. et Callon M. (1988), « A quoi tient le succès des innovations ? »,.

Armatte M. (2008), « Histoire et Préhistoire de l'Analyse des données par J.P. Benzécri : un cas de généalogie rétrospective. », *Journ@l Electronique d'Histoire des Probabilités et de la Statistique*, vol. 4, n°2, pp. 1‑22.

Beaudouin V. (2002), *Mètre et rythmes du vers classique - Corneille et Racine*, Paris, Champion, coll. Lettres numériques.

Beaudouin V. et Lahlou S. (1993), *L'analyse lexicale : outil d'exploration des représentations*, Paris, CRÉDOC, Cahier de Recherche, 146 p.

Beaudouin V. et Pasquier D. (2016), « Forms of contribution and contributors' profiles : An automated textual analysis of amateur on line film critics »,.

Benzécri J.-P. (1964), « Linguistique mathématique », Rennes.

Benzécri J.-P. (1968), « La place de l'a priori, "Organum" », *in Encyclopaedia Universalis*, pp. 11‑24.

Benzécri J.-P. (1980), *Pratique de l'analyse des données. Analyse des correspondances & classification. Exposé élémentaire.*

Benzécri J.-P. (1982), *Histoire et préhistoire de l'analyse des données*, Paris, Dunod.

Benzécri J.-P. (1992), *Correspondence Analysis Handbook*, New-York, Basel, Hong Kong, Marcel Dekker, Inc;

Benzécri J.-P. et coll. (1973a), *L'analyse des données. 1 La taxinomie*, Paris, Bordas.

Benzécri J.-P. et coll. (1973b), *L'analyse des données. 2 L'analyse des correspondances*, Paris, Bordas.

Benzécri J.-P. et coll. (1981), *Pratique de l'analyse des données, Linguistique et lexicologie*, Paris, Dunod.

Biber D. (1989), « A typology of English texts », *Linguistics*, vol. 27, pp. 3‑43.

Bloomfield L. (1973), *Language (1st ed 1933)*, London, Allen and Unwin.

Bourdieu P. (1979), *La distinction. Critique sociale du jugement.*, Les éditions de Minuit.

Brian E. (1986), *Techniques d'estimation et méthodes factorielles, exposé formel et application aux traitements de données lexicométriques*, Thèse de doctorat, Orsay.

Brunet E. (1988), *Le vocabulaire de Hugo*, Slatkine-Champion, 484 p.

Brunet E. (2009), *Comptes d'auteurs - Tome 1. Etudes statistiques, de Rabelais à Gracq.*, Paris, Honoré Champion.

Brunet É. (2011), *Ce qui compte. Ecrits choisis, tome II. Méthodes statistiques*, Paris, Honoré Champion.

Diday E. et Lebart L. (1977), « “L’analyse des données” », *La Recherche*, n°n°74, pp. 15‑25.
Fisher R.A. (1940), « The precision of discriminant function », *Annals of Eugenics*, vol. 10, pp. 422‑429.
Greenacre M. et Blasius J. (2006), *Multiple Correspondence Analysis and Related Methods*, Boca Raton, Chapman & Hall/CRC.
Guiraud P. (1954), *Les caractères statistiques du vocabulaire*, Paris, PUF.
Habert B., Illouz G., Lafon P., Fleury S., Folch H., Heiden S. et Prevost S. (2000), « Profilage de textes : cadre de travail et expérience . », *JADT’2000. 5èmes Journées Internationales d'Analyse statistique des Données Textuelles, Lausanne, 9-11 mars 2000*.
Habert B., Nazarenko A. et Salem A. (1997), *Les linguistiques de corpus*, Paris, Armand Colin/Masson, 240 p. p.
Harris Z.S. (1954), « Distributional structure », *Word*, vol. 10, n°23, pp. 146‑162.
Hill M.O. (1974), « Correspondence Analysis: A Neglected Multivariate Method », *Journal of the Royal Statistical Society*, vol. 23, n°3, pp. 340‑354.
Husson F., Lê S. et Pagès J. (2009), *Analyse des données avec R*, Rennes, Presses Universitaires de Rennes, 224 p.
Kendall M.G. et Stuart A. (1961), *The Advanced Theory of Statistics, Volume 2: Inference and Relationship.*, Hafner Publishing Company.
Lafon P. (1984), *Dépouillements et statistiques en lexicométrie*, Genève-Paris, Slatkine-Champion.
Lahlou S. (1992), *Si/alors : « bien manger » ? - Application d’une nouvelle méthode d’analyse des représentations sociales à un corpus constitué des associations libres de 2000 individus.*, Paris, CRÉDOC.
Lahlou S. (1995), « Vers une théorie de l’interprétation en analyse statistique des données textuelles » L Lebart S. Bolasco A. Salem (eds) (dir.), *JADT 1995 : III Giornate internazionali di Analisi Statistica dei Dati Testuali*, vol. vol. I, pp. 221‑228.
Lahlou S. (1998), *Penser manger. Alimentations et représentations sociales*, Paris, PUF, 241 p.
Lebart L. et Morineau A. (1982), « SPAD: Système Portable pour l’Analyse des Données. »
Lebart L., Morineau A. et Bécue Bertaut M. (1989), « Spad.T : Système portable pour l’analyse des données textuelles. »
Lebart L., Morineau A. et Tabard N. (1977), *Méthodes et logiciels pour l’analyse des grands tableaux*, Paris, Dunod.
Lebart L. et Salem A. (1988), *Analyse statistique des données textuelles*, Paris, Dunod, xxx p.
Lebart L. et Salem A. (1994), *Statistique textuelle*, Paris, Dunod, 342 p. p.
Lebart L., Salem A. et Berry L. (1998), *Exploring Textual Data*, Dordrecht, Boston, Kluwer Academic Publisher, 246 p. p.
Malrieu D. et Rastier F. (2001), « Genres et variations morphosyntaxiques », *TAL*, vol. 42, n°2, pp. 547‑577.
Martin O. (1997), « Aux origines des idées factorielles », *Histoire & Mesure*, vol. Vol. 12, n°n°3-4, pp. 197‑249.
Mayaffre D. (2000), *Le poids des mots. Le discours de gauche et de droite dans l’entre-deux guerre.*, Paris-Genève, Slatkine-Champion.
Muller C. (1992), *Principes et méthodes de statistique lexicale*, Larousse, 1977, réimpression Champion-Slatkine, 1992, 211 p.
Murtagh F. (2005), *Correspondence Analysis and Data Coding with Java and R*, Boca Raton, Chapman & Hall/CRC.
Rastier F. (2011), *La mesure et le grain*, Paris, Honoré Champion.
Reinert M. (1983), « Une méthode de classification descendante hiérarchique : application à

l'analyse lexicale par contexte », *Les cahiers de l'analyse des données*, vol. Vol VIII, n°2, pp. 187‑198.
Reinert M. (1987), « Classification descendante hiérarchique et analyse lexicale par contexte : application au corpus des poésies d'Arthur Rimbaud », *Bulletin de Méthodologie Sociologique*, vol. n°13, n°1, pp. 53‑90.
Reinert M. (1990), « ALCESTE : Une méthodologie d'analyse des données textuelles et une application : Aurélia de Gérard de Nerval », *Bulletin de Méthodologie Sociologique*, n°n°26, pp. 24‑54.
Reinert M. (1993), « Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars », *Langage et société*, n°n°66, pp. 5‑39.
Rouanet H. (2006), « Chapter 5 The Geometric Analysis of Structured Individuals × Variables », *in* Michael Greenacre et Jörg Blasius (dir.), *Multiple Correspondence Analysis and Related Methods*, Boca Raton, Chapman & Hall/CRC, pp. 137‑160.
Salem A. (1987), *Pratique des segments répétés*, Paris, Klincksieck.
Salem A. (1995), « La lexicométrie chronologique. L'exemple du Père Duchesne d'Hébert », *in Langages de la Révolution (1770-1815) (Actes du 4ème Colloque international de lexicologie politique)*, Paris, Klincksieck.
Schonhardt-Bailey C., Yager E. et Lahlou S. (2012), « Yes, Ronald Reagan's Rhetoric was Unique—But Statistically, How Unique? », *Presidential Studies Quarterly*, vol. 42, n°3 (September), pp. 482‑513.
Spearman C. (1904), « "General intelligence" objectively determined and measured », *American Journal of Psychology*, vol. 15, pp. 201‑292.
Tournier M. (2010), « Mots et politique, avant et autour de 1980. Entretien », *Mots. Les langages du politique*, vol. 94, pp. 211.
Yvon F. (1990), « L'analyse lexicale appliquée à des données d'enquête : états des lieux », Paris, CREDOC, Cahier de recherche.