RODOLPHE GHIGLIONE
AGNÈS LANDRÉ
MARCEL BROMBERG
PIERRE MOLETTE

CD-ROM INCLUS

L'ANALYSE AUTOMATIQUE DES CONTENUS

BIBLIOTHEQUE DU CERIST

DUNOD

P S Y C H O S U P

L'ANALYSE AUTOMATIQUE DES CONTENUS

Rodolphe Ghiglione
Agnès Landré
Marcel Bromberg
Pierre Molette

DUNOD

TABLE DES MATIÈRES

INTRODUCTION

Grammaire de discours, analyse du (de) discours, analyse de contenu, autant de termes différents, de théories différentes, de cloisonnements, autant de fascinations. Ainsi que le dit Jenny (1997) : « Le paysage de l'analyse textuelle française semble aussi varié et diversifié que nos paysages, nos vins et nos fromages ; aussi compartimenté que nos vieux cantons ; aussi séduisant par certains côtés que notre langue chérie. » Nous disions déjà, pour notre part, en 1985 : « La proximité des objets "discours", "langue" et "parole" oblige à une réflexion sur les passages possibles d'un domaine à l'autre [...] La problématique à laquelle répond la formation d'analyse de discours est fondamentalement celle d'une recherche des conditions de possibilité de la stabilité du langage. » (Ghiglione, Matalon, Bacri, 1985, p. 10).

La grammaire de discours, issue de la grammaire générative, a adopté la phrase comme unité d'analyse, « c'est-à-dire une structure munie des traces laissées par son histoire, par les transformations qui l'ont produites à partir d'une forme symbolique minimale : la forme logique » (Ghiglione, Matalon, Bacri, *op. cit.*, p. 12).

Les analyses de discours, quant à elles, ont proposé comme unité d'analyse la proposition ou l'énoncé, permettant ainsi de développer une étude des composants de phrase d'une part, et de réintroduire l'énonciateur et la situation dans l'analyse d'autre part. On reprendra ici ce que nous disions d'une des mises en œuvre les plus abouties, en France, de l'analyse de discours, celle de M. Pêcheux : « Le travail s'effectue, selon Pêcheux et Fuchs (1975) sur la proposition "en tant qu'unité indépendante énonçable" et va porter sur deux systèmes imbriqués l'un dans l'autre : le système des énoncés et celui des relations inter-énoncés, de telle sorte

que les objets du premier système servent d'éléments pour la construction du second. En second lieu [...] substituer la phrase à l'énoncé, aux séquences de phrases le discours [...] permet de s'éloigner du problème des relations strictement syntaxiques entre composants d'une phrase ou d'une séquence de phrases pour aborder la question du rapport du discours à celui qui le tient, à ce dont il parle. La spécificité de l'analyse de discours tient à ce que cette personne qui parle n'est plus le sujet universel des linguistes, sujet supposé savoir sa langue. Elle est inscrite dans une formation idéologique. » (Ghiglione, Matalon, Bacri, *op. cit.*, p. 13).

Les analyses de contenu, quant à elles, n'ont ni l'ambition de construire une métalangue, ni celle de reproduire une typologie des formes discursives. L'analyse de contenu a comme objectif « la compréhension de la structure et de la cohérence interne, ou des incohérences marginales d'un entretien, d'un discours [...] L'analyse de contenu cherche à étudier une parole, une personne, ce qu'elle dit et non pas les conditions idéologiques de la reproduction/transformation des rapports de production » (Pêcheux, 1975 ; Ghiglione, Matalon, Bacri, *op. cit.*, pp. 18-19).

Il est clair que ces divergences théoriques et méthodologiques auront des incidences sur la pratique des analystes et – c'est là l'objet de cet ouvrage – sur l'informatisation des procédures d'analyses.

Nous nous appuierons dans cette introduction sur l'« inventaire » réalisé par J. Jenny (1997) des « pratiques d'analyse textuelle informatisée » qui renvoient à quelques grands types d'approches :

- *lexicométrique*, « qui consiste à comparer des profils lexicaux (distributions relatives des occurrences lexicales, sans nécessité de lecture préalable) » ;
- *socio-sémantique*, « par segmentation du corpus en unités de significations pertinentes et par catégorisation multidimensionnelle conforme aux grilles d'analyses conceptuelles spécifiques de chaque recherche (dans une optique classique de codage a posteriori) » ;
- *par réseaux de mots associés* « qui visent à re-présenter des configurations cognitives liées à un ou plusieurs thèmes, considérées comme cachées sous la surface textuelle » ;
- *propositionnelle* et *prédicative* qui visent à décrire « les logiques de construction progressive de tout univers référentiel cohérent... ainsi que les finalités ou intentions de chaque mise en scène langagière particulière » ;
- *d'ingénierie textuelle*, « à visée d'audit textuel ou à dominante de documentation-communication » et de systèmes experts « dédiés à des

problématiques de recherche sociologique particulière », que nous ne citerons ici que pour mémoire.

Chacune de ces approches méritant d'être interrogée tant dans ses présupposés que dans ses outils, nous allons nous y employer brièvement, au moins pour ce qui est des trois premières.

1. LES APPROCHES LEXICOMÉTRIQUES

D'inspiration benzecriste, elles fonctionnent sur la fréquence d'occurrence des mots d'un corpus. Jenny (*op. cit.*) identifie quatre logiciels [1] susceptibles de traiter les corpus dans une optique lexicométrique et décrit les procédures informatisées d'analyse en trois étapes principales :

- « dresser l'inventaire de toutes les "formes graphiques brutes" (ou "lexèmes", équivalents de "mots") du corpus à analyser, dans un double classement : par ordre alphabétique et par ordre de fréquence d'occurrence » ;
- « vient ensuite la construction du tableau lexical entier (TLE) de ce corpus, composé d'autant de lignes (ou de colonnes) qu'il y a de "mots", classés en rang de fréquence décroissante... et d'autant de colonnes (ou de lignes) qu'on aura préalablement partitionné ce corpus en parties distinctes » ;
- « les calculs ultérieurs consistent à comparer les "profils" lexicaux (exprimés par les fréquences différentielles des mots, dans les colonnes ou lignes du tableau) des différentes parties du corpus ».

Outre les problèmes de traitement des ambiguïtés sémantiques que posent de telles pratiques, il convient de s'interroger de façon fondamentale sur ce que signifient les résultats obtenus et leur interprétation.

De fait, à notre sens, ce type d'analyse fait directement dériver l'interprétation de la fonction référentielle, seule traitée par l'analyse lexicométrique. En effet, un texte renvoie à trois questions fondamentales : quoi (que dit-il ?), comment (le dit-il ?), pourquoi (le dit-il ?). Seule la première de ces questions renvoie à la référence, mais rien ne garantit que celle-ci permette de répondre automatiquement – c'est-à-dire sans autres ingrédients théoriques, méthodologiques et d'outillages – aux deux autres questions. C'est cependant la pratique courante.

1. Alceste, Hyperbase, Lexico 1, Spad T.

Lorsque le Centre de communication avancée (CCA) « inventa » le *lexico'styl*, il partit de l'idée qu'il existait une correspondance entre les « styles de vie » des Français [1] et leur façon de parler. Ainsi, à un style de vie donnée correspondait un sous-vocabulaire spécifique. Vieille idée que celle qui consiste à dire que les groupes sociaux se servent du langage comme d'un marqueur d'identité de groupe. Toutefois les choses ne sont ni aussi simples, ni aussi linéaires. Afin d'illustrer ce propos, imaginons que là où « l'activiste » parle des « gens » de façon permanente, le « décalé » parle des « mecs » de façon tout aussi permanente, ce qui pourrait occasionner ce type de texte.

Activiste	Décentré
Bon alors les gens il font ce qu'ils ont à faire, ils mènent leur barque comme ils peuvent. Ce n'est pas toujours simple pour les gens que d'être confrontés aux problèmes économiques, à la pollution, à l'encombrement des villes, etc. Mais bon ils font ce qu'ils peuvent pour s'en sortir les gens.	Les mecs ils font comme ils peuvent, ils font ce qu'ils ont à faire, mais ce n'est pas toujours facile quant on est un mec. On est confronté aux problèmes économiques, à l'encombrement des villes, au stress, etc. Mais bon quand on est un mec on fait ce qu'on peut pour s'en sortir.

Un document publicitaire, édité par le CCA, annonçait à l'époque : « Comment parler le décalé, comment parler l'activiste », etc. [2] et fondait son argumentaire sur l'analyse lexicométrique d'un nombre impressionnant de corpus analysés dont les émetteurs étaient les dits décalés, égocentrés, activistes et autres hommes sandwichs à étiquette.

Revenons à nos trois questions fondamentales : quoi ? Comment ? Pourquoi ?

Si l'on suit les conseils du CCA, le « quoi » renvoie à un individu générique qui englobe les humains (adultes ?) sans distinction de sexe, de classe, etc. ; le « comment » renvoie à une appartenance catégorielle qui détermine l'équivalent paradigmatique choisi ; le « pourquoi » renvoie à l'expression de cette appartenance.

On retrouve là des idées que Bernstein (1971) a, en son temps, utilisées d'une certaine façon, à travers les notions de code restreint – réservé

1. On trouvait en 1985 : l'activiste, le recentré matérialiste, le recentré rigoriste, le décentré, l'égocentrique.
2. Afin de mieux vendre tel ou tel produit naturellement. *Cf.* Annexe 1.

aux *lower class* – et de code élaboré – accessible aux seules *middle-upper* et *upper class*... et que Labov (1966, 1972) a contestées de façon convaincante, à notre sens (Ghiglione, Beauvois, 1981). En effet, Labov a clairement montré que les variations situationnelles mettaient à mal les prédictions qui pouvaient être faites quant à l'enfermement des gens dans un code quelconque.

Si l'on ne suit pas les conseils du CCA :

- le « quoi » renvoie à un énonciateur situationnellement inscrit et dont l'énoncé, émis ici et maintenant, s'insère dans un tissu discursif doté d'une histoire et d'un avenir ;
- le « comment » renvoie à un jeu interlocutoire inscrit dans un contrat de communication spécifique ;
- le « pourquoi » renvoie à une visée d'influence sur un auditoire ou un interlocuteur, présent ou absent.

Autrement dit, lorsque « l'activiste » énonce « les gens », il le fait non parce qu'il est un activiste, mais parce qu'il est placé dans une situation de communication qui lui fait préférer ce substantif à un autre (mec, par exemple) pour des raisons d'opportunité.

Bien sûr si l'on se contentait du « quoi ? », du « que dit-il ? », cela ne poserait aucun problème... mais serait peu informatif. Or, l'analyse lexicométrique n'est guère fondée, en droit, à faire autre chose qu'à décrire le nombre d'occurrences d'un mot ou s'il est co-occurrent de façon significative avec tel mot, etc. Si parler, c'est transmettre un sens et une intention, alors l'outil lexicométrique ne peut traiter que du sens, c'est-à-dire du contenu informatif et non de l'intention.

2. LES APPROCHES SOCIO-SÉMANTIQUES

Plus « nouvelles », elles sont peut-être plus claires quant à leur méthodologie.

> « Priorité est donnée à l'élaboration "nouvelle" d'une grille d'analyse des contenus thématiques, en fonction du cadre théorique de l'enquête, et si on recourt à des traitements statistiques "automatiques", ce n'est d'abord que pour mesurer l'extension empirique des catégories établies a priori, puis pour valider des hypothèses concernant le système des relations qui en font une organisation structurée [...] Ainsi

> dans le cas d'une étude récemment publiée sur les représentations du handicap [...] c'est la problématisation préalable des notions de représentation et de handicap qui a conduit à constituer un échantillonnage précis de locuteurs contrastés [...] à construire une grille d'analyse à facettes comportant des catégories thématiques ("de quoi/de qui parle-t-on ?"), des modalités fonctionnelles ("comment en parle-t-on ?"), des processus d'énonciation ("qui fait-on parler ?"). Quant aux logiciels [1] [...] ils procèdent à la segmentation du corpus en "unités de significations" (l'équivalent des propositions grammaticales) et à la classification de ces unités, puis à l'aide d'un algorithme d'analyse discriminante pas à pas, ils calculent "l'arbre hiérarchique" de classification... » (Jenny, *op. cit.*).

Ce genre d'approche présente le mérite d'établir clairement l'analyse sur des savoirs préalables qui serviront à étayer l'interprétation, c'est-à-dire la réponse au « pourquoi dit-il ça ? » dont on remarquera qu'il n'y est fait aucune référence dans le descriptif ci-dessus. En effet, c'est là le domaine de l'analyste interprétant.

Ainsi que le dit Jenny par ailleurs, les deux types d'analyse – lexicométrique et socio-sémantique – sont complémentaires. De fait, car l'analyse lexicométrique est de type *bottom-up* et paradigmatique alors que l'analyse socio-sémantique est de type *top-down* et syntagmatique. Dans le premier cas, ce sont les traitements des données qui guident l'interprétation et la production de savoirs, dans le second cas ce sont les savoirs *a priori* qui guident le traitement des données. Partant de là, on peut très bien imaginer que quelques pistes interprétatives nouvelles naissent du traitement des données dans le mode *bottom-up*, de même qu'on peut très bien imaginer que les savoirs *a priori* évitent quelques dérives interprétatives dans le mode *top-down*.

Toutefois plusieurs questions restent en suspens : comment construit-on les grilles d'analyses préalables, sur quelles bases théoriques ? Quelles sont les théories de référence servant à interpréter les modalités fonctionnelles ? Quelle est la théorie du sujet qui sous-tend le « qui fait-on parler ? » Autant de questions qui méritent des réponses précises bien difficiles à trouver, même si les tentatives d'intégration de divers logiciels vont dans ce sens (voir notamment la « plate-forme informatique dénommée Aladin... capable de supporter les processus cognitifs humains impliqués dans les opérations d'analyse et de lecture de textes »).

---

1. « AC2 » et « Alice », mais aussi les modules « Interviews » du logiciel « Modalisa » et « Lexica » du logiciel « Sphinx ».

3. LES ANALYSES PAR RÉSEAUX DE MOTS ASSOCIÉS (RMA)

Fondées sur l'extraction des « configurations cognitives cachées sous la surface textuelle » (Jenny, *op. cit.*), elles utilisent les occurrences des termes et leur proximité. Mais là, comme dans l'approche lexicométrique, « les deux conceptions – fréquentiste et "intuitionniste" – sont également possibles, selon qu'on confie l'élaboration de ces réseaux soit au "calcul informatique" des relations de proximité attestées entre les termes lexicaux du corpus, soit à l'expertise explicite et formalisée des dictionnaires usuels (compétents dans la connaissance ordinaire et partagée des univers référentiels), ou des sociologues enquêteurs (qualifiés pour leur connaissance "savante"), soit encore à l'expertise explicite et "naïve" des enquêté(e)s eux/elles-mêmes... »

La critique est ici tout à la fois simple et courante. Qu'en est-il de la représentation du réel pour un sujet humain ? Le statut cognitif de la représentation du réel est-il celui de mots substantialisés inscrits dans des réseaux de proximité plus ou moins hiérarchisés ? Ou bien la représentation du réel est-elle une construction permanente d'actions transitoirement énonçables en termes de *frame*, scénario ou script ?

La réponse à cette question porte en elle l'acceptation ou le rejet des analyses par « réseaux de mots associés ». En effet, si l'on considère que la représentation du réel est construite à partir d'une articulation de référents qui dénotent le réel, les analyses RMA sont satisfaisantes, mais dans le cas inverse, si l'on considère que le réel est une suite d'actions, alors les analyses RMA sont insuffisantes. D'ailleurs, à y regarder de près, ainsi que le fait Jenny, « l'autre courant d'analyse des réseaux de mots associés [qui] s'est développé parallèlement au précédent, et sans interférence apparente avec lui », ne fait rien de plus que de réintroduire l'action au cœur du dispositif. On peut en juger sur la base de ce qu'en dit Jenny : « Sans entrer dans le détail des procédures de traitement et d'analyse, on peut dire que la principale originalité de ce programme [1] [...] c'est sans doute de définir les "acteurs/actants" précisément par leur "profil d'association", c'est-à-dire la liste des mots auxquels ils sont associés [...] et de définir le contenu textuel comme "le réseau des asso-

1. Il s'agit du logiciel *Candide*.

ciations opérées par le texte entre les acteurs qu'il met en scène" (Teil, 1991, p. 225). Système dynamique (chaque acteur créant ou détruisant des relations avec d'autres acteurs, les réseaux se transformant sans cesse) et système ouvert (à tout nouvel acteur controverse, consensus ou conflit) ».

En somme, que ce soit l'analyse lexicométrique ou l'analyse RMA, l'essentiel de la critique porte sur un mode d'analyse fondé sur un traitement des données du type *bottom-up*, non informé par un savoir préalable, et sur la seule prise en compte de l'axe paradigmatique qui, outre le fait qu'il est porteur d'une visée substantialiste, ne permet pas une interprétation actantielle, cependant toujours produite au terme des analyses.

On aura d'ailleurs remarqué que les deux autres analyses présentées ont un caractère « correcteur ». La première, socio-sémantique, tente de nuancer le caractère *bottom-up* des traitements qui pourrait conduire à des interprétations arbitraires ; la seconde, variante du RMA, tente de pallier le défaut structurel qui rend illégitimes les interprétations actantielles.

4. LES ANALYSES PROPOSITIONNELLES ET PRÉDICATIVES

Regroupées dans l'analyse cognitivo-discursive (ACD) (Ghiglione, Kekenbosch, Landré, 1995), elles tentent, de répondre aux critiques précédentes. Fondées sur quelques principes ayant donné lieu à des hypothèses et à des expérimentations (voir notamment Ghiglione, Trognon, 1993), elles essaient de pallier les différents manques théoriques constatés et les critiques méthodologiques habituelles.

Les principes qui fondent l'ACD peuvent être résumés simplement :

- tout discours s'inscrit dans un contrat porteur d'enjeu et a une visée d'influence ;
- tout discours est inscrit dans un inter-discours mais est également le produit d'un « ici et maintenant » qui actualise certains des paramètres du Contrat de Communication (Ghiglione, 1986) ;
- tout discours met en scène des mondes inscrits dans une histoire, construits selon des règles de cohésion, de cohérence, de consistance et causalement liés ;

- tout discours porte en lui les traces des opérations cognitives effectuées par un locuteur qui met en scène, dans un certain but, un certain sens et une certaine intention ;
- compte-tenu de ce qui précède, tout discours peut être interrogé en tant que tel, si toutefois on s'intéresse au sens qu'il véhicule et à l'intention qu'il manifeste.

Ces principes entraînent certaines questions :

- si le discours est la mise en scène de mondes causalement liés et s'il est porteur des traces des opérations cognitives effectuées, quelle est l'unité de cette mise en scène ?
- si le discours manifeste un sens et une intention, comment sont-ils accessibles ?
- quelles sont les marques de cohésion, cohérence, consistance, etc., et que signifient-elles au-delà de leurs fonctions textuelles ?

Le logiciel [1] construit sur la base des réponses apportées à ces questions :

- identifie les propositions (de forme grammaticale), privilégiant ainsi l'axe syntagmatique et la mise en scène d'actants, d'actés et d'actes ; les actants et actés principaux, n'ignorant pas de ce fait l'axe paradigmatique ;
- met à jour le « chemin causal » qui construit l'histoire mise en scène dans l'énoncé, exhibant de ce fait le noyau générateur de ce produit discursif (et retrouvant les théories du sujet développées dans le champ psycholinguistique) en réduisant ainsi le corpus à l'essentiel, sans opérations interprétatives ;
- traite les jeux de prise en charge et de modalisation du discours, exhibant ainsi le rôle de l'énonciateur dans l'énoncé.

Pour peu que les savoirs préalables soient disponibles, ceux du discours politique par exemple (voir Ghiglione, 1989 ; Ghiglione, Bromberg, 1998), ces éléments issus d'une analyse *bottom-up* permettent de réinscrire un discours particulier dans une interdiscursivité qui permet, quant à elle, des interprétations sécurisées par un traitement *top-down*. De plus, en cas de questionnement *top-down* d'un corpus, sur la base de savoirs constitués préalablement, le logiciel (par le jeu des scénarios ouverts, nous y reviendrons) s'y prête.

Au terme de cette introduction, il nous semble que l'analyse de contenu informatisée, tout comme l'analyse de contenu en général (et pour-

1. *Tropes*, que nous aurons l'occasion de présenter longuement par la suite.

quoi en serait-il autrement ?) doit répondre, pour être valide ou le plus valide possible, à deux questions essentielles :

– sur quelle théorie du sujet repose-t-elle ?
– sur quelle théorie de l'interlocution s'appuie-t-elle ?

Prétendre que la réponse à ces questions garantit contre toute dérive interprétative serait abusif, tout au plus peut-on tenter de s'approcher au plus près du sens et de l'intention que le locuteur a voulu rendre manifestes, ou qu'il a manifestés sans les contrôler complètement.

Dire que l'analyse informatisée des contenus est suffisante serait tout aussi abusif. Les analyses fines supposent et supposeront encore longtemps, selon nous, des compléments manuels… et puis quelques ambiguïtés résistent à l'ordinateur, de même que quelques métaphores et bien d'autres choses encore.

Quoiqu'il en soit, nous allons tenter de répondre à l'essentiel des questions essentielles et de donner à voir – et à tester – un logiciel (*Tropes*) qui, issu de nos recherches, se montre et nous expose. Mais, ainsi qu'il est dit là où la foi s'exhibe : *Ite missa est*.