Les logiciels d'analyse du discours, une tension entre sociologie et linguistique

Pierre Vergès

Les logiciels d'analyse du discours, une tension entre sociologie et linguistique

L'analyse de discours a déjà un long passé et l'interrogation de Google vous renverra près de 6000 références. Le domaine est donc largement balisé et nous nous limiterons ici à envisager les seuls usages sociologiques. Il faut avant tout revenir sur l'histoire de ce domaine de recherche. Jusqu'aux années soixante, il est dominé par l'analyse de contenu dont le fameux article de B.Berelson en 1952[1] était la référence. Les premières critiques en France viennent d'un groupe de chercheurs normaliens qui dénoncent les limites d'une définition de la catégorisation des mots par la subjectivité de l'analyste. Ils vont montrer l'intérêt d'une analyse syntaxique[2] qu'ils trouvent dans les débuts de l'analyse documentaire automatisée. Ils vont proposer une analyse automatique du discours[3] en s'appuyant sur l'intérêt de la démarche linguistique de Harris[4]. Ces premières années de l'analyse du discours à la française furent effervescentes et passionnèrent la communauté des chercheurs sociologues et psychosociologues. On a l'impression quarante ans après que les avancées et les interrogations de cette période sont bien oubliées, peut être parce qu'elles étaient trop liées à une vision marxiste des sciences sociales ou à une vision un peu utopiste de ce que l'informatique pouvait apporter au traitement du langage, de la pensée sociale.

Depuis cette époque des précurseurs, on a pu explorer les possibilités et les limites du traitement automatique du discours. On a surtout vu l'intérêt pour le sociologue de ne pas ignorer les avancées de la linguistique et les démarches recherchant une argumentation dans le discours[5]. En effet, les linguistes se sont de plus en plus intéressés aux formes orales et non seulement littéraires, écrites. Ils ont mis en évidence l'inscription du locuteur dans son propos en développant une théorie de « l'énonciation ».

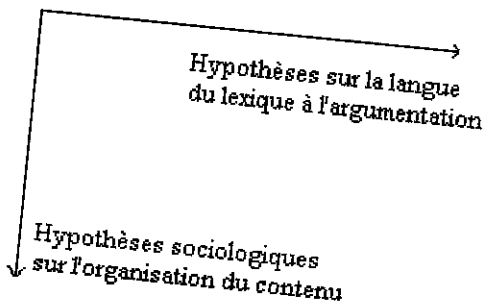
On doit retenir des tentatives de la période des précurseurs et des développements permis par l'usage de plus en plus complexe de l'informatique que l'analyse de discours est fondamentalement inscrite dans une double dimension : celle de la langue utilisée et celle de la science sociale convoquée, ici de la sociologie. Chacune de ces dimensions est construite à partir d'hypothèses propres à chaque discipline. Les hypothèses sociologiques sont bien souvent présentées, car elles ont conduit la démarche d'enquête : entretien, questionnaire à réponses libres, textes (de journaux, d'élèves…). Par contre, on fait trop souvent l'impasse sur les hypothèses linguistiques, elles sont souvent implicites ou carrément ignorées. Or fondamentalement le discours fait appel à toutes les ressources d'une langue, d'une parole qui utilisera des mots, une syntaxe et une mise en argumentation. Se limiter à l'un de ces niveaux relève soit d'une volonté consciente justifiée par la démarche de recueil des données soit d'une véritable erreur épistémologique. Notre propos cherche à baliser un peu cette double

---

[1] Berelson, B., 1952, *Content analysis in communication research*, The Free Press, Glencoe.
[2] Cros, R.C., Gardin,J.C., Lévy, F., 1964, *L'automation des recherches documentaires, un modèle général, le « SYNTOL »*, Gauthier Villars, Paris
[3] Pêcheux, M., *Analyse automatique du discours*, Dunod, 1969.
[4] Harris, A., 1952, *Discourse analysis*, Language, 28, 1-30.
[5] Grize, J.B., ed., 1984, *Sémiologie du raisonnement*, P. Lang, Berne, Francfort/M., New York.

dimension, à y situer les procédures informatiques que l'on propose actuellement au sociologue.

Les différents logiciels qui sont à leur disposition doivent être placés sur un double axe pour comprendre ce qu'ils peuvent apporter et surtout ce qu'ils laissent de coté.



1. Quand le sociologue ne prend en compte que le lexique.

1.1 Les mots n'ont-ils qu'un sens ?

Bien souvent le sociologue fait comme s'il y avait une absence totale de la dimension langagière dans les données qu'il analyse ; ainsi il ne s'intéresse qu'aux termes, qu'aux mots. Il cherche à organiser le lexique dans une démarche qui tend simplement à mettre de l'ordre, souvent statistique, dans le contenu des textes. En effet, certains logiciels ne font que dépasser la simple collation des mots du texte et leur classification par la seule volonté de l'analyste. Il faut voir que la recherche d'une simplicité d'utilisation et d'une possibilité d'interprétation presque immédiate des résultats fournis par les programmes informatiques, tend seulement à déplacer l'intervention du sociologue d'une interprétation *a priori* (Content analysis) à une interprétation *a posteriori* tout aussi sujette à critique ou à lectures multiples peu contrôlées.

Dans ce domaine les logiciels se sont complexifiés au cours du temps. Les plus simples ne procurent qu'une comptabilité des mots (lexicométrie), à un degré plus élevé ils recherchent les syntagmes dits « figés » : suite de mots que l'on retrouve plusieurs fois dans le même ordre (tel « analyse du discours »). Une telle forme est supposée renvoyée à une fossilisation du langage propice au fonctionnement idéologique (cas en particulier des slogans). On voit que même à ce niveau le plus élémentaire on peut associer une hypothèse sociologique aux formes lexicales. Il en est de même dans l'analyse des textes littéraires où on recherche la fréquence des termes pour définir des styles ou la signature d'un auteur. L'inconvénient de se limiter à l'identification du mot, au contenu qu'il est censé représenter, est double. En premier un même terme peut exprimer des contenus différents. En second on fait abstraction de la possibilité pour l'auteur, d'utiliser des synonymes ou des périphrases en tenant lieu pour éviter les répétitions.

La prise en compte du contexte de l'apparition des termes permet de complexifier un peu l'hypothèse sociologique implicite de l'analyse. En effet, cette dernière postule alors qu'un terne n'acquiert son sens que dans un contexte. Certains logiciels établissent une comptabilité des cooccurrences des mots dans un contexte donné : les

phrases, le paragraphe… Le logiciel ALCESTE est de ce type tout en complexifiant cette démarche par l'utilisation d'une procédure de classification des termes. Ici l'analyste se trouve investie du pouvoir de donner une interprétation à ces classes produites automatiquement. L'analyste propose alors soit une synthèse des mots de la classe par une notion englobante, soit il privilégie certains des mots de la classe considérés comme plus caractéristiques, plus illustratifs. Cette démarche ne peut échapper à passer sous silence une partie des termes de la classe. Le caractère subjectif de l'interprétation est repoussé à ce moment de l'analyse, ce qui est déjà une avancée certaine au regard de la simple analyse fréquentielle.

1.2 La sélection de portions de discours.

L'analyste peut toujours faire un découpage *a priori* du texte en tenant compte des ruptures dans la linéarité du texte, par exemple en tenant compte des interventions de l'enquêteur. Dans ce cas on ajoute seulement un codage tenant compte de la nature de l'intervention.

Plus intéressante est la recherche de zones du discours autour de mots sélectionnés à l'avance pour leur pertinence au regard des préoccupations du chercheur. On définit alors des mots-pivots et on ne garde du texte que le contexte de ces mots-pivots. On voit ici que l'intervention du chercheur ne porte pas sur des propriétés linguistiques, mais seulement sur des hypothèses sociologiques, sur la pertinence de ces mots et de leur contexte. On fait appel ici à la propriété paradigmatique du langage qui permet de mettre en évidence des cooccurrences entre termes et qui peut utiliser des propriétés syntaxiques si on pousse plus loin l'analyse de la place des termes, comme l'identification des sujets ou celle des classes des verbes de ces phrases. Certes avec cette technique on peut passer à coté de portions de discours particulièrement pertinentes pour le sociologue, car non sélectionnées par les mots-pivots. Cependant, elle peut permettre le traitement de très gros corpus tout en donnant à l'analyste la possibilité d'intégrer non seulement une analyse de vocabulaire éliminant une partie du « bruit » mais aussi un début d'interrogation sur le rôle de la syntaxe par l'analyse des places dans la phrase.

1.3 L'utilisation d'hypothèses sociologiques.

Jusqu'à présent, les hypothèses sociologiques sont assez rudimentaires, elles se réduisent à privilégier certains termes ou classe de contenu. On peut aller un peu plus loin sans pour autant prendre en compte les caractéristiques linguistiques du discours. En effet, il suffit de définir des catégories de contenus et d'établir des rapports entre ces catégories. La plus simple est celle de leur cooccurrence dans un contexte donné dont l'hypothèse sociologique est bien ténue. On peut aller plus loin en enrichissant le texte d'hypothèses telle la définition de thèmes *a priori* et la recherche des portions de texte qui peuvent les exprimer. Pour ne pas rester unidimensionnel, on peut définir parallèlement des raisons, des motivations de l'utilisation de ces thèmes par le locuteur. Il est alors possible de construire le croisement de ces deux dimensions pour

définir des « thèmes - motivés » et de s'interroger sur les effectifs des différentes cases ainsi définies.

On peut aussi faire l'hypothèse de l'identification de mots ou syntagmes (ainsi que de certains verbes) qui sont privilégiés par une idéologie donnée. Ici les mots sont considérés comme le support, le représentant, d'une thématique définie a priori dont on mesure l'importance et dont on définit le contexte d'utilisation. La mise en évidence des relations entre ces thématiques est alors possible et conduit à la construction d'un réseau sémantique représentant les différentes facettes de l'idéologie recherchée ou la mise en évidence des oppositions entre représentations sociales concurrentes.

2. La prise en compte de la syntaxe du texte.

L'évolution de l'analyse de discours depuis ses débuts a par ailleurs tenu compte des avancés de la linguistique. Celle-ci s'est en effet intéressée à l'analyse automatique du langage soit en direction de la documentation soit pour permettre des traductions assistées sur ordinateur. Il y a eu parallèlement des avancées théoriques importantes telles celles de Chomsky ou celles des différents courants de recherche français, celui autour de Culioli en particulier. De manière plus pragmatique, on a vu se développer des recherches sur la classification du vocabulaire en entités, mots outils, qualités… Parallèlement, il a été défini des classes de verbes. Le sociologue peut associer à ces classifications, peut-être un peu rapidement, des propriétés sociales.

Deux concepts importants ont émergé de ce travail théorique : la performativité et l'énonciation. Ils visent tous deux l'identification de l'intentionnalité du locuteur. Bien longtemps les sociologues ont refusé de prendre en compte les motivations de celui qui s'exprime, mettant seulement l'accent sur l'organisation du contenu de son texte et sur ce qu'il révèle de l'idéologie du locuteur. Ils ont cependant pris conscience qu'ils ne pouvaient pas ignorer le degré d'engagement de l'acteur social dans ce qu'il avance. La mise en évidence de la différence entre un énoncé performatif et un énoncé descriptif a ouvert toute une série d'interrogation sur l'efficacité du discours : « Quand dire c'est faire » fut le titre d'un livre qui permit d'identifier dans le discours des zones particulières dont le propos est d'articuler texte et action. Il en est de même des recherches sur l'énonciation qui ont conduit à repérer la place que se donne le locuteur dans son discours. Celles-ci ont mis à jour tout un ensemble de marqueurs qui définissent ces places (l'engagement du locuteur). Pour le sociologue, ces deux conceptualisations lui permettent d'identifier des zones de discours dont il peut qualifier l'importance en terme d'implication du locuteur dans son propos. Ces zones peuvent être distinguées de celles qui concernent la description de processus ou d'actions.

Ce travail théorique a mis en évidence qu'au-delà des classifications qui peuvent être établies sur la base de propriétés linguistiques il fallait tenir compte de la syntaxe employée. On passe alors du simple comptage de mots ou de portions de texte à la prise en compte de la syntaxe. Celle-ci permet de distinguer « ce dont on parle » de

« ce qui en est dit ». Le premier peut être envisagé comme plus important pour le locuteur car source d'actions ou de définitions. Le second est composé de ce qu'il est possible de faire ou d'éléments souvent qualificatifs. On trouve ici une possible connexion avec la logique des prédicats. Des tentatives de traduire logiquement le discours ont été faites, mais elles s'appliquent seulement sur des textes très particuliers. Dans leur prolongement des recherches ont donné lieu à la constitution d'une « logique naturelle » sur laquelle nous reviendrons plus loin.

Les sociologues qui ont exploré ce mode de traitement du discours se sont trouvés devant la difficulté d'interpréter les rapports entre les thèmes du discours et leur place dans les énoncés. Le fait de trouver un mot particulier (ou plus largement un objet de discours[6]) en position « thématique » ou inversement dans la partie du texte qui décrit l'action, les propriétés de cet objet (qui est souvent appelé « rhématique »), n'est pas sans conséquence sur l'intention que veut exprimer le locuteur. C'est ainsi que l'analyse d'un compte rendu de Comité d'Entreprise a montré que le discours de la direction a, pour justifier sa position, bien mis en valeur l'importance de l'évolution des salaires (en position thématique – sujet) et a passé presque sous silence les contraintes liées à l'internationalisation de la production. Celles-ci sont situées systématiquement en position rhématique.

Rare sont les logiciels qui prennent en compte cette dimension essentielle du discours. Le cas de PROSPERO est assez singulier pour qu'on s'y arrête un moment. Il peut proposer plusieurs niveaux de complexité dont le premier relève de cette prise en compte de la syntaxe dans l'analyse de discours. Il le fait à l'aide de la multiplication des clefs d'entrée : identification des personnes, de catégories, de collections (classe existante dans le monde). Il le complète par la mise en évidence du réseau de ces entités et par la recherche de séquences particulières (configuration discursive, ou portion de texte défini par une formule). Aussi sont convoquées deux propriétés des textes : le contexte des énoncés, les transformations des arguments au long du texte. Les auteurs de cette méthode ne veulent pas se limiter à la seule analyse de discours, ils visent plus largement une pragmatique des transformations sociales par l'identification de configurations socio-politiques dans lesquelles baignent le locuteur. Ici nous ne pouvons pas les suivre sur ce point, mais nous pouvons relever la grande qualité de ce logiciel qui permet d'articuler les trois niveaux que nous avons précédemment distingués : le vocabulaire, son organisation par la syntaxe et la définition de catégories formelles, enfin l'utilisation de modèles a priori, ici séquentiels ou historiques.

3 Vers une analyse argumentative.

Plusieurs tentatives de définitions de formes ou de procédés argumentatifs ont vu le jour. Elles visent soit à la recollection des formes que la philosophie a léguées à la logique soit à la construction d'une procédure calculatoire sur des entités formelles rarement associables aux énoncés d'un discours social. Il existe cependant une école dont l'initiateur fut J.B. Grize à Neuchâtel qui, s'intéressant à l'analyse logique des

[6] Un thème particulier bien identifié tel « le rôle de l'Etat » ou une personne tel « le locuteur ».

discours, en vient à abandonner, pour décrire les textes, les procédures de la logique formelle et à proposer une logique particulière dite « logique naturelle ». Cette formalisation est complètement adaptée à l'analyse des textes, car elle permet de tenir compte d'une grande partie des questions que l'on peut se poser sur un discours[7].

La première innovation de cette démarche est le remplacement de la classique comptabilité des mots d'un texte par l'identification « d'objets de discours ». En effet, tout discours peut être considéré comme une démarche visant à élaborer un micro-univers composé par les objets de discours. Ces objets sont à la fois une construction sémiologique où les mots ont une référence(s) dans le monde (notons en passant que ces objets peuvent être des personnes) et une construction cognitive constituée de fragments de connaissance visant à donner à voir une représentation. Tout objet de discours est ancré dans une entité sociale et associé à un faisceau d'aspects ou à un domaine de mise en perspective. Ainsi, l'objet du discours ne peut se résumer à un mot, à un syntagme, il traîne (pourrait-on dire) avec lui tout un ensemble de déterminations, de qualités, d'autres termes auxquels il fait référence. Aussi ces objets sont fondamentalement des « classe-objets » constitué d'un objet et de ses ingrédients. On peut en donner un exemple tiré d'un article du magazine municipal de Marseille (O1 se décline de manière arborescente en divers ingrédients O_{11}, O_{12}...)

O_1 le commerce traditionnel en centre ville
- *$O_{1\text{-}1}$ sa vocation de lieu d'échange*
- *$O_{1\text{-}2}$ commerce marseillais*
- *$O_{1\text{-}3}$ commerce de luxe*
  - *$O_{1\text{-}3\text{-}1}$ L'enseignes nationales aussi prestigieuses que Cartier et Vuitton*
- *$O_{1\text{-}4}$ un potentiel commercial*
  - *$O_{1\text{-}4\text{-}1}$ encore insuffisamment exploité*

Les objets sont ancrés dans des représentations et des pratiques sociales très largement prédéterminées avant même sa mise en discours. Ils sont donc indissociables de ce qu'on peut appeler un « pré-construit ». La définition de ce pré-construit ne se fait pas *ex nihilo* mais par la seule volonté de l'analyste : seulement par ses traces dans le discours analysé, traces que décrit l'arborescence. Le locuteur ne fait pas que décrire des objets, il les inscrit dans des propositions tenues à son propos. Le deuxième niveau d'analyse est alors celui des prédicats. Le prédicat est une mise en rapport qui recouvre en fait deux formes de relation. On doit distinguer les prédicats « unaire » des prédicats complexes. Le prédicat unaire exprime par sa partie rhématique la propriété d'un objet x. Il propose en fait, sous une autre forme linguistique, un ingrédient de l'objet. Les prédicats complexes sont de deux types : soit, ils sont relationnels ; c'est à dire mettent en relations deux ou plusieurs objets définis dans le texte ; soit, ils utilisent une relative qui développe l'objet. On voit l'intérêt d'une telle classification, d'un tel codage pour identifier des textes plus ou moins complexes.

---

[7] Ici le texte est redevable particulièrement à J.B. Grize, M.J. Borel et D.Mieville dont on trouvera certaines productions dans les « Travaux du Centre de Recherche de Neuchâtel ». Le logiciel qui devrai y être associé a eu quelques tentative de développement mais n'a pas trouver ses sponsors !

Ces deux premiers niveaux d'analyse débouchent sur une description relationnelle du propos du locuteur qui est déjà une argumentation par la simple construction des objets et des prédicats. On peut prolonger la démarche en s'appuyant sur les marqueurs de l'argumentation. En effet, le texte vise à élaborer un micro-univers complexe et cohérent. En premier il faut rendre compte de la dimension argumentative des enchaînements entre énoncés. Ces enchaînements ont une fonction : celle "d'étayer". On exprime par là un mode de relation minimal entre les énoncés. Cette relation se repère le plus souvent par la présence d'un connecteur comme « Parce que ». Mais ce n'est pas toujours le cas dans la mesure où le texte peut être constitué par des successions d'énumérations ou par l'existence implicite de connecteurs non marqués. Le discours est alors organisé par cette relation entre les segments de texte dont l'un est « étayé », et l'autre « étayant ». Cette relation minimale peut se développer sous la forme d'une arborescence. Après avoir décrit une représentation sociale par les objets et les prédicats on atteint ici la construction d'un monde (propre au locuteur) ayant sa propre logique : celle d'établir une vraisemblance.

4 Le futur des logiciels d'analyse de discours.

La mise en perspective que nous venons de faire montre les difficultés de l'analyse automatique du discours. Elle doit tenir compte des volontés du sociologue d'introduire dans la procédure même d'analyse ses hypothèses, comme celles sur les polémiques. Elle doit ne pas ignorer que le discours s'exprime dans une langue c'est-à-dire dans des formes linguistiques et argumentatives.

Les procédures d'analyse automatique des textes doivent privilégier la mise en évidence de la démarche de la recherche. Or bien souvent les logiciels se présentent comme des boîtes noires ne permettant que de manière marginale l'intervention du chercheur. Certes il faut que la recherche ait un temps de neutralisation des données, c'est-à-dire un moment où les techniques statistiques ou formelles ne soient pas dépendantes de la lecture du chercheur. Mais on doit toujours pouvoir établir les rapports entre les hypothèses sociologiques et ces procédures formelles, ces algorithmes informatiques. Il n'y a pas d'automatisme possible, car il y a toujours une « lecture » du texte. Le seul comptage des mots, leur classification *a priori* ou par un algorithme passe à côté de la manière dont l'auteur a construit son discours, son argumentation. Le repérage des marqueurs d'énonciation peut dire beaucoup sur l'inscription de l'auteur dans son texte et par la même indiquer les zones les plus intéressantes ou problématiques, mais ils ne créent pas la possibilité de multiples lectures.

Mon propos ici n'est pas de montrer les limites de telle ou telle manière de faire une analyse de discours, mais de mettre le sociologue devant des choix conscients. Il ne peut se laisser séduire par un logiciel sans établir les implications épistémologiques de son choix et les possibilités de son protocole d'étude. La situation actuelle montre que toute procédure est limitée à une classe de questionnement sociologique. Il n'y a pas, malgré les années passées depuis les prémisses de l'analyse de discours, de solution

idéale qui propose de formaliser les deux dimensions sociologique et linguistique qu'il convient de prendre en compte.

Pierre Vergès, LAMES, MMSH, Aix en Provence