Logométrie et modélisation des interactions discursives. L'exemple des entretiens semi-directifs

Julien Bonneau [1], Anne Dister [2]

[1] Laboratoire BCL – Université Nice Sophia-Antipolis – CNRS UMR 6039 – MSH de Nice – 98 bd E. Herriot – 06200 NICE – France

[2] FUSL et Université de Louvain – 43, Bd du Jardin Botanique – 1000 Bruxelles– Belgique

Résumé

Cet article présente les résultats d'une analyse logométrique sur des entretiens semi-directifs. La problématique tient à la comparabilité des entretiens au travers des interactions et des influences réciproques mises en œuvre entre intervieweurs et interviewés. Nous répondrons notamment aux questions suivantes : Peut-on faire émerger une classe intervieweurs et une classe interviewés ? Un même intervieweur a-t-il les mêmes mises en œuvre de conduite d'entretiens ? Peut-on construire des classes de conduite d'entretiens ? Quelles sont les particularités externes (sociales) et internes (linguistiques) des classes ? Y a-t-il des classes d'interviewés ? Si oui, existe-t-il une corrélation entre les différents profils construits ?

Abstract

This article presents the results of a logometrical analysis of semi-directive interviews. We wonder about our capacity to compare interviews in the point of view of interactions and reciprocal influences between interviewers and interviewees. In particular, we will answer to these following questions: Are there interviewers and interviewees statistical significant classes? For a given interviewer, is there only one way to carry out interviews? Can we find carrying out interviews classes? What are the externals (socials) and internals (linguistics) classes' particularities? Are there interviewees' classes? If so, can we find correlations between the different constructed profiles?

Keywords: logometry, discursive interaction, linguistic insecurity, Alceste, Lexico

1. Introduction

Au-delà d'une lexicométrie traditionnelle (Lebart and Salem, 1994), la logométrie prétend se donner tous les moyens statistiques d'accéder aux textes (Mayaffre, 2005), acceptant toutes ses dimensions et, par là, toutes les unités linguistico-statistiques cohérentes nécessaires à leur description. Ce positionnement est d'autant plus nécessaire quand l'objet d'étude nous est méconnu ou étranger. C'est notamment le cas des textes oraux dont les traitements au travers des unités lexicométriques classiques (forme, lemme, catégorie grammaticale) ne sont pas toujours pleinement satisfaisants. En effet, les énoncés oraux contiennent un certain nombre de particularités liées principalement au mode de production de l'oral : les énoncés ne sont pas planifiés, et le locuteur fait d'incessants aller-retours sur l'axe syntagmatique. Dans le discours en train de se construire sont présentes un certain nombre de marques typiques des productions orales (répétitions, mots amorcés, etc.) et qui sont souvent regroupées sous le terme de « disfluences » (cf. point 2).

Ces disfluences permettent la construction de nouveaux observables logométriques. L'observation des unités statistiques ainsi construites s'appuie sur la méthodologie décrite dans Bonneau (2008) : définition d'une partition induite du corpus ; articulation et description du corpus selon un axe intervieweur/interviewés transversal à l'ensemble des parties du corpus, propre à faire valoir les interactions mises en jeu.

Dans cet article, nous utiliserons les logiciels Alceste (Reinert, 2002) et Lexico (Salem et al., 2003) afin d'étudier des données textuelles orales transcrites, en particulier pour l'analyse des phénomènes de disfluences et des marques de résonnance thématique présentes dans le corpus. Nous porterons ainsi un double regard : sur les rapports entre interaction et influence discursive dans le corpus ; sur la capacité des disfluences à enrichir cette description.

2. Les données

2.1. Enquête sur l'insécurité linguistique

Les données sur lesquelles nous travaillons sont constituées d'entretiens semi-directifs issus d'une vaste enquête sociolinguistique sur l'insécurité linguistique des Belges en Communauté française de Belgique (Francard et al., 1993). L'insécurité linguistique, largement étudiée dans les communautés périphériques (qu'on pense aux nombreuses études sur le sujet réalisées en Belgique ou au Québec), se rencontre chez les locuteurs qui, d'une part, ont conscience qu'il existe une variété de langue légitime et qui, d'autre part, estiment ne pas maitriser cette variété légitime. Dans ces cas, comme le dit Bourdieu (1979), il y a « reconnaissance sans connaissance ». Par contre, la sécurité linguistique se manifeste dans deux cas : lorsque le locuteur évalue sa pratique comme confirme à la norme, ou lorsque qu'il n'a pas conscience de l'existence d'une norme et des écarts entre sa propre pratique et cette norme.

Ainsi, lors des entretiens, les personnes interviewées ont été amenées à se prononcer sur la perception qu'elles ont de l'accent, sur le fait de savoir dans quel pays on parle le mieux le français, sur la « qualité » du français des médias, etc. Toutes ces questions avaient pour but de faire émerger les attitudes et représentations linguistiques des belges francophones, différenciés selon certaines catégories socioprofessionnelles (cf. 2.4.).

2.2. La transcription des données orales

Les enregistrements de cette enquête ont été transcrits selon des conventions explicites (Dister et al., 2006) et sont consignés dans la banque de données textuelles orales VALIBEL [1]. Ces transcriptions orthographiques non ponctuées (voir Blanche-Benveniste and Jeanjean, 1987, pour une justification de ces choix) consignent les « particularités » des données orales, en accordant une attention particulière à la transcription des phénomènes que la littérature regroupe classiquement sous le terme de *disfluences* : *euh*, ponctuants, répétitions, amorces de morphèmes, interruptions, autocorrections, chevauchements de parole, etc.

2.3. Les disfluences

Certains auteurs (Benzitoun et al., 2004, par exemple) réfutent le terme de *disfluence*, qui implique notamment une comparaison implicite avec l'écrit : si l'oral est disfluent, ce serait par rapport à un écrit qui lui, évidemment, ne le serait pas. La disfluence de certains énoncés est

[1] http://www.uclouvain.be/valibel-corpus.html.

donc opposée à la fluence de certains autres, avec les jugements de valeur que cela implique bien souvent. Comme le constate Habert (2005 : 57) :

> (…) on manque ainsi de termes positifs pour décrire les régulations de l'oral, parfois fâcheusement dénommées *disfluences* par transfert de l'anglais *disfluencies.*

Nous conservons néanmoins ce terme, partagé par la communauté scientifique (cf. les colloques DISS), en attirant l'attention du lecteur sur le fait qu'il n'est associé pour nous à aucun jugement de valeur.

Les disfluences brisent la linéarité des énoncés et sont, de ce fait, un point d'achoppement majeur pour les analyses automatisées des données textuelles orales. Dans l'exemple suivant, la locutrice corrige son énoncé, en répétant une construction amorcée (*qui s'en/*) et en la complétant (*qui s'enfonce*)

> ilePA2 or une trémie euh grammaticalement c'est une chose qui s'en/ qui s'enfonce plutôt dans la terre.

On voit bien l'aller-retour sur l'axe syntagmatique dont nous parlions, avec un entassement sur une même place syntaxique, qu'illustre clairement la notation suivante :

> ilePA2 or une trémie euh grammaticalement c'est une chose qui s'en/
> qui s'enfonce plutôt dans la terre.

Nous avons développé un système (Dister et al., in press) qui balise de manière totalement automatisée les disfluences pour ne garder dans le texte que la réparation de l'énoncé (le *reparandum,* pour reprendre la terminologie de Shriberg, 1994). L'exemple ci-dessus devient alors :

> ilePA2 or une trémie {euh,.IGN+Euh} grammaticalement c'est une chose {<r> qui s'en/ </r>,. IGN+Rep} qui s'enfonce plutôt dans la terre-

C'est sur des données ainsi balisées [2] que nous nous ferons l'analyse de 3 marques de disfluence (section 4), la répétition, l'autocorrection immédiate et l'amorce de morphème, que nous présentons brièvement ici.

Nous entendons par **répétition** la reprise à l'identique, dans le contexte direct, d'un mot ou d'un groupe de mots, comme c'est le cas de *sans* et de *la* dans l'exemple suivant :

> **ilrMS1** je sais pas / parler sans accent pour moi c'est sans // sans // sans bafouiller sans / sans sans se tromper de mots quoi sans sans sans que la la langue fourche quoi [ilrMS1r].

Les **autocorrections immédiates** (Dister, 2008) constituent une variante de la répétition. Dans les autocorrections, l'un des traits morphologiques de l'élément répété varie, comme l'illustre l'exemple suivant où le déterminant défini est répété, en changeant le trait 'singulier' par le trait 'pluriel' :

> **ileFN1** et le journalisme et puis euh le les études de journalisme en soi ne me plaisaient pas [ileFN1r].

Nous appelons **amorce** le phénomène langagier qui consiste en « une interruption de morphèmes en cours d'énonciation » (Pallaud, 2002 : 79). L'exemple suivant est un cas typique d'amorce. Le morphème interrompu – symbolisé par une barre oblique collée directement à la droite de celui-ci, au lieu de l'interruption – est corrigé plus loin dans l'énoncé, où il est repris sous sa forme pleine [3] :

> **ilrPC1** (…) j'aimerais bien moi ouvrir un **ma/** un petit magasin (…) [ilrPC1r].

---

[2] Nous renvoyons le lecteur à Dister et al. (in press) pour le détail des balises.

[3] Notons que l'exemple que nous donnions ci-dessus, balisé comme une répétition ({<r> qui s'en/ </r>,. IGN+Rep}), présente aussi un phénomène d'amorce.

2.4. Le corpus et les sous-corpus

Les différents textes qui composent notre corpus de travail sont issus d'entretiens semi-directifs consacrés à l'insécurité linguistique. Chaque entretien met en présence un intervieweur et un interviewé. Un questionnaire écrit sert d'amorce à chacun des entretiens, suivi d'échanges libres, où la conduite d'entretien de l'intervieweur se doit d'être la plus neutre possible. La population interviewées se répartie en 5 profils professionnels différents, dont voici la répartition :

- des journalistes de la presse écrite (ile [4]) : 5 textes ;
- des journalistes de l'audiovisuel (ilj) :6 textes ;
- des cadres d'entreprise (ilc) : 5 textes ;
- des hommes politiques (ilp) : 7 textes ;
- des étudiants en dernière année de l'enseignement secondaire (ilr) : 11 textes.

Ces données totalisent plus de 255.170 occurrences (9.248 formes) et correspondent approximativement à 20 heures de parole.

De ces données initiales ont été extraits différents corpus de contenus et de formats différents (cf. section 3.7) selon les hypothèses mises à l'épreuve de nos traitements et les logiciels utilisés pour nos analyses. Pour permettre l'accès à la production textuelle de chacun des locuteurs, les tours de paroles de ceux-ci ont été décroisés et individualisés. Les caractères utilisés pour la description externe des sous-parties ainsi constituées sont les suivants : le locuteur est-il un intervieweur ou un interviewé ? Quel est le milieu socioprofessionnel du locuteur ? Quel individu parle ? Quels sont les individus en présence durant une même interview (interaction) ? Le corpus est donc segmenté en sous-partie de manière à obtenir une réponse univoque à chacune de ces questions.

Le corpus textuel étudié comporte *in fine* des observables de nature hétérogène : d'une part les occurrences de mots, d'autre part les occurrences de disfluences. Dans nos analyses, nous traitons l'ensemble de ces unités comme des objets statistiques équivalents, considérant la disfluence comme un signe linguistique au même titre que le mot et traitant donc celle-ci comme une forme de plus de l'oralité en supplément des formes lexicales. Une occurrence de mots ou de disfluences possédera donc le même poids statistique. Nos données textuelles orales correspondent donc à un corpus d'occurrences textuelles « classiques » enrichit de formes nouvelles, les disfluences, décrites par un codage occurrenciel univoque pour chaque forme (et donc hors du vocabulaire « classique » du corpus), l'ensemble des formes des corpus ainsi obtenus étant considérées comme des unités statistico-linguistiques de même niveau.

3. Partition des données et premiers résultats

Nous réalisons une première analyse dont le but est de construire une ou des partition(s) endogène(s) du corpus et de déterminer les caractères externes pertinents pour décrire cette/ces partition(s). A cette fin, nous utilisons le logiciel Alceste (Reinert, 2002). Celui-ci peut être décrit comme un traitement des co-occurrences généralisé d'un corpus multiparamétré. L'intégralité du texte est lemmatisée puis découpée en segments de longueur comparable, appelés « unités de contexte élémentaire » (uce). Ces segments sont ensuite comparés à l'aide d'une classification hiérarchique descendante (CDH), le traitement portant sur la distance lexicale de ceux-ci selon la distance du Ki2, en fonction de la présence ou de l'absence des occurrences qui les constituent.

[4] Ce code de 3 lettres est le nom du sous-corpus initial dont ont été tirés les différents textes qui ont servi à notre étude : *ile* pour *insécurité linguistique de la presse écrite*, etc.

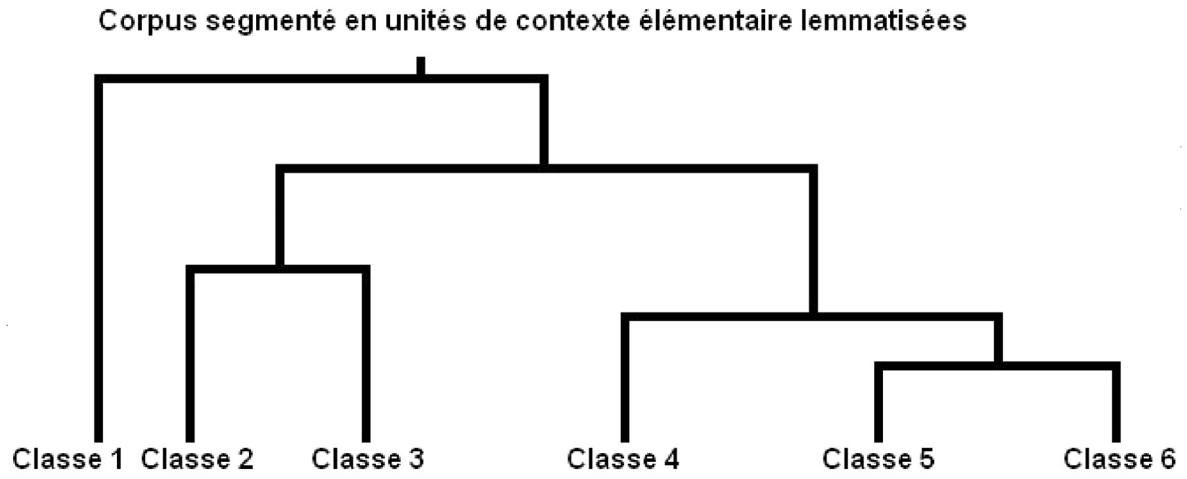
Les classes construites (*Graphique 1*) par cette méthode correspondent à des profils co-occurrenciels convergents, des régularités, associées à des univers lexicaux communs. Alceste projette ensuite sur ces classes les divers caractères utilisés pour la description externe des parties du corpus. La granularité des partitions que nous proposerons *in fine* doit pouvoir rendre compte des différents univers lexicaux pointés par la CDH d'Alceste s'ils sont caractérisés par des descriptions externes, des caractères qui les spécifient par rapport au reste du corpus.

De ces résultats émergent des indices de partitions monoparamétrées cohérentes pour la constitution de corpus. Le logiciel Lexico ne permettant pas l'accès synchronique aux caractères externes des sous-parties des corpus, cette première étape est nécessaire pour confirmer la validité statistique du partitionnement en divers corpus d'études proposé à celui-ci.

convergents, des régularités associées à des univers lexicaux communs. Alceste projette ensuite sur ces classes les divers caractères utilisés pour la description externe des parties du corpus.

De ces résultats émergent des indices de partitions monoparamétrées cohérents pour la constitution de corpus. Le logiciel Lexico ne permettant pas l'accès synchronique aux caractères externes des sous-parties des corpus, cette première étape est nécessaire pour confirmer la validité statistique du partitionnement en divers corpus d'études proposé à celui-ci.

Nous détaillons dans la suite les résultats de la classification hiérarchique réalisée par Alceste sur l'ensemble du corpus et des descriptions externes. Le *graphique 1* présente l'arbre construit par ce traitement.



Graphique 1: Classification hiérarchique Alceste

Nous avons étudié les 25 premières unités textuelles (lemmatisées), disfluences et caractères externes les mieux corrélés à ces classes selon la distance du Ki2. Le rang de chaque unité (ou caractère) correspond à sa position dans un classement par ordonnancement décroissant de Ki2 dans les classes hiérarchiques construites. On a donc observé les rangs 1 à 25 de chaque classe, le rang 1 étant associé à l'unité textuelle ou au caractère externe le mieux corrélé à celle-ci. Nous détaillons maintenant les caractères externes fortement corrélés à ces classes et donc descriptifs de celles-ci.

3.1. Classe 1 (Rang(s) du/des caractère(s) observé(s) ; Ki2 respectif(s) associé(s))

- Le groupe socioprofessionnel des étudiants (1 ; 939) ;
- le groupe socioprofessionnel intervieweur d'étudiant + l'identité de l'unique intervieweur d'étudiant (3, 4 ; 192) ;
- l'interaction des différents étudiants et de l'intervieweur d'étudiants (5, 9, 15, 17 ; 191, 140, 105, 100) ;
- l'identité des étudiants (6, 13, 19, 22 ; 169, 116, 96, 84).

Cette partie du corpus correspond donc à l'univers discursif spécifique constitué par les locuteurs étudiants et leur intervieweur. Du fait de possibles phénomènes de réappropriation thématique (Bonneau, 2008), c'est à dire de sélection lexicale d'au moins un locuteur dans le discours de l'autre, on peut faire l'hypothèse, à la constitution d'une telle classe, d'une influence, réciproque ou non, entre l'intervieweur et les interviewés.

Cette classe est, de plus, caractérisée par le « tu » (Ki2 156), des auxiliaires verbaux « être », « avoir », « aller » et « faire » et le vocabulaire attaché à la thématique du loisir : « aimer », « année », « vacance ».

3.2. Classe 2 (Rang(s) du/des caractère(s) observé(s) ; Ki2 respectif(s) associé(s))

- Le groupe socioprofessionnel journalistes de presse écrite (1 ; 292) ;
- 2 identité de journalistes de presse écrite (9, 21 ; 118, 69) ;
- 3 interaction de journalistes de presse écrite et de leur intervieweur (12, 17, 25 ; 109, 91, 50).

Comme pour la classe 1, cette classe porte la marque de l'interaction entre intervieweur et journalistes de presse écrite. Le fait que l'identité de l'intervieweur de journaliste n'apparaisse pas dans ce classement montre néanmoins l'émergence de thématiques spécifiques aux interviewés.

Le vocabulaire caractéristique de cette classe, selon la distance du Ki2, est « écrit », « phrase », « orthographe », « écrire », « lecteur », « vocabulaire », « langage », « presse », « journal », « article », « média », « écriture », « information ». La co-occurrence de ce vocabulaire évoque la sémantique du langage et des médias. De plus, on voit apparaitre au rang 26 l'une des marques de disfluence les plus fréquentes en corpus (Dister, 2007) : la répétition.

3.3. Classe 3 (Rang(s) du/des caractère(s) observé(s) ; Ki2 respectif(s) associé(s))

- Le groupe socioprofessionnel des cadres (3 ; 149) ;
- une identité de cadre (4 ; 114) ;
- l'interaction entre un cadre et son intervieweur (5 ; 100) ;
- le groupe socioprofessionnel des hommes politiques (11 ; 71) ;
- l'interaction entre un homme politique et son intervieweur (16 ; 60) ;
- l'identité d'un homme politique (21 ; 53) ;
- l'interaction entre un journaliste de l'audiovisuel et de son intervieweur (23 ; 50) ;
- un journaliste de l'audiovisuel (25 ; 48).

Classe aux caractères hétérogènes, la classe 3 est associée aux groupes socioprofessionnels cadres et hommes politiques. Elle porte de plus les marques de possibles interactions entre intervieweurs et interviewés.

Le vocabulaire associé à cette classe est celui de la communication (« social », « exprimer ») et du travail (« profession », « formation », « métier »).

3.4. Classe 4 (Rang(s) du/des caractère(s) observé(s) ; Ki2 respectif(s) associé(s))

- Le groupe des interviewés (22 ; 38).

Le vocabulaire associé à cette classe est le suivant : « accent », « liégeois », « bruxellois », « entendre », « parisien », « terroir », « marseillais », « connotation », « voix », « juge », « supposer ». Ces co-occurrences évoquent les particularités régionales, géographiques des langues et leur évaluation par les locuteurs.

3.5. Classe 5 (Rang(s) du/des caractère(s) observé(s) ; Ki2 respectif(s) associé(s))

- Le groupe des interviewés (25 ; 30).

Les classes 4 et 5 correspondent à des profils co-occurenciels distincts spécifiques aux interviewés.

Le vocabulaire associé à la classe 5 rappelle fortement celui associé à la classe 4, avec un centrage sur la géographie linguistique belge ; il diffère par des références individuelles, familiales : « wallon », « français », « parler », « dialecte », « flamand », « wallonne », « grand-père », « vieille », « rural ». Cette classe est, de plus, caractérisée par la marque du silence (rang 25, Ki2 31) et par la particule de négation « ne », dont on connait par ailleurs la tendance très forte à ne pas être présente dans les productions orales ordinaires (Berit Hansen and Malderez, 2004).

3.6. Classe 6 (Rang(s) du/des caractère(s) observé(s) ; Ki2 respectif(s) associé(s))

- Le groupe des intervieweurs (3 ; 542) ;
- le groupe socioprofessionnel intervieweur de cadres + l'identité de l'unique intervieweur de cadres (9, 10 ; 273) ;
- le groupe socioprofessionnel intervieweur d'hommes politiques (23 ; 126).

La classe 6 correspond au vocabulaire co-occurrenciel spécifique aux intervieweurs, en particulier envers les cadres et les hommes politiques.

Le vocabulaire associé à cette classe est « belgique », « pays », « ailleurs », « francophone », « france », « région », « suisse », « français », « parler », « canada », « expressif ». Le profil co-occurrenciel de cette classe met en avant l'aspect géographique francophone en termes généraux. Apparaissent aussi des termes liés spécifiquement à la fonction des intervieweurs qui ont comme objectif explicite de poser des questions aux locuteurs [5]. On a ainsi les termes « opinion », « vous », « votre » et la particule discursive « mm », qui a clairement pour fonction, dans les entretiens semi-directifs, d'encourager l'interlocuteur à poursuivre (rang 7, Ki2 342).

3.7. Synthèse pour la constitution de corpus monoparamétrés

L'articulation globale des classes constituées et de leurs caractères permet d'ores et déjà d'affiner les arguments statistiques pour un découpage du corpus et, au-delà, nos hypothèses de travail : on a bien une dichotomie entre intervieweurs et interviewés, mais on voit aussi apparaitre l'influence forte du caractère socioprofessionnel. En parallèle, des différences de positionnement semblent apparaitre de la part des intervieweurs, les entretiens les plus interactifs étant *in fine* ceux qui n'apparaissent pas corrélés avec leur discours le plus spécifique. C'est là la marque d'une forte présomption de différences dans l'interaction mise en jeu et la direction d'entretiens par l'intervieweur, notamment mis en valeur par des phénomènes de réappropriations thématiques. On a des profils co-occurenciels soit très proches dans l'interactivité – et donc peu

[5] Rappelons que dans le cas de nos données, les intervieweurs mènent une enquête sociolinguistique et se positionnent explicitement comme des enquêteurs.

d'apports thématiques nouveaux –, soit des profils co-occurrenciels proches du discours de l'intervieweur, ce qui marque des influences directionnelles attendues, mais aussi des apports nouveaux de la part des interviewés.

Au vu de ces résultats et pour observer plus finement les marques d'influences et d'interactions dans le corpus, nous avons retenu trois strates sociolinguistiques différentes : intervieweur/interviewé, socioprofessionnelle et identité de l'intervieweur. Nous poursuivons maintenant nos analyses à l'aide du logiciel Lexico (Salem et al., 2003) sur les sous-corpus suivants :

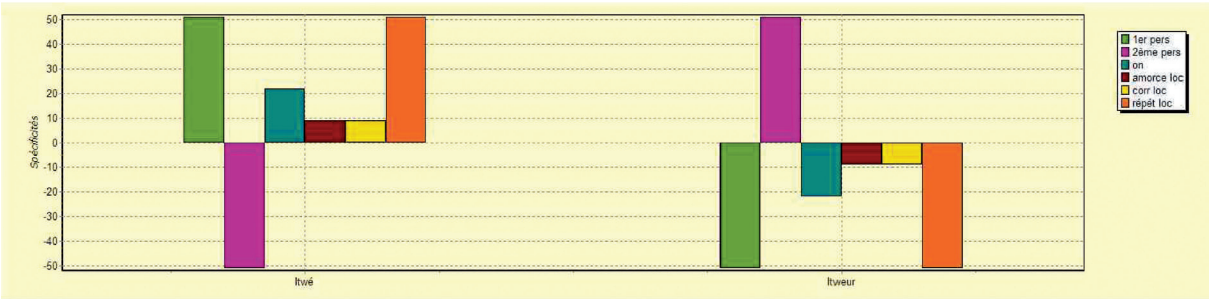
a) un corpus complet partitionné selon la dichotomie intervieweurs et interviewés, sans différenciation d'individus (255 170 occurrences, 9 248 formes) ;
b) un sous-corpus partitionné en fonction du profil socioprofessionnel de l'interviewé et contenant en parallèle (en alterné) d'une part la production langagière des interviewés et d'autre part celle des intervieweurs ;
c) un corpus alterné de la production langagière des intervieweurs différenciés par leur identité et de celle de leurs interviewés différenciés par leur profil socioprofessionnel.

4. Mise en évidence des interactions

Dans la suite, nous ne prolongerons plus le travail de description des interactions réalisé pour la construction du partitionnement du corpus, mais nous nous concentrerons sur des phénomènes morphosyntaxiques plus en rapport avec les particularités de l'oralité, à savoir, d'une part l'usage des pronoms personnels comme marqueurs de l'implication des locuteurs (Bonneau, 2008 ; Biber, 1993) et, d'autre part, l'analyse de 3 marques de disfluence : l'autocorrection immédiate, la répétition et l'amorce de morphème. Ce travail s'appuie sur l'analyse des spécificités de ces unités linguistiques dans les différents sous-corpus.

4.1. Analyse logométrique des pronoms personnels

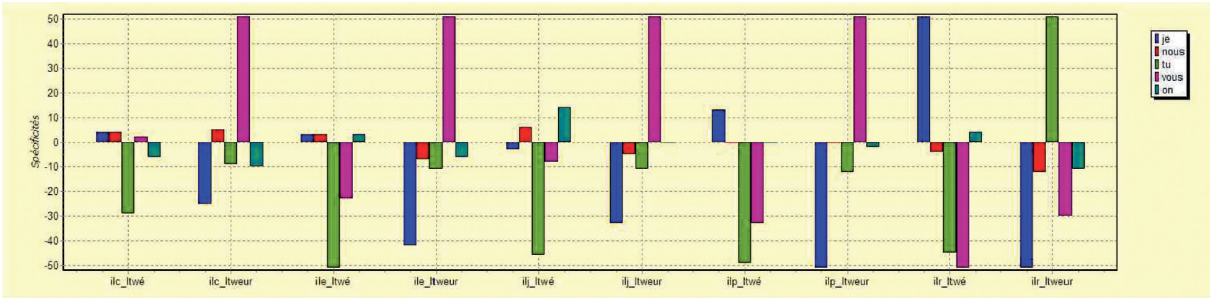
Le *Graphique 2* montre un sur-emploi de la première personne et du « on » (respectivement aux 1re et 3e colonnes) chez les interviewés et un sur-emploi de la seconde personne chez les intervieweurs (2e colonne), avec les sous-emplois réciproques qu'ils impliquent. C'est un résultat attendu, déjà décrit dans Bonneau (2008), associé à la forme des entretiens semi-directs : un questionnement des intervieweurs et des réponses des interviewés à la première personne.



Graphique 2 : Spécificités des pronoms personnels et des disfluences sur la partition intervieweurs / interviewés

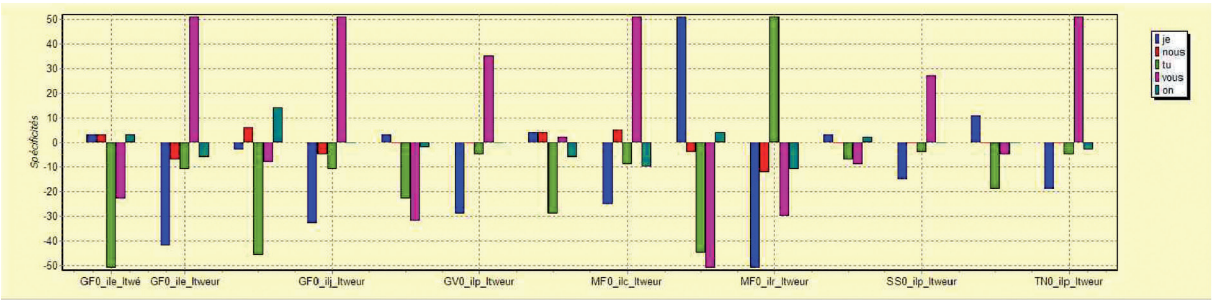
Pour aller plus loin, nous nous appuyons sur le *Graphique 3* dans lequel la catégorie socioprofessionnelle des interviewés est différenciée et où nous avons affiné la description

des pronoms personnels en intégrant le nombre en plus de la personne. On s'aperçoit à ce niveau de description de particularités de certains entretiens. En effet, la description proposée précédemment ne s'applique en fait qu'avec les interviews d'étudiants, particularisés par le « tu » (3e colonne) aux dépends du « vous » (4e colonne) chez l'intervieweur et le « je » (1re colonne), ainsi que le « on » (5e colonne) chez les étudiants. Le « tu » de l'intervieweur semble donc appeler le « je » des interviewés, effet que ne provoque pas le « vous » utilisé dans les autres groupes socioprofessionnels. À ce niveau, la conduite d'entretiens semble néanmoins homogène si l'on exclut la question du nombre et les résultats des autres groupes sont relativement comparables. Sur notre échantillon, il semble donc que c'est une particularité de la situation intervieweur institutionnel/étudiant que de produire ce type de différenciations.



Graphique 3 : Spécificités des pronoms personnels sur la partition catégorie socioprofessionnelle des interviewés - intervieweurs / interviewés (alternés)

Dans le *Graphique 4* ci-dessus, nous nous intéressons enfin à l'identité des intervieweurs. À ce stade, nous voyons nettement apparaitre le phénomène attendu : l'intervieweuse MF0, qui est en fait l'unique intervieweuse à avoir rencontré des étudiants, a bien une conduite d'entretien très différente envers eux par rapport à l'autre groupe socioprofessionnels auprès duquel elle mène son enquête, à savoir les cadres. On pourrait émettre l'hypothèse que ces résultats sont dus au fait que l'intervieweuse est relativement jeune (23 ans) et qu'elle s'adresserait ainsi, en quelque sorte, à des pairs. S'il est vrai qu'en termes d'âge, elle est plus proche des étudiants du secondaire que du groupe des cadres, elle ne peut en aucun cas apparaitre comme l'une d'entre eux : sa position est clairement celle d'une chercheuse universitaire, détentrice du savoir, et à aucun moment les étudiants ne s'autorisent à être familiers avec elle.



Graphique 4 : Spécificités des pronoms personnels sur la partition identité de l'intervieweur – catégorie socioprofessionnelle des interviewés - intervieweurs / interviewés (alternés)

4.2. Analyse logométrique des disfluences

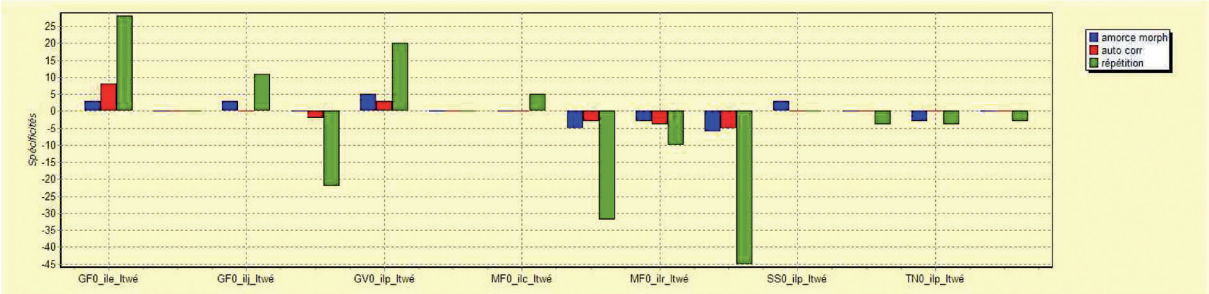
Le *Graphique 2* présentait une forte sur-représentation de la répétition chez les interviewés par rapport aux intervieweurs et, d'une manière générale, une plus forte propension aux autres marques de disfluence que sont l'amorce de morphème et l'autocorrection.

L'observation à un niveau sociodescriptif plus fin des données, en ajoutant la variable groupe socioprofessionnel de interviewé, avait modifié notre lecture des résultats sur l'étude des pronoms personnels (cf. *supra*). Nous renouvelons l'expérience en limitant notre étude aux 3 marques de disfluences suivantes : les amorces de morphème (col. 1), les auto-corrections (col. 2) et les répétitions (col. 3).

À l'exception du groupe des étudiants, le *Graphique 5* décrit des tendances d'interaction similaires dans les entretiens, avec des traits sur- à moyennement représentés chez les interviewés et sous- à moyennement représentés chez les intervieweurs, même si les écarts entre spécificités des interviewés et des intervieweurs peuvent varier selon le groupe socioprofessionnel étudié. Ceci est différent pour les entretiens avec les étudiants où les spécificités des répétitions sont remarquables chez les étudiants et très fortes chez les intervieweurs. Ce groupe socioprofessionnel, déjà à part lors des analyses des pronoms personnels, semble subir ici (aussi ?) l'influence de l'intervieweur en éliminant les répétitions, c'est-à-dire en gommant cette fois-ci un trait typique de l'oralité.

Graphique 5 : Spécificités des disfluences sur la partition catégorie socioprofessionnelle des interviewés - intervieweurs / interviewés (alternés)

Nous précisons ce résultat à l'aide du *Graphique 6* où l'identité de l'intervieweur est discriminante pour la partition du corpus.



Graphique 6 : Spécificités des disfluences sur la partition identité de l'intervieweur – catégorie socioprofessionnelle des interviewés - intervieweurs / interviewés (alternés)

On constate ici que la conduite d'entretien de l'intervieweuse en cause, MF0, reste similaire d'un groupe socioprofessionnel à l'autre. Mais les interviewés réagissent différemment selon leur groupe socioprofessionnel. Là où les étudiants réagissent cette fois-ci avec une forme de symétrie discursive vers un sous-emploi, le groupe des cadres reste lui parfaitement neutre par rapport à l'intervieweur comme à l'ensemble des parties du corpus. On peut interpréter ce résultat comme la marque d'une différence de représentation de l'intervieweur et de la situation d'énonciation pour les deux groupes socioprofessionnels mis en cause.

Conclusion

Dans cet article, nous avons avant tout montré l'intérêt, du point de vue opératoire, de l'interaction envisagée comme unités linguistiques communes de groupes d'individus ou de leur articulation ; nous avons également tenu compte des disfluences comme traits linguistiques pertinents des corpus textuels oraux.

Dans toutes ces analyses, et sous ce point de vue, on a pu voir que les interviews des étudiants étaient les plus riches en interactions à tous les niveaux étudiés et selon des schémas interactifs différents : reprises, représentations différentes des intervenants, etc.

À ce stade, le premier prolongement de l'analyse serait donc de répéter celle-ci en excluant les entretiens des étudiants afin d'augmenter les contrastes entre les autres groupes socio-professionnels.

Dans un second temps, le commentaire de ces résultats par les individus responsables de ces entretiens, le regard critique qu'il leur permet sur leur propre performance d'intervieweur, doit donner une profondeur, une réalité quantitative évaluative de leur direction d'entretien. Gageons que les analyses, qui mettent en valeur les particularités des entretiens envers les étudiants, peuvent permettre, sous réserve d'autres prolongements (cf. paragraphe précédent), un regard plus objectif sur la réalité de leur comparabilité, dépassant l'idée naïve qu'un entretien comparable est le résultat d'un même questionnement, mais celui d'une même influence sur la production discursive des interviewés.

Références

Benzitoun Chr., Campione E., Deulofeu J., Henry S., Sabio Fr., Teston S., Valli A. and Véronis J. (2004). L'analyse syntaxique de l'oral : problèmes et méthode. In *Journée d'étude de l'ATALA sur l'annotation syntaxique de corpus* (15 mai, Paris).

Berit Hansen A. and Malderez I. (2004). Le *ne* de négation en région parisienne : une étude en temps réel. *Langage et société*, 107 : 5-30.

Biber D. (1988). *Variaton across speech and writing*. Cambridge: Cambridge University Press.

Blanche-Benveniste Cl. and Jeanjean C. (1987). *Le Français parlé. Transcription et édition*. Paris : Didier Érudition.

Bonneau J. (2008). Outils d'aide à l'exploitation d'entretiens semi-directifs : étude de l'interaction entre intervieweur et interviewés sur un corpus ethnoécologique. In *Actes des JADT 2008*, Lyon.

Bourdieu P. (1979). *La distinction*. Paris : Ed. de Minuit.

Dister A. (2007). *De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelles orales* VALIBEL. Thèse de doctorat, Université de Louvain.

Dister A. (2008). L'autocorrection immédiate en français parlé : le cas des déterminants. In *Actes des JADT 2008*, Lyon.

Dister, A., Francard M., Geron G., Hambye P., Simon A.C. and Wilmet R. (2006). *Conventions de transcription régissant les corpus de la banque de données VALIBEL*. http://valibel.fltr.ucl.ac.be/, rubrique corpus oraux, conventions de transcription.

Dister A., Constant M. and Purnelle G. (in press). Normalizing speech transcriptions for Natural Language Processing. In *Actes du colloque international Spoken Communication*, Université de Naples.

Francard M, Lambert J. and Masuy Fr. (1993). L'insécurité linguistique en Communauté française de Belgique. In *Français et Société* 6, Service de la langue française, Bruxelles.

Grangé D. and Lebart L. (1993). *Traitements statistiques des enquêtes*. Paris : Dunod.

Habert B. (2005). *Instruments et ressources électroniques pour le français*. Paris : Ophrys.

Lebart L. and Salem A. (1994). *Les statistiques textuelles*. Paris : Dunod.

Mayaffre D. (2005). De la lexicométrie à la logométrie. In *Astrolabe*. http://www.uottawa.ca/academic/arts/astrolabe/articles/art0048/Logometrie.htm.

Pallaud B. (2002). Les amorces de mots comme faits autonymiques en langage oral. *Recherches sur le français parlé*, 17 : 79-101.

Reinert M. (2002). *Alceste, Manuel de référence*. Université de Saint-Quentin-en-Yvelines, CNRS.

Salem A., Lamaille C., Martinez W. and Fleury S. (2003). *Manuel Lexico 3*. version 3.41, http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/team.htm.

Shriberg E. (1994). *Preliminaries to a Theory of Speech Disfluencies*. Thèse non publiée, Université de Berkeley.