Analyse de discours et informatique

par

Michel PECHEUX

L.I.S.H. - C.N.R.S. - PARIS

Je commencerai par une remarque terminologique concernant l'expression "analyse de discours", en soulignant que l'évidence de sa traduction littérale par "discourse analysis" dissimule une profonde équivoque.

En effet, dans l'espace anglophone, la notion de "discourse analysis" semble surtout renvoyer à l'étude des processus interactifs de la conversation et de la parole ordinaires. Il s'agit donc essentiellement d'études psycho-linguistiques (mettant en jeu des notions telles que codage/décodage, niveaux de mémoire et systèmes cognitifs individuels) ou sociolinguistiques (concernant les variations d'usages langagiers, à travers l'analyse des tours de parole et des actes de langage).

Quant au domaine des études spécifiquement textuelles, il est principalement occupé par les méthodes d'analyse de contenu, mettant en oeuvre, sur des corpus textuels de dimension variable, une lecture qu'on peut appeler "artificielle", dans la mesure où cette lecture suppose le détour par un certain nombre d'opérations systématiques de lemmatisation, extraction, comptage, comparaison, etc...
Le caractère artificiel de cette lecture devient seulement plus évident quand le recours à l'informatique impose d'expliciter ces opérations à travers des algorithmes. Dans cette perspective, le but général de l'analyse textuelle informatisée serait de construire des procédures automatiques de *lecture-traduction*, allant de la surface des textes vers une représentation formalisée non-ambiguë susceptible de se prêter à divers calculs (logiques, sémantiques etc...) que ne supportent pas les langues naturelles :en somme, il s'agirait de "nettoyer" les textes pour en extraire le sens univoque, comme si on voulait se délivrer des embarras (ambiguïtés, glissements etc...) du langage naturel, afin de se retrouver le plus vite possible dans ces espaces logiquement stables qu'il est convenu d'appeler des"langages de représentation". N'est-ce-pas, d'ailleurs -dira-t-on- ce que fait tout sujet humain quand, entendant ou lisant une séquence, il s'en construit une représentation utilisable (c'est-à-dire un schéma, un modèle simplifié et manipulable) ?

Or, c'est précisément cette évidence logico-empirique de la lecture qui s'est trouvée mise en question, à travers l'existence de ce que, depuis une quinzaine d'années, la tradition francophone désigne sous le terme d' "analyse de discours" : il est bon de rappeler que, historiquement, cette problématique s'est formée (en ce qui concerne la France) autour de la question de l'idéologie, et en particulier de celle de la lecture des discours idéologiques.

Cette problématique de la lecture idéologique, qui au début des années 60 était en train de se condenser sous le nom de "structuralisme" autour de quelques noms comme ceux de Lévi-Strauss, Foucault, Barthes, Lacan, Althusser . . . était en fait un dispositif polémique contre les conceptions dominantes de l'époque, tout autant qu'un programme de travail.

Les conceptions dominantes de l'époque, c'est-à-dire : à la fois l'herméneutique littéraire spiritualiste lisant les thèmes à travers les oeuvres comme les traces visibles d'un créateur invisible, mais aussi les formes sécularisées, plus quotidiennes, de cette pratique spontanée de la lecture qui, sous les

multiples formes empirico-logiques de l'analyse de contenu, que je viens d'évoquer, commençait à envahir les sciences humaines et sociales. Et enfin l'objectivisme quantitatif, réagissant par une référence au sérieux des sciences, et d'abord, en l'occurrence, aux théories de l'Information et aux statistiques, avant de faire appel à la logique mathématique comme c'est le cas aujourd'hui.

Face à ces diverses formes (spontanées ou savantes) d'évidence empirique de la lecture, le mouvement structuraliste européen des années 60 ouvrait la question de savoir ce que c'est que parler, écouter et lire, à travers des concepts comme ceux de "lecture symptômale" et d' "effet de discours", et des mots d'ordre théoriques comme celui du "repérage de l'efficace d'une structure sur ses effets, à travers ses effets" : "C'est depuis Freud, écrivait Louis Althusser au début de *Lire le Capital*, que nous commençons à soupçonner ce qu'écouter, donc ce que parler (et se taire) veut dire; que ce "veut-dire" du parler et de l'écouter découvre, sous l'innocence de la parole et de l'écoute , la profondeur assignable d'un double-fond, le "veut-dire" du discours de l'inconscient - ce double fond dont la linguistique moderne, dans les mécanismes du langage, pense les effets et conditions formels"(1).

Ainsi, l'appui stratégique sur le structuralisme linguistique était clairement revendiqué : s'il était question d'analyser le "discours inconscient" des idéologies, la linguistique structurale, science moderne de l'époque, apparaissait comme le moyen scientifique privilégié d'un changement de terrain. Si les discours idéologiques étaient bien les mythes propres à nos sociétés, comparables à ceux qu'avaient étudiés, dans leur domaine particulier, des théoriciens comme Vladimir Propp, puis Claude Lévi-Strauss, il devait être possible de construire des procédures effectives capables de restituer la trace de leur structure invariante (le système de leurs "fonctions") sous la série combinatoire de leurs variations superficielles, "empiriques" : donc de reconstituer quelque chose de cette "structure présente dans la série de ses effets".

La mise au point du programme d'Analyse Automatique du Discours (publié en 1969 et informatiquement opérationnel à partir de 1971) constitue une tentative parmi d'autres de réaliser cet objectif en s'efforçant de prendre "la linguistique moderne" au sérieux, en particulier F. de Saussure, et les travaux du linguiste américain Z. Harris, auteur d'un texte providentiellement intitulé *"Discourse Analysis"*, qui servit pendant toute une période de référence concrète aux linguistes, historiens et philosophes travaillant dans le champ de l'analyse de discours, sur la lancée des travaux de Jean Dubois.

De ce point de vue, la spécificité de AAD 69 dans l'espace francophone des travaux d'analyse de discours, ce fut d'abord de pousser la linguistique harrissienne jusqu'au bout de ses conséquences, du point de vue théorique que je viens de rappeler, en ignorant plus ou moins délibérément, aussi bien la linguistique générative-transformationnelle que la sémiotique, les grammaires de texte et les études analytiques du "langage ordinaire" qui se développaient pendant ce temps-là dans la sphère anglophone.

Je reviendrai tout à l'heure sur les conséquences rétrospectives de ces ignorances, mais j'indique d'abord comment *Discourse Analysis* de Harris s'est trouvé incorporé, transformé (et peut-être

défiguré ?) dans cette entreprise : si l'on pose, selon la perspective structuraliste, que le sens d'une surface textuelle existe dans le jeu des rapports (d'équivalence, commutation, paraphrase . . .) qui s'établissent nécessairement entre elle et d'autres surfaces textuelles spécifiques, il en résulte que l'étude des processus discursifs (inhérents à la structure sous-jacente à étudier) suppose la référence à des *ensembles de surfaces* (ou "corpus discursifs") que le dispositif informatique aura pour effet de mettre en état d'auto-paraphrase potentielle, pour l'interroger sur sa structure, en généralisant à des corpus ainsi repérés par leurs "conditions (socio-historiques) de production", les procédures que Harris avait conçues et appliquées sur certaines séquences très particulières, marquées par des répétitions et des stéréotypies internes, dont le fameux "Millions can't be wrong" reste l'exemple princeps.

L'ordre et la disposition de la procédure AAD 69 se trouvaient par là même déterminés, dans une forme qui a été effectivement appliquée à différents corpus socio-historiques, le plus souvent référés à des doctrines idéologiques homogènes (2).

Je n'exposerai pas ici le détail de cette procédure (on peut se reporter sur ce point au document annexe mis à la disposition des participants), mais simplement son principe, à savoir :

1) Une phase de construction socio-historique du système de corpus soumis à l'analyse, chaque corpus étant constitué d'un ensemble de "séquences discursives autonomes" (SDA) de dimension généralement supérieure à la phrase, et pouvant atteindre la taille d'un paragraphe.

2) Une phase de délinéarisation syntaxique (manuelle) des SDA de chaque corpus, dégageant des énoncés élémentaires (munis d'une forme fixe, énonciative et grammaticale, remplie d'éléments lexicaux) et des connecteurs (de détermination, subordination et coordination) entre ces énoncés; chaque SDA est ainsi restructurée sous la forme d'un graphe dont les énoncés constituent les noeuds, les connecteurs constituant des arcs valués entre les noeuds.

   Les données du programme informatique sont donc constituées par une liste d'énoncés, et une liste de relations binaires entre les énoncés.

3) Une phase de traitement informatique, justifiant la prétention automatique de AAD 69, et comportant :
   - un algorithme de comparaison des relations binaires deux à deux, sur la base de leurs contenus lexicaux identiques ou différents, à des places morpho-syntaxiques données, et aboutissant à une liste des couples de relations binaires déterminées comme lexicalement proches (à partir d'un calcul également indiqué dans le document annexe);
   - un algorithme construisant, à partir de ces couples de relations (ou "quadruplets") des chaînes de proximité, elles-mêmes regroupées par transitivité en "domaines sémantiques", qui constituent ainsi des points de rassemblement des sous-séquences (portions de SDA) liées entre elles par des relations de synonymie, métonymie ou paraphrase;

- enfin, un algorithme calculant les rapports de dépendance entre les domaines sémantiques, sur la base des relations amont/aval entre les sous-séquences à l'intérieur du corpus, et réalisant ainsi une reconstitution des trajets micro-argumentatifs propres à ce corpus (le document annexe fournit également des exemples de résultats concrets à ces différents niveaux).

Je conclurai cette présentation rétrospective de AAD 69 par quelques remarques liées à l'état actuel des travaux du groupe de recherche ADELA ("Analyse de discours et lectures d'archive") dont plusieurs orateurs de cette session sont partie prenante, sous diverses formes qu'ils préciseront eux-mêmes.

Plus de quinze ans après l'épisode structuraliste que j'ai évoqué avec ses *a priori* et ses ignorances délibérées, il est temps pour nous de faire le point, sur les différents aspects philosophiques, socio-historiques, linguistiques et informatiques engagés dans cette entreprise interdisciplinaire.

Ma première remarque concerne le rapport entre variation de forme (syntaxique et lexicale) et variation de sens. Nous avons désormais les moyens de soutenir de manière argumentée sur le terrain de l'informatique la thèse selon laquelle les ambiguïtés, métaphores et glissements propres aux langues naturelles sont des propriétés incontournables du champ de l'analyse de discours, qui se différencie par là même de toute perspective strictement informationnelle, documentaire ou "intellectique" (3). Un corpus d'archive textuelle n'est pas une "banque de données".

Simultanément, je soulignerai combien les procédures AAD 69 restent loin de compte quant à l'appréhension de ce jeu entre le même et l'autre, qui caractérise l'hétérogénéité contradictoire de tout champ d'archive : tant par les méthodes de calcul des proximités que par la rigidité pesante de l'analyse syntaxique (manuelle de surcroît) qu'elles supposent, et aussi l'obstination à reconstruire des identités paraphrastiques, les procédures AAD 69 demeuraient bien plus proches que je ne pouvais le supposer à l'époque des évidences empirico-logiques de la lecture. Encore une fois : l'équivoque du rapport à Harris !

Quant au refus historique de tout langage logique de représentation *a priori*, il apparaît de plus en plus justifié dans le domaine de l'informatique en sciences humaines, face à l'élargissement prévisible de l'emprise des langues logiques à référents univoques, importées du domaine des sciences de la nature, des technologies industrielles ou des dispositifs de gestion-contrôle administratifs. Mais tenir cette position n'implique pas nécessairement que l'analyse de discours informatisée doive tendre à réaliser une auto-lecture de la structure des corpus par les corpus eux-mêmes, comme AAD 69 le sous-entendait : ce ne serait finalement qu'une nouvelle théologie, une *théologie de la structure* étayée sur une conception orthopédique de la connaissance; pour tout dire, l'informatique comme prothèse de la lecture, machine à laver les textes, ou appareil à rayons X !

L'ignorance des recherches de la "philosophie du langage ordinaire" semble avoir eu pour conséquences de surestimer, en analyse de discours, le principe de l'homogénéité socio-historique des corpus

discursifs, en restant aveugles sur le rôle théorique que doivent y jouer l'événement, la question, la réplique, l'interruption et l'irruption.

C'est une situation assez intéressante, pour un francophone comme moi, de pouvoir reconnaître de tels défauts, sans tomber immédiatement dans l' "empirisme anglophone", tout en s'adressant à lui. De manière plus générale, c'est même, à mon avis, une condition pour que l'analyse de discours puisse aujourd'hui continuer à suivre son propre chemin.

NOTES

1) L. Althusser, *Lire le Capital*, t. I, Paris, Maspéro, 1968, p. 14-15.

2) Cf. en annexe la liste non exhaustive des travaux réalisés à l'aide de AAD 69.

3) De ce point de vue, la manière dont Maurice Gross et Marcel-Paul Schutzenberger présentent les recherches menées actuellement en ce domaine apparaît partielle, partiale et quelque peu tendancieuse : "Les méthodes (maintenant traditionnelles) d'analyse du discours ou de documentation automatique reposent, sans exception, sur l'utilisation de mots-clés". Suivent des remarques sur les tentatives de raffinements méthodologiques pour remédier à ce déplorable état de fait dans les sciences humaines, et les deux auteurs poursuivent : "Toutes ces méthodes mettent en jeu un langage documentaire particulier, c'est-à-dire un système formalisé dans lequel il est nécessaire de traduire textes et questions." (Compléments sur le traitement des langues naturelles, in *Les enjeux culturels de l'informatisation*, La Documentation Française, 1980, p. 136-137).

BIBLIOGRAPHIE SUR L'ANALYSE DE DISCOURS EN FRANCE.

COURTINE Jean-Jacques, "Quelques problèmes théoriques et méthodologiques en analyse du discours, à propos du discours communiste adressé aux Chrétiens", *Langages*, 62, juin 1981, pp. 9-128.

DUBOIS, Jean, SUMPF Joseph, "L'analyse du discours", *Langages*, 13, mars 1969.

FUCHS Catherine, "Référentiation et paraphrase; variation sur une valeur aspectuelle", *DRLAV*, 21, novembre 1979, pp. 32-41.

GADET Françoise, PECHEUX Michel, *La langue introuvable*, Paris, Maşpéro, 1981.

GUILHAUMOU Jacques, MALDIDIER Denise, "Courte critique pour une longue histoire", *Dialectiques*, 26, 1979.

HAROCHE Claudine, HENRY Paul, PECHEUX Michel, "La sémantique et la coupure saussurienne : langue, langage, discours", *Langages*, 24, décembre 1971, pp. 93-106.

HAROCHE Claudine, PECHEUX Michel, "Manuel pour l'utilisation de la méthode d'analyse automatique du discours (AAD)", *T.A. Informations*, 1, 1972, pp. 13-55.

HENRY Paul, *Le mauvais outil*, Paris, Klincksieck, 1977.

LEON Jacqueline, TORRES-LIMA Maria Emilia, "Etude de certains aspects du fonctionnement de l'AAD; traitement des syntagmes nominaux complexes en expressions figées et segmentation d'un corpus en Séquence Discursives Autonomes", *T.A. Informations*, 1, 1979, pp. 25-46.

MARANDIN Jean-Marie, "Problèmes d'analyse du discours. Essai de description du discours français sur la Chine", *Langages*, 55, septembre 1979, pp. 17-88.

PECHEUX Michel, *Analyse automatique du discours*, Paris, Dunod, 69.

PECHEUX Michel, FUCHS Catherine, "Mises au point et perspectives à propos de l'analyse automatique du discours", *Langages*, 37, mars 1975, pp. 7-80.

PECHEUX Michel, *Les vérités de la Palice*, Paris, Maspéro, 1975.

PECHEUX Michel, HENRY Paul, POITOU Jean-Pierre, HAROCHE Claudine, "Un exemple d'ambiguïté idéologique : le rapport Mansholt", *Technologies, Idéologies et Pratiques*, vol. II, 2, avril-juin 79, pp. 3-83.

CONEIN Bernard, COURTINE Jean-Jacques, GADET Françoise, MARANDIN Jean-Marie, PECHEUX Michel, *Matérialités Discursives I.*, (actes du colloguqe de Nanterre, Paris X, avril 1980), Lille, PUL, 1981.

ROBIN Régine, *Histoire et Linguistique*, Paris, Armand Colin, 73.

LISTE NON-EXHAUSTIVE DE TRAVAUX REALISES A L'AIDE DE AAD 69.

BONNAFOUS, S. 1980.
Les motions du congrès de Metz (1979) du parti socialiste : processus discursifs et structures lexicales. Thèse de IIIème cycle en linguistique, Université de Paris X, Nanterre, ronéo, 259 p. + annexes.

COTE, P. 1981.
L'analyse automatique du discours (AAD) de Michel Pêcheux, Documents généraux du GRIDEQ (Université du Québec à Rimouski), n. 8, 78 p.

GAYOT, G. 1981.
Du pouvoir et des lumières dans la fraternité maçonnique au XVIIIème siècle, in *Peuple et Pouvoir*, Essais de lexicologie, Lille, PUL.

GAYOT, G. et PECHEUX, M. 1971.
Le "Portrait" de Claude de Saint-Martin. *Annales*, XXVI (3-4) : 681-704.

PECHEUX, M. 1978.
Are the masses an animate object ? in D. Sankoff (ed), *Linguistic Variation : models and methods*, New York : Academic Press, 251-266.

PECHEUX, M., HENRY, P., POITOU, J.-P. et HAROCHE Cl., 1979. Un exemple d'ambiguïté idéologique : le rapport Mansholt, Technologies, Idéologies et Pratiques, I (2) : 3-83.

POITOU, J.-P., 1978.
La dynamique des groupes : une idéologie au travail. Paris, Editions du CNRS, 257 p.