

Dialogue méthodologique autour de l'utilisation du logiciel Alceste en sciences humaines et sociales : "lisibilité" du corpus et interprétation des résultats

Angeline Aubert-Lotarski¹, Valérie Capdevielle-Mougnibas²

¹ CREFI – Université Toulouse-le-Mirail – 31058 Toulouse cedex – France
DFC – Université Marc Bloch – 67000 Strasbourg – France

² PCS – Université Toulouse-le-Mirail – 31058 Toulouse cedex – France

Abstract

Based on the exchanges of two lecturers in Education Sciences and Clinic psychology, this contribution intends to expose and question the bases of two research practices in human and social sciences, using the software Alceste for the analysis of their data. We insist on the need for articulating the methodological, theoretical and epistemological options of the software to those of the research practice concerned. We will be particularly interested in the procedures, above the analysis, which allow to improve the "legibility" of the results, and in the pros and cons of the various possible strategies of interpretation starting from this analysis software.

Résumé

Cette contribution, écrite à partir des échanges de deux enseignants-chercheurs en Sciences de l'Education et en Psychologie clinique, a pour objectif d'exposer et d'interroger les fondements de deux pratiques de recherche en sciences humaines et sociales utilisant le logiciel Alceste pour l'analyse de leurs données. L'accent est mis sur la nécessité d'articuler les options méthodologiques, théoriques et épistémologiques du logiciel à celles de la démarche au service de laquelle il est employé. On s'intéressera en particulier aux procédures, en amont de l'analyse, permettant d'améliorer la "lisibilité" des résultats ainsi qu'aux avantages et limites des différentes stratégies d'interprétation possibles à partir de cet outil d'analyse.

Mots-clés : logiciel Alceste - interprétation - énonciation - pratiques de recherche

1. Introduction

Cette contribution, écrite à partir des échanges de deux enseignants-chercheurs en Sciences de l'Education et en Psychologie clinique, a pour objectif d'exposer et de questionner les fondements de deux pratiques de recherche en sciences humaines et sociales utilisant le logiciel Alcesteⁱ pour l'analyse de leurs données. Jusqu'à présent, nous avons surtout utilisé l'outil Alceste à partir d'une connaissance fonctionnelle et statistique du logiciel qui laisse à un second plan les travaux linguistiques qui ont été à son origine. Ainsi nous avons peu à peu construit un ensemble de routines, de "trucs" - voire de trucages ? - permettant de "faire tourner" et de "faire parler" un corpus. En parallèle, s'est développé le sentiment que ces pratiques d'analyse de données sont peut-être sur-déterminées par des exigences informatiques liées au fonctionnement propre du logiciel ou au cadre théorique qui le sous-tend. Il nous a

ⁱ Ce logiciel d'analyse de données textuelles a été conçu par M. Reinert à partir des travaux de JP. Benzécri. Le terme Alceste était originellement présenté comme l'acronyme de *Analyse lexicale par contexte d'un ensemble de segments de texte* ; plus récemment lors d'une conférence donnée le 7 mai 1999 à l'Institut d'études doctorales de l'Université Toulouse-le-Mirail, son auteur l'a défini comme l'*Analyse des lexèmes cooccurrents dans les énoncés simples d'un texte*. Le logiciel, sous licence CNRS/UTM, est commercialisé par la société Image (Toulouse).

alors semblé nécessaire d'interroger l'adéquation entre cet usage - productif du point de vue des résultats de recherche - et les repères théoriques et épistémologiques qu'il implique nécessairement. Cette communication marque donc une étape dans la poursuite de nos échanges et de nos discussions sur nos usages respectifs. Il ne s'est pas agi pour nous de formuler une problématique unique mais d'exposer, à partir de deux exemples issus de travaux s'inscrivant dans des paradigmes de recherche très différents, deux des grands pôles de questionnements récurrents dans notre dialogue méthodologique. Après une présentation de la méthodologie Alceste et de ses objectifs, chacune d'entre nous s'attachera à présenter à l'aide d'exemples les principales caractéristiques de son mode d'utilisation du logiciel. Il s'agira d'insister sur les questionnements issus de ces pratiques. A partir d'un corpus composé de 27 entretiens semi-directif menés auprès d'étudiants et de professionnels du conseil en management, Angeline Aubert-Lotarski traitera plus particulièrement des aspects méthodologiques liés à l'utilisation de cette technique. Elle présentera, en particulier, certaines des procédures qu'elle emploie pour faciliter la "lisibilité" du corpus par le logiciel et augmenter la pertinence des résultats ainsi obtenus. Valérie Capdevielle-Mougnibas, traitera pour sa part des fondements linguistiques du logiciel. A partir d'une recherche sur le diagnostic prénatal des pathologies fœtales urinaires, elle insistera sur certaines difficultés soulevées par l'interprétation des résultats fournis par Alceste. Enfin, il s'agira en conclusion de lancer des pistes de discussion entre chercheurs-utilisateurs et chercheurs-concepteurs de logiciels d'analyse de données textuelles pour peut-être parvenir à une utilisation plus éclairée de ces outils extrêmement précieux quel que soit le paradigme de recherche envisagé.

2. La méthodologie Alceste

2.1. Fonctionnement du logiciel

L'objectif général de cette méthodologie vise, comme le précise son auteur, à "déterminer comment sont organisés les éléments qui constituent [un texte]" (Reinert, 1992). Il s'agit de "réduire l'arbitraire de la description du corpus en mettant en évidence ses régularités, ses symétries cachées" (Thom, 1974, 21). Alceste met l'accent sur les ressemblances et dissemblances du vocabulaire et rend compte de sa distribution dans les propositions qui constituent le texte étudié. Il construit une modélisation "des lois de distribution du vocabulaire dans un corpus à l'aide d'un tableau à double entrée croisant unités de contexte (u.c.) et vocabulaire retenu" (Reinert, 1990). Il faut donc noter que les mécanismes statistiques mis en oeuvre sont indépendants du sens. A ce moment de l'analyse, "le discours est vu comme une combinaison de phrases, une suite linéaire, et l'objet de l'analyse n'est pas d'en chercher le sens mais de déterminer comment sont organisés les éléments qui le constituent, sans faire appel à la connaissance que l'analyste peut avoir du sens spécifique de chaque morphème, de l'intention de l'auteur ou de sa situation." (Marchand, 1998 ; 52) Le logiciel établit un classement statistique des subdivisions (proposition, phrase ou paragraphe) du corpus en fonction de la distribution des mots dans ces subdivisions. Deux analyses sont principalement effectuées :

- *Mise en évidence du commun* : Une **Classification hiérarchique descendante (CHD)** établit des "profils-types" de réponses. On peut parler également de tendances, ou de typologie. Dans le cas d'un corpus texte, on emploiera les termes "classes", "mondes lexicaux" ou "zones d'insistances" du discours pour désigner une partie composite du corpus caractérisée par le sur-emploi – la présence significative – de certaines formes. En d'autres termes, chaque classe de la CHD, regroupe des formes présentant des environnements (des cooccurrents) similaires. D'un point de vue technique, le logiciel propose de procéder à deux analyses successives du corpus, selon un découpage légèrement différent. Ainsi la classification construite se base sur

les éléments stables des deux analyses, limitant alors les aléas du découpage en segments de texte.

- *Mise en évidence des différences* : Une **Analyse factorielle des correspondances (AFC)** permet de dégager des logiques (les facteurs) dans les différences de prises de position. Ainsi, l'AFC renseigne sur les rapports de proximité ou d'éloignement des différentes parties du corpus caractérisées par la CHD, les unes par rapports aux autres (Marchand, 1998 ; 52). Ces relations sont représentées graphiquement dans un espace à deux dimensions (plan factoriel).

Comme on peut le voir, Alceste nous donne ainsi accès à plusieurs niveaux d'analyse de la structure du discours du sujet. Il apporte, notamment, des indications sur le type de vocabulaire le plus fréquemment utilisé, la cooccurrence de certains termes, la présence d'énoncés répétés, l'importance de certaines formes comme les déictiques par exemple, etc... Pour autant, l'analyse ne se limite pas à l'étude des caractéristiques du vocabulaire qui constitue l'énoncé. En construisant des classes d'unités de contexte regroupées sur la base de la distribution différenciée et de la cooccurrence des mots qui les composent, le logiciel offre la possibilité de décrire ces classes à partir de différentes informations telles que le relevé du vocabulaire le plus spécifique ou encore la sélection des unités de contexte les plus représentatives du vocabulaire caractéristique de la classe. Ainsi, il tient compte des trois plans principaux de structuration sémantique d'un corpus : le vocabulaire, les propositions, et le texte en lui-même. Les classes sont appelées "mondes lexicaux". Elles sont considérées comme présentant un reflet, une "image" des représentations du sujet. L'intérêt des classes qui rendent compte de l'organisation formelle du corpus réside dans les possibilités d'interprétation sémantique qu'elles offrent. *"Même si les formes sont reconnues à l'aide d'opérations formelles, leur utilité est de pouvoir servir de support au sens"* (Reinert, 1990). Chaque classe renvoie, selon Reinert, non pas uniquement à la représentation d'un objet mais aussi *"à la manière dont un sujet l'appréhende en fonction de sa propre identité, en fonction aussi de son intention"* (Reinert, 1990). Elles traduisent le point de vue du sujet. *"Notre hypothèse principale, écrit l'auteur, consiste [...] à considérer le vocabulaire d'un énoncé particulier comme une trace pertinente de ce point de vue. Il est à la fois la trace d'un lieu référentiel et d'une activité cohérente du sujet énonciateur"* (Reinert, 1993). La "trace linguistique" de l'énoncé du sujet est envisagée, dans cette perspective, comme reflétant l'acte d'énonciation qui le supporte au sens linguistique du terme. Les classes mises en évidence ont, de ce fait, le statut d'un signe qui renvoie à un référent défini par l'auteur comme "une image mentale". Même si cette "icône" apparaît globale et approximative, elle permet de rendre compte des cadres référentiels qui structurent le discours du sujet, de sa position, et de son attitude envers le monde qui l'entoure. Face aux possibilités offertes par le logiciel et aux objectifs présentés par son inventeur nos échanges nous ont conduit à discuter différents points, dont le lien existant entre la théorie de l'énonciation sous-jacente aux différentes techniques d'analyse et les pratiques d'interprétation des résultats qui en résulte.

2.2. Les instances de l'énonciation

L'intégration de la problématique de l'énonciation à l'analyse du discours ne va pas sans passer par une critique radicale de la notion de sujet parlant, soit de « l'actant » du discours. Que faut-il entendre par ce que certains appellent la subjectivité énonciative ? Pour les non-spécialistes en linguistique que nous sommes, il nous semble que l'on peut distinguer deux grandes dimensions de l'énonciation dans le champ de l'analyse du discours.

2.2.1. Le sujet linguistique

Selon Benveniste, la subjectivité dont il s'agit dans le champ de la linguistique renvoie à la capacité du locuteur à se poser comme sujet. *"Elle se définit non comme le sentiment que*

chacun éprouve d'être lui-même (ce sentiment, dans la mesure où on peut en faire état, n'est qu'un reflet), mais comme l'unité psychique qui transcende la totalité des expériences vécues qu'elle assemble, et qui assure la permanence de la conscience. Or nous tenons que cette subjectivité qu'on la pose en phénoménologie ou en psychologie, n'est que l'émergence dans l'être d'une propriété fondamentale du langage. Est ego qui dit ego. Nous trouvons là le fondement de la subjectivité qui se détermine par le statut linguistique de la personne" (1966, 260). On repère ici que, selon Benveniste, la subjectivité trouve son fondement, son principe dans le langage. L'acte d'énonciation révèle le sujet qui le pose avant de dire quelque chose du monde. Elle est la condition sine qua non de l'individuation. Ce sont les formes linguistiques de la subjectivité qui déterminent la possibilité de se reconnaître comme sujet, et non la subjectivité qui précède la possibilité de son expression. En d'autres termes, l'expérience de la subjectivation se trouve tout entière liée à celle de ses formes linguistiques. On repère ici une première application de l'analyse du discours dans le champ des sciences humaines. Il revient à la linguistique dite de l'énonciation voire à la pragmatique d'étudier le matériel linguistique qui permet l'expression de la subjectivité (au sujet de se poser et de se situer dans et par le langage). Ces marqueurs de la subjectivité sont dans cette perspective envisagées du point de vue de leur fonction, étudiée qu'à partir de l'activité des locuteurs. Ainsi l'acte d'énonciation par lequel tout sujet énonce sa position de locuteur est tout à la fois un acte de conversion et un acte d'appropriation de la langue en discours. Le fait que par cet acte le locuteur mobilise la langue pour son propre compte détermine une situation d'énonciation dans laquelle émergent les énoncés.

2.2.2. *Les places d'énonciation*

De nombreux théoriciens de l'analyse du discours mettent en cause le postulat de l'originalité du sujet partant à la fois de son unité et de son autonomie. L'accent est mis, par ce qu'on appelle l'école française de l'analyse du discours, sur le système de contraintes socio-discursives qui pèse sur toute prise de parole. Ainsi, Maingueneau (1994) insiste sur la différence existant de ce point de vue entre l'analyse du discours telle qu'il la définit et la pragmatique incompatible quant à leurs présupposés théoriques surtout en ce qui concerne la question de la subjectivité énonciative. Selon lui, la pragmatique s'intéresse à l'intention de sujets parlants dont la conscience serait transparente à elle-même et l'identité stable par delà les différents rôles qu'ils jouent. Les objets de l'analyse du discours correspondent donc assez bien à ce qu'on appelle le plus souvent des formations discursives en se référant plus ou moins directement à Michel Foucault. Pour Foucault, le discours est envisagé comme un ensemble de règles anonymes, historiques, toujours déterminées dans le temps et l'espace qui ont défini à une époque donnée, et pour une aire sociale, économique, géographique ou linguistique donnée les conditions d'exercice de la fonction énonciative. Dans cette perspective, il ne s'agit plus d'envisager un corpus en tant qu'il a été produit par tel sujet mais en tant que son énonciation est le corrélat d'une certaine position socio-historique pour laquelle les énonciateurs apparaissent interchangeable. Du coup, cette école préfère formuler les instances d'énonciation en terme de place de manière à insister sur la prééminence et la préexistence de la topographie sociale sur les sujets parlants qui viennent s'y inscrire ; Foucault le dit clairement : "*Les discours doivent être traités comme des ensembles d'événements discursifs (...) des séries homogènes mais discontinues les unes par rapport aux autres*" (1971, 59-60). Il existe donc plusieurs types de discours qui s'opposent les uns aux autres, et il est possible de rendre compte du principe de rangement et du point d'origine de chacun d'entre eux. Nous pourrions définir, dans un premier temps, le discours comme l'ensemble de ce qui se dit, de ce qui s'écrit sur un domaine envisagé, chaque type de discours prenant corps par rapport à un ensemble de contraintes et, notamment, en référence à "l'épistémé" de son époque. Ainsi, un discours résulte de l'homogénéité des paroles et/ou des écrits d'une communauté de locuteurs

ou d'écrivains indépendamment de la communauté linguistique à laquelle ils appartiennent. De ce point de vue, l'auteur du discours n'est plus un individu parlant ou écrivant, mais "*un principe de groupement du discours comme unité et origine de leur signification, comme foyer de leur cohérence*" (Foucault, 1971). Par exemple, le discours médical ne se réduit pas au discours de l'institution médicale même si ce dernier en fait partie. Ainsi, chaque type de discours fonctionne à partir d'un principe de régularité qui lui est propre et d'un certain nombre de règles (qui ne lui sont pas imposées par la seule discipline dont il relève) qui déterminent son contenu, sa nature mais aussi ses formes de diffusion et de circulation. Donc, à partir du principe de classement des faits linguistiques qui le fondent, il est possible de rendre compte de ce qui le constitue. Il existe, en effet, pour reprendre les termes de Foucault un "*ordre du discours*". Il devient ainsi possible d'envisager le discours centré sur un domaine particulier à partir de ses conditions de production socio-historiques. Il existe de ce fait un nombre limité de discours. On le constate : si le discours est soumis à un certain nombre de contraintes, il ne va pas sans produire des effets sur ceux qui participent à son existence. "*Les discours religieux, judiciaires, thérapeutiques et pour une part aussi politiques, ne sont guère dissociables de cette mise en œuvre d'un rituel qui détermine pour les sujets parlants à la fois des propriétés singulières et des rôles convenus*" (Foucault, 1971). Dans cette perspective, chaque type de discours apparaît comme une forme qui détermine le lien social et impose une position et un rôle définis à chacun des partenaires en présence. L'exercice du discours présuppose une place d'énonciation affectée de telle ou telle capacités que tout individu dès lors qu'il l'occupe est censé détenir. La parole médicale ne peut pas venir de n'importe qui, sa valeur, son efficacité, ses pouvoirs thérapeutiques eux-mêmes, et d'une façon générale son existence comme parole médicale ne sont pas dissociables du personnage statutairement défini qui a le droit de l'articuler. La théorie du discours n'est pas ici une théorie du sujet avant qu'il énonce mais une théorie de l'instance d'énonciation qui est en même temps et intrinsèquement un effet d'énoncé. L'accent est mis ici sur le genre du discours. Face à ces deux grandes conceptions de l'énonciation et de la subjectivité énonciative, il ne nous a pas semblé si facile que cela d'identifier la théorie de l'énonciation qui sous-tend le logiciel. En effet, si l'on examine les travaux contemporains qui font cas de cette technique, il existe deux façons d'envisager un corpus :

- selon qu'on considère qu'il a été produit par tel sujet
- selon qu'on considère qu'il a été produit par des sujets différents.

Qu'en est-il de la signification des résultats obtenus à partir d'un discours émanant de sujets différents ? N'y a-t-il pas une contradiction à considérer le discours comme le reflet de l'acte d'énonciation et à traiter un corpus composé du discours de plusieurs personnes ? La dimension de l'énonciation est-elle systématiquement, pour les travaux qui font ce choix, envisagée comme le corrélat d'une certaine position socio-historique pour laquelle les énonciateurs apparaissent substituables, interchangeable ? Une tierce voie est-elle possible, comme l'affirment certains psychologues sociaux utilisant cette démarche méthodologique ? En parallèle, comment cet outil statistique fondé sur des critères formels et quantitatifs peut-il être mis au service d'une démarche de recherche qualitative ? A quoi Alceste, au travers de la structuration du corpus qu'il révèle, nous donne-t-il accès ? En présentant de manière plus précise deux de nos travaux, nous souhaitons densifier et contextualiser ces interrogations. Le premier exemple insistera plus particulièrement sur l'intérêt de préparer le corpus pour faciliter l'interprétation des résultats. Nous mettrons l'accent, dans une seconde partie sur les difficultés à effectuer cette interprétation sans un solide cadre théorique.

3. "Lisibilité" du corpus et interprétation des résultats

Pour développer cet aspect de notre dialogue méthodologique, je m'appuierai sur l'analyse d'un corpus composé de 27 entretiens semi-directifs menés auprès de professionnels du conseil en management (13 sujets) et d'étudiants se destinant à cette même profession (14 sujets, niveau Bac + 5, en formation initiale ou continue)ⁱⁱ. Le logiciel Alceste a ici été employé comme un outil d'aide à l'interprétation d'un corpus textuel composite au sein duquel il va permettre de dégager des lignes de force, des polarités sur lesquelles le chercheur va pouvoir étayer son interprétation. La recherche évoquée est la phase quantitative d'un processus de recherche plus global. Généralement, cette première partie à visée exploratoire est associée à une phase de type qualitatif où des monographies permettent de mettre au jour des processus psychologiques individuels. Me situant dans le cadre théorique de la psychologie sociale, le discours est abordé comme le reflet de l'objet sur lequel il porte mais aussi de l'activité d'un sujet – qui cherche à assurer une cohésion entre le réel et ses expériences antérieures – et de son insertion sociale. Dans cette première phase exploratoire, l'objectif poursuivi était de repérer différents rôles attribués à l'expérience professionnelle dans ce secteur d'activité. Une seconde hypothèse de travail établissait l'existence de discours différenciés selon l'expérience professionnelle du sujet (étudiants en formation initiale vs. autres interviewés) et, en particulier, selon l'expérience professionnelle du conseil (étudiants en formations initiale et continue, débutants dans le conseil vs. professionnels expérimentés du conseil). Enfin, je m'intéressais à la place prise par le terme expérience dans les discours. Ainsi, on considère classiquement que la fréquence d'un terme ou d'un thème dans un propos est indicatrice de son importance pour l'auteur. On retrouve notamment ce principe dans la définition par leur centralité des éléments constitutifs du noyau des représentations socialesⁱⁱⁱ. Cette hypothèse apparaissait comme cohérente avec la méthodologie Alceste qui met en avant des zones de récurrence dans les discours. Les résultats de cette recherche ne seront pas discutés dans leur contenu, mais après avoir précisées mes attentes quant à l'utilisation de ce type d'analyse, je présenterai des procédures techniques mises en place en réponse à des limites ou biais fonctionnels identifiés, mais aussi les questions qu'elles ne manquent pas de soulever. Dans le cadre de cette recherche, trois attributs d'une analyse Alceste m'apparaissaient particulièrement pertinents :

- Un gain de temps significatif lorsque l'on traite un grand nombre de données dans une perspective exploratoire et sans la prétention de procéder à une validation d'hypothèses.
- La possibilité d'une approche des discours différente d'une analyse de contenu sémantique. Ainsi l'analyse pratiquée avec Alceste vise "*à établir des procédures méthodologiques permettant de réduire et de classer les segments de texte, c'est-à-dire d'en donner un condensé parfaitement formalisé et d'en construire la structure.*" (Marchand, 1998 ; 52). Lorsque le chercheur a pour objectif l'analyse du sens, comme c'est le cas dans les études de représentations, l'utilisation d'un tel logiciel constitue à la fois un détour et un outil méthodologique pour contrer les a priori du chercheur. Il est alors fait l'hypothèse que la structure formelle du discours implique des relations de sens établies par le sujet.
- La possibilité de se centrer sur les réponses des sujets, qui seront catégorisées en fonction de leur ressemblance (CHD), puis opposées (AFC). La recherche d'une

ⁱⁱ A. Aubert, *Valeurs, validité, valence de l'expérience professionnelle et activités du conseil*, thèse : Sciences de l'éducation, UTM, Toulouse, 2000.

ⁱⁱⁱ Cf. pour exemple les travaux de "l'école d'Aix" avec P. Moliner (1996) et J-C. Abric (1994).

correspondance avec un individu ou des groupes d'individus caractérisés par certaines variables extra-textuelles se faisant dans un second temps^{iv} (Marchand, 1998 ; 53). Cette fonctionnalité était particulièrement utile pour aborder la seconde hypothèse de travail précédemment énoncée. Avec cet outil de recherche, mon intention était d'éviter le risque de se focaliser sur des catégorisations et des attributions prédéterminées.

Ainsi, afin de mettre au jour des relations inattendues entre variables (intra et/ou extra-textuelles) et des hypothèses nouvelles, la distance est marquée par rapport aux sujets réels interviewés. Le résultat de l'analyse est considéré comme une modélisation heuristique, à partir des énoncés transcrits, des discours recueillis. Chaque classe de discours, ou monde lexical, constitue un sujet épistémique synthétisant de manière homogène une part des discours de un ou plusieurs sujets réels. Ma connaissance du fonctionnement statistique du logiciel et ma pratique sur différents corpus m'a amenée à élaborer un panel de solutions techniques permettant de réduire certains biais de l'analyse. Globalement, il s'est agi en constituant le plan d'analyse et en modifiant la transcription fidèle des entretiens de lever les ambiguïtés possibles lors de l'analyse textuelle informatisée et de donner un cadre à cette même analyse pour que les éléments pris en compte soient en cohérence avec les hypothèses de travail. On passera rapidement sur l'illusion technique, la "magie de la machine" pour souligner le double écueil tant d'une simplification extrême que d'une trop grande complexification. Ainsi, les résultats chiffrés doivent être entendus et pensés en termes d'indicateurs partiels et non de preuve d'une compréhension globale et absolue (Marchand, 1998 ; 137). Par ailleurs, de part le fonctionnement du logiciel Alceste, on ne peut échapper à "*la prise en compte de variables non pertinentes, qui n'ont pas de propriétés liées à l'objet étudié, de relations impropres, artefactuelles, produites par les algorithmes eux-mêmes (la prédominance de certains calculs peut masquer d'autres phénomènes), ou au contraire l'omission de variables importantes*" (Marchand, 1998 ; 136). La possibilité d'erreurs, notamment des contresens et des ambiguïtés liés à la non prise en compte du sens, incite à la vigilance, notamment par un retour au corpus. C'est également ce fait qui m'a amenée à "m'autoriser" à agir sur le corpus pour le rendre plus lisible. La "lisibilité" du corpus est ici entendue comme sa capacité intrinsèque à pouvoir laisser voir et mettre au jour le sens qu'il recèle. Augmenter la lisibilité d'un corpus doit donc, sans en altérer le sens, tenir compte des caractéristiques fonctionnelles du logiciel employé. Il ne s'agit donc pas d'une transformation du corpus pour le faire correspondre aux résultats attendus. Là résident l'ambiguïté et la subtilité du rôle du chercheur. Ainsi, pour préparer le corpus de cette recherche, j'ai dû modifier certains termes pour éviter que les différenciations entre les classes se basent sur des différences de niveaux de langage ou sur l'usage de synonymes non signifiants pour la recherche. Par exemple, certains locuteurs employaient le terme *société*, alors que d'autres parlaient de *cabinet* ou de *boîte de conseil*. Ceci aurait pu être à l'origine d'une différenciation de discours non pertinente par rapport à mes hypothèses de travail, j'ai donc choisi de toujours utiliser le code "cab_cons" pour désigner une organisation pratiquant le conseil. De même, j'ai parfois lié ensemble des locutions composées qui présentaient une unité de sens à respecter. Enfin, j'ai dû modifier la transcription de certains termes afin d'éviter une réduction à la racine préjudiciable, les capacités du logiciel dans la levée des ambiguïtés morpho-syntaxiques n'étant pas suffisantes. Ainsi, *s_savoir* distingue le nom du verbe ; *c_consultant* distingue le consultant du participe présent du verbe consulter. D'autre part, pour obtenir un plan d'analyse cohérent avec mon corpus, j'ai modifié deux paramètres par rapport à l'analyse standard proposée par défaut. J'ai en premier lieu procédé à plusieurs essais pour trouver un découpage en u.c.e. qui respecte au mieux le rythme global de

^{iv} On utilisera alors pour enrichir l'analyse des discours, des éléments illustratifs (variables illustratives, le plus souvent des éléments socio-biographiques mais également thématique de l'entretien, questions, locuteur, source du document...) et supplémentaires (rires, silences, questions de l'intervieweur...).

l'élocution^v. Le chercheur a ainsi la possibilité de faire varier la longueur des u.c.e. et de procéder à une classification double (avec 2 longueurs d'u.c.e. différentes) qui permet, en donnant les résultats stables communs aux 2 CHD opérées, de diminuer les aléas du découpage. La seconde modification concerne la liste des "clés et valeurs d'analyse" qui permet d'indiquer le type de mots (adjectif, mots outils, noms...) que l'on souhaite voir pris en compte dans l'analyse (codé 1), ignoré (codé 0), ou rajouté dans un deuxième temps en tant qu'élément supplémentaire ou illustratif (codé 2)^{vi}. Les modifications que j'ai apportées dans ce cas sont principalement relatives aux verbes modaux et aux formes non reconnues. En effet, à la "lecture flottante" des entretiens, les thématiques abordées faisaient apparaître un usage significatif des verbes modaux : savoir, vouloir, pouvoir, falloir... Enfin, l'utilisation de codes (c_consultant, etc.) et de locutions composées rendait indispensable la prise en compte des formes non reconnues, qui étaient bien souvent les mots-clés de mon travail. C'est par une relecture et un travail, coûteux en temps, que l'on arrive progressivement à corriger corpus, dictionnaires et plan d'analyse afin d'obtenir un résultat respectueux du corpus et en cohérence avec les orientations fixées par les hypothèses de travail. Quelle que soit leur efficacité, ces différentes procédures techniques ne manquent pas de poser des questions, aussi bien méthodologiques que théoriques, quand vient le moment de l'interprétation. Ainsi on peut se demander si deux classes différentes relatent véritablement des prises de position différentes. En effet il peut s'agir de modes d'expression différents au niveau du vocabulaire, mais qui pourtant concernent les mêmes objets. Au niveau de la forme, les classes font état de mondes lexicaux différenciés. Le fond peut être exprimé par des synonymes, des paraphrases, on peut donc obtenir deux classes différentes quant à la forme, et proches, voire identiques, quant au fond. Il faut donc veiller à raisonner en termes de monde lexicaux et non en termes d'opinions. Le recours à des lexiques différents selon la formation, l'insertion professionnelle, le domaine d'exercice... est logique et c'est *a minima* ce que l'on constate dans un corpus. Mais quand on consulte les contextes d'utilisation des mots au sein d'une même classe, on constate, dans certains cas, que des opinions différentes sont exprimées en employant pourtant les mêmes termes. Par exemple, dans la classe 4 de la recherche évoquée, on retenait les mots-clés *méthode* et *expérience*. Dans les entretiens, les propos pouvaient être du type "*la méthode ne vaut rien sans expérience*", "*avec de la méthode, pas besoin d'expérience*", "*avec de l'expérience, pas besoin de méthode*". Du point de vue du logiciel il s'agit toujours de cooccurrences entre l'item *méthode* et l'item *expérience*. Ceci amène la question suivante : Le logiciel Alceste met-il en évidence des différences de prises de position au sens où le chercheur l'entend généralement ? Je dirais que si une partie des discours dit "blanc" et l'autre "noir", oui, effectivement, une analyse des lexèmes cooccurents peut le caractériser car il s'agit d'un emploi de mots différents. Cependant, dans les cas où une partie s'exprime à l'affirmative et l'autre dans un mode négatif, cela est bien plus délicat. Ainsi, le logiciel réunit au sein d'une même classe des termes parce qu'ils apparaissent souvent, significativement, ensemble dans une même unité d'analyse. Le cas de l'affirmation et de la négation est particulier car on a des cooccurrences semblables, seuls les marqueurs de modalisation (ne, pas...) indiquent la différence de sens. Si l'on effectue une analyse standard, par défaut la liste des clés d'analyse utilisée ne se base pas sur ces marqueurs de modalisation pour établir la CHD. Afin de vérifier

^v L'u.c.e. est construite dans un compromis entre la forme syntaxique (respect de la ponctuation) et les contraintes statistiques (les u.c.e. doivent être de grandeur comparable). Pour définir les unités de contexte qui seront à la base des statistiques effectuées, le logiciel procède par concaténation des u.c.e. successives (d'une même u.c. initiale, c'est-à-dire d'un même sujet) jusqu'à ce que son nombre de mots atteigne le seuil fixé par le chercheur. Dans certains cas, il est possible de fixer arbitrairement des u.c.e. de taille inégale, en utilisant le symbole \$ comme séparateur. Malgré plusieurs inconvénients, l'avantage principal est une prise en compte bien plus importante du corpus pour l'analyse, notamment intéressante lors de dialogues très courts.

^{vi} Si précieux qu'ils soient lors de l'interprétation, les indicateurs morpho-syntaxiques sont parfois sources d'erreurs dans ce logiciel en raison des ambiguïtés de certains termes.

la prise en compte de ces modalisateurs lorsqu'ils sont des variables actives, j'ai fait un test avec un corpus de texte établi de la manière suivante. Deux entretiens (A et B) étaient fabriqués au moyen d'extraits d'entretiens divers (α , β , χ et δ). Les propos tenus par A et B étaient rigoureusement identiques si ce n'est que A s'exprimait toujours à l'affirmative et B toujours à la négative. J'espérais que le logiciel fasse une première partition en fonction de ce mode d'expression affirmatif ou négatif. En fait, les classes de discours correspondaient aux différents extraits d'entretiens originels (α , β , χ et δ) car le vocabulaire était sensiblement différent. Les formes de négation ont été "éliminées" de l'analyse que j'avais demandée car elles étaient trop fréquentes^{vii}, c'est du moins ce que j'ai déduit des fichiers résultats. Plusieurs tests supplémentaires seraient nécessaires afin de cerner le niveau d'expression de la négation qui peut être pris en compte par Alceste. A mon sens, on ne peut pas pour autant parler d'un défaut du logiciel. Il s'agirait plutôt d'une mauvaise utilisation par un chercheur non conscient des limites et des conséquences du mode de fonctionnement. Consciente de ces possibles biais, je ne peux qu'appeler à la prudence dans l'interprétation des différences de discours. Même si un corpus est quantitativement suffisamment important pour que l'on puisse établir des liens statistiquement valides, il me paraît délicat, voire dangereux, d'aller beaucoup plus loin que le constat de l'existence de récurrences particulières et de mondes lexicaux différenciables par les cooccurrences qui les constituent. On peut ainsi s'interroger sur la place que peut prendre une analyse Alceste dans une démarche d'administration de la preuve de type nomothétique.

4. Subjectivité et interprétation des résultats

La seconde recherche présentée par Valérie Capdevielle-Mougnibas (2000 ; 1998 ; 1997) vise à préciser ce qu'il en est de l'impact et des conséquences de la découverte d'une pathologie foetale sur le vécu de la grossesse par la mère et sur l'investissement de l'enfant à naître. L'ensemble de la partie empirique de ce travail est basé sur l'étude approfondie de cas individuels. Elle s'inscrit dans un paradigme de recherche qualitatif qui met l'accent sur l'approche clinique des phénomènes. Dans ce cadre, l'auteur a rencontré cinq jeunes femmes dont le fœtus était atteint d'une pathologie urinaire^{viii} découverte à l'occasion d'une échographie. Ces pathologies peuvent être considérées comme mineures sur le plan médical au sens où elles ne seront pas à l'origine d'un handicap ou d'une difficulté particulière pour l'enfant qui en est atteint. Pour analyser les cinq entretiens (deux pendant la grossesse, trois pendant la phase postnatale : 15 jours, 1 mois et demi, 3 mois) effectués avec chacun des sujets, j'ai choisi d'associer plusieurs méthodes d'analyse du discours (analyse de contenu, analyse de l'énonciation au sens psychanalytique du terme) dont une analyse statistique des données textuelles effectuée à l'aide du logiciel Alceste. Il devient possible en procédant ainsi de réduire l'arbitraire de la description du corpus en mettant en évidence ses régularités, ses symétries cachées. Cette technique, bien que fondée sur des critères quantitatifs permet de prendre en compte dans l'explication clinique des indicateurs réutilisables par d'autres. Ils pourront ainsi vérifier la validité des résultats et le cas échéant les comparer avec d'autres études. L'analyse effectuée selon ces différentes modalités montre que dans tous les cas, les représentations qui concernent l'enfant ont un rapport étroit avec les circonstances et les termes associés au diagnostic prénatal. J'ai choisi pour étayer mon propos de vous présenter le cas d'une jeune femme (28 ans, secrétaire dans un établissement d'enfant) que j'ai rencontrée pour la première fois au cours de son septième mois de grossesse. J'ai analysé dans un premier temps les cinq entretiens à l'aide du logiciel Alceste. Cette analyse aboutit ainsi à une

^{vii} Il est classique qu'un terme trop fréquent, ne présentant donc pas de cooccurrences particulières, ne soit pas distribué dans une classe particulière. C'est la source d'une erreur d'interprétation très répandue qui amène le chercheur à considérer ce terme comme non significatif pour les sujets alors qu'il n'est que non spécifique.

^{viii} Pour quatre d'entre elles, les malformations en cause se traduisaient par une dilatation excessive des bassinets rénaux. Le cinquième bébé n'avait qu'un seul rein.

modélisation du corpus qu'il recompose en différentes classes regroupées sur la base de la cooccurrence des mots qui les définissent. Alceste m'a permis d'étayer mes interprétations, sur une série de marques textuelles significatives pour le sujet.

4.1. Premier indicateur : la fréquence d'un vocabulaire caractéristique

Dans le cas de Mme L, l'analyse réalisée avec Alceste montre comment les significations induites par le diagnostic et le suivi médical qui se met en place par la suite détermine la façon dont l'enfant est imaginé, parlé par la mère. Par exemple, la classe 7 - le logiciel offre une modélisation du corpus en 8 classes - rassemble tout un ensemble de vocables désignant le corps de l'enfant et notamment ses organes internes. Il est : "corps", "membre", "organe", "muscle", "rein". L'enfant est envisagé à partir de l'image offerte par l'appareil échographique qui, on le voit ici, ne va pas sans influencer l'imaginaire maternel.

4.2. Deuxième indicateur : l'hétérogénéité du vocabulaire

L'hétérogénéité du vocabulaire de la classe 2 par rapport à la classe 7 rend compte de l'importance de l'amniocentèse dans la grossesse de Mme L. Il n'est pas anodin que ce thème de l'amniocentèse se détache par rapport à l'ensemble du suivi médical. C'est à partir du moment où un médecin lui a proposé de faire cet examen que tout a basculé pour Mme L. Il se révèle être synonyme de gravité : *"Pourquoi, pourquoi d'un coup on me propose l'amniocentèse. A partir du moment où on me touchait, où on la touchait, où on touchait le bébé, alors là. Alors que moi, je me sentais très bien (...) je ne me sentais pas du tout inquiète par rapport à ce problème. C'était pas flagrant pour moi. Voilà c'est ça"*. Si l'on examine le vocabulaire caractéristique de ces deux classes, il est possible d'opposer deux ensembles de vocables. Un certain nombre sont relatifs aux caractéristiques de la situation sur le plan médical alors que d'autres rendent plus particulièrement compte du vécu de la patiente : ainsi il apparaît que les termes de "colère" et de "peur" sont uniquement associés à la question de l'amniocentèse tandis que des vocables tels que "alarmer", "inquiéter", "rassurer" rendent compte du vécu en particulier de l'examen échographique : *"Ce qui m'a fait le plus peur c'était aussi l'amniocentèse. (...) En fait, c'est l'amniocentèse qui m'a fait peur parce qu'on touchait quelque part au bébé (Classe 2)." C'est véritablement l'amniocentèse et non pas l'annonce de la pathologie urinaire qui joue un rôle déterminant dans le vécu de Mme L. Cet exemple montre que la découverte d'une pathologie foetale détermine considérablement les projections et les anticipations dont l'enfant est l'objet aussi bien pendant la grossesse que pendant la phase postnatale. Ainsi, la place de l'enfant apparaît singulièrement marquée par ce savoir qui pèse sur son être. Le diagnostic a interrompu pour un temps le roman familial de l'enfant imaginaire. Selon l'inventeur d'Alceste, l'intérêt des classes qui rendent compte de l'organisation formelle du corpus réside dans les possibilités d'interprétation sémantique qu'elles offrent. "Même si les formes sont reconnues à l'aide d'opérations formelles, leur utilité est de pouvoir servir de support au sens" (Reinert, 1990). Comment faire cette interprétation ? Quel sens est-il possible de donner à ces classes ? Il s'agit là d'une étape délicate qui ne va pas sans poser problème au regard de ce que nous savons des registres de l'énoncé et de l'énonciation. Il convient, en effet, d'être prudent et de ne pas verser dans la tendance - pas toujours évitée - à donner trop rapidement une signification à ce qui se présente d'abord comme un ensemble de signifiants regroupés sur la base de leur cooccurrence. Les linguistes montrent bien que la genèse du sens ne dépend pas de la somme des significations qui caractérisent les mots qui composent un énoncé. Il n'est donc pas possible d'interpréter les classes en se souciant uniquement des significations apparentes auxquelles renvoient les mots qui lui sont spécifiques. Il importe de tenir compte de la syntaxe et de replacer chaque terme dans son contexte. Le logiciel permet de le faire très facilement. Il construit, en effet, un concordancier qui remplace chacune des formes spécifiques de la classe dans les u.c.e. qui les contiennent. L'ensemble des auteurs qui utilisent Alceste étayent, d'abord, leur interprétation sur le relevé du vocabulaire le plus spécifique des classes retenues, ordonné selon des critères formels comme par exemple les différentes*

catégories grammaticales (Zaouche-Gaudron, 1995) auxquels s'ajoutent souvent des critères sémantiques (Reinert, 1990, 1992 ; Cascino, 1992 ; Bourçois, 1993). L'accent est mis, dans un second temps, sur les u.c.e. les plus représentatives ordonnées par Chi 2 décroissants. Si cette démarche nous semble tout à fait pertinente et intéressante, le choix des critères sémantiques (Capdevielle, 2000 ; Hermet, 2001) permettant d'organiser les différentes formes retenues ne nous semble pas avoir été suffisamment explicité. Il repose souvent, semble-t-il, sur l'intuition du chercheur, sa sensibilité clinique et sa connaissance des entretiens. Ce parti ne nous semble pas irrecevable à condition qu'il soit clairement affirmé. Pour notre part, nous avons essayé de nous doter de critères rigoureux pour organiser l'ensemble de ces formes caractéristiques des classes. Nous les avons ordonnées en fonction de la force du lien du Chi 2 qui les rassemble mais aussi en tant qu'elles étaient susceptibles de renvoyer à une même thématique à l'intérieur de la classe. Nous avons donc effectué une analyse thématique des u.c.e. les plus caractéristiques de la classe. L'association de ces deux techniques qui consiste à combiner une analyse statistique et une analyse sémantique nous semble particulièrement féconde dans le domaine de la recherche en psychologie clinique. Néanmoins la question de l'interprétation des indicateurs obtenus à l'aide d'Alceste nous semble fondamentale. L'analyste du discours vient apporter sa contribution aux herméneutiques contemporaines. Comme tout herméneute, il suppose qu'un sens doit être atteint et que ce sens est caché, inaccessible sans une technique adaptée. L'analyste du discours ne prétend pas s'instituer en spécialiste de l'interprétation maîtrisant le sens des textes mais seulement construire des procédures exposant le regard-lecteur à des niveaux opaques de l'action stratégique d'un sujet. Il s'agit d'extraire des contenus ou une structure pour répondre à des questions précises. Le pas que nous effectuons, nous utilisateurs de ces méthodes, consiste à considérer le discours en tant que déterminé par l'énonciation comme reflétant ces marqueurs de l'énonciation comme les traces d'opérations psychiques, des stratégies d'interlocution en psychologie, de positions sociales, de faits culturels et sociaux en sociologie, ethnologie, de conjonctures historiques, des processus idéologiques. Dans cette perspective, le discours est envisagé comme une production du sujet qui porte sa marque, sa signature (cf. modélisateur). Le message est envisagé comme la "*trace d'une subjectivité, d'une pensée ou plus précisément d'une intentionnalité*" (Blanchet, 1997, 13). Mais de quel sujet s'agit-il ? Ce sujet est défini par les différents auteurs comme "la personne qui parle", le sujet locuteur, le sujet communicant avec autrui. Selon eux, il ne s'agit plus du sujet universel des linguistes mais d'un sujet particulier dont le discours reflète les opérations mentales. Il importe de repérer au delà de leur définition linguistique le "statut psychologique" (Blanchet, 1997) de ces marqueurs. On repère ici qu'au delà des aspects techniques qui limitent déjà le type d'informations apportées par ces logiciels la question de l'interprétation de ces indices renvoie finalement à la théorie des rapports du langage (au sens large) et de la pensée adoptée par l'utilisateur. Ses interprétations ont à être étayées par un solide cadre théorique.

5. Conclusion

Alceste est présenté par son auteur et ses utilisateurs comme un instrument qui permet l'accès et la description des représentations du sujet dans leur singularité. Il constitue un outil original qui propose une aide à l'interprétation qu'il ne remplace en aucun cas. Il facilite l'analyse d'entretiens ouverts et il permet de rendre compte de la spécificité des représentations mises en évidence. Il donne accès aux représentations propres du sujet concernant l'objet étudié. On perçoit l'intérêt d'un tel outil pour éviter certains biais introduits par des techniques plus classiques comme le questionnaire qui impose des rubriques préétablies et influence ainsi les réponses des sujets. Néanmoins, pour nous simples utilisatrices de ces techniques les questions restent nombreuses. Nous aurions souhaité avoir votre avis.

Références

- Aubert A. (1997). L'audit : des représentations éclatées. *Education permanente*. 132. 97-107.
- Aubert A. (2000). *Valeurs, validité, valence de l'expérience professionnelle et activités du conseil*. Thèse : Sciences de l'éducation. Toulouse II.
- Benveniste E. (1966). *Problèmes de linguistique générale*. tome 2. Paris. Gallimard.
- Blanchet A. (1997). *Recherches sur le langage en psychologie clinique*. Paris. Dunod.
- Blanchet A. (1991). *Dire et faire dire. L'entretien*. Paris. Armand Colin.
- Bourçois V. (1993). *Du père aux pairs: ou l'influence des modes d'engagement du père sur le développement affectif et social de l'enfant*. Thèse pour le doctorat nouveau régime sous la direction de Jean Le Camus. Toulouse II.
- Capdevielle V., Laterrasse C. (2000). L'incidence du diagnostic anténatal sur l'enfant via le discours maternel. Propositions de méthode. *Apprentissage et Développement*. Volume VIII. 29/30. 193-203.
- Capdevielle V. (1998). Représentations et images de l'enfant pendant la grossesse. In Fine A., Laterrasse C., Préteur Y. *A chacun sa famille : approches pluridisciplinaires. Tome 2*. Toulouse. Editions Universitaires du Sud. 147-160.
- Capdevielle V., Laterrasse C. (1997). Incidences du diagnostic prénatal sur la place de l'enfant à naître. In Préteur Y., De Léonardis M. *Milieus, groupes et développement socio-personnel de l'enfant*. Toulouse. Editions Universitaires du Sud. 163-167.
- Cascino N. (1992). *Répétition d'une perturbation et récurrence du chômage. Le rôle des schèmes adaptatifs construits*. Thèse pour le doctorat nouveau régime sous la direction de Jacques Curie. Toulouse II.
- Foucault M. (1971). *L'ordre du discours*. Paris. Gallimard.
- Ghiglione R., Landré A., Bromberg M., Molette P. (1998). *L'analyse automatique des contenus*. Paris. Dunod.
- Hermet I., Laterrasse C., Capdevielle V. (2001). Rapport au savoir et importance des caractéristiques disciplinaires dans les choix d'études des doctorants en histoire et mathématiques. *L'orientation scolaire et professionnelle*. 30/4. 447-482.
- Mainueneau D. (1994). *L'énonciation en linguistique française*. Paris. Hachette.
- Mainueneau D. (1991). *L'analyse du discours : introduction aux lectures de l'archive*. Paris Hachette.
- Marchand P. (1998). *L'analyse du discours assistée par ordinateur*. Paris. Armand Colin.
- Reinert M. (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 M. (1992). *Manuel d'utilisation du logiciel Alceste*. Inédit
- Reinert M. (1993). Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars. *Langage et société*. 60. 5-39.
- Thom R. (1974). *Modèles mathématiques de la morphogenèse*. Paris. Presses de la Cité.
- Zaouche Gaudron C. (1995). *Analyse des processus de subjectivation et de sexualisation au travers de la relation père- bébé*. Thèse de doctorat nouveau régime sous la direction de Jean Le Camus. Toulouse II.