Linguistique et algorithmes textuels

par

Jacqueline LEON
J.-M. MARANDIN
L.I.S.H. - C.N.R.S. Paris - FRANCE

I. Introduction

L'analyse de discours renvoie idéalement pour nous à une *morphologie discursive*, ou description des formes regroupant ou distribuant les éléments hétérogènes de toutes séquences discursives, et à une *lecture*, quand l'analyste de discours règle et écrit cette description ou qu'il s'autorise de cette description pour interpréter le corpus de discours décrit. Si notre projet s'inscrit historiquement dans la problématique d'analyse de discours issue en France de *Discourse Analysis* (et en particulier d'AAD 69), nous pointons, en donnant cette définition, un nouvel élargissement de la problématique et un changement de perspectives.

II. Présentation générale

2.1 En employant "morphologie", nous soulignons qu'une analyse distributionnelle est insuffisante pour décrire une suite discursive (1). Nous prenons en compte plusieurs propriétés formelles tant au plan paradigmatique (où seule était prise en compte l'identité de sens définie par l'identité totale ou partielle de la distribution) qu'au plan syntagmatique ou séquentiel (où seule était prise en compte la disposition relative sur l'axe linéaire des classes d'équivalence); sur ce dernier point, AAD 69 avait développé une problématique originale de recherche de chemins argumentatifs profonds.

Nous soulignons d'autre part, que la suite discursive en tant qu'elle est une séquence énoncée à un moment donné par un locuteur a été insuffisamment décrite et principalement en référence aux thèses qui définissent notre problématique, thèses concernant la formation discursive, l'autonomie de la syntaxe, l'interpellation idéologique.

En employant "lecture", nous ne faisons que reconnaître la place occupée par l'analyste de discours face au texte qu'il décrit. Trop longtemps, les analystes de discours ont pris pour un modèle de production leur grille de description : toute lecture découpe le texte, en privilégie certains éléments pour en occulter d'autres, rapproche ce qu'elle a dispersé, disperse ce qui était conjoint. Nous renvoyons aux premières pages de *La seconde Main* de A. Compagnon (Seuil, 1979) pour la description de ce geste. Notre pari est de faire de ces interventions, opérées en quelque sorte de façon "sauvage ou inconsciente" dans la "lecture spontanée", des interventions réglées triturant l'objet à lire selon les axes mêmes qui le structurent. L'Analyse du discours ne serait plus une prothèse de la lecture, mais une provocation à la lecture.

2.2 Dans cette perspective, la construction d'algorithmes n'est qu'une partie de la problématique d'ensemble. Les algorithmes que nous présentons ici sont en cours d'élaboration : il s'agit donc d'un exposé programmatique. A cette construction, nous voyons deux raisons :

a) Une raison interne : la construction d'algorithme oblige à une approche constructiviste et à la définition la plus claire possible des objectifs et des objets. Elle joue le rôle de dispositif

d'expérimentation; dans la phase actuelle, il faut procéder à la description d'un grand nombre de discours pour tester la validité de nos hypothèses et non, comme cela a trop souvent été le cas, hypostasier une description comme modèle de discours.

b) Une raison externe : les algorithmes que nous présentons correspondent à une philosophie différente de celle qui préside à la constitution de bases de données ou aux systèmes d'intelligence artificielle. Les textes dans leur instance historique ne sont pas de pures séquences d'items encodant une information ou une fonction sémantique, et manipulables par un sujet (ou une machine simulant le sujet humain) sur le mode des choses. On ne peut ignorer que les textes sont le produit d'une histoire et qu'ils circulent : c'est ce mode d'être que nous désignons par le terme de discours.

2.3

2.3.1 Notre description débute, une fois la phase de constitution de corpus achevée (et non décrite ici), par une analyse syntaxique : en effet les éléments manipulés par les algorithmes textuels sont repérés syntaxiquement. A cette position trois raisons :

a) Nous refusons de traiter les suites discursives dans la grille d'un langage de représentation sémantique a priori. Il ne s'agit pas pour nous de reconnaître le contenu véhiculé par les suites discursives (une série d'informations accrochées au train des mots) mais de décrire ces suites.

b) Les suites discursives rassemblées en corpus ne sont pas amorphes : elles sont organisées comme des énoncés et des séquences, i.e. elles sont syntaxiquement organisées et ont été énoncées. Nous sommes dans la position de l'archéologue face à un rassemblement de tessons au cours d'une fouille. C'est par la description de ces tessons qu'il pourra construire un savoir sur les vases et peut-être même ce qui a présidé à leur fabrication : leur matière, leur forme, leur couleur ne sont pas du tout indifférentes. C'est bien sûr reprendre là, et la pousser à bout, l'opposition désormais célèbre de M. Foucault dans *L'Archéologie du Savoir* : les textes ne sont pas des documents mais des monuments.

c) Enfin, nous tenons le niveau syntaxique comme une structure irréductible de la langue. C'est reprendre la thèse de l'autonomie de la syntaxe telle qu'elle a été soutenue par la linguistique transformationnelle, même si cette thèse est problématique, en particulier en ce qui concerne la définition de la notion de règle et donc de la forme et de la définition de la syntaxe. Nous dirons, en adoptant une position contradictoire, que le système syntaxique est irréductible à un calcul sémantique mais qu'il n'est pas un système de règles consistant et plein, indifférent au travail de la signification.

2.3.2 Techniquement, il s'agit du redoutable problème de la construction d'un analyseur syntaxique (rappelons que jusqu'à présent l'analyse syntaxique du programme AAD 69 était réalisée manuellement). Pour la construction des algorithmes, nous avons choisi de travailler avec la Grammaire de

Surface (GDS) de Pierre Plante (qu'il exposera ici-même); elle construit ce que Plante appelle des "Relations de Dépendance Contextuelle", relations syntaxiques entre constituants (GN, GV, GP). Qu'une telle grammaire analyse de façon semblable des phrases telles que :

- Pierre donne une pomme à Marie.
- Pierre donne le cornet à piston.
- Pierre donne un concert à Paris.

ne nous apparaît pas rédhibitoire, au contraire. La délimitation des constituants et leur interrelations ne sont évidentes que dans ces phrases pour linguistes. Elles dépendent dans un énoncé, de la façon dont sont construits les objets de discours, de ce que la linguistique range sous l'étiquette informe de savoir extra-linguistique. Autrement dit, il ne peut s'agir d'une décision purement syntaxique. Il est illusoire de dissocier radicalement, théoriquement et dans la procédure, la phase d'analyse syntaxique et la phase d'analyse discursive.

C'est pourquoi nous avons choisi la GDS programmée en DEREDEC comme syntaxe minimale en raison de sa "faiblesse" entendue comme sa capacité à ne pas résoudre les ambiguïtés "à tout prix". Il faut cependant prévoir, à long terme, une stratégie interactive revenant sur cette analyse minimale pour l'affiner : le corpus serait alors soumis à une seconde analyse syntaxique prenant en compte les éléments discursifs repérés dans les résultats fournis par une première série d'algorithmes. On pourra ensuite appliquer à ce corpus des algorithmes de second niveau.

Cette syntaxe plus fine articulerait deux systèmes : un système de règles hiérarchisées sous forme d'arbres et un système de règles séquentielles travaillant sur l'énonciation et les liens interphrastiques. Les énoncés ne seraient plus forcés dans une forme propositionnelle unique, comme dans AAD 69; on considèrera comme unités d'analyse à part entière des énoncés non propositionnels tels que des phrases sans verbe (ex. : "une bière, et je m'en vais") ou des groupes nominaux complexes.

Nous nous proposons donc d'élaborer une procédure d'Analyse du Discours articulée sur deux niveaux de syntaxes et sur deux niveaux d'algorithmes, assurant l'interaction entre le linguistique et le discursif. A l'heure actuelle, seuls les algorithmes de premier niveau sont à l'étude. Ce sont ceux que nous allons présenter maintenant.

III. Les algorithmes de premier niveau

3.1 Présentation.

Pour décrire des suites discursives, analysées syntaxiquement par la grammaire de surface précédemment évoquée, nous construisons deux espaces de description que nous appelons vertical et horizontal, et, dans chaque espace, plusieurs algorithmes répondant à des définitions d'objets différents.

a) L'espace vertical renvoie à la dimension historique du discours et commande des algorithmes de regroupement d'unités extraites des suites discursives : tout énoncé est pris dans une série d'énoncés appartenant à d'autres séquences discursives émises antérieurement ou de façon contemporaine et qui constituent sa condition d'existence (ce que Foucault nomme une formation discursive ou Pêcheux l'inter-discours (2), et Courtine un domaine de mémoire).

b) L'espace horizontal renvoie au "fil du discours", à cette unité complexe où se séquentialisent les suites discursives (ce que Pêcheux nomme l'intra-discours) : tout énoncé est pris dans un enchaînement d'énoncés organisé par rapport à un système de places énonciatives et selon plusieurs systèmes de disposition (3).

Ces deux espaces ne sont pas absolus; ils ne sont pas sans rapports l'un avec l'autre : les différents systèmes de séquentialisation ne sont pas l'invention sans cesse renouvelée d'un sujet d'énonciation; ils sont réglés par une formation discursive. Dans la procédure, les algorithmes verticaux, en contractant le fil du discours, permettent de confronter chaque point de ce fil à son ensemble : les algorithmes verticaux fonctionnent alors comme une mémoire.

3.2 Les algorithmes verticaux

Les algorithmes verticaux opèrent des regroupements sur la base du même (à distinguer de l'identique : le même mot peut avoir plusieurs sens), en privilégiant ce qui est répété dans la séquence discursive. Ils auront pour tâche de dégager des points de stabilité réalisés lexicalement en regroupant des items lexicaux ayant même base morphologique, et des points de stabilité syntaxiques en regroupant des structures syntaxiques (lexique, proposition ou phrase) appartenant à une même famille de paraphrase.

3.2.1 Algorithme 1 : variation syntaxique d'un item lexical

Cet algorithme est réalisé en trois étapes :

a) Il s'agit de produire, pour chaque item lexical du corpus, la liste de ses catégorisations (N ou V pour "lever" par exemple) et de ses positions syntaxiques (SN ou SV, etc. . .). On ne prendra en compte que les items lexicaux pleins (N, V, adj. et les adverbes en -ment) et non les mots fonctionnels de la langue. Cette catégorisation sera opérée sur les structures syntaxiques issues de l'analyse syntaxique.

b) On regroupe les items ayant la même base morphologique. Ainsi, les formes : nation, national, nationalisation, nationaliser seront considérées comme un seul et même item lexical.

c) On regroupe les items affectés de la même liste ou de listes dont la similitude est jugée pertinente par l'analyste.

L'objectif de cet algorithme est d'explorer le traitement syntaxique du lexique dans un corpus. Il repose sur l'hypothèse que la signification d'un item lexical dépend de sa définition syntaxique catégorielle et/ou contextuelle et que la fonction discursive d'un item lexical est liée à sa définition syntaxique. Par exemple un item lexical fonctionnera comme thème de discours en apparaissant régulièrement dans des places syntaxiques induisant l'interprétation "tête de phrase" ou détachement à gauche. Il faut prendre garde cependant à ne pas associer systématiquement une fonction discursive et une liste de places syntaxiques. Il n'est en effet pas du tout certain que dans tout discours les thèmes de discours soient réalisés par les mêmes fonctionnements syntaxiques; il faut procéder discours par discours en travaillant sur le lexique syntaxiquement défini de chacun.

3.2.2 Algorithme 2 : constellation

Cet algorithme a pour objectif de construire des regroupements lexicaux définissant des référentiels discursifs. Un référentiel de discours étant pris non comme une situation réelle d'un monde préexistant se constituant comme condition d'un énoncé, mais comme les constructions discursives qui permettent cet énoncé. Le référentiel forme le lieu d'émergence, les possibilités d'apparition de l'énoncé.

La procédure de regroupements des constellations est la suivante : si deux ou n structures syntaxiques ont un item lexical commun, on regroupe les items lexicaux restants de ces structures. Ce regroupement est appelé "constellation" et l'item lexical commun est appelé "tête de constellation".

Dans une seconde étape, si deux constellations (construites séparément) ont un item commun, on formera une seule constellation, par transitivité, qui aura alors plusieurs têtes.

Cette procédure est un algorithme de cooccurrence fondé sur la syntaxe et non sur un calcul sémantique. Les items cooccurrents sont en effet regroupés dans une structure syntaxique élémentaire et c'est ce lien syntaxique entre des items lexicaux qui forme l'indice d'un référentiel de discours dans lequel ces items prennent sens et référence.

Les résultats de cet algorithme peuvent servir d'entrée à d'autres algorithmes, tel que le troisième algorithme vertical dit de "paraphrase".

3.2.3 Algorithme 3 : paraphrase

Cet algorithme repose sur une notion de la paraphrase prise dans son sens strict de paraphrase syntaxique. Notre idée est qu'il y a stabilité référentielle, par construction discursive entre des phrases entretenant une relation de paraphrase syntaxique. Elles sont équivalentes par rapport à un référentiel de discours, mais elles n'ont pas le même sens. On ne doit pas en effet confondre, lorsqu'on parle de paraphrase, référent extra-discursif, identité sémantique et référent construit

discursivement; nous appellerons "miroitement syntaxique" la différence de sens liée à des différences syntaxiques conservant l'identité d'un référentiel discursif (4)

Ainsi, nous dirons que deux structures syntaxiques en paraphrase sont équivalentes non parce que de l'une à l'autre est conservée la valeur logique de vérité, mais parce qu'est conservé le référentiel de discours.

La procédure de repérage des paraphrases syntaxiques et l'étude des miroitements nécessitent un certain nombre de conditions :

- elle suppose une liste, la plus complète possible, des relations de paraphrase syntaxique (passif/actif, adverbe de phrase/circonstanciels, etc.). Cette liste restant à construire;
- les paraphrases syntaxiques peuvent mettre en jeu des structures plus complexes que des propositions simples : ainsi des enchâssements de propositions ou même des phrases. La recherche de paraphrases s'effectuera donc dans le cadre de la phrase, définie comme suite discursive comprise entre deux marques d'arrêt;
- on l'a vu à l'instant, on ne peut rechercher des paraphrases qu'à l'intérieur d'un référentiel discursif donné : le repérage des paraphrases et des miroitements sera donc limité aux phrases ayant du lexique appartenant à une même constellation, et en particulier aux phrases ayant du lexique commun.

A partir de ces indications, la procédure consistera à comparer toutes les phrases construites sur du lexique appartenant à une même constellation, à la liste des paraphrases syntaxiques entrée en donnée et à regrouper les phrases qui déclenchent un effet de paraphrase.

Il n'est pas encore certain qu'un algorithme de paraphrase soit maintenu : c'est à titre expérimental qu'il sera testé sur des corpus afin d'essayer de savoir si la problématique de la paraphrase peut réellement faire l'objet de l'analyse de discours ou si elle n'est pas simplement un artefact issu des problématiques de la linguistique. Ce qui reste par contre intéressant c'est l'idée de miroitement sémantique combinant les deux types de miroitements précédemment évoqués; ces miroitements ne peuvent se concevoir comme des différences de sens stables, propriétés de la langue, mais comme des phénomènes particuliers au processus discursif qui doivent être interprétés pour chaque discours.

3.3 Les algorithmes horizontaux

Les algorithmes horizontaux étudient la séquence discursive par elle-même contrairement aux algorithmes verticaux qui la découpent en unités syntaxiquement définies.

L'objectif des algorithmes horizontaux consistera à repérer dans la succession des suites discursives les marques du fil du discours (telles que le système modalité-aspect-temps-déterminants-connecteurs);

On considèrera cependant que ce qui est implicite, non marqué, est aussi important que les marques de l'enchaînement des séquences discursives. Etant donné la difficulté de mise en oeuvre du repérage de ces marques, nous ne pouvons dès maintenant qu'avancer des esquisses d'algorithmes horizontaux et non présenter des procédures comme pour les algorithmes verticaux.

Deux algorithmes sont envisagés :

3.3.1 Algorithme "dynamique lexicale"

A partir de l'algorithme 1 vertical produisant une liste de définitions syntaxiques du lexique (catégorie + position syntaxique), on pourrait étudier la transformation de ces définitions dans le déroulement du fil du discours. Il serait ainsi possible d'approcher la notion de fonction discursive d'un item lexical, évoquée précédemment.

3.3.2 Algorithme de découpage de la séquence

Cet algorithme aurait pour but de découper automatiquement un corpus en séquences discursives sur des critères intrinsèques (le système MATCD), contrairement à la procédure AAD 69 qui opérait ce découpage à partir de critères extrinsèques tels que la construction d'unités autonomes autour d'un thème sélectionné par l'analyste.

En ce qui concerne ces algorithmes, le travail reste à faire. Il n'est pas possible, à l'heure actuelle, d'envisager précisément la structure de ces algorithmes. Le dernier en particulier est très programmatique : il permet de se donner un cadre pour étudier les séquences. En effet, on ne peut définir les séquences par bornage, mais on peut se donner des critères pour les borner et construire ainsi un objet d'étude empirique.

Cette série de cinq algorithmes, appelés algorithmes de premier niveau, ont pour objectif de définir un objet discursif. L'application d'algorithmes, dits de second niveau, au corpus reconsidéré en fonction de cet objet et analysé par une syntaxe plus fine, permettront d'opérer une certaine lecture, ni "objectivante" ni "sauvage", mais assimilée à une réécriture de l'objet à lire.

Ainsi se trouvent progressivement réinvestis dans l'analyse de discours des thèmes qui lui étaient originellement étrangers, tels que la grammaire générative transformationnelle, au niveau de surface et au niveau plus profond de la manipulation de structures de base, la grammaire de texte et l'analyse de l'énonciation.

NOTES

1) Nous appelons *"suite discursive"* un fragment de discours tel qu'il apparaît, qu'il est lu dans une appréhension naïve; nous appellerons *"séquence discursive"* le système construit par la description faisant d'une suite un tout; c'est dire que la construction des séquences discursives ne se résume pas à une segmentation des suites et que la même suite est susceptible d'appartenir à plusieurs séquences.

2) Nous appelons *phrase*, une suite discursive entre deux marques d'arrêt fortes; *proposition*, une unité syntaxique simple correspondant à l'axiome P chomskyen ou à la proposition de l'analyse logique scolaire; *proposition complexe*, un enchâssement ou une coordination de propositions; *énoncé*, le système syntaxique et énonciatif construit par la description faisant d'une phrase ou d'une suite de phrases un tout.

3) En reprenant ce terme à la rhétorique, nous désignons la distribution d'éléments définis par le fil du discours; les récits où le discours argumentatif ne sont que des formes particulièrement codifiées de disposition.

4) On oppose à ces miroitements syntaxiques, les miroitements lexicaux émergeant entre deux énoncés à structure syntaxique constante mais présentant une variation au niveau du lexique.