La variation comme principe d'exploration de
corpus :
Intérêts et limites de l'analyse lexicométrique
interdisciplinaire pour l'étude de discours

Nathalie Garric et Valérie Capdevielle-Mougnibas

Édition électronique

URL : <http://journals.openedition.org/corpus/1681>

ISSN : 1765-3126

Éditeur

Bases ; corpus et langage - UMR 6039

Édition imprimée

Date de publication : 15 novembre 2009

Pagination : 105-128

ISSN : 1638-9808

Référence électronique

Nathalie Garric et Valérie Capdevielle-Mougnibas, « La variation comme principe d'exploration de corpus :

Intérêts et limites de l'analyse lexicométrique interdisciplinaire pour l'étude de discours », *Corpus* [En ligne], 8 | 2009, mis en ligne le 01 juillet 2010, consulté le 07 septembre 2020. URL : <http://journals.openedition.org/corpus/1681>

**La variation comme principe d'exploration de corpus :
Intérêts et limites de l'analyse lexicométrique
interdisciplinaire pour l'étude de discours**

Nathalie GARRIC

Laboratoire Ligérien de Linguistique (LLL - Tours)

Valérie CAPDEVIELLE-MOUGNIBAS

Laboratoire psychologie du développement et processus de
socialisation (PDPS - Toulouse)

Nous aborderons dans cet article, co-écrit par une linguistique et une psychologue, l'Analyse de discours (AD) comme un cheminement vers un lieu de rencontre interdisciplinaire dans les Sciences Humaines et Sociales. Nous discuterons l'intérêt de l'utilisation interdisciplinaire de la pratique de corpus et de l'analyse textuelle des discours.

Notre réflexion s'appuiera sur une recherche qualitative menée conjointement entre 2006 et 2008 à partir d'une commande de la Chambre régionale des Métiers et de l'Artisanat et le Conseil régional Midi-Pyrénées. Cette étude visait à repérer et analyser les processus à l'origine des ruptures de contrat d'apprentissage dans le parcours de formation d'apprentis de niveau V. Nous souhaitons, plus particulièrement, évaluer les enjeux de la socialisation professionnelle soit l'importance de la relation apprenti/maître d'apprentissage et des conditions de travail et de formation au sein de l'entreprise dans la genèse de ce phénomène qui concerne depuis le début des années 90, plus de 25 % des contrats.

Notre démarche relève d'un projet herméneutique dont l'objectif est la compréhension d'un fait psychosocial à partir de l'une de ses manifestations sémiotiques, le discours. Il s'agit de formuler ensemble des hypothèses de lecture de la textualité, susceptibles, par l'analyse de processus discursifs, de révéler

Corpus n°8 « Corpus de textes, textes en corpus »
(2009), 105-128

des positionnements socio-idéologiques qui permettront de d'étudier les facteurs et processus en jeu.

1. Une réflexion renouvelée

La volonté de définir les fondements d'une herméneutique est, depuis quelques années, soulignée par de nombreux travaux (Adam et Heidmann, 2005). L'AD a été, dès son origine, animée par la recherche d'outils et de méthodes pour démasquer les pouvoirs de l'idéologie. Toutefois, si la recherche d'un protocole d'interprétation des textes, en lien avec la définition d'une sémantique discursive, appartient à l'histoire de l'AD, la notion de *corpus* l'actualise sous un jour nouveau lié à la dimension applicative à laquelle sont soumises les analyses de corpus. Les commandes sociales, impulsées par la disponibilité des données numériques, sont plus nombreuses, en direction de disciplines variées, favorisant elles-mêmes, sur le terrain, la circulation d'outils théoriques et applicatifs à définir en vue d'une exploitation interdisciplinaire et d'une diffusion sociale. Dans ce contexte, deux risques guettent « le champ des sciences discursives » : « une refondation totalisante » et « une réduction utilitariste » (Viprey 2005 : 51). Pour échapper à ces écueils, la théorisation d'une sémantique discursive doit être le véritable enjeu de ce questionnement herméneutique (Rastier 2005). Cet objectif peut toutefois être atteint autrement que par de nouvelles théorisations. L'innovation théorique peut aussi consister en l'articulation, voire la redéfinition, originale d'un certain nombre de concepts présents dans le champ des analyses de discours.

2. L'étude du sens comme lieu de rencontre interdisciplinaire

Au début de notre collaboration, nous souhaitions apparemment étudier le même objet : le sens des discours / textes recueillis. Mais, s'agissait-il d'atteindre le sens d'un fait social, d'une réalité subjective et/ou d'une pratique ? Avions-nous véritablement le même objectif, la même conception du *discours*, du *texte*, du *corpus*, du *sens* ou encore du *sujet* ?

Pour le linguiste...

Atteindre la réalité sociale par le langage implique une mise en relation de ces deux dimensions. Cette articulation est matérialisée par et dans le discours défini comme pratique sociale signifiante (Boutet, Fiala & Simonin 1976). Le discours fait partie de la réalité sociale, il en est une construction dynamique qui lui donne forme et sens. Dès lors, l'ambition de comprendre les phénomènes sociaux devient possible par l'analyse de leurs manifestations langagières si toutefois cette analyse tient compte de la discursivité du matériau. Il faut se donner les moyens de saisir la sociabilité des textes. C'est ce que nous visons par l'AD et l'actualisation de certains présupposés : la *formation discursive* et le *genre*.

La formation discursive rend compte de régularités qui délimitent des familles d'énoncés caractéristiques d'une pratique discursive (Foucault 1969 : 53). Ces régularités fonctionnent comme des normes qui règlent le discours. Pour Pêcheux (1990), la formation discursive relève de conditions de production particulières, une conjoncture sociohistorique donnée, et se matérialise par des indices identifiables au niveau « de la nature des mots employés, mais aussi (et surtout) des constructions dans lesquelles ces mots se combinent » (1990 : 148). Elle est indissociable de la notion d'*idéologie* et de celle d'*interdiscours* : une formation discursive détermine ce qui peut et/ou doit être ou ne pas être dit dans telles conditions de production. Il en résulte, dans une perspective herméneutique, que « le lien qui relie les "significations" d'un texte aux conditions socio-historiques de ce texte n'est nullement secondaire, mais constitutif des significations elles-mêmes » (Pêcheux 1990 : 141).

La notion de *genre* apparaît notamment dans les travaux de Bakhtine et Volochinov (1984). Le genre désigne des régularités formelles liées à une pratique sociale : « Tout énoncé particulier est assurément individuel, mais chaque sphère d'usage du langage élabore ses *types relativement stables d'énoncés* [énonciations], et c'est ce que nous appelons les *genres discursifs* [du langage] » (Bakhtine, cité par Adam 1999 : 87). L'étude des genres, systèmes de contraintes disposant d'autonomie et de stabilité, relève donc

essentiellement d'une analyse énonciative. Ce lien avec l'énonciation est également posé par Achard *et alii* (1996 : 12), « il est possible de mettre en rapport les activités sociales (registres) et des contraintes de formes (les genres discursifs) que l'on peut décrire en s'appuyant sur une linguistique de l'énonciation ». Le *registre* désigne l'organisation sociale et institutionnelle d'une formation discursive : des rôles socio-langagiers manifestés dans et par l'accumulation de textes partageant une même indexicalité (Achard 1995). Le *genre* est le pendant linguistique du registre, il renvoie à l'organisation des formes.

La discursivité est indissociable de l'interdiscursivité qui définit le sujet en tant que subjectivité construite par des rapports de codétermination socio-idéologiques (places sociales) et génériques (place énonciatives). Le sujet discursif n'est ni tout individuel, ni tout intentionnel, ni tout surdéterminé. Relevant du registre, il en reçoit les déterminations identitaires, idéologiques et collectives ; relevant d'une formation discursive, il est sujet (social), singulier et critique, construit dans la différenciation. Cette conception du discours, du sens et du sujet est compatible avec celle du psychologue d'orientation socio-constructiviste.

Pour le psychologue...

Comme le linguiste, le psychologue à condition qu'il travaille à partir de ce paradigme épistémologique envisage l'accès à la réalité comme relevant d'une construction cognitive, sociale et culturelle complexe qui a une histoire, s'inscrit dans des contextes sociaux et est reliée à des pratiques. Il ne peut y avoir d'images fidèles de la réalité ou de vérité révélée. Toute activité humaine relève d'une appropriation de systèmes symboliques socialement construits, fournis par la culture (Vygotsky 1997). Elle s'inscrit toujours dans un contexte et des rapports sociaux portés par des institutions, dans des conditions socio-historiques particulières. Pour ce modèle de la socialisation active, définie à partir des deux versants que sont l'acculturation et la personnalisation (Malrieu 1973), l'étude de l'extériorité du discours structurée par l'AD – en tant que déterminant les conditions de production d'un texte – permet de prendre en

compte les idéologies, les normes, les contextes et le lien social dans lesquels évoluent les sujets. Néanmoins, parce que le chercheur en psychologie vise avant tout à rendre compte de l'activité d'un sujet, contrairement peut-être au linguiste, il mettra l'accent sur la part qui lui revient dans la construction textuelle.

L'individu n'est pas un simple produit de la société : il est l'acteur qui interprète et analyse son environnement, qui participe activement à son existence et donc à son changement. La construction du sens apparaît dans cette perspective comme une des conditions de la détermination qui consiste avant tout pour l'humain en une détermination signifiante. Quel que soit le déterminant envisagé, l'ordre de causalité dont il relève (social, culturel, organique), il produit pour une part son effet à partir des significations que le sujet lui accorde. Il s'agit par conséquent par l'étude des productions langagières d'atteindre une réalité intrinsèquement subjective et sociale.

3. La variation comme stratégie de recherche

La recherche qualitative menée met en œuvre plusieurs procédures de triangulation méthodologique visant à croiser les sources (apprentis, lycéens professionnels, artisans / maîtres d'apprentissage), les modes de recueils (entretiens semi-directifs, textes écrits) et les modes d'analyse (analyses thématiques ou lexicométriques menées par des chercheurs différents) (Capdevielle-Mognibas *et al.* 2008). Cette stratégie de recherche vise à renforcer la validité et la rigueur de notre démarche de construction de savoirs en permettant aux chercheurs de se doter d'une méthodologie robuste mais aussi et surtout d'espérer trouver des résultats comparables et/ou cohérents à partir d'études de nature différente.

Nous avons établi deux corpus de textes, l'un à partir de 48 entretiens semi-directifs recueillis auprès de 24 dyades apprentis/maîtres d'apprentissage, l'autre à partir de 346 textes rédigés par 229 apprentis et 117 lycéens professionnels de niveau V. Ces corpus, organisés en question / réponse, délimitent des unités textuelles à part entière. Ils ont été convoqués successivement pour, à l'étape finale de l'analyse,

parvenir à une lecture informée par les différentes configurations sélectionnées.

Ces choix méthodologiques ont été guidés par notre définition du discours. N'existant que dans l'interdiscursivité, le discours ne peut être qu'un construit visé par l'analyste à partir d'un matériau. Ce matériau est lui-même construit de textes à structurer en corpus. Les textes s'apparentent à une structuration linguistique complexe, une matérialité langagière observable, fondée sur des relations et des agencements formels de différents niveaux. Il définit une totalité créant des dépendances et déterminations internes mais dépendant et déterminé lui-même par cette extériorité que vise la mise en corpus. Le corpus n'est, quant à lui, qu'un point de vue (Lethier & Viprey 2008 : 285), qui, par des mises en relations textuelles, trouve sa dimension sociale ou discursive. La construction de différents corpus constitue autant de points de vue pour une lecture productive qui résulte alors de la confrontation de différentes hypothèses d'interprétation.

Contrairement à certains corpus de ce numéro, ceux que nous exploitons ne témoignent pas des innovations permises par le numérique. Mais, recueillis par le chercheur en psychologie à partir d'une consigne précise, ils relèvent d'une énonciation « non spontanée » et apparaissent ainsi singuliers dans le paysage de la Linguistique de corpus et de l'AD de tradition française.

4. Recueil textuel et corpus

Nous nous intéressons au corpus de textes écrits, dits « bilans de savoir » produits par les élèves, à partir d'une consigne¹ écrite (Charlot, Bautier & Rochex 1992). La passation a été réalisée en classe, sans limite de temps, auprès de 346 jeunes de 15 à 25 ans, au début de la deuxième année de formation. Nous avons obtenu des textes relativement courts se résumant parfois à un mot (par ex. : « Éducation ») et pouvant aller jusqu'à une

¹ « Depuis que tu es né, tu as appris beaucoup de choses, chez toi, dans ton quartier, à l'école et ailleurs... Qu'as-tu appris ? Avec qui ? Qu'est-ce qui est important pour toi dans tout ça ? Et maintenant, qu'est-ce que tu attends ? ».

quinzaine de lignes. Certains jeunes n'ont pas du tout répondu, d'autres ont produit des énoncés (par ex. : « Je ne comprends pas la question posée » ou bien encore « Je n'ai rien à dire ») exprimant leur refus de participer à l'acte discursif par incompetence linguistique ou rejet de la thématique et/ou du protocole. Tous les énoncés ont été pris en compte dans l'analyse.

Le choix de ce type de corpus ne peut être sans conséquence sur le matériau : on soulignera, en particulier, le poids de la normalisation scolaire liée aux conditions de production et notamment à la situation de l'écrit. Nous posons toutefois que le texte conserve son statut de médiateur vers le discours et la réalité sociale. Le dispositif sous-jacent au recueil est envisagé comme un ensemble de contraintes génériques participant à la discursivisation textuelle. Or, toute production discursive s'inscrit dans un genre, dont les déterminations sont plus ou moins fortes. La présence d'une consigne donne nécessairement lieu à une activité d'interprétation. Mais celle-ci est propre à tout discours du fait de son interdiscursivité, laquelle engendre systématiquement une interlocutivité (ou interactivité) explicite ou implicite, *in situ* ou virtuelle. L'interprétation de la consigne doit suggérer plusieurs registres que l'analyse identifiera dans la matérialité énonciative en lien avec les différentes structurations du corpus.

Ces textes – suggérés mais authentiques – sont représentatifs de plusieurs registres du discours sur la formation professionnelle initiale et reçoivent leur homogénéité du protocole de recueil qui s'appuie sur des conditions de production identiques. Les bilans de savoirs ne sont pas, comme tout texte, des discours, ils n'en deviennent que par le travail de structuration des données.

5. Principales caractéristiques génériques

Certaines caractéristiques linguistiques résultent du dispositif propre au bilan de savoirs. Elles ne constituent pas notre objet de recherche mais doivent être identifiées pour ne pas fausser le processus de construction du sens auquel elles participent cependant. Ces contraintes délimitent deux figures énonciatives.

La première met en scène un comportement délocutif marqué par la non-personne, de nombreux syntagmes verbaux à l'infinitif et des syntagmes nominaux abstraits.

La deuxième correspond à un comportement élocutif pris en charge par « je ». Ce comportement est suscité par l'allocutivité de la consigne qui interpelle directement le sujet par « tu » et la modalité interrogative : « Depuis que je suis né, il est clair que j'ai appris pas mal de choses et ce qui est le plus important pour moi est de m'épanouir de jour en jour dans mon métier ! Depuis que je travaille, j'ai appris à m'exprimer, à moins garder les choses pour moi. ». Toutefois cette figure, essentiellement déterminée par la reprise thématique des prédicats de la consigne, est insuffisante pour être maintenue dans toute la séquence textuelle. La délocutivité est plus représentative puisque d'une part les sujets peuvent limiter leur réponse à l'expression du rhème qui s'apparente alors à un syntagme infinitival, (« À vivre avec les autres à rigoler, s'amuser, travailler. »), ou à un syntagme nominal (« Le respect de l'autre, les valeurs humaines, l'humilité ») et d'autre part, s'ils entament la réponse par « j'ai appris », ils glissent rapidement vers une dépersonnalisation du dire : « j'ai appris à être autonome, à se méfier des gens, ne faire confiance qu'à soi-même, à se démerder tout seul ». Cette faible prise en charge discursive est confirmée par les marqueurs de qualification, les adjectifs et adverbes sont très peu nombreux.

6. La lexicométrie comme principe de variation

Le principe de variation va au delà du protocole de recueil des données. Il est également intégré à notre méthode et à nos outils d'analyse : la lexicométrie et l'utilisation de deux logiciels, *Alceste* et *Lexico3*.

La lexicométrie repose sur « toute une série de méthodes qui permettent d'opérer des réorganisations formelles de la séquence textuelle et des analyses statistiques portant sur le vocabulaire à partir d'une segmentation » (Salem 2006). Ces méthodes s'appuient sur un principe comparatif qui produit des sorties quantifiées portant sur le vocabulaire d'un ensemble de textes (sous-corpus ou partie) par rapport au vocabulaire de la

La variation comme principe d'exploration de corpus

totalité du corpus. La procédure permet ainsi une mise en discours du matériau. La comparaison délinéarise les textes en établissant des réseaux interdiscursifs systématiques entre eux : elle est un point de vue dynamique sur le corpus.

Au delà de ce principe commun, chaque logiciel offre des procédures et outils spécifiques.

Le logiciel *Alceste*² rend compte de la distribution du vocabulaire dans les propositions qui constituent le texte à partir d'une procédure de lemmatisation. Il permet de repérer la présence de différents ensembles d'énoncés (classes) qui se ressemblent du point de vue de la cooccurrence significative (Chi2) du vocabulaire qui les compose. Le logiciel établit ainsi une modélisation du corpus reconstruit en classes d'énoncés présentant un profil différent.

Alceste repose sur une démarche inductive : il catégorise d'abord les réponses des sujets en fonction de leur ressemblance grâce à une classification hiérarchique descendante (CHD) et en fonction de leur opposition grâce à une analyse factorielle des correspondances (AFC). Ce n'est que dans un deuxième temps qu'il cherche à définir une correspondance avec un individu ou des groupes d'individus caractérisés par certaines variables extra-textuelles.

La démarche de *Lexico*³ est hypothético-déductive : le traitement statistique est effectué sur le corpus préalablement partitionné. Après la segmentation, il construit un dictionnaire des formes, chacune reçoit une fréquence. Il propose, ensuite, une AFC et une analyse des spécificités. La première figure la plus ou moins grande proximité (et distance) lexicale entre les parties et indique, en même temps, leur conformité au vocabulaire de base, représenté par l'intersection des axes. La seconde permet de savoir si une forme est caractéristique ou non d'une partie du corpus à partir d'un modèle probabiliste. *Lexico* distingue ainsi les formes non-spécifiques ou banales,

2 *Alceste*, sigle pour « Analyse des Lexèmes Cooccurrents dans un Ensemble de Segmentation du Texte Étudié », est développé par Max Reinert (1990 ; 1993).

3 *Lexico3* est développé par l'équipe SYLED-CLA2T sous la responsabilité d'André Salem, (Lebart & Salem 1994).

constituant le vocabulaire élémentaire, normalement distribué dans toutes les parties, des formes spécifiques qui peuvent être sur-représentées (spécifiques positives), ou sous-représentées (spécifiques négatives) dans certaines parties par rapport à un seuil de probabilité d'apparition défini.

7. *Aceste* et *Lexico* à l'épreuve des corpus

L'utilisation de deux logiciels et surtout les différentes possibilités de structurations des données qu'ils offrent sont adoptées comme dispositif participant à la variation du matériau. Les variables retenues sont : *sexe du sujet ; *lycéen / apprenti ; *rupture / non rupture de contrat ; *échec / réussite à l'examen ; *filière professionnelle et *établissement.

Alceste fournit une représentation globale du corpus qui catégorise 71% des 468 unités de contexte en trois classes.

Figure 1 : Dendrogramme des classes stables

Les calculs de tris croisés (synthétisés dans le tableau ci-dessous), et l'AFC, montrent principalement l'existence d'une spécificité du vocabulaire en fonction de la filière professionnelle, de l'établissement et de l'âge des jeunes.

Variable illustrative	Chi2	Effectif uce/classe	Effectif uce total	%
CLASSE 1				
Mécanique	8	29	49	59 %
Plomberie	3	29	56	52%
CFA M	3	20	37	54 %
Age : 17	2	30	60	50 %
Lycée pro. B	2	19	36	53 %
Réussite à l'examen	2	2	112	43 %
Lycéens	2	56	120	47 %

La variation comme principe d'exploration de corpus

CLASSE 2				
Age : 20	20	17	26	65 %
Lycée pro. B	12	17	31	55 %
Peinture	14	10	22	45 %
Maçonnerie	4	28	77	36 %
Age : 15	3	36	104	35 %
CFA A	3	14	36	30 %
CLASSE 3				
Métiers du bois	4	16	35	46 %
CFA J	4	14	30	47 %
Age : 17	3	24	60	40 %
CFA P	3	17	41	41 %
Echec à l'examen	3	3	28	38 %
Apprentis	3	72	212	34 %
Age : 18	2	9	20	45 %

Figure 2 : Synthèse du calcul des tris croisés

A ce stade, l'analyse apporte peu d'informations relatives aux questions posées par la recherche. La variable *rupture / non rupture n'est pas corrélée à une classe particulière. On repère bien que 47% des uce rédigées par les lycéens sont rassemblés dans la classe 1 alors que 34% des énoncés des apprentis est réuni dans la classe 3 mais les Chi2 associés ne montrent pas un niveau de corrélation très élevé.

Ici, la méthodologie *Alceste* laisse le chercheur relativement démuni pour espérer véritablement répondre aux questions qu'il se pose. Tributaire d'une structuration du corpus qui s'impose à lui, il ne peut plus espérer obtenir d'autres informations sur ces variables essentielles au regard de ses objectifs. En comparaison, *Lexico3* est moins contraignant du moins vis-à-vis de la relation au corpus. Il offre un système de balisage qui autorise l'introduction de plusieurs variables pour sous-catégoriser le corpus de référence. Si la base a été suffisamment renseignée, de multiples mises en série textuelles sont envisageables au cours du traitement.

Dans une première étape, nous avons formulé, l'hypothèse que les variables *sexe ; *apprenti / lycéen ; *rupture / non rupture et *échec / réussite à l'examen pouvaient

déterminer des registres discursifs. Nous obtenons l'AFC suivante :

Figure 3 : Représentation factorielle globale

Cette représentation donne lieu à différentes pistes d'interprétation. Le positionnement, par exemple, plus ou moins périphérique des sous-corpus <féminin>, <masculin>, <apprenti>, <rupture> et <échec> invite à poursuivre les mises en série textuelles pour interroger la validité de ces hypothétiques registres. *Lexico* autorise cette variation *a posteriori* du corpus en permettant au cours de l'AFC de neutraliser la participation d'une ou plusieurs partie(s). Il est possible également de modifier la partition initiale pour le calcul des spécificités en regroupant, ponctuellement, des sous-corpus en un seul ensemble de textes. On accède ainsi aux spécificités d'un nouveau sous-corpus associant les textes <échec> et <rupture> et reconfigurant l'ensemble du corpus en un corpus inédit.

Parmi les multiples essais de partition envisageables, nous pouvons opter pour la suspension des variables *sexe et *apprenti / lycéen (de couleur différente sur l'AFC). Nous

obtenons la représentation suivante qui montre une proximité textuelle entre <maintien> et <réussite> alors que les deux autres sous-corpus semblent relever de positionnements distincts.

Figure 4 : Représentation factorielle réduite

Ces différentes procédures de partition constituent autant de pistes d'« exploration textuelle » (Viprey 2006) correspondant à autant d'hypothèses de registres. En effet, les calculs statistiques de *Lexico* reposent sur le croisement des formes graphiques de parties présélectionnées par l'analyste, alors que ces parties (ou hypothèses de registres) sont une résultante automatisée de la CHD d'*Alceste*.

8. *Alceste* et *Lexico* à l'épreuve des pratiques interprétatives

Tout logiciel repose sur des procédures statistiques afférentes à des choix de son concepteur. Ces procédures doivent être comprises par les utilisateurs pour toute interprétation mais elles ne définissent pas un mode d'emploi pour l'exploitation des sorties informatiques. Dans notre collaboration, chacune a, dans un premier temps, mis en œuvre ses modes d'utilisation et

d'interprétation habituels. La confrontation n'est advenue que dans un second temps, lors de la mise en commun des résultats.

Pour le psychologue, l'analyse des bilans de savoir vise à repérer les apprentissages qui font sens pour les sujets et non pas les apprentissages effectivement réalisés.

L'étude du contenu s'effectue ici, à partir d'unités d'analyse construites sur des critères formels avec pour postulat que la distribution du vocabulaire et ses répartitions différentielles définissent des « mondes lexicaux » qui sont considérés comme des « traces » (Reinert 1990 ; 1993) de positionnements socio-idéologiques et énonciatifs différents sur l'objet discursif. L'utilisation d'*Alceste* permet d'accéder à certaines dimensions saillantes en évitant la lourdeur des méthodes de validation inter-juges. En effet, d'une part *Alceste* fonde le « dépouillement » sur un critère distributionnel ; d'autre part il effectue les calculs statistiques sur la totalité cohérente que constitue le corpus saisi comme un tout homogène. Conformément à la littérature (Charlot, Bautier, Rochex, 1992), l'analyse *Alceste* montre que la sphère relationnelle et affective occupe une place centrale pour ces élèves d'une origine sociale majoritairement populaire.

CLASSE 1

u.c.i. : 163 *num_334 *statut_1 *cfa_avey *cap_1 *sexe_1 *age_16
*rupture_2 *réussite_1 *K_1

u.c.e. : 206 Classe : 1 Khi2 : 14

le plus **important** c est le **sport**, etre toujours la pour la **famille** et les **amies**. et de faire la **fete**!

u.c.i. : 288 *num_550 *statut_2 *cfa_bay *cap_2 *sexe_2 *age_15
*rupture_2 *réussite_1 *K_1

u.c.e. : 387 Classe : 1 Khi2 : 8

J ai appris a **respecter** autrui, et a ne pas me laisser **marcher** sur les pieds. les **choses** les plus **importantes** a mes yeux sont ma **famille** et mes **amis**.

u.c.i. : 326 *num_613 *statut_2 *cfa_mont *cap_6 *sexe_2 *age_16
*rupture_2 *réussite_1 *K_1

u.c.e. : 439 Classe : 1 Khi2 : 8

J ai appris a etre **poli**, a **respecter** les autres, a aimer les autres. le plus **important** pour moi sont mes **amis**, ma **famille**, les **filles**.

La variation comme principe d'exploration de corpus

CLASSE 2

u.c.i. : 111 *num_252 *statut_1 *cfa_avey *cap_1 *sexe_1 *age_18
*rupture_2 *réussite_2 *K_2

u.c.e. : 125 Classe : 2 Khi2 : 15

J ai appris que l on etait tous egaux, dans le sens physique et religion, je ne supporte pas le **racisme**. les **gens** qui n **acceptent** pas les **personnes différentes** devraient etre rejetes.

u.c.i. : 144 *num_305 *statut_1 *cfa_avey *cap_4 *sexe_2 *age_2+
*rupture_2 *réussite_1 *K_2

u.c.e. : 177 Classe : 2 Khi2 : 15

A cela on peut rajouter le respect des autres. le pardon est aussi important. il faut **savoir** pardonner les **gens**. ce-qui est important aussi c est de ne pas juger les **personnes** sans les **connaître**, ne pas se fier aux apparences et etre tolerant.

u.c.i. : 166 *num_337 *statut_1 *cfa_avey *cap_1 *sexe_1 *age_15
*rupture_2 *réussite_1 *K_2

u.c.e. : 210 Classe : 2 Khi2 : 11

J ai appris a **vivre** en famille, en collectivite, a respecter des regles, a-savoir **parler differemment** aux **personnes** a qui je m adresse, **personnes agees, enfants**,

Figure 5 : Sélection de quelques uce caractéristiques

Les classes 1 et 2 regroupent des uce qui introduisent une dimension altruiste forte et évoquent des valeurs relatives au respect et à la politesse. Néanmoins, alors que la classe 1 met en évidence l'importance que ces jeunes accordent à leur « famille » et leurs « amis », la classe 2 désigne des acteurs sociaux ou relationnels, des « gens », des « personnes » en général. La classe 3 présente des co-occurrences plus hétérogènes qui mettent l'accent sur le thème du développement personnel. Ces jeunes insistent sur leur volonté d'arriver à faire quelque chose de leur vie, d'apprendre un métier qui leur plaît et les intéresse. Même si la dimension professionnelle est évoquée, on repère que les savoirs scolaires et techniques sont loin de constituer un centre d'intérêt prédominant : apprendre passe avant tout par la nécessité de pouvoir gérer sa vie sociale et ses aléas.

Centré sur l'étude de la référence, ce compte rendu d'analyse utilise *a minima* les différents indicateurs statistiques

destinés à décrire le système de catégorisation construit par le logiciel. Il laisse volontairement de côté l'ensemble des informations permettant d'étudier la dimension de l'énonciation qui relève d'une véritable analyse du discours pour laquelle la psychologue a besoin de l'aide du linguiste.

Lexico, comme *Alceste*, propose divers outils pour délinéariser le texte. Il n'existe pas non plus pour cet outil de procédure prédéfinie pour exploiter les sorties statistiques. Nous considérons cependant que

Les caractéristiques lexicométriques ne constituent pas un cadre rigide d'analyse (au sens où nous devrions expliquer la raison de tel ou tel classement) mais nous aident à construire des cohérences (Garric & Léglise 2005 : 107).

La position très périphérique des sous-corpus <échec> et <rupture> appelle l'identification de certains marqueurs linguistiques susceptibles de confirmer l'hypothèse de registres distincts pour ces deux groupes de locuteurs. Nous identifions des indices énonciatifs originaux : notamment la spécificité positive de « je » (<échec> +30 ; <rupture> +13), associée à celle de « il » (<échec> +7 ; <rupture> +3). La prise en compte des verbes de modalité permet de reconnaître en « il » des occurrences essentiellement impersonnelles exprimant l'obligation. Toutefois, celles-ci ne se répartissent pas uniformément : les segments répétés « il faut » et « il fallait » sont spécifiques positifs pour les locuteurs en échec, et le déontique au présent est sous-employé par les locuteurs en rupture qui sur-utilisent « il fallait » (+5).

La variation comme principe d'exploration de corpus

Figure 6 : Ventilation des marqueurs du déontique

Les locuteurs en rupture rejettent la contrainte portée par l'obligation dans un autre espace temporel qui n'appartient plus à leur actualité et qui occulte, comme le signale la fréquence du lemme (« falloir+ », l'emprise contraignante. Dans ce contexte, la spécificité (+3) du volitif à l'infinitif produit un effet discursif complémentaire. Elle modalise des procès, également à l'infinitif (« pouvoir commencer la vie » ; « pouvoir se débrouiller seul » ; « pouvoir avancer dans la vie »), présentés comme des possibles virtualisés et absolus, coupés de toute condition actancielle.

Ces sous-corpus partagent également l'occurrence spécifique de la forme pronominale « s(e) » (<échec + rupture> +5) et « m(e) » (<échec + rupture> +10). On relève les syntagmes verbaux : « s'en sortir, se débrouiller, se contrôler, se bouger, se faire respecter ». La construction pronominale est un opérateur de diathèse qui intervient dans la distribution des participants au procès. L'énonciateur est construit comme agent et patient, ces rôles pouvant être plus ou moins prégnants selon le verbe. Cette caractéristique, associée à « je », délimite un énonciateur saisi dans son individualité, voire dans sa solitude, comme en témoigne la forme « seul » (+3).

Reflet du sujet, le pronom réflexif opère en quelque sorte une clôture du procès verbal en limitant son domaine d'application à la seule sphère du sujet. Ainsi, le tour pronominal peut rapporter un procès interne, opposé au procès

externe qui caractérise le tour transitif; en effet, la présence d'un objet distinct du sujet crée deux domaines que le verbe relie. (Melis 1990 : 64)

Les élèves en échec à l'examen et, plus encore, ceux en rupture utilisent un dire égocentré qui exprime, par la clôture du procès, une coupure vis-à-vis des acteurs évoqués (famille, amis, gens, par exemple). Cette dernière reçoit deux représentations en fonction de cooccurrences divergentes : alors que « on » est sur-représenté (+5) dans le sous corpus <échec>, il est faiblement occurrent dans la classe <rupture> qui introduit un acteur indéterminé exprimé par « autres » (+3). La forme « gens » dispose d'une fréquence élevée et comparable dans les deux sous-corpus.

Figure 7 : Ventilation des marqueurs actanciels

Les locuteurs en échec construisent avec « on », une entité collective. Ils expriment des règles comportementales, auxquelles ils estiment devoir se conformer. Les sujets se font énonciateurs de bonnes résolutions qui, toutefois, en raison du déontique et de l'indétermination de « on », semblent s'exercer comme une contrainte externe. Ces règles ne sont pas prises en charge par l'énonciateur ce que confirme l'emploi de « autres » qui, sur-représentatif des locuteurs en rupture, participe à l'effet discursif de repli sur le sujet qui extériorise tout autre acteur

possible. Deux dimensions actanciennes sont dissociées : le sujet et les autres.

9. Vers une complémentarité des résultats dans l'interdiscours

Sans entrer dans le détail de l'analyse (Garric & Capdevielle-Mognibas 2006), il convient de remarquer, tout en tenant compte des contraintes génériques différentes, la proximité des résultats obtenus avec ceux de l'analyse *Lexico* des entretiens semi-directifs. L'analyse des spécificités de ce corpus, sous catégorisé selon les mêmes variables, avait permis de repérer dans le sous corpus des apprentis en rupture les marqueurs caractéristiques d'une énonciation débrayée. Nous avons construit deux réseaux de cohérence distincts quant à la nature du groupe identitaire convoqué par les sujets. Nous observions que « les apprentis sans rupture se construisent discursivement comme représentants d'un groupe dont ils partagent tout ou partie, [...], des croyances », alors que ceux en rupture ne semblent pas « parler au nom de ce groupe dont ils n'assument pas les croyances puisqu'elles sont rejetées à la responsabilité d'autres énonciateurs et dont ils refusent de se poser en terme de représentants » (Garric & Capdevielle-Mognibas 2006 : 446).

Convoquées successivement, ces configurations textuelles, offrent l'opportunité de repérer dans l'interdiscursivité de la formation discursive une forme de rapport à soi et aux autres qui témoigne d'un positionnement caractéristique du registre discursif de cette population.

10. Vers une complémentarité des logiciels et des pratiques

Les sorties d'*Alceste* et l'analyse discursive des bilans de savoir nous avaient permis d'établir des corrélations faibles entre les classes construites et les variables importantes pour le chercheur. Nous avons dénoncé la démarche inductive du logiciel, celle-ci n'autorisant pas une nouvelle structuration des données.

Pourtant, à l'issue de notre expérimentation, cette utilisation conjointe mais raisonnée des logiciels nous est

apparue méthodologiquement complémentaire. Nous pouvons constater que les différentes interprétations obtenues avec *Lexico* entrent en résonance avec la classe 3 construite par *Alceste*. Les indicateurs utilisés, parce que spécifiques, se trouvent presque tous inscrits dans les présences ou absences significatives de cette classe.

CLASSE 3

Présences significatives				Absences significatives	
Formes réduites	Khi2	Formes réduites	Khi2	Formes réduites	Khi2
debrouill+er	34	dire.	6	famille+	-22
seul+	29	confi+ant	6	import+ant	-11
pouvoir+	23	apprenti<	6	chose+	-10
ger+er	18	entreprise	6	vivre.	-8
se	16	compte	6	respect+er	-7
gagn+er	16	toute	6	plein+tous	-7
faire.	15	vie+	5	gens	-6
trouv+er	14	avoir	5	différ+ent	-6
metier+	14	on	5	sœur	-5
argent	14	qu+	5	ai	-5
arriv+er	12	ils	4	ami+	-5
plaire.	11	tard	4	moment+	-5
me	10	savoir.	4	parent+	-5
vouloir.	8	personne	4	quartier+	-4
etude+	8	nous	2	aid+er	-4
facile+	7	son	2	petit+	-4
falloir.	7	si	2	moi	-4
futur+	6	et	2		
amus+er	6				

La découverte, dans l'après-coup, des classes d'*Alceste* confirme les cohérences produites à l'aide des spécificités de *Lexico*. Elle fait aussi apparaître de nouveaux marqueurs que le chercheur n'avait pas initialement pris en compte. On interprètera, par exemple, très différemment la présence significative des formes réduites « debrouill+er ; ger+er ; trouv+er ; arriv+er » de la classe 3. Toutes, dans l'unité textuelle, peuvent être glosées ainsi : *s'extraire d'une situation sans compter sur quiconque*, ce que confirme l'absence significative de « aid+er ». On découvre également avec intérêt

la présence de segments répétés tels que « me débrouill+er », « il falloir », « respect+ autre+ », « je vouloir ».

On repère ici que l'étude des classes construites par *Alceste* qu'elle soit préalable ou postérieure à l'analyse des spécificités peut orienter le regard de l'analyste vers l'identification de cohérences. En revanche, du fait de la lemmatisation et de l'occultation des mots outils, elle ne peut suffire à elle seule à construire les indicateurs nécessaires à l'interprétation des données. Si l'analyse des classes d'*Alceste*, informée par un retour aux textes et par les différents outils, a favorisé l'identification de différentes sphères relationnelles à relier avec des registres, l'analyse des sorties de *Lexico* révèle que les locuteurs en rupture sont précisément en rupture avec les acteurs auxquels ils s'adressent.

11. Conclusion

Lors des premiers résultats interprétatifs, la psychologue a été séduite et convaincue par les analyses proposées par la linguiste au point de quelque peu dévaloriser ses propres résultats et la procédure *Alceste*. De son côté, la linguiste, par l'expérimentation et la collaboration, a pu formuler, ses interprétations d'un lieu de savoir indispensable, la psychologie. Ces constats interrogent sur la réelle valeur ajoutée de notre collaboration, puisque la plupart des observables linguistiques permettant d'étayer les interprétations avaient été produits par la linguiste.

Pourtant, progressivement, malgré (ou grâce à) de nombreux questionnements méthodologiques et théoriques, les conditions d'un véritable cadre herméneutique, indissociable d'une certaine forme d'interdisciplinarité qui adopte la variation comme principe systématisé, se sont imposées. Ce cadre doit, d'abord, s'inscrire dans une réflexion commune, distincte d'un simple partage ou d'une addition de compétences, qui problématise le texte, le discours et le sens comme objet unifié des Sciences Humaines et Sociales. Il doit, ensuite, poser l'analyse des formes langagières au centre de l'activité interprétative : les hypothèses de lecture sont conditionnées par la description des observables qui, dans une dynamique entre

les fonctionnements textuel, discursif et social délimite des réseaux de cohérence socio-discursifs.

Références bibliographiques

- Achard P. *et alii.* (1996). *Rapport Slade* (Sociologie du Langage, Analyse de Discours, Énonciation). Présenté au CNRS.
- Achard P. (1995). « Registre discursif et énonciation : induction sociologique à partir des marques de personne. Le Congrès des députés du peuples d'URSS en 1989 », *Langage et société* 71 : 5-34.
- Adam J-M. & Heidmann U. (éd.) (2005). *Sciences du texte et analyse de discours. Enjeux d'une interdisciplinarité*. Genève : Slatkine Érudition.
- Adam J-M. (1999). *Linguistique textuelle. Des genres de discours aux textes*. Paris : Nathan.
- Bakhtine M. & Volochinov V-N. (1984). *Esthétique de la création verbale*. Paris : Gallimard.
- Boutet F., Fiala P. & Simonin J. (1976). « Sociolinguistique ou sociologie du langage », *Critique* 344 : 68-85.
- Capdevielle-Mougnibas V., de Léonardis M., Garric N., Préteur Y., Rossi-Neves P., Rossi-Gensane, N. Besses M-O & Courtinat A. (2008). Etude des effets des dispositifs pédagogiques de l'enseignement professionnel sur les savoirs, les compétences et les représentations des apprenants. Genèse des ruptures dans les parcours de formation des lycéens professionnels et des apprentis de niveau V. (Rapport terminal. ACI PIREF « Education et formation : contextes et effets». PROJET EF 0020). Toulouse.
- Charlot B. Bautier E. & Rochex Y. (1992). *École et savoir dans les banlieues et ailleurs*. Paris : Armand Colin.
- Foucault M. (1969). *L'archéologie du savoir*. Paris : Gallimard.
- Garric N., Capdevielle-Mougnibas V. & Besses M-O. (2006). « Intérêts et limites de l'analyse du discours pour la recherche interdisciplinaire et la coproduction de

- connaissances scientifiques. Le cas d'une analyse lexicométrique d'entretiens avec Lexico3 », in *Actes des 8^{ème} JADT*. Besançon : Presses Universitaires de Franche-Comté, pp. 439-450.
- Garric N. & Léglise I. (2005). « La place du corpus, de l'analyste, du logiciel », in G. Williams (éd.), *La linguistique de corpus*. Rennes : Presses Universitaires de Rennes, pp. 101-114.
- Léglise I., Canut E., Desmet I. & Garric N. (2006). *Applications et implications en sciences du langage*. Paris : L'Harmattan.
- Lethier V. & Viprey J-M. (2008). « Lecture de l'archive et linguistique de corpus : presse écrite du XIX^è siècle », *Texte et corpus* 3 : 285-306
<http://web.univ-ubs.fr/corpus/jlc3/jlc2003.pdf>
- Melis L. (1990). *La voie pronominale : la systématique des tours pronominaux en français moderne*. Paris : Louvain-la-neuve : Duculot.
- Malrieu, P. (1973). « La socialisation », in H. Gratiot-Alphandéry & R. Zazzo (éds), *Traité de psychologie de l'enfant*, T 5 : La formation de la personnalité. Paris : PUF, pp. 5-234.
- Pêcheux M. (1990). *L'inquiétude du discours. Textes choisis et présentés par D. Maldidier*. Paris : Éditions des Cendres.
- Pêcheux M. (1969). *Analyse automatique du discours*. Paris : Dunod.
- Rastier F. (2005). « Enjeux épistémologiques de la linguistique de corpus », in G. Williams (éd.), *La linguistique de corpus*. Rennes : Presses Universitaires de Rennes, pp. 31-46.
- Rastier F. (2001). *Arts et Sciences du texte*. Paris : Presses Universitaires de France.
- 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.
- Reinert M. (1990). « Alceste. Une méthodologie d'analyse des données textuelles et une application : Aurélia de Gérard

N. GARRIC, V. CAPDEVIELLE- MOUGNIBAS

- de Nerval », *Bulletin de méthodologie sociologique* 26 : 24-54.
- Salem A. (1986). « Segments répétés et analyse statistique des données textuelles », *Histoire et mesure* 1-2 : 5-28.
- Viprey J-M. (2006). « Quelle place pour les sciences des textes dans l'Analyse de discours ? », *Semen* 21 : 167-182.
- Viprey J-M. (2005). « Philologie numérique et herméneutique intégrative », in J-M. Adam & U. Heidmann (éds), *Sciences du texte et analyse de discours. Enjeux d'une interdisciplinarité*. Genève : Slatkine Érudition, pp. 51-68.
- Vygotsky L.S. (1997³). *Pensée et Langage*. Paris : La dispute.