# Lexique et syntaxe en analyse du discours : propositions d'analyse automatique

par

J.-J. COURTINE

*Université de Grenoble II - FRANCE*

Ce travail s'inscrit dans le champ de l'analyse du discours, discipline qui s'est développée, en France notamment, sur les marges de la linguistique, en se donnant comme objet privilégié *le discours politique*. Nous considérerons ici le discours politique comme point de contact entre la matérialité de la langue, au sens que les linguistes donnent à ce terme, et l'espace historique des contradictions où des forces politiques s'affrontent : lieu d'une rencontre, donc, entre le linguistique et le champ politique comme extérieur de la langue : champ de forces, champ de luttes politiques qui visent à transformer, le rapport des forces caractérisant une conjoncture historique donnée, étant entendu que des discours y circulent, s'y affrontent, s'y recouvrent . . . à partir de positions déterminées.

Ce point de départ entraîne les conséquences suivantes :

1) Analyser des discours, c'est analyser une matérialité linguistique, c'est-à-dire les formes d'organisation lexico-sémantiques *et* syntaxiques de séquences discursives données (ou discours concrets) *sans dissociation possible* de l'organisation lexicale et de l'organisation syntaxique de ces séquences.
2) Il conviendra d'établir le rapport entre la matérialité linguistique d'une séquence discursive ainsi définie et les contradictions du champ historico-politique qui dominent les conditions dans lesquelles cette séquence est produite.
3) L'adoption d'une telle démarche rend selon nous nécessaire la mise en oeuvre d'une *procédure automatique* d'analyse du discours qui garantisse l'univocité des manipulations effectuées sur le matériau discursif, en même temps que l'exhaustivité de la description et que la reproductibilité de l'analyse; et cela en tenant compte des possibilités nouvelles de description de discours offertes par l'existence de logiciels de traitement de texte permettant la description et l'exploration de vastes corpus de données discursives : c'est le sens de notre recours au logiciel DEREDEC (PLANTE, 79), dont la description est donnée dans ce même volume dans le texte de P. PLANTE. Nous y renvoyons pour plus de détails.

Nous allons ainsi examiner tout d'abord la manière dont le rapport entre syntaxe et lexique est posé en analyse du discours, en montrant que ces deux éléments y sont généralement dissociés; puis nous formulerons quelques propositions d'analyse automatique sur la base d'un exemple, en réalisant la construction de lexiques à partir de contraintes syntaxiques.

## 1.- LE RAPPORT ENTRE LEXIQUE ET SYNTAXE EN ANALYSE DU DISCOURS

a) En lexicométrie :

Il s'agit ici de procédures d'investigation contextuelles basées sur le calcul statistique. Ces recherches fréquentielles de co-occurrences, même si elles ont pu donner lieu à des travaux historiques pertinents quant à la description du vocabulaire politique de telle ou telle époque, s'inscrivent dans une perspective différente de celle que nous présentons ici, dans la mesure où elles ne prennent pas en compte la structuration syntaxique des discours qu'elles analysent. Effacement de la syntaxe donc; et pourtant,

un discours, *ce n'est pas une simple concaténation de mots.*

b) En analyse harrissienne :

L'application en analyse du discours du modèle de HARRIS (52) aboutit à la constitution de classes d'équivalence distributionnelle sur un énoncé suivi. Cette constitution suppose une manipulation transformationnelle des énoncés qui vise à les normaliser afin d'obtenir des classes homogènes de propositions. Une telle perspective est dominée par le postulat de neutralité sémantique des transformations : les manipulations transformationnelles opérées (réduction des passifs en actifs, par exemple) ne changent pas le sens des phrases. Cette conception se fonde ainsi sur le présupposé selon lequel le contenu lexico-sémantique des énoncés peut être séparé de la forme syntaxique des énoncés : elle opère une *dissociation* forme du discours/contenu du discours, à l'intérieur d'une position lexicaliste qui pose *l'indifférence* du contenu lexico-sémantique des discours à la forme syntaxique de ces mêmes discours.

La démarche tend donc vers une position proche de celle de la lexicométrie sous la forme non pas d'un effacement pur et simple de la syntaxe, mais d'une *neutralisation de la syntaxe* conçue comme espace d'homogénéisation des discours. Le discours y est pensé sur le modèle du dictionnaire de la langue.

## 2.- L'ORGANISATION LEXICALE D'UN DISCOURS A PARTIR DE CONTRAINTES SYNTAXIQUES

a) Le logiciel DEREDEC :

Le DEREDEC est un logiciel consacré au traitement linguistique ainsi qu'à "L'analyse de contenu" des textes; sa conception générale en fait un instrument particulièrement adéquat aux buts que nous visons ici. Outre ses qualités au plan strictement informatique, il a permis l'écriture d'une *grammaire de reconnaissance* du français, indexant automatiquement des descriptions syntaxiques arborescentes sur les séquences d'entrée d'un corpus discursif. La réalisation de cette condition (donner une base syntaxique automatisable à la description linguistique d'une séquence discursive) constitue de notre point de vue une condition primordiale à tout traitement d'analyse du discours.

La grammaire de surface élaborée est une grammaire récursive à réseaux de transition, ascendante, sensible au contexte et non-déterministe. Elle permet de construire sur chacune des phrases d'une séquence discursive des structures syntaxiques arborescentes incluant des *relations de dépendance contextuelle* entre certains éléments des structures (en particulier les relations : *thème/propos* et *déterminant/déterminé* à l'intérieur du groupe nominal, dont nous nous servirons seules ici); on appellera *description de texte* (DDT) le résultat de l'application de la grammaire à une séquence discursive.

Les *modèles d'exploration* programmables en DEREDEC constituent différentes manipulations des DDT construites, qui correspondent à différents types de dépistage à éléments dans les structures.

Nous allons en donner un aperçu en fournissant un exemple d'application d'un modèle d'exploration à une DDT produite à partir d'une séquence discursive extraite d'un corpus que nous avons traité ailleurs (COURTINE, 81) : il s'agit d'un corpus de discours du Parti Communiste Français adressé aux chrétiens (de 1936 à 1976); la séquence discursive à laquelle nous avons appliqué la grammaire de surface DEREDEC consiste en une adresse de G. MARCHAIS aux chrétiens, à Lyon, en juin 1976.

Nous nous sommes servis d'un nombre restreint de modèles d'exploration par rapport aux possibilités offertes dans ce domaine par le système : l'objectif retenu ici est de tester les possibilités d'exploration de *l'organisation lexicale* de la séquence discursive analysée.

b) La construction de lexiques :

Les modèles d'exploration utilisés ont donc conduit à la construction de lexiques à partir de la DDT obtenue.

On appellera *lexique* un tri alphabétique d'expressions atomiques (ou unités minimales de description syntaxique) réalisé selon certaines contraintes et indiquant le nombre d'occurrences dans la DDT de l'expression atomique en fonction de la contrainte choisie.

6 lexiques ont été construits :

| | | | | |
|---|---|---|---|---|
| $L_1$ | : | Lexique des formes pleines (N, V et Adj.) | = | 3.521 entrées |
| $L_2$ | : | Lexique des formes pleines thématisées | = | 303 entrées |
| $L_3$ | : | Lexique des formes pleines du propos | = | 2.134 entrées |
| $L_4$ | : | Lexique des nominaux thématisés | = | 248 entrées |
| $L_5$ | : | Lexique des nominaux déterminés | = | 437 entrées |
| $L_6$ | : | Lexique des formes pleines qui déterminent les nominaux | = | 1.236 entrées |

Les contraintes admises dans la construction des lexiques sont donc définies dans la grammaire, qu'il s'agisse simplement de *catégorisations* (formes pleines pour L1), ou bien de contraintes plus fortes, combinant la présence d'une catégorie et de *relations de dépendance contextuelle* (thème/propos pour $L_2$, $L_3$, $L_4$ ; déterminant/déterminé pour $L_5$ et $L_6$) comme condition de tri des expressions atomiques classées et recensées.

Cela nous ramène aux critiques adressées plus haut aux procédures d'investigation contextuelle ignorant la syntaxe : les lexiques constitués ici nous semblent échapper à de telles critiques : leur élaboration intervient en effet après l'indexation des structures syntaxiques à la séquence, ce qui revient à donner un environnement linguistique aux comptages opérés; les DDT s'inscrivent en effet comme contraintes dans les modèles d'exploration qui construisent les lexiques : cela permettra d'interpréter à partir de

*relations structurelles* définies dans les fonctions de description les résultats de l'application des fonctions d'exploration. La séquence analysée n'est plus une concaténation aléatoire d'objets quelconques, mais une suite d'éléments ordonnés et hiérarchisés.

De telles relations structurelles peuvent être dégagées de la comparaison entre certains des lexiques construits pris deux à deux. Par exemple :

1) *Relation d'inclusion* entre les expressions classées dans deux lexiques : c'est le cas de $L_4$ dont les expressions forment un sous-ensemble (nominaux thématisés) des expressions classées dans $L_2$ (formes pleines thématisées).

2) *Relation de distribution sur une relation de dépendance contextuelle* (déterminant/déterminé) *de deux catégories* : c'est le cas des expressions classées en $L_5$ (nominaux déterminés), par rapport aux expressions classées en $L_6$ (formes pleines qui déterminent les nominaux).

3) *Relation de distribution d'une catégorie sur deux relations de dépendance contextuelle* : c'est le cas des expressions classées en $L_4$ (nominaux thématisés) par rapport aux expressions classées en $L_5$ (nominaux déterminés).

Attardons-nous sur ce dernier exemple, afin d'observer si des comptages pratiqués sur une telle base permettent des interprétations qui correspondent à des observations ou des intuitions sur ce type de discours.

Le lexique $L_4$ renferme les expressions atomiques classées comme "nominaux thématisés" : il comporte 248 entrées dans le détail desquelles nous n'irons pas. Nous extrairons simplement de ce lexique le sous-ensemble des nominaux thématisés ayant la plus forte occurrence, et ceci au-dessus d'une borne arbitrairement fixée à 5 occurrences : la liste obtenue est de 8 expressions. Reportons-nous à présent au lexique $L_5$ des nominaux déterminés, et observons le comportement de ces 8 expressions dans $L_5$ en y relevant leur fréquence d'apparition; puis mesurons l'écart entre le nombre d'occurrences du même item dans $L_4$ et $L_5$. Le résultat est donné au tableau I.

TABLEAU I : occurrence des nominaux les plus fréquemment thématisés (borne d'occurrence $\geqslant$ 5)

| N | Dans $L_4$ | Dans $L_5$ | ECART |
|---|---|---|---|
| Chrétiens | 9 | 2 | 7 |
| Communistes | 8 | 3 | 5 |
| Crise | 6 | 3 | 3 |
| Français | 5 | 2 | 3 |
| France | 11 | 4 | 7 |
| Parti | 5 | 4 | 1 |
| Pays | 5 | 4 | 1 |
| Peuple | 13 | 4 | 9 |

Si l'on conserve la même borne ($\geqslant 5$) pour décider du caractère significatif des écarts calculés, on s'aperçoit que certains nominaux, souvent thématisés, sont inversement faiblement déterminés; ce sont dans l'ordre décroissant à partir du plus fort écart : *peuple, Chrétiens, France, Communistes.*

Réalisons à présent l'opération inverse en observant le comportement des expressions apparaissant le plus fréquemment en position de nominal déterminé (en $L_5$) dans le lexique $L_4$, et donnons l'écart observé au tableau II.

TABLEAU II : Occurrence des nominaux les plus fréquemment déterminés (borne d'occurrence $\geqslant 5$).

| N | Dans $L_4$ | Dans $L_5$ | ECART |
|---|---|---|---|
| Action | 5 | 3 | 2 |
| Classe | 8 | 0 | 8 |
| Démocratie | 7 | 2 | 5 |
| Hommes | 8 | 4 | 4 |
| Monde | 6 | 1 | 5 |
| Union | 8 | 3 | 5 |
| Vie | 7 | 1 | 6 |

On observe ainsi qu'à l'inverse, certains nominaux, souvent déterminés, sont faiblement thématisés; il s'agit, en conservant le même seuil d'écart et la même présentation d'ordre que précédemment, de *classe, vie, monde, union, démocratie.*

Ce qui apparaît ainsi, c'est bien *une relation de distribution complémentaire* des expressions nominales classées en $L_4$ et $L_5$ sur les deux relations de dépendance contextuelle de thématisation et de détermination, en ce qui concerne tout au moins certaines occurrences fortes en $L_4$ et $L_5$ : certains nominaux, souvent thématisés, sont peu déterminés; des notions comme *le peuple, les chrétiens, la France, les communistes*, qui figurent fréquemment en position thématique, se passent de détermination : "on sait ce que c'est", on peut en parler, en faire un thème de son discours, parce que "cela va de soi". Ces notions, fréquentes dans le discours politique en général, et ici dans le discours communiste, sont *saturées par le consensus idéologique qui stabilise leur référence* : elles réalisent, dans le discours politique français, une véritable *intersection lexicale* entre les formes d'organisation lexicale et de construction de la référence des mots propres à des discours différents, et éventuellement antagonistes. Elles tendent vers le statut linguistique du *nom propre* ou vers la forme logique de la *tautologie* ("la France c'est la France", "le peuple, c'est le peuple" . . ., définitions qui peuvent d'ailleurs se croiser en : "le peuple c'est la France". . .).

La séquence discursive analysée en porte quelques traces éloquentes.

Laissons parler G. MARCHAIS (ces exemples sont extraits de la séquence discursive analysée) :

> "Je ne veux pas ce soir prétendre donner une définition scientifique de ce qu'est le peuple. *Chacun sait ce que parler veut dire*"

. . . et plus loin :

> "(Le peuple) ce sont ceux et celles qui ont fait de notre pays ce qu'il est".

D'un côté donc, les notions qui vont de soi et de l'autre, du côté des nominaux fréquemment déterminés et peu souvent thématisés ce qui doit être défini, déterminé, expliqué : les *concepts*, les mots du "vocabulaire de parti", toujours à définir; la *classe* ("ouvrière", "exploitée", . . .), l'*union* ("du peuple de France", "des communistes et des chrétiens", "de tous les travailleurs", . . .), la *démocratie* ("politique", "économique", "moderne", "socialiste", . . .). Et également *les notions à réinterpréter*, les mots à arracher à leur sens commun : le *monde* ("meilleur", "de demain", . . .), la *vie* ("plus belle", "plus juste", "plus heureuse", "plus libre" . . .).

L'organisation lexicale de la séquence analysée, interprétée à travers les fonctions descriptives et exploratrices qui ont été exposées, laisse ainsi paraître la manière dont "les mots changent de sens en fonction des positions de ceux qui les emploient"; on y repère la trace des zones de *neutralisation discursive*, où les mots sont pris dans le consensus du "même sens pour tous", celle aussi des zones de *constitution et de clôture d'un savoir*, où les concepts reçoivent leur définition, celle enfin des zones où la *contradiction affleure*, zones où les mots sont des enjeux.

Notons pour conclure les limites de l'analyse produite : pour rendre explicite l'existence de ces différentes zones de stabilité ou d'instabilité des expressions, il convient de faire référence à l'espace historique où des discours s'affrontent. Cet espace, que nous nommons : *interdiscours* (COURTINE, 81) implique la mise en rapport dans un corpus discursif de plusieurs séquences discursives antagonistes; il est absent de l'expérience limitée que nous réalisons ici.

Limites également dues à l'optique de comptage sur laquelle les lexiques sont bâtis : de simples décomptes fréquentiels ne sauraient se substituer à l'analyse du fonctionnement linguistique de la séquence discursive. Ils peuvent cependant être précieux dans une phase préparatoire à un travail d'analyse du discours, opérant un débroussaillage empirique antérieurement à l'application d'une procédure linguistique.

BIBLIOGRAPHIE

COURTINE, J.-J. (1981) : "Analyse du discours politique", dans *Langages*, n. 62, juin 1981.

HARRIS, Z.S. (1952) : "Discourse Analysis", dans *Language*, vol. 28.

PECHEUX, M. (1969) : "*L'analyse automatique du discours*", Dunod, Paris.

PLANTE, P. (1979) : "*DEREDEC, un logiciel pour le traitement linguistique et l'analyse du contenu des textes*", Thèse de Doctorat, Université du Québec à Montréal.