Compte rendu paru dans le numéro 75 de la revue *MOTS*.

Pascal Marchand, *L'Analyse du Discours Assistée par Ordinateur*, Armand Colin, Paris, 1998, 222 p.

Le titre de cet ouvrage se veut résolument généraliste. Pascal Marchand, membre du Groupe de Recherche sur la Parole de l'Université de Paris 8 – Saint-Denis, propose une manière de recension des applications de l'outil informatique à l'analyse du discours et, plus largement à l'ensemble des sciences humaines et sociales qui s'intéressent à l'étude de données verbales.

Le volume présente dans un premier mouvement les " concepts, méthodes et outils ", qui sont illustrés au travers d'un exemple d'application concret dans la deuxième partie.

Les cinq chapitres de la première partie sont autant d'axes de l'analyse discursive sur lesquels l'outil informatique peut apporter des éléments de réponse. En l'espèce, la part belle est faite à la statistique lexicale, le chapitre s'y rapportant représentant près de la moitié de ce premier mouvement.

Le premier chapitre développe succinctement ce que l'auteur nomme les " analyses para-verbales ". Il y aborde quelques aspects de l'analyse conversationnelle, en termes de verbalisation, de successions de tours de parole, de comportements verbaux, puis présente rapidement quelques problématiques relevant de la phonologie, citant des travaux menés sur l'intonation ou des recherches en psychologie sociale et en sociolinguistique comme l'exemple bien connu de Labov et son analyse de la prononciation dans les différentes couches socioprofessionnelles new-yorkaises.

Le deuxième chapitre sur les analyses lexicales, s'appuie sur de nombreux exemples issus de travaux de divers domaines, qui fournissent quelques repères conceptuels, abondamment nourris de citations de Muller, Benveniste, Bronckart, Bally, Saussure, Harris… Ce chapitre qui s'inspire largement de l'ouvrage de Lebart et Salem, revient sur les différentes approches du lexique, psychologique, pragmatique, linguistique puis définit les notions élémentaires d'unité lexicale, d'index, de formes graphiques, avant d'aborder des aspects tels que la richesse du vocabulaire, la connexion et la distance lexicale, les concordances, les segments répétés, les spécificités. Quelques exemples d'application à l'indexation documentaire sont évoqués. Le sous-chapitre intitulé " statistique textuelle ", présente les deux familles de statistique multidimensionnelle que sont les analyses factorielles et les classifications automatiques, fournissant des aides à l'interprétation mais aussi un approfondissement des algorithmes utilisés.

L'auteur clôt ce chapitre en introduisant la problématique des unités à considérer concernant les flexions d'une forme, les mots composés, les locutions et séquences figées. Il aborde ainsi le débat entre " lemmatiseurs " et " formalistes " et les procédés de réductions de formes, notamment ceux utilisés par Alceste. Il revient aussi sur les différentes conceptions du mot en termes d'unités de pensée avec Bally, d'unité de sens et de morphème avec Sapir, en évoquant les travaux de G. Gross, de Bloomfield, et des approches plus récentes.

Consacré aux analyses morpho-syntaxiques, le troisième chapitre montre comment les traitements informatisés peuvent repérer les relations de syntaxe dans la phrase et les parties du discours. Les statuts de l'article et du pronom, des verbes et des adjectifs, des opérateurs et des connecteurs sont notamment exposés et mis en relation avec des études relevant de plusieurs disciplines.

Le chapitre suivant pose les grands principes de l'analyse sémantique, en distinguant une approche consistant à synthétiser le contenu d'un texte pour en livrer une compréhension a posteriori par la classification des énoncés qui le composent, d'une approche a priori consistant à projeter les catégories générales d'une langue sur plusieurs corpus. Les règles ontologiques, dictionnaires thématiques, constructions de catégories sont exposées aux travers de travaux

couvrant le champ des sciences humaines. Une présentation du logiciel Tropes et de ses classes d'équivalence, illustrée d'un exemple d'application clôt le chapitre.

L'auteur achève cette première partie en abordant rapidement la dimension pragmatique du discours, et en soulignant la difficulté de coder une telle approche.

Embrassant les différents niveaux discursifs, il dresse ainsi un état des lieux de l'analyse automatisée du discours, présentant les caractéristiques et les fonctionnalités de quelques-uns des logiciels les plus répandus en sciences humaines, donnant des exemples d'application et d'interprétation, replaçant chaque notion dans son contexte théorique. La terminologie employée pourrait, sur certains points être discutée, si elle ne renvoyait à un vaste champ d'application. Les logiciels Tropes, Hyperbase, Sphinx Lexica, Intex, Spadt… sont présentés de façon pédagogique, ce qui concourt à conférer à l'ouvrage plusieurs niveaux de lecture. Le lecteur rompu aux techniques de la statistique lexicale y trouvera quelques approfondissements concernant les calculs mis en œuvre par telle ou telle application, tandis que le chercheur néophyte y puisera un large panorama méthodologique, mais aussi des liens bibliographiques vers les spécialistes de chaque approche.

Dans la seconde partie la méthodologie laisse place à l'analyse ce qui achève de donner à l'ouvrage sa dimension concrète. A partir des approches évoquées dans la première partie, les auteurs (Pascal Marchand, Brigitte Lecat et Janine Larrue) étudient les interventions de 29 journalistes politiques interrogeant les candidats aux élections européennes de 1994 dans le cadre d'émissions radiodiffusées, posant l'hypothèse que ceux-ci mettent en place des stratégies discursives différentes selon les invités qu'ils ont à interroger.

L'analyse des temps de parole montre notamment que les candidats des listes traditionnelles (P.S, P.C, Génération Ecologie mais plus encore UDF/RPR et FN) investissent l'espace discursif de façon plus importante que les représentants des listes nouvelles.

L'analyse lexicale, appliquée aux seules interventions des journalistes, confirme ces disparités. Les indices de longueur (occurrences), de richesse et de banalité, obtenus au moyen de Sphinx Lexica, puis l'analyse des spécificités, mettent en évidence des comportements sensiblement différents des journalistes face à leurs invités. Une analyse factorielle des correspondances réalisée sur les formes graphiques met en évidence un système d'opposition entre les interventions tenues face aux listes nouvelles comme " Autre Europe " ou " Radicale " et les listes plus traditionnelles (P.S, P.C, RPR/UDF, Verts, FN). Cette hypothèse est confirmée par l'examen des emplois et distributions des pronoms, verbes et prédicats, modalisateurs et joncteurs, adjectifs qualificatifs mené au moyen de Tropes. Une analyse factorielle réalisée sur les 50 catégories " logico-syntaxiques " recensées par le logiciel confirme les clivages mis en lumière par l'analyse factorielle des formes graphiques. L'analyse s'attache enfin à étudier les grands thèmes abordés par les journalistes, tels que déterminés par Tropes. [1]

Cette étude, dont quelques-unes des conclusions ont été présentées notamment lors des Journées Internationales d'Analyse Statistique des Données Textuelles tend à démontrer que le traitement réservé aux candidats se fait moins en fonction de l'idéologie politique que de la notoriété de ceux-ci et de l'attente supposée de l'opinion publique.

Le lecteur aura ainsi pris conscience de l'intérêt des interventions de procédures informatisées, aussi bien au niveau de la verbalisation, du lexique, de la syntaxe, de la sémantique et, pourrions-nous ajouter, de l'énonciation. On y verra aussi la démonstration que l'opposition du quantitatif et du qualitatif n'est pas pertinente, tant il est vrai, ainsi que le souligne l'auteur, que ces deux aspects ne constituent, tout au plus, que deux phases, deux moments de l'analyse.

---

[1] P.Marchand est co-auteur avec Laurence Monnoyer-Smith d'une étude publiée dans Mots n°62 et Lexicométrica portant sur les discours de politique générale prononcés par les Premiers ministres français entre 1974 et 1997. Il souligne notamment, par le biais d'un traitement informatisé mené au moyen de Tropes et Lexico, et suivant une démarche proche de celle évoquée ici, un net recul de la dimension idéologique dans les interventions gouvernementales.

Il n'en reste pas moins que l'outil informatique ne saurait suppléer à une démarche méthodologique rigoureuse, l'établissement d'hypothèses préalables, l'indispensable connaissance du matériau étudié. Il appartient au chercheur, d'acquérir une compétence minimale quant aux procédures de traitement, de constituer un corpus susceptible de permettre une interprétation fiable et d'éviter l'écueil qui consisterait à prêter aux résultats issus de l'ordinateur un caractère de vérité absolue. Car au final, c'est à l'analyste qu'il revient de mener l'interprétation et de valider ses hypothèses ; Comme le souligne l'auteur, " on peut faire dire aux mots autant d'inepties qu'aux chiffres ".