# Contenu des discours et approche statistique

Max Reinert[1]

*" L'écrivain est lui-même comme un nouvel idiome qui se construit… "*
Merleau-Ponty

[in « Analyse statistique de données textuelles en sciences de gestion », coordonné par Claire Gauzente et Dominique Peyrat-Guillard, Editions Management & Société, 2007]

Dans " l'écriture et la différence ", J. Derrida insiste sur cet " étrange mouvement " de l'écriture, qui vise l'unité, et vit de la différence qu'impose chaque instant, reportant dans un horizon toujours repoussé, ce beau rêve unitaire. Celui-ci ne se situe dans aucun futur ; Il voile plutôt une origine : *" L'infini (ment) (autre) ne peut être objet puisqu'il est parole, origine du sens et du monde "* (Derrida). L'inaccessibilité de ce qui cherche à s'écrire n'est-il pas également dans cette expression évanescente d'un contenu échappant à toute maîtrise, à tout effort pour le contenir ? Progressivement un texte finit par articuler une sorte de logique d'auteur, qui a défaut de retenir ce contenu infini préalable à toute visée permet du moins à une communauté de retrouver ses marques, les traces de ce qui inlassablement revient, et à travers elle l'espérance d'une parole perçue comme vraie, promesse d'un monde enfin sien.
Dans cet essai sur l'approche des discours par des méthodes statistiques, nous proposons d'aborder ce débordement perpétuel du sens par un contenu immédiat dans son mouvement même, dans son renouvellement de chaque instant, dans son brisement incessant. Le discours s'ancre d'abord dans une dynamique temporelle bien avant d'être enchâssé dans une logique du sens et des expressions linguistiques régulières. En ce sens, chercher à étudier un texte d'abord par ce qu'il porte en lui comme différences peut permettre de rendre compte d'une dynamique à l'origine même de ce qui peut faire sens pour un interprète, celui-ci n'existant pas en soi, mais s'appuyant cependant sur des traces textuelles pouvant être observées statistiquement.

Cet essai comprend six courtes variations sur cette recherche du sens, qui abordent les discours d'abord comme la trace d'activités à la fois répétitives et conflictuelles, répétitives dans leur recherche d'unité et conflictuelle par l'impossibilité de trouver le prédicat capable de recouvrir le contenu initialement perçu. C'est cependant à travers ce conflit et cette répétition que se stabilisent peu à peu des formes régulatrices des sens et des activités. On retrouve par là que l'auteur d'un discours, dans sa consistance, est ce que le discours produit, ce que chacun peut revivre par la lecture ou l'écriture, d'où la citation en exergue.

## 1. La tresse du sens[2]

[1] CNRS – Laboratoire PRINTEMPS-UMR8085 - Université de Versailles - St-Quentin-en-Yvelines
Courriel : max.reinert@printemps.uvsq.fr

Ce dont on parle à travers les discours apparaît à une conscience avec tant d'évidences qu'on ne doute pas de sa transmissibilité par les mots. Cette transparence du langage a fait croire qu'un énoncé se définissait avant tout comme l'expression même des choses. Certes les linguistes et logiciens se sont occupés à montrer que cette belle transparence devait être opacifiée pour préciser dans l'énoncé même la position de celui qui parle, et sous quelles conditions, ce dont il parlait pouvait être encore entendu comme vrai ou faux. Cette approche de l'opacification repose cependant encore sur l'idée qu'un locuteur a ce pouvoir omniscient, de contrôler son dire jusqu'à y définir lui-même la place qu'il y occupe et à partir de laquelle il introduit des propositions sur son objet. Le discours scientifique cherche justement à établir des protocoles précis pour fixer les points de vue. Mais, dans cette opération, on distingue deux langages : celui où l'on parle des objets et celui où l'on parle des points de vue. L'un joue le rôle d'un métalangage par rapport à l'autre.

On sent dans cette procédure une régression infinie car elle pose le problème du point de vue sous lequel un point de vue pourra être défini, etc. En fait, il n'en est rien. Il y a seulement un langage : le langage de notre naissance, qu'on appellera "naturel" et à l'intérieur de celui-ci des "langages spécialisés" qui permettent de parler des "choses" à partir de points de vues sans cesse renégociés à l'aide du langage "naturel". La caractéristique la plus importante de ce langage bien vue par Lacan est justement de ne pas avoir de métalangage et ceci pour une raison simple : nous sommes immergés dans le langage depuis la naissance et notre conscience des choses a été formée avec son acquisition. Cette formation de la conscience n'est pas un processus dépassé. A chaque nouvelle énonciation, quelque chose de cette formation se poursuit. D'où des doutes sur cette capacité de l'homme à maîtriser a priori son énonciation. Cela a été un enjeu de la psychanalyse de montrer la place de l'inconscient dans la formation des énoncés. Mais que signifie l'analyse d'un énoncé si l'on n'est assuré ni de son auteur ni de son objet ? C'est pourtant bien de cela qu'il s'agit.

*En tant que conscience de quelque chose, on est, soi-même, un produit du langage et le "je" d'un discours ne peut être préalable à celui-ci.* Une fois l'énoncé terminé, le comprendre consiste le plus souvent à en énoncer un autre et à cumuler ainsi les énoncés dans une suite sans fin répétant toujours diversement un même et insaisissable objet. Comment pourrait-on étudier cet objet du discours autrement qu'en entrant dans ce cycle sans fin des énonciations ?

Dans ces conditions, que peuvent apporter des méthodes d'analyse de discours ? Ce qui se joue dans chaque discours effectif ne dépasse-t-il pas de loin le simple problème d'une transmission d'informations ou d'une simple représentation de quelque chose ? Comme le dit

[2] Ce chapitre reprend avec quelques modifications substantielles une communication présentée aux cinquièmes Journées d'Analyse des données textuelles de Lausanne (9-11 mars 2000) sous le titre *" La tresse du sens et la méthode Alceste "*

Robert-Dany Dufour (1990), à chaque énonciation, c'est un rituel ancien qui se rejoue : *« Lorsqu'un sujet parle, il dit "je" à un "tu" à propos d'un "il" »*. Cette structure trinitaire de base de toute énonciation ne donne-t-elle pas également un statut singulier à l'objet puisque celui-ci ne se définit ni en soi ni dans son rapport à un sujet précis, mais comme pure circulation entre ces différents moments d'une conscience : "je-tu-il". De ce point de vue, un énoncé ne prend pas sa valeur dans un tableau de vérité, mais dans sa capacité à être de nouveau proférer, repris, traduit, trahi, à travers cette chaîne sans fin des énonciations (Dufour).

Si ce que l'on cherche ne peut être qu'énoncé en reprenant à son compte ce qui a toujours été dit, si l'objet d'un discours est insaisissable, voire indéfinissable, voire purement imaginaire, que peut apporter une méthode d'analyse fut-elle informatisée ? Certainement pas la saisie de l'objet, on en conviendra…

Une méthode statistique d'analyse ne peut se substituer au langage pour dire les choses. Elle peut seulement remettre à plat le cycle des répétitions, éventuellement le cycle des ratages à énoncer un impossible à dire (le Réel). Mais en dévoilant la répétition, elle peut aider à prendre conscience des formes de ce qui, dans le discours, insiste. En prendre conscience, c'est déjà le nommer, et c'est faire de ce moment d'échec un lieu de circulation avec ses régularités et ses écarts, c'est le reconnaître comme paysage, et peut-être le rendre habitable. Ainsi on accède, à partir de la conscience d'un manque, à une certaine reconnaissance d'un monde qu'il soutient. C'est en tout cas à travers ce parcours que le sujet parlant se construit également comme conscience de quelque chose même s'il ne cesse de trébucher à chaque pas.

Notre hypothèse (sous l'influence de Peirce) est justement que ces écarts, ces trébuchements, ces discontinuités dans le cheminement du sens se déploient et oscillent entre des positions très archaïques d'un système ternaire. Car la ternarité n'apparaît pas uniquement dans les pronoms personnels qui n'en sont qu'un symptôme, mais à travers notre manière même de nous poser à chaque instant dans le sens. Le mot "sens" garde d'ailleurs en lui la diversité de ces positions à partir desquelles il est constitué, tressé. Comme l'écrit Frédéric François (1998) : *« Il y a une forte affinité entre le sens comme ressentir, le sens comme s'orienter et le sens comme signification »* Ces trois acceptions - sens comme sensation ou intuition ; sens comme direction ou mouvement ; sens comme signification ou démonstration - sont trois aspects du sens qui font tresse dans chaque énoncé.

Réduits à l'analyse d'un énoncé ces trois brins de la tresse peuvent être rapidement glosés[3] Le premier brin dépend de la conscience sensitive, intuitive. Un énoncé est d'abord perçu comme

[3] en affinité avec la phénoménologie peircienne (que Peirce nomme phanéroscopie : voir la traduction des écrits de Peirce par Deledalle, 1978).

« contenu ». Il est plein. Il a la consistance d'un monde, même si celui-ci échappe à la représentation. On retiendra le rôle particulier des mots "pleins" dans cette perception de l'énoncé. On peut d'ailleurs se demander si cette plénitude ne déborde pas les mots, si elle n'est pas l'attribut de l'acte de parole ou de lecture même. Ce brin du sens perçu immédiatement, comme débordement, nous l'appellerons la fibre de l'Imaginaire par laquelle un regard s'émerveille ou glisse simplement sur un monde qu'il entrevoit.

Le second brin dépend d'un mal être, d'un trouble, d'une résistance : l'énoncé s'opacifie, résiste au contenu. On avait cru le saisir et le vertige du vide soustrait brutalement la conscience au charme, *car le contenu promet plus qu'il ne donne*, et, dans son reflux, l'écart avec ce que l'on croit avoir perdu oblige à percevoir une altérité. Alors que la plénitude d'un « je » renaissant se confondait avec la plénitude du monde (mythe du paradis perdu), la conscience s'éveille à présent étrangère à elle-même. « Je » devient autre. Ce qui semblait perdu devient demande, adresse à l'autre. L'énoncé s'offre comme question, comme ouverture à l'altérité.

Si le premier brin de la tresse ouvre le sens à la plénitude des sensations, le second brin de la tresse l'ouvre donc à l'autre par la demande ou à la frustration. Par ce moment de déchirement s'introduit la conscience d'une altérité.

Le troisième brin prolonge ce déchirement en le constituant « lien », « symbole ». Le déchirement devient écart de soi avec soi. N'est-ce pas dans la nature de la conscience réflexive que de se voir elle-même ? Mais en se voyant, elle se perdrait dans sa propre vacuité. Cet écart n'est possible que par l'acceptation de l'altérité. Pour pouvoir dire « je », il faut accepter ce signifiant qui n'est pas soi, et par lequel on se fait représenter. Cet accueil de l'altérité pour prendre place dans la représentation est donc un partage qui n'est pas « compromis » mais ouverture et assujettissement. Le « il » exprime cette position où chacun semble prendre sa juste place. Le partage ne diminue pas la part, puisque cette part n'existe que par le fait des autres parts, qui lui donnent sa valeur. Le « Il » est l'expression de cette position de médiation. Se représenter, c'est accepter l'autre, c'est donc accepter sa propre altérité, comme constitutive de son être. Mais il ne s'agit là que d'un moment, et non pas d'une finalité… Car ce nouvel être, si plein des espérances dont on s'est ébloui, peut, le moment suivant, apparaître leurre dont il s'agit de s'extraire à nouveau…

Aucun des brins de la tresse ne tient seul et ce mouvement sans fin se nourrit de la répétition, répétition depuis toujours dans les récits et les mythes, répétition renouvelée, revivifiée par la diversité de ses versions, par ses interprétations.

L'objectif d'une analyse statistique des textes ne peut être l'étude d'un objet particulier qui se trouverait enfoui en eux, mais d'étudier comment un sujet[4] se constitue à travers son propre tressage, à travers ses ancrages, ses écarts, ses insistances, ses redites, ses échappements, le

[4] avec son ambiguïté comme on le verra, entre sujet-thème et sujet-acteur, qui recouvre un même indicible à l'origine de l'énonciation.

mot "sens" n'exprimant, selon ce point de vue, que ce par où un sujet est passé au fil de l'énonciation.
Nous chercherons à montrer dans les paragraphes suivants comment un tableau de nombres peut être interprété comme modélisant, d'une infime manière, cette tresse du sens, le texte analysé étant supposé pris par l'analyste, comme support du contenu de sa recherche.

## 2. De l'analyse des données à l'analyse de discours

L'analyse des données dite « à la française » commence à la fin des années soixante avec la création de l'*Analyse Factorielle des Correspondances* par Jean-Paul Benzécri comme outil d'analyse des *« données linguistiques »* (1973,1981,1982). Cette approche a été rapidement étendue aux traitements des données d'enquêtes par questionnaires[5].

Dans la préface du livre de Lebart et Salem *« statistique textuelle »* (1994), Christian Baudelot remarque *« Avec ses graphes d'analyse factorielle, J.P. Benzécri a rendu les individus à la statistique : longtemps ignorés à force d'être confondus dans de vastes agrégats ou pulvérisés dans des formules inférencielles qui s'intéressent d'abord aux relations entre grandeurs abstraites (revenu et consommation, salaire et diplôme…), les individus effectuent leur rentrée sur la scène statistique… ».* Baudelot pensait à l'analyse factorielle appliquée à des données d'enquêtes par questions ouvertes, données se présentant alors sous forme de tableau à double entrée, croisant les différentes réponses avec le vocabulaire.
L'intérêt de Benzécri pour le langage est indéniable mais son questionnement est plus philosophique[6] que sémiotique : Comment remonter par induction et synthèse des « faits élémentaires » aux lois. Il dit notamment dans sa conférence d'Honolulu : *« la notion de forme ou de modèle devrait émerger d'une mer de données, non par des postulats nominalistes ou des axiomes a priori, ni par des mesures trop fragmentaires de faits isolés, en eux-mêmes dénués de sens puisqu'ils dépendent du milieu ambiant et se réorganisent sans cesse, mais par la synthèse simultanée(…) d'un bon nombre de faits élémentaires qui nous aide à gravir les échelons de la hiérarchie des causes. »* (Benzécri, 1973).
Cette orientation a été mal perçue, car elle donnait le primat à l'observation pure, à un

[5] D'abord par questionnaires fermés, puis ouverts (Lebart, 1982)
[6] avec notamment l'influence des écrits d'Aristote.

moment où il était plutôt de bon ton d'être « constructiviste ». Pourtant, appliquée aux individus, et à leur production langagière, cette orientation méthodologique a déplacé la problématique de la description vers une problématique de l'interprétation. L'approche des « individus » soulignée par Baudelot est un premier pas vers une approche que nous préconisons nous-même, et que nous pourrions appeler une approche interprétative des « points de vue ». Comment représenter la complexité des « points de vue » mis en oeuvre à travers un discours sans les annihiler individuellement ? Mais n'allons pas trop vite, car la notion de « point de vue » est également trop réductrice et nous lui préférons celle de « mondes lexicaux ». Remarquons simplement qu'une analyse de données « individus par vocabulaire » ne peut prouver une interprétation sur un discours, puisqu'elle n'implique aucune hypothèse a priori ; elle peut seulement être une aide pour la formuler, éventuellement l'étayer ou la nuancer ou même la critiquer. L'analyse devient un outil rhétorique dans le cadre d'une élaboration exploratoire des contenus.

Aussi l'approche Benzécriste, malgré ses illusions descriptivistes, a été et reste novatrice. Et si l'analyse statistique des discours ne peut permettre d'avoir des réponses définitives sur les contenus, elle peut permettre de construire des espaces sémiotiques de manière rigoureuse à partir desquels les configurations de signes relevés sont susceptibles d'être interprétés et discutés. Le débat lui-même est déjà une victoire (de la vérité), car il nécessite un lieu partagé où les places sont distribuées selon une loi acceptée. En ce sens, l'analyse de discours peut avoir besoin des ressources de l'analyse des données textuelles pour élaborer des positions de négociation — sans référence à des échelles de valeur a priori ou des influences d'école — dans un espace construit selon des procédures explicites et surtout indépendantes des enjeux débattus. Il ne s'agit donc pas, à travers l'analyse statistique, de chercher des synthèses à partir des faits élémentaires, mais plutôt de donner un cadre sémiotique acceptable à une communauté, pour débattre des objets *dont elles ont l'expérience par ailleurs.*

*L'approche distributionnelle*

Pour introduire les travaux dédiés à la linguistique et la lexicologie et regroupés dans le tome 3 de *« pratique de l'analyse des données »*, Benzécri (1981) se réfère aux études de Zellig Sabbetai Harris sur l'analyse distributionnelle. Cette forme d'analyse avait d'abord été élaborée par Leonard Bloomfield, dans les années 1930, sous la double influence du

béhaviorisme (John B. Watson) et des premiers succès de la phonétique notamment avec Edward Sapir (1968).

Bloomfield pensait qu'il était possible de déduire les différents niveaux d'une langue (phonologie, morphologie, syntaxe) en analysant les distributions des séquences récurrentes dans un corpus suffisamment grand. En définitive, *étudier une langue, consistait donc avant tout à réunir un corpus, c'est-à-dire un ensemble aussi varié et aussi exhaustif que possible d'énoncés émis par les utilisateurs d'une langue donnée, à une époque donnée. Toutes les questions de sens d'un énoncé étant mises à l'écart, la tâche du linguiste consistait à faire apparaître des régularités dans la langue étudiée pour que la description ait un caractère systématique.* (Pottier, 1973).

Cette approche a ensuite été développée et systématisée dans les années 1950 par Harris. Pour cet auteur, *les énoncés linguistiques sont des suites d'éléments discrets, ordonnées linéairement. Un morphème peut être défini comme une séquence de phonèmes, un mot comme une séquence de morphèmes, une phrase une séquence de mots, et un discours une séquence de phrases*. (o.c.). On pensait ainsi pouvoir dériver les lois du langage des lois de distribution de ces différents « éléments ». Si ce premier objectif s'est révélé trop ambitieux, Z.S. Harris eut l'idée d'utiliser cette approche distributionnelle non plus pour étudier les lois du langage mais les lois du discours (Harris, 1954), la *distribution d'un élément* étant entendue comme *l'ensemble de ses environnements* dans un corpus particulier.

C'est cette définition de la distribution qui attira J.P. Benzécri, car son objectif était le même : *non pas chercher le sens d'un texte mais déterminer comment sont organisés les éléments qui le constituent*. Notons cependant que les préoccupations d'Harris étaient plus d'ordre logique et transformationnel que statistique (et la notion de « fait linguistique » n'était pas la même).

Cela dit, l'approche distributionnelle pouvait être entendue par un statisticien dans la mesure où les notions d'élément et d'environnement étaient précisées opérationnellement... A ce propos, J.P. Benzécri remarque : *« le terme de contexte, ou celui d'environnement qu'emploie Z.S. Harris, offre la même ambiguïté, que celui d'unité ou d'élément.. »* (Benzécri, 1981)

Cela étant dit, Harris est également connu en France pour avoir rendu pensable une approche formelle d'analyse des discours (voir les travaux de Michel Pêcheux : Maldidier, 1990). Dans la version française de l'article de Harris (1969) sur l'analyse du discours, on peut lire : *« Cet article présente une méthode d'analyse de l'énoncé suivi (écrit ou oral) que nous appellerons discours. C'est une méthode formelle qui ne se fonde que sur l'occurrence des morphèmes*[7] *en*

[7] On distingue les morphèmes grammaticaux et lexicaux. Par exemple « chantera » est composé de deux morphèmes : « chant », morphème lexical, « era », morphème grammatical. Cette distinction n'est pas toujours aussi simple : par exemple « va », dans « il va »…

*tant qu'éléments isolables ; elle ne dépend pas de la connaissance que le linguiste qui analyse peut avoir du sens spécifique de chaque morphème, et elle ne nous apprend rien de nouveau sur le sens particulier de chacun des morphèmes qui figurent dans le discours en question. Mais ceci ne signifie nullement que nous ne puissions pas découvrir autre chose que la manière dont la grammaire s'illustre dans ce discours. Car bien que nous usions de procédures formelles, proches de celles de la linguistique descriptive, nous pouvons obtenir sur le texte étudié des renseignements que cette dernière ne fournissait pas ».* Harris ajoute plus loin : *« Nous avons soulevé deux questions : celle des rapports distributionnels entre phrases, et celle de la corrélation entre langue et situation sociale ».* C'est ce dernier aspect qui a conduit toute une école à s'intéresser au rapport entre la forme distributionnelle des discours et leurs conditions de production.

Pour mettre en évidence ces traces, il ne s'agit pas d'étudier la syntaxe ou la sémantique, mais de mettre en évidence des faits distributionnels : *« L'analyse distributionnelle à l'intérieur d'un seul discours, considéré individuellement, fournit des renseignements sur certaines corrélations entre la langue et d'autres formes de comportement. »* (o.c., p11). Autrement dit, une analyse formelle d'un discours semblait pouvoir être tenté, en se restreignant uniquement à son analyse interne : *« Il résulte de tout ce qui précède que notre méthode devra établir les occurrences d'éléments et en particulier les occurrences relatives de tous les éléments d'un discours dans les limites de ce seul discours. »*(o.c., p14).
L'approche Harrissienne a en définitive été abandonnée par la difficulté de donner des critères précis pour définir les unités d'analyses (morphème, constituant, environnement, contexte).

## 3. Analyse de discours et sémiotique peircienne[8]

*Dans ce paragraphe, nous présentons quelques aspects « théoriques » qui vont permettre de se familiariser avec l'orientation « pragmatique » suggérée dans le premier paragraphe : étudier ce que fait le discours, plutôt que ce qu'il est ; étudier le « sens » comme mouvement plutôt que comme signification.*

Retenons de l'école Harrissienne et de sa reprise par les analystes du discours qu'ils se sont fortement opposés à l'approche des contenus. L'intérêt pour le sens a été en quelque sorte inversé par eux. Au lieu de chercher à interpréter le sens en aval de l'énoncé (dans sa signification), il s'agit de remonter à sa source aux « conditions de production ». Cette

orientation conduit à un intérêt pour ce qui engendre l'énoncé, un intérêt pour la situation, la référence, du moins dans le sens où l'emploie Benveniste dans cette citation : *“ Si le sens de la phrase est l'idée qu'elle exprime, la référence de la phrase est l'état de choses qui la provoque ”*[9] Benveniste, qui a été très critique vis-à-vis de la théorie sémiotique de Peirce[10], se trouve ici très Peircien. La référence est d'abord dans ce qui s'impose comme sujet du discours : elle détermine du moins dans une certaine conscience le signe et son sens. Cela n'a plus grand-chose à voir avec l'idée d'un signe qui dénoterait arbitrairement son objet.

En tant que sémioticien, le point de départ de Peirce pour définir le signe n'est pas la linguistique mais la logique. Il cherche à définir le signe à partir de la notion d'inférence. Sa question : comment passe-t-on d'un signe à l'autre ? Une de ses définitions du sens est la suivante : *“ Le sens d'un signe est le signe dans lequel il doit être traduit ”*. [11]

Ce point de départ inférenciel de la sémiotique pourrait faire penser qu'elle ne constitue qu'une partie de la logique mais, à l'inverse, c'est à son dépassement que Peirce aboutit… Et il l'appelle la *sémiotique.* La déduction n'est qu'un aspect trivial de l'inférence d'où le temps est exclu. Une déduction ne crée pas de nouveau signe, tout est déjà dans le premier signe.

Par exemple déduire de « Socrate est un homme », « Socrate est mortel » sachant par ailleurs que « les hommes sont mortels » ne vient que d'une autre manière de lire le même schéma où l'ensemble des mortels contient l'ensemble des hommes dont un élément est Socrate.

Peirce distingue deux autres types d'inférences, *l'induction* et *l'abduction*. *L'induction* est liée à la capacité d'abstraction, de généralisation. Trouver la loi générale à partir de cas particuliers. Si l'induction est moins triviale que la déduction, le temps, comme moteur de la sémiose, ne sert qu'à renforcer — ou éventuellement, à l'anéantir — une conviction déjà constituée. Tant qu'elle se vérifie, l'induction n'implique ni histoire, ni parcours, ni sujet.

Notre expérience quotidienne montre au contraire la capacité créative du temps. Par quel type d'inférence en arrive-t-on à poser une hypothèse nouvelle, par exemple ? Pour Peirce, il s'agit bien d'une inférence et il l'appelle *« abduction »* Sa nature est à rechercher dans la notion d'habitude. Elle ressemble à *l'induction* par le fait qu'elle se constitue en fonction d'une série d'expériences antérieures, mais s'en sépare par le fait qu'elle ne présuppose pas une identité

---

[8] Cette partie est tirée principalement de notre article de 2001 « Approche statistique et problème du sens dans une enquête ouverte », paru dans le *Journal de la Société Française de Statistique*, 142, 59-71

[9] Problèmes de Linguistique générale, page 226 du tome 2 (Benveniste, 1966)

[10] *Op.cit.*, page 45 : *“ La difficulté qui empêche toute application particulière des concepts peirciens, hormis la tripartition bien connue (icône, indice, symbole), est qu'en définitive le signe est posé à la base de l'univers entier, et qu'il fonctionne à la fois comme principe et définition pour chaque élément et comme principe d'explication pour tout ensemble, abstrait et concret. L'homme entier est un signe, sa pensée est un signe, son émotion est un signe. Mais finalement ces signes étant tous signes les uns des autres, de quoi pourraient-ils être signe qui ne soit pas signe ? ”*

[11] Traduit et cité par Everaert-Desmedt (1990) des "Collected Papers " (4.132). Notre présentation de Peirce est très orientée par nos objectifs. Pour une introduction à la pensée peircienne on peut se référer avec profit au livre de G. Deledalle « Ecrits sur le signe ». Voir également son article dans l'encyclopeadia Universalis.

des cas dans la série qui mène à la loi. Dans ce qui advient, ici et maintenant, il y a seulement comme un air de famille, une ressemblance avec du déjà connu. C'est par cet intermédiaire que Peirce en vient à s'intéresser aux usages dans leurs aspects créatifs.
Par la notion d'habitude active, d'usage, le temps devient une composante à part entière, une composante créative de l'inférence, chaque parcours dépendant maintenant d'une histoire avec ses régularités, ses échecs, et sa singularité. Mais une habitude, une pratique, n'est pas une représentation, ce qui ne signifie pas qu'on ne cherche pas à la représenter… A défaut d'y arriver, on cherche des métaphores… jusqu'à ce qu'elle se laisse apprivoiser dans un nouveau signe. Entre ce que l'on cherche à voir et ce qu'on est déterminé à faire par l'habitude une sorte de dialogue s'instaure avec beaucoup de plasticité pour créer de la ressemblance, du même. Dans la définition de Peirce :*" Le sens d'un signe est le signe dans lequel il doit être traduit »*. Le *« il doit »* dépasse donc la simple déduction logique, pour exprimer une détermination vécue, voire un choix (éthique) de sens. Du moins c'est là notre interprétation de Peirce.

Dans cette perspective dynamique, ce qu'on appelle signe n'est pas à définir en soi mais en rapport avec ce mouvement du sens. Un signe est signe pour une conscience, dans le mouvement de celle-ci, au cours d'un processus de transformation des signes que Peirce appelle "sémiose". Plus précisément Peirce distingue trois manières de "faire signe" : l'icône, l'indice et le symbole que nous développerons comme les trois *moments* caractéristiques d'une sémiose[12] :

1) *Le moment iconique de la sémiose*, liée au temps de l'abduction L'objet dynamique en situation, l'histoire du sujet, l'usage, le désir, font qu'une apparence s'impose pour quelqu'un comme sujet de son intérêt. L'apparence, en tant qu'elle recouvre un sujet, est une forme d'abduction en ce sens qu'elle se donne comme conscience de quelque chose. Il y a un aspect sensible, et immédiat de l'apparence[13]. Mais cet aspect premier pour la conscience s'inscrit dans un mouvement antérieur vécu dynamiquement comme la résultante d'une série des expériences passées, que l'accrochage présent actualise cependant en la réordonnant en fonction de la singularité du sujet actuel.
*Au plan du discours,* ce moment de prise d'une conscience d'un contenu immédiat, ce moment iconique de la saisie d'un sens se retrouve dans toute lecture. S'agit-il du contenu

[12] Notre propos est d'exposer nos hypothèses le plus simplement et nous ne prétendons pas exposer la théorie peircienne. Cependant, il nous semble évident qu'elles en découlent très directement même si nous ne cherchons pas à le montrer ici.
[13] notion de "primeité" chez Peirce

d'un énoncé ou du contenu des mots pleins ? Il s'agit du contenu d'un même acte, à la fois singulier dans son actualité et répétition comme forme de vie. Un mot n'est plein pour quelqu'un que d'être constitutif de l'histoire de sa propre conscience. Par cette rencontre des mots pleins dans un même énoncé, quelque chose de la singularité de l'acte par ce qui est visé est restituée comme marque (la cooccurrence multiple) à partir d'une expérience commune des mots… Le fait que ce qui est visé est aussi celui qui vise, implique un amalgame entre les deux significations du mot « sujet » : sujet-thème et sujet-acteur, car l'acteur est assujetti au thème qu'il poursuit, et le thème n'existe qu'à l'horizon toujours repoussé du parcours des acteurs qui cherchent à le viser. Ce fait marque la notion même de « sujet » d'une ambivalence fondamentale[14].

2) *Le moment indexical de la sémiose* : Dans le premier moment de la sémiose, l'objet[15] apparaît à la conscience sous la forme d'une icône (temps de l'abduction). S'agit-il d'ailleurs d'une forme ? Il s'agit bien davantage d'un contenu, dont on prend conscience plus tard avec cependant des effets réels par les métaphores choisies dans un corps réel, au niveau des désirs, des affects ou des actes, ici et maintenant. Le second moment de la sémiose est relatif à la prise de conscience de cette pression de la référence (ou du sujet…) sur ce qui se produit comme conséquence. Cette pression oblige de la même manière qu'au plan logique, la déduction oblige[16]. Je crois voir un brouillard gris : moment iconique de la sémiose. Je me cogne dans un mur : moment indexical de la sémiose. Je prends conscience d'une propriété de ce brouillard gris avec le mot « mur » : moment symbolique et réflexif de la sémiose que l'usage social m'aide à représenter et fixer comme « réalité » à travers ses symboles. C'est par ce cheminement qu'un contenu initial relativement libre, au fil des expériences, prend peu à peu forme…

Au plan du discours, cette rupture dans la possibilité de dire peut être traduite par un silence, une ponctuation. Elle peut être traduite, linguistiquement, par un certain nombre de mots qui n'ont de sens qu'en situation, appelés déictiques, shifters ou embrayeurs : *je, tu, ici, là, là-bas, maintenant, hier, demain, aujourd'hui, voilà, dedans,* etc...
Au moment iconique de la sémiose, le signe est perçu uniquement en lui-même (notion de priméité chez Peirce). Au moment indexical, le signe semble se dédoubler, il se révèle par ses

[14] Ce que Bachelard exprime par cette mise en garde dans son introduction de la psychanalyse du feu **: *"Il suffit que nous parlions d'un objet pour nous croire objectifs. Mais par notre premier choix, l'objet nous désigne plus que nous ne le désignons."*.**
[15] ou le sujet en tant que ce qui est visé…
[16] Si Peirce pense que le réel est totalement représentable, il est normal qu'il associe alors cette obligation « réelle » à la déduction qui en est une obligation logique.

failles dans son rapport au référent ou au « sujet », en tant qu'il affecte directement le cheminement de la sémiose. Peirce relie ce moment à la notion d'effort, de résistance, de choc. Dire que cela affecte directement, c'est prendre conscience d'un écart entre un intérieur affecté et un extérieur apparent. Ainsi le sujet réel ne se révèle à la conscience que sous la forme d'une coupure séparant cet extérieur et cet intérieur, perçu également comme coupure de soi avec soi, de soi avec l'autre, de soi avec son objet.

3) Si la sémiose s'ouvre sur des apparences multiples, fluctuantes, se prolonge comme coupure, comme rupture ; elle se réfléchit également en elle-même dans un concept de ce qui se produit. Dire qu'il y a « mouvement » c'est déjà représenter la coupure comme passage entre deux moments. Ce ne sont pas des choses qui se représentent mais les actes qui nous ont amenés à les voir ou à les construire. Une recette de cuisine représente la préparation d'un plat. Nous utilisons le mot « concept » en ce sens. Un concept de plat, de logement. Ce sens n'est pas opposé au sens du concept comme prédicat, comme propriété, au contraire, il le restitue dans son dynamisme, dans son pragmatisme. Ce point est important car il permet de passer de la notion d'acte à la notion de structure[17]. En se représentant ses actes, on passe des actes particuliers aux concepts généraux, aux idées, aux lois (notion de tiercéité chez Peirce). Prendre conscience d'un sens, de ce point de vue, c'est aussi, par là même, sortir de l'expérience immédiate en se situant soi-même comme assujetti à des lois, à un cycle des répétitions, dans un ordre social ou mythique. On appellera ce moment, *le moment symbolique de la sémiose*[18].

*Au plan du discours*, l'objet dynamique (son *sujet*) est lui aussi, à la fois à l'origine du discours, qu'il provoque, et à l'horizon de tout nouvel énoncé, puisque, en situation, il nous apparaît comme ce qu'on cherche justement à dire comme sens. Mais le *sens* n'est pas dans le texte, le *sens* était dans le temps de cette circulation dynamique de la parole.

[17] Comment ne pas évoquer Piaget…

[18] Ce moment est associé à l'induction comme inférence d'une loi générale, à partir de cas particuliers. Un symbole, un mot, n'existe qu'en tant que loi induite. La différence avec l'abduction tient au fait que l'induction présuppose la loi de représentation, contrairement à l'abduction, qui la crée à travers une nouvelle manière de *percevoir* les choses.

***Fiche sur Peirce*** *(extrait résumé de l'article de G. Deledalle dans l'Encyclopaedia Universalis)* **:**

Charles Sanders Peirce, fils du mathématicien Benjamin Peirce, naquit à Cambridge (Massachusetts), le 10 septembre 1839. Il fut éduqué par son père qui l'initia très tôt à la chimie, aux mathématiques, à la logique et à la philosophie.

Il entra à Harvard en 1855 où il suivit des cours de mathématiques, philosophie et aussi de physique et de chimie. Après son diplôme en 1860, il travaille au Service géodésique des États-Unis (United States Coast Survey). Il y resta jusqu'en 1891. Il se retire ensuite à Milford (Pennsylvanie) où il y vécut péniblement de sa plume jusqu'à sa mort (1914).

On distingue trois orientations dans les travaux de Peirce : l'une relative aux sciences expérimentales ; l'autre relative à la logique mathématique et la dernière relative à la philosophie. Cette dernière se subdivise également en trois : on distinguera le phénoménologue, le sémioticien et le métaphysicien

# 4. L'exploration automatique des discours par la méthode *ALCESTE*

Nous proposons un modèle dans lequel est calculé un aspect de la signification non pris en compte par le sémanticien. Le discours y est conçu non pas par ce qu'on s'en représente mais par ce qui s'y inscrit comme activité temporelle.

On s'intéresse donc au discours dans un de ses aspects les plus archaïques avec l'hypothèse suivante : La signification d'un discours ne se forme pas essentiellement dans ce qui s'élabore à partir d'une représentation, puisque celle-ci ne peut être préalable à celui-ci ; Elle s'élabore à partir d'une activité rythmique plus primitive, conflictuelle ou dialogique, qui laisse une place centrale à la *répétition*. C'est par la *répétition* qu'une stabilisation de l'activité discursive peut se concevoir. Mais cette *répétition* ne se dévoile dans sa complexité que dans un jeu ternaire.

Reprenons rapidement nos trois modes de *répétition* pour l'analyse statistique[19]. On considère un discours dans sa temporalité comme une succession de moments discursifs sans donner pour l'instant à cette notion un statut bien précis. Formellement, on découpe le texte en segments disjoints successifs.

Chaque segment est supposé recouvrir une faible durée. Chaque segment recouvre *possiblement* le moment d'un *contenu* (immédiat), *donc la cooccurrence multiple des mots pleins dans un segment est une trace formelle possible*[20] *de ce contenu...*

La rupture entre deux segments de texte modélise, quant à elle, la faillite de toute énonciation à installer durablement une représentation, le temps se renouvelant sans cesse dans un nouveau présent.

Certes le texte n'est plus dans le présent du discours, mais cela montre surtout qu'il n'y a aucune manière de savoir ce que sera le présent du discours dans une lecture particulière, ce

[19] Présenté dans notre article de 2003 " Le rôle de la répétition dans la représentation du sens et son approche statistique dans la méthode Alceste ". Ces trois types reprennent les catégories peirciennes du le paragraphe précédent.

[20] Notre hypothèse est la suivante : le contenu serait premier, antérieur à l'élaboration linguistique, et l'association verbale est un effet du contenu (et non sa cause). C'est par elle qu'un tout premier niveau d'organisation pré-linguistique se met en place. Aussi, les mots pleins d'un même « segment de texte », c'est-à-dire, en association dans un même « moment », sont susceptibles, selon cette hypothèse, d'être dépendant d'un « même contenu » associatif.

qui, pour un lecteur, fera contenu, et donc rupture entre un avant et un après. *La rupture arbitraire, nous a semblé une solution raisonnable pour affirmer cette impossibilité de saisir, pour un algorithmique, ce qui est de la prérogative du sujet.* Aussi elle ne peut *modéliser une altérité réelle qu'en acte ;* Elle ne modélise ici que la simple *possibilité* d'une telle altérité, coextensive du temps du discours. On ne peut rien en dire a priori, sinon qu'à chaque instant, un évènement peut perturber le cours du discours.

Au plan de la modélisation, on remarquera que le découpage du texte en segments permet de passer de la notion de *possibilité* à la notion de *probabilité.* En effet, ce n'est qu'une fois que le découpage a été effectué, qu'un calcul des cooccurrences est rendu possible. J'entends ici par cooccurrence, non pas simplement la cooccurrence entre deux mots, *mais la cooccurrence généralisée entre tous les mots appartenant à un même segment de texte* Ce segment sert justement d'*Unité de Contexte*. La simultanéité des mots présents dans un segment est une marque probable du contenu, en tant que trace possible d'un acte subjectif (propre au lecteur, à l'auteur ou au locuteur).

Formellement, ce découpage du texte, par l'alternance des moments pleins et des possibles instants de rupture, modélise également le rythme d'une lecture quelconque, dans sa possibilité. C'est un aspect, sur lequel je n'ai pas insisté, mais qui exprime bien le fait qu'un moment de lecture, avant d'être un énoncé, est déjà intégré à une pré-forme qui est le rythme. La lecture, aussi bien que l'écriture, présuppose un rythme, et notre hypothèse est qu'un texte ne peut s'interpréter sans une scansion particulière, propre au lecteur ou au locuteur, en tant qu'il est *sujet.*

*Le principe de la méthode* est simple : le corpus à analyser est découpé en une suite de segments de texte et l'on observe la distribution des *mots pleins* dans ces segments… d'où le nom de la méthode : *" Analyse des Lexèmes Cooccurrents dans un Ensemble de Segments de Texte "*.

Comme on l'a déjà évoqué, ces deux types d'*unités – mots pleins* (plutôt que lexèmes) et *unités de contexte* (plutôt que énoncés) - sont problématiques. Un *mot plein* n'est pas un *lexème.* De même un segment de texte n'est pas un énoncé. Nous avons vu que les notions de mots pleins et de coupures sont relatives à l'expérience d'un *sujet*, en fonction de sa propre perception des *contenus*. Aussi *il y a un aspect fondamentalement contradictoire dans cette problématique de l'unité textuelle* aussitôt que l'on veut la saisir en dehors d'une expérience subjective. Cela n'implique cependant pas une impossibilité de calculer. Par exemple, la possibilité de représenter un objet concret n'implique pas d'avoir des *unités* précises pour cela.

Prenons l'exemple d'une plaque photographique ou d'une photo numérique. L'image se forme à partir d'un grain. Le grain de la photo est défini de manière encore plus arbitraire que les *coupures* que nous proposons entre énoncés. *Et cela importe peu dans la mesure où ce grain permet à un sujet de retrouver une stabilité dans les formes de son expérience*. La définition du grain est arbitraire et pourtant la possibilité de *représenter quelque chose* dépend de lui. Dans notre modèle, les grains sont les *segments de texte* ou *unités de contexte* et la coloration des grains dépend des *mondes lexicaux.*

Une fois les *unités* approximativement retenues, on cherche à étudier la structure des cooccurrences des *mots pleins* dans les différentes *unités de contextes (Cf. Tab.1)*

| Tableau de données | mot 1 | mot 2 | mot 3 | mot 4 | mot 5 | mot 6 |
|---|---|---|---|---|---|---|
| *u.c. 1* | *0* | *1* | *1* | *1* | *0* | *1* |
| *u.c. 2* | *1* | *1* | *0* | *0* | *1* | *1* |
| *u.c. 3* | *1* | *0* | *1* | *0* | *1* | *0* |

*Tab. 1. Le tableau modélisant l'activité du locuteur*

Ce tableau est une modélisation possible des trois aspects de la *Répétition* associés à l'*objet* d'un discours pour un lecteur : (1) par ce qui est en lignes ; (2) par ce qui est en colonnes ; et (3) par ce qui se construit comme ordre à travers le tableau numérique proprement dit :

*(1) Ce qui lie les colonnes entre elles est la première expression de la plénitude du contenu.* Elle est modélisée par la cooccurrence multiple. En effet, la présence simultanée de *mots pleins* dans une même *unité de contexte* est la trace *possible* d'un même acte de *contenu*. La cooccurrence multiple est donc une marque formelle possible de ce contenu. Mais de ce contenu, on ne saisit, par ce modèle, non pas le thème, mais sa deixité : la cooccurrence est une marque possible d'une position à l'origine de l'énonciation.

*(2) Ce qui sépare les lignes entre elles est une expression possible du dialogisme, d'un changement de fondement.* Ainsi, par la succession des lignes du tableau, celui-ci modélise la succession temporelle. Cette succession est donc une trace d'un changement de position. La *coupure*(39) entre deux segments exprime la possibilité d'une rupture des contenus entre deux moments successifs.

*(3) Les coupures* entre les segments de textes, en tant qu'elles peuvent permettre de discriminer des *mondes lexicaux,* permettent d'inscrire un aspect du processus énonciatif dans un tableau numérique. Elle donne un sens calculable à la répétition. Il se peut en effet que le système des *mondes lexicaux dégagé par le calcul tombe juste pour l'analyste,* autrement dit, *que ce système exprime ce que l'analyste perçoit comme contenu du fait de sa propre*

*expérience de lecteur et de sa propre expérience du monde à partir de laquelle cette lecture est possible*. En ce cas, il y a une forme de *résonance* entre ce qui est calculé et un usage vécu, entre une forme symbolique et une expérience particulière. C'est par cette résonance, que le calcul symbolique devient *signifiant, support d'un nouveau contenu* pour l'analyste. La forme calculée en devenant signifiante, devient transparente pour lui, car il y voit son monde.

*La Classification Descendante Hiérarchique (C.D.H.)*

Le mode de calcul appliqué au tableau de données pour faire apparaître le système des répétitions a été conçu dans le cadre de ma thèse avec J. P. Benzécri (1979). L'algorithme proposé a été ensuite amélioré en 1986 (Reinert 1983, 1986) pour le traitement de grands tableaux très vides (jusqu'à 99 % de zéros). Cette méthode est dérivée de l'analyse factorielle des correspondances. Elle permet depuis 1999 (version 5) d'analyser des tableaux ayant au maximum 40 000 lignes et 3 000 colonnes avec un nombre de " uns ", toutefois, qui ne dépasse pas 1 500 000[21]. Ceci permet de traiter des corpus allant jusqu'à 60 millions de caractères.

*Principe de la méthode.* Comme on l'a déjà évoqué, l'objectif est d'approcher les *mondes lexicaux* d'un corpus sans avoir à poser préalablement le problème de la définition des catégories de contenu. C'est la dynamique même du discours qui conduit à distinguer des ancrages topiques différents. Cet aspect différentiel de l'analyse statistique est fondamental au niveau sémiotique, car *c'est par lui que s'introduit l'idée seconde d'une proximité pour les éléments d'une même classe. Autrement dit, la représentation d'un contenu ne peut être que troisième (en aucun cas elle est première).* Elle nécessite en effet d'avoir d'abord expérimenter la notion de *Différence.* Ainsi les fluctuations répétées de ce qui diffère permettent l'accès au niveau symbolique, ce niveau n'exprimant rien d'autre qu'une *logique des ruptures* dont les classes donnent justement une première représentation[22].
À ce propos, il y a également *oscillation* entre ce qui se stabilise au plan statistique, et ce qui s'interprète au plan de la prise en charge des résultats par un analyste. Il peut en effet exister une fluctuation statistique que seul l'interprète peut lever, par son choix de lecture. Cela ne signifie pas que l'interprète peut lire n'importe quoi. Mais les frontières entre classes statistiques ne peuvent prendre consistance, qu'à partir du moment où ces classes peuvent être nommées. *Interpréter, c'est trouver un ajustement qui ne prouve rien en lui-même sinon qu'il*

[21] En 1999 les modifications de l'algorithme ont été minimes (mais délicates du point de vue informatique). C'est surtout l'amélioration des moyens de calcul qui a permis de traiter des tableaux de très grandes dimensions.

*permet d'apprécier concrètement ce qui tombe juste* entre ce qui est interprété et ce que montre l'analyse. *Cette appréciation, comme on l'a dit, n'est pas de l'ordre d'une preuve mais de l'ordre d'une abduction, d'une hypothèse.*

*Principe du calcul.* Le principe de la classification utilisée peut être présenté très simplement. L'objectif est de classer les segments de texte en fonction des mots pleins qui y sont présents. Ce classement est indépendant de l'ordre entre segments ou de l'ordre entre mots. Autrement dit, si l'on change l'ordre des mots ou l'ordre des segment, le tableau de données reste le même du point de vue de sa structure pertinente.

Au moins mentalement, on peut alors envisager toutes les manières possibles d'ordonner les lignes et les colonnes. Supposons que l'on en trouve une s'approchant du tableau [Tab. 2]. On obtient alors la discrimination de deux classes de segments très contrastés quant à la distribution des mots. Il existe même dans ce cas deux classes de mots complètement différenciées. Obtenir cela à partir de l'analyse d'un discours conduit à mettre en évidence, dans ce discours, deux régimes différents de production des énoncés (si l'on assimile, pour l'instant l'unité de contexte à un petit énoncé). Il est clair que, même si la méthode de recherche des classes est purement formelle, pour l'analyste ayant choisi le corpus, cela va faire sens...

| | *Vocabulaire 1* | *Vocabulaire 2* |
|---|---|---|
| *Classe 1 de segments* | *Tableau 1* | |
| *Classe 2 de segments* | | *Tableau 2* |

*Tab. 2. La distribution recherchée " idéale " pour l'obtention de la première dichotomie*

*Principe de l'algorithme.* Appelons maintenant « Tableau 1 », le tableau croisant la classe 1 de segments avec tout le vocabulaire (et idem pour le tableau 2). On peut alors comparer le profil des marges[23] en lignes de ces deux tableaux pour mesurer leur contraste. Le critère choisi dans cette méthode est le $\chi 2$. L'objectif peut être maintenant exprimé clairement : quelle est la partition en deux classes de segments qui maximise le $\chi 2$ des marges (en lignes) des deux sous-tableaux associés ? La méthode algorithmique utilise le premier facteur de l'analyse factorielle des correspondances pour obtenir une solution (Benzécri 1973 ; Reinert 1983, 1986). C'est en effet l'extraction du premier facteur qui permet de réduire considérablement le nombre de partitions à consulter pour trouver la partition optimale en 2 classes, à chaque pas. Une fois cette partition obtenue, on peut déplacer à la gauche du tableau les mots relativement plus présents dans la classe 1 et à la droite du tableau les mots les plus présents dans la classe 2. D'où la représentation proposée dans [Tab. 2]

[22] La différence structurelle est l'expression d'un différé en oeuvre dans chaque instant d'un discours. Cette idée « alcestienne » est si proche de ce qu'écrit Derrida dans « l'écriture et la différence » (1967), qu'il m'est impossible de ne pas le citer.

[23] Un tableau du type présenté ici a deux marges, dans les quels sont inscrits les totaux en lignes (par segment) et en colonnes (par mot). Ici nous ne considérons que les totaux en colonnes pour chaque sous-tableaux, totaux qui correspondent pour un mot donné, au nombre de fois où il apparaît dans la classe 1 ou la classe 2.

Après ce premier calcul, il est possible de recommencer le même algorithme sur le plus grand des sous-tableaux restants à traiter, après élimination des colonnes presque vides ou liées à des mots très spécifiques de l'autre classe, etc. D'où le nom de la méthode : *Classification Descendante Hiérarchique.*

# 5. Quelques questions à propos de la méthode « Alceste »

Si l'on peut comprendre qu'*il ne peut y avoir de contenu stabilisé sans différence*[24], cette notion de *« mondes lexicaux »* reste encore problématique par bien des aspects. Trois questions les concernant hantent cette recherche depuis son origine.

• *Dans quelle mesure, les mondes lexicaux obtenus sont dépendants du découpage en unités de contexte ?* La réponse à cette question peut sembler en partie résolue, puisque, comme on l'a suggéré, dans la procédure standard d'analyse, on procède à plusieurs analyses successives (au moins deux) avec des unités de longueur légèrement différentes, que l'on compare ensuite. Mais la réponse n'est pas si simple. Certains discours produisent des mondes lexicaux différents quand on fait varier la longueur de ces unités de contexte. D'autres y sont relativement insensibles. Le volume de texte joue également un rôle. Pour de petits textes, l'effet de la taille des unités peut être sensible[25]. Dans l'analyse d '» Aurélia » de G. de Nerval, on a pu montrer, par exemple, des ressemblances entre l'analyse de ce texte découpé en 1000 « unités de contexte », et du texte découpé en 19 unités, constituées par les différents grands paragraphes[26]. Cela n'empêche pas que dans des analyses particulières, il puisse y avoir des modifications dans les configurations des mondes lexicaux[27] du fait de sa petite taille[28].

• *Le second problème concerne l'objet même d'une telle analyse en mondes lexicaux.* En effet, la méthode *Alceste* est purement algorithmique. La seule compétence présupposée pour le calcul est celle qui permet de différencier les mots pleins des mots outils. Cette analyse est cependant souvent utilisée comme substitut à l'analyse de contenu. Comme nous l'avons dit, c'est un fait fréquent, que des utilisateurs interprètent les classes comme des catégories de contenu. Que faut-il en penser ?

---

[24] Le contenu suture la faille creusé par l'instant. S'il n'y a pas contenu, il y a angoisse ou affect. C'est l'affect, en tant qu'il mobilise une mémoire de l'acte, qui permet de transformer l'angoisse en contenu.

[25] Voir notre étude dans JADT 1993, 539-550, Actes des secondes journées internationales d'analyse des données textuelles, Montpellier, 21 et 22 octobre 1993, Editeur S. J. Anastex (TELCOM Paris 93 S 003)

[26] L'analyse est cependant moins claire sur ce découpage, du reste approximatif, la seconde partie d'Aurélia étant restée vraisemblablement inachevée.

[27] Du fait de l'inversion de certains facteurs (au sens de l'analyse factorielle des correspondances)

[28] Dû à des inversions dans l'ordre des facteurs, du fait que l'ordre des valeurs propres de l'analyse des correspondances est moins stabilisé sur de petits corpus.

• *Le troisième problème prolonge le second. La spécificité des contenus traités dans chaque interprétation n'empêche pas de trouver des mondes lexicaux semblables dans l'analyse de corpus qui peuvent être de nature très différente*, comme, par exemple, un corpus de réponses d'écoliers à une question concernant leur avenir professionnel ou sentimental, un corpus constitué par une oeuvre littéraire, comme « Aurélia » de G. de Nerval, ou « Les rêveries du promeneur solitaire » de J.J. Rousseau, ou encore le corpus de l' ensemble des articles des six numéros de la revue « le Surréalisme au Service de la Révolution », voire celui du dictionnaire Larousse, ou encore celui d'un corpus de récits de cauchemars… Il est surprenant de trouver des ressemblances entre des analyses si diverses, et en même temps, de trouver dans chaque analyse particulière une expression singulière du contenu des oeuvres analysées que l'expérience de lecteur de l'analyste reconnaît comme spécifique de l'oeuvre ![29]

Ce sont ces questions qui m'ont engagé dans la conception de la version 5 du logiciel. J'ai, en effet, tenté de constituer un étalonnage de ces mondes lexicaux stabilisés de façon à faciliter la comparaison entre les analyses. Au delà de cet objectif, cette procédure permet de distinguer au moins deux types de corpus, ceux dont l'analyse conduit à des mondes lexicaux stabilisés et les autres. Ce simple fait conduit à s'interroger d'une autre manière à la notion de conditions de production des discours. Il dévoile non pas des différences de classes, mais des différences de positions des locuteurs selon le type de problématique qui les anime. Plus précisément, on doit distinguer les discours dont l'objet peut être plus ou moins représenté socialement, des discours qui s'imposent à leurs auteurs par une nécessité impérative de dévoilement d'une question qui ne finit jamais de se poser. Selon notre hypothèse, c'est ce deuxième type de discours qui conduit l'analyse à des mondes lexicaux stabilisés (Reinert, 1999).

*Pour l'instant, et nous conclurons là-dessus, cet air de ressemblance entre mondes lexicaux obtenus à partir de l'analyse d'oeuvres différentes montre que le concept de l'analyse « Alceste » n'est pas basé sur ce que l'auteur de l'œuvre (et encore moins le lecteur) cherche à se représenter.*

[29] voir Reinert, 1990, 1993, 1997, 2000,

*Les Unités de contexte dans la méthode « Alceste »*
Il s'agit d'opérer automatiquement un découpage du corpus en segments de texte de longueur relativement arbitraire, mais de même « grandeur », afin d'apprécier ce que l'on peut entendre par une « cooccurrence multiple ». Pour cela le programme effectue successivement les constructions suivantes :
a) Calcul des *segments de texte calibrés (S.T.C.) :* il s'agit de suite de *formes graphiques (les mots)* terminée par une ponctuation ou un séparateur, à défaut, d'au plus une certaine longueur (définie en nombre de caractères). L'ensemble des segments de texte calibrés constitue une partition.
b) Calcul des *unités de contexte élémentaire* (U.C.E.) par concaténation des S.T.C. successifs. Leur « longueur » varie en fonction de la grandeur du corpus. Elle est calculée en nombre d'occurrences de formes dans l'unité. *L'U.C.E. est l'unité statistique de base* pour le calcul des *cooccurrences*.
c) Calcul des *Unités de Contexte* (U.C.) de longueur variable pour tester la stabilité des classes (dans la procédure d'analyse standard). Ces U.C. sont composées par un nombre entier d'*Unités de Contexte Elémentaires* de manière à ce que le nombre de *mots analysés différents* par unité soit inférieur à un seuil fixé. L'ensemble des *Unités de Contexte* retenues pour un calcul est supposé constituer une partition du corpus (aux aléas du découpage et du classement près).

*Le Problème de la stabilité des classes obtenues par la C.D.H..*
Dans la procédure standard d'analyse, une première appréciation de la stabilité des résultats est basée sur le calcul suivant. On effectue deux classifications sur des unités de contexte de grandeur légèrement différente. Chaque U.C. de l'une ou l'autre analyse étant composée par un nombre entier d'U.C.E., on peut considérer les classes de ces deux classifications comme des classes d'U.C.E. : il suffit en effet de remplacer chaque U.C. classée par les U.C.E. qui la composent.
La comparaison entre les deux classifications s'effectue ensuite de la manière suivante : On construit un tableau de cooccurrences entre les classes de la première classification et les classes de la seconde classification — qu'elles soient terminales ou non — en inscrivant dans chaque case, le nombre d'U.C.E. classées dans telle classe dans la première analyse et dans telle classe dans la seconde. On calcule ensuite un $\chi^2$ entre chaque couple de classes en correspondance afin d'apprécier la significativité du lien et l'on retient les couples les plus significatifs, constituant une partition comportant le plus de classes possibles.

# 6. Une application illustrative

Voici, à titre d'illustration, une présentation rapide de l'analyse d'« Aurélia » de G. de Nerval (à l'aide de la version 5). Nous partirons du texte d'Aurélia tel qu'il est accessible sur le Web[30].. Voici quelques phrases du début et de la fin de cette oeuvre pour illustration.

Le Rêve est une seconde vie. Je n'ai pu percer sans frémir ces portes d'ivoire ou de corne qui nous séparent du monde invisible. Les premiers instants du sommeil sont l'image de la mort; un engourdissement nébuleux saisit notre pensée, et nous ne pouvons déterminer l'instant précis où le moi, sous une autre forme continue l'oeuvre de l'existence. C'est un souterrain vague qui s'éclaire peu à peu et où se dégagent de l'ombre et de la nuit les pâles figures gravement immobiles qui habitent le séjour des limbes.

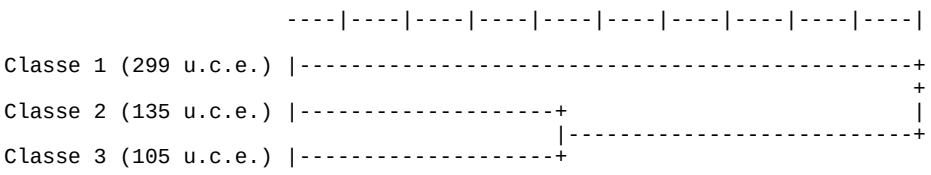
……

Telles sont les idées bizarres que donnent ces sortes de maladies; je reconnus en moi-même que je n'avais pas été loin d'une si étrange persuasion. Les soins que j'avais reçus m'avaient déjà rendu à l'affection de ma famille et de mes amis, et je pouvais juger plus sainement le monde d'illusions où j'avais quelque temps vécu. Toutefois, je me sens heureux des convictions que j'ai acquises, et je compare cette série d'épreuves que j'ai traversées à ce qui pour les anciens, représentait l'idée d'une descente aux enfers.

[30] Extrait du site de Pierre Pernoud : http ://hypo.ge-dip.etat-ge.ch/athena/nerval/nerv_aur.htlm

Cette oeuvre a été écrite par Nerval dans les dernières années de sa vie, à partir de souvenirs, de récits, de notes, datant de sa première hospitalisation, plus de 10 ans auparavant. Nous ne présupposerons aucune composition à l'oeuvre (qui est composée de deux parties[31]).

Le texte analysé[32] comprend environ 120 000 caractères[33]. Nous avons appliqué la procédure d'analyse standard. Rappelons que cette procédure consiste à construire un tableau de données comprenant en lignes, les segments de texte composant le corpus (924 unités de contexte élémentaire ou U.C.E.), et en colonnes, les mots pleins (585 mots de fréquence supérieure ou égale à 4). Le tableau soumis à l'analyse statistique est binaire. La valeur « zéro » correspond à l'absence d'un mot dans l'U.C.E., et la valeur « un », sa présence (quelque soit le nombre d'occurrence).
Ainsi n'est retenu du discours dans l'analyse statistique que la co-présence des mots dans les mêmes unités de contexte. L'approche standard est un peu complexe du fait d'une double analyse pour tester les effets d'un changement de longueur des unités de contexte. Mais le principe reste simple : il s'agit d'extraire des classes d'unités de contexte élémentaires stabilisées, très contrastées quant à la distribution de leur vocabulaire, à l'aide de la classification descendante hiérarchique. La double classification impliquant le rejet des unités de contexte non stabilisées, la partie stabilisée ne comprend pas toutes les unités. Ici, elle s'élève à 58% des unités de contexte élémentaires du corpus (u.c.e.). Elle peut être présentée à partir de l'arbre suivant :




Ainsi, la classe 1 recouvre la classe des u.c.e. la plus contrastée du reste. Les deux autres classes correspondent à une différenciation seconde.

Voici le vocabulaire caractéristique de chaque classe[34] :

[31] La première fut publiée quelques jours avant sa mort, et la seconde, vraisemblablement inachevée, quelques mois après.
[32] Comme transformation, nous avons ajouté une ligne d'introduction nécessaire pour le programme et supprimé les étoiles du texte (en les remplaçant par des espaces)
[33] L'espace, la ponctuation, les titres, les notes, les numéros de pages, certaines lettres comme le « oe » sont susceptibles de traitement diverses, selon les programmes qui rendent le nombre de caractères fluctuant d'un programme à un autre, même si l'ordre de grandeur reste inchangé. Quand la précision est illusoire, il est trompeur de la donner.

*Vocabulaire caractéristique de la classe 1 (299 u.c.e.) :* figure+, vis, lumiere+, porte+, fleur+, cru+, jardin+, montagne+, nuage+, couleur+, mit, lune+, forme+, jour+, trait+, mur+, fille+, long+, grand+, maison+, voir, enfant+, instant+, vue+, arbre+, lit+, plein+, cote+, entrai, feu+, haut+, longue+, pied+, voyais, creation+, elanc+er, jeune+, profond+, promen+er, vetement+, blanc+, color+er, escalier+, feuille+, trouvai, angle+, elev+er, guide+, peint+, retraite+, revet+ir, spectacle+, trace+

*Vocabulaire caractéristique de la classe 2 (135 u.c.e.) :* faire, ami+, joie+, lettre+, aimais, retrouver, cimetiere+, dire+, mort, idee+, seul+, tombe+, larme+, aller, papier+, route+, disais, vouloir., mort+, perdu+, senti+, memoire+, mourir, voyant+, poete+, pretre+, moment+, amour+, fatal+, rendre, pari+, evenement+, avoir, sembla, hasard+, espoir+, parole+, portait, rencontr+er, desordre+, environs, devoir.,

*Vocabulaire caractéristique de la classe 3 (105 u.c.e.) :* ame+ , cabal+, esprit+, monde+, etude+, reproduire., religion+, pouvoir., desespoir+, divis+er, venaient, pere+, science+, mauvais+, mere+, compte+, ennemi+, vie+, dieu+, pensee+, eternel+, humain+, sait, eloim, univers, dernier+, image+, histoire+, etre, chretien+, nature+, oriental+, passe+, harmonie+, force+, vague+, nombre+, raison+, err+er, fils, frem+ir, nom+, art+, christ+, gloire+, souvenirs, entoure, existe, mour+ir, certitude+, generation+, immortel+, perle+, puissant+, subir, talisman+, vaincu+

Ajoutons que l'analyse standard effectuée pour cette présentation met en évidence trois classes proches de celles que nous avions publiées en 1990[35] avec une autre retranscription du corpus et une autre version du logiciel. Nous appelons justement *« mondes lexicaux »*, cette tendance du vocabulaire à se distribuer dans des classes stabilisées. De plus, *dans le cas de cette analyse, on a pu montrer que les mondes sont à la fois relatifs à l'oeuvre particulière de Nerval… et ce sont également des mondes stabilisés que l'on retrouve dans d'autres analyses*[36].

Cette oeuvre « Aurélia » a pris corps à partir des notes et souvenirs de Nerval recueillis lors de son internement en 1841 ; Nerval avait vécu sa crise comme une révélation de l'existence d'un autre monde, crises cependant qui se sont répétées de manière plus douloureuses jusqu'à la mort du poète[37] retrouvé pendu rue de la Vielle-Lanterne, au matind du 26 janvier 1855.

Nous pourrions conclure l'analyse de cette oeuvre de la même manière qu'en 1990 :

---

[34] Les classes obtenues sont des classes d'unités de contexte élémentaires. Le vocabulaire présenté correspond aux mots pleins caractérisant plus spécifiquement ces classes (au sens du 2). Il est ordonné par 2 décroissant. Ont été retenus tous les mots pleins de 2 supérieur ou égal à 4.

[35] voir notre article de 1990.

[36] Evidemment le vocabulaire dépend des corpus, mais nous avons constitué pour la version 5 un corpus d'étalonnage d'environ 26 millions de caractères, pour permettre de répartir le vocabulaire usuel dans des classes stabilisée. Cela permet de comparer la distribution du vocabulaire d'une analyse particulière avec celle de l'analyse d'étalonnage…

[37] Au matin du 26 janvier 1855, Gérard de Nerval a été retrouvé pendu, rue de la Vieille Lanterne à Paris avec, dit-on, le manuscrit de la seconde partie d'Aurélia en poche.

*« au vu des résultats, trois types de "monde" semblent se dessiner dans cette oeuvre :*

*- le monde imaginaire, celui des rêves, lié à l'évocation de la nature et des "forces végétantes" (pour reprendre un terme de Bachelard), monde des sensations (visuelles surtout), lieu d'un désir premier qui, chez Gérard, prend le nom d'Aurélia.*

*- Le monde réel : Paris et ses environs ; les amis, les parents, les inconnus ; les rues où Gérard erre des nuits durant en proie à l'ivresse ou la dépression.*

*- Enfin le monde symbolique, à la fois mystique et rationnel, celui à qui Gérard confie ses doutes et ses interrogations sur la vie et la religion, sur le sens des rêves, de ses rêves et de cet espoir fou, une nuit pressenti, qu'il pourra retrouver, à la fin des épreuves de cette vie, tous les êtres chers qui l'ont définitivement quitté. »*

Mais ces commentaires ne sont pas plus *validés* aujourd'hui qu'hier par l'analyse présentée ! Celle-ci cependant les étayent, car elle montre bien une stabilisation d'un quelque-chose (dont on prend conscience). Mais cela ne peut cependant s'évaluer sans discours. C'est d'ailleurs par le discours que cette prise de conscience s'opère vraiment. Aussi la persistance de ce « quelque chose » est fondamentale. C'est par elle que se met en ordre notre propre expérience de lecture de l'oeuvre. L'analyse rend seulement possible ce processus. Elle rend possible une mise en forme des contenus que les classes évoquent. C'est justement en cela que l'analyse est exploratoire.

Mais une question persiste : Pourquoi une analyse formelle permettrait-elle une telle démarche ? N'est-ce pas une simple illusion ? Notre réponse consiste à dire que *cette analyse formelle ne porte pas sur ce que Nerval cherche à se représenter, mais sur les traces de son activité d'écrivain qui mémorisent ses redites, ses retours.* Les classes expriment ici une tension impossible à résoudre dans son activité d'écrivain (durant plus d'une dizaine d'années). Cette tension repose sur un deuil impossible à résoudre :

> Aimez qui vous aima du berceau dans la bière ;
> Celle que j'aimai seul m'aime encor tendrement :
> C'est la Mort — ou la Morte… O délice ! ô tourment !
> La rose qu'elle tient, c'est la *Rose trèmière.*[38]

Cet impossible du sens ne peut se décliner qu'en rapport avec la tresse évoquée au premier paragraphe : Car cet impossible du sens est également l'impossible de tout désir… Et la démarche de Nerval rejoint la démarche commune, avec sans doute une intensité qui la rend plus impérative, plus émouvante. Mais, pour chacun, le sens se déploie dans un temps indéterminé. Le temps n'exprime-t-il pas concrètement cette impossibilité même de trouver, ici et maintenant, ce que l'on cherche sans cesse à unifier ? La méthode statistique employée ne fait que détresser cette tresse du sens en séparant dans des mondes distincts ce qui la

[38] Extrait du poème « Artémis » du recueil « Les chimères » (LaPléiade, 1974)

constitue comme trame, en démêlant ainsi les postures successives prises par l'auteur pour donner sens à sa souffrance et à ses contradictions.

## *Bibliographie*


Achard, Pierre (1993) *La Sociologie du Langage,* Paris : P.U.F.

Achard, Pierre (1991) Une approche discursive des questionnaires : l'exemple d'une enquête pendant la guerre d'Algérie, *Langage et Société,* 55, 4-40.

Authier-Revuz, Jacqueline (1982) Hétérogénéité montrée et hétérogénéité constitutive : élément pour l'approche de l'autre dans les discours, *DRLAV* 26, 91-151.

Bachelard, Gaston (1942) *L'eau et les rêves.* Paris : Librairie José Corti

Bachelard, Gaston (1949) *La psychanalyse du feu,* Gallimard

Bakhtine, Michael (1979, v.f., 1984) *Esthétique de la création verbale,* Paris : Gallimard.

Balat, Michel (2000) *Des fondements sémiotiques de la psychanalyse,* Paris : L'harmattan.

Benveniste, Emile (1966) *Problèmes de linguistique générale II.* Paris : Gallimard.

Benzécri, Jean-Paul & col (1973) *Analyse des Données,* Tomes 1 et 2. Paris : Dunod.

Benzécri, Jean-Paul & col (1981) *Pratique de L'Analyse des Données* – Linguistique & Lexicologie, Dunod

Benzécri, Jean-Paul (1982) *Histoire et Préhistoire de L'Analyse des Données*, DUNOD

Blanchot, Maurice (1969) *L'entretien infini,* Gallimard

Charaudeau, Pierre et Maingueneau, Dominique (2002) *Dictionnaire d'analyse du discours.* Paris : Seuil

Derrida, Jacques (1967) *L'écriture et la différence,* Editions du Seuil

Dufour, Dany-Robert . (1990) *Les mystères de la trinité,* Éditions Gallimard

Dufour,Dany-Robert (1988) *Le bégaiement des Maîtres, Lacan, Benveniste, Lévi-Stauss….* Strasbourg : Arcanes.

Everaert-Desmedt, N. (1990) *Le processus interprétatif : Introduction à la sémiotique de Peirce,* Pierre Mardaga Editeur, Liège.

Foucault, Michel (1969) *L'archéologie du savoir.* Paris : Gallimard.

Foucault, Michel (1971) *L'ordre du discours.* Paris : Gallimard.

Frédéric François (1998) *Le discours et ses entours,* L'Harmattan

Freud, Sigmund (1950, v.f. 1956, A. Berman) *La naissance de la psychanalyse.* Paris : P.U.F.

Guitart, René (1999) *La pulsation mathématique.* Paris : L'Harmattan.

Harris, Zellig S., (1954, v.f. 1969, Dubois-Chalier) Analyse du discours, *Langage*, N° 13, pp 8-45.

Lacan, Jacques (1964, v.p. 1973, J.-A. Miller) *Le Séminaire, Livre XI (Les quatre concepts de la psychanalyse)* Paris : Seuil (Points)

Lacan, Jacques (1972-73, v.p.1975, J.-A. Miller) *Le Séminaire, Livre XX (Encore)* Paris : Seuil (Points)

Lebart, Ludovic (1982) L'analyse statistique des réponses libres dans les enquêtes socio-économiques, *Consommation*, 1, 39-62, Dunod

Lebart, Ludovic et Salem, André (1994) *Statistique Textuelle,* Paris : Dunod.

Noël-Jorand M. -C., Reinert M. & al. (1997) A new approach to discourse analysis in psychiatry, applied to a schizophrenic patient's speech. *Schizophrenia Research Elsevier Science* 25, 183-198.

Noël-Jorand, M.C., Reinert, M., & al. (1995) Discourse analysis and psychological adaptation to high altitude hypoxia, *Stress Medecine,* vol 10, 27-39

Pêcheux, Michel (1969) *Analyse automatique du discours.* Paris : Dunod. [voir égagalement Maldidier (1990)]

Maldidier, Denise (1990) *Linquiétude du discours : Textes de Michel Pêcheux,* Editions des cendres

Peirce, Charles S. (1978, traduit et commenté par G. Deledalle) *Écrits sur le signe,* Éditions du Seuil

Peirce, Charles S. (1931-1935) *Collected Papers of Charles Sanders Peirce*, 8 vols., Charles Hartshorne, Paul Weiss, and A. W. Burks (eds.) Cambridge, MA : Harvard University Press.

Pottier, Bernard (ed.) (1973) *Le Langage,* Paris : centre d'étude et de promotion de la lecture

Ramos, J.-M., Reinert M. (2004) " Les inscriptions du temps dans les devises solaires. Analyse d'un corpus ancien par la méthode Alceste ", *Temporalités,* 1, 19-36

Reinert, M. (1997) « Les Mondes lexicaux des six numéros de la revue "Le Surréalisme au Service de la Révolution" », *Cahiers du Centre de Recherche sur le Surréalisme (Mélusine),* L'Age d'Homme, XVI, 270-3

Reinert, M. (1998) « Mondes lexicaux et Topoï dans l'approche ALCESTE», in *Mots chiffrés et déchiffrés, Mélanges offerts à Etienne Brunet*, SLATKINE, Genève, 289-303

Reinert, M. (1998) « Processus catégorique et co-construction des sujets et des mondes dans différents récits », *Linguistique et Psychanalyse,* Colloque de Cerisy-la-salle, septembre 1998;

Reinert, Max (1983) Une méthode de classification descendante hiérarchique. *Cahiers de l'Analyse des Données* 3 : 187-198.

Reinert, Max (1990) ALCESTE, une méthodologie d'analyse des données textuelles et une application: Aurélia de Gérard de Nerval, *Bulletin de Méthodologie Sociologique* 26, 24-54.

Reinert, Max (1993) Les ‘mondes lexicaux’ et leur ‘logique’ à travers l'analyse statistique d'un corpus de récits de cauchemars. *Langage et Société* 66, 5-39.

Reinert, Max (2001) *ALCESTE*, une méthode statistique et sémiotique d’analyse de discours – Application aux ‘ Rêveries du promeneur solitaire ’, *Revue Française de Psychiatrie et de Psychologie Médicale* n° 49, 32-36

Reinert, Max (2001) Processus catégorique et co-construction des sujets et des mondes à travers l'analyse statistique de différents corpus. In *Linguistique et Psychanalyse*, [colloque de Cerisy, Sept. 1998], M. Arrivé et Cl. Normand (eds.), 379-392. EDITIONS IN PRESS.

Reinert, Max (2003) " Le rôle de la répétition dans la représentation du sens et son approche statistique dans la méthode Alceste ", *Semiotica*, 147, 389-420

Sapir, Edward (1968, présentation de Jean-Elie Boltanski) *Linguistique*, Les Editions de Minuit.

Saussure, Ferdinand (1972) *Cours de linguistique générale.* France : Payot.

Wald, Paul (1999) Classes d'énoncés, dimensions modales et catégories sociales dans ALCESTE, *Utinam,* 1999-1/2, 303-24.

Wittgenstein Ludwig (1961, v.f. Klossowski) *Tractatus logico-philosophicus.* Paris : Gallimard

Zapata, Luz. et Sauret, Marie-Jean, (2001) Analyse du discours et méthode psychanalytique : la question du style dans la clinique des névroses, *Revue Française de Psychiatrie et de Psychologie Médicale* 49, 57-65