Etude d'application des méthodes et des outils statistiques sur les données du corpus ESLO : cas de la question sur mai 68

Athéna Dupont, Iris Eshkol, Laurent Delsol

Etude d'application des méthodes et des outils statistiques sur les données du corpus ESLO : cas de la question sur mai 68

Athéna Dupont[1], Iris Eshkol-Taravella[1], Laurent Delsol[2]

[1]LLL-Université d'Orléans-France

[2] MAPMO-Université d'Orléans-France

Abstract

Our study focuses on the corpus ESLO1, the Socio-Linguistic Survey in Orleans, realized in 1968-69 by British researchers and digitized, then transcribed by the LLL team as part of ANR Variling project. We interested specifically in the subcorpus extracted using XSLT style sheets and consisting of responses to a question posed to various speakers on the events of may 68. The transcriptions on which we have worked are XML files of Transcriber where certain information such as the time of the utterance, the metadata or the speech events are annotated with XML tags. Each speaker is described in a database containing his sex, age, level of education, etc. All these information were extracted and processed with the statistical software R. We used the methods of descriptive statistics to highlight the relationship between digital variables like time of silence or duration of response and sociological caracteristics on speakers.

Keywords : oral corpus, statistical analysis, sociological variable, may 68, multiple correspondence analysis, time of silence, duration of response

Résumé

Notre étude porte sur le corpus ESLO1 (Enquête Socio Linguistique à Orléans) constitué en 1968-69 par des chercheurs britanniques et numérisé, transcrit ensuite par l'équipe du LLL dans le cadre du projet ANR Variling. Nous nous sommes intéressés plus particulièrement au sous-corpus extrait à l'aide de feuilles de style XSLT et composé des réponses à une question posée aux différents locuteurs sur les événements de mai 68. Les fichiers de transcription sur lesquels nous avons travaillé sont des fichiers XML de Transcriber où certaines informations comme le temps de l'énoncé, les métadonnées ou les événements du discours sont annotés à l'aide de balises XML. Chaque locuteur est renseigné dans une base de données contenant son sexe, âge, niveau d'études, etc. Toutes ces informations ont été extraites et traitées avec le logiciel statistique R. Nous avons utilisé l'analyse des méthodes de statistiques descriptives : boîtes à moustaches, analyse des composantes principales, analyse des correspondances multiples, pour mettre en évidence les relations entre des variables numériques comme temps de pause ou durée de la réponse avec des caractéristiques sociologiques sur des locuteurs.

Mots-clefs : corpus oral, analyse statistique, variable sociologique, mai 68, boîtes à moustaches, analyse des composantes principales, analyse des correspondances multiples, durée de pause, durée de section, durée de la réponse

1. Introduction

Le travail présenté dans cet article porte sur les méthodes actuelles d'exploitation des corpus oraux et des données sociolinguistiques. Nous nous intéressons plus particulièrement sur les outils informatiques et statistiques pouvant prendre en compte la variation. A la différence de l'écrit, un corpus oral associe la parole collectée à la transcription. Notre objectif est de mettre

en relation les transcriptions et les informations sociologiques liées aux locuteurs afin d'observer d'éventuels champs d'influence. Ce qui nous intéresse véritablement dans cette étude est la variation diastratique, liée au sujet social ou individuel (impacts du fait du sexe, de l'âge, de la profession, de la position sociale, du niveau d'études...).

Beaucoup de travaux quantitatifs d'analyse des variations de la langue en fonction de différentes variables comme la position sociale des locuteurs, la situation de communication, le niveau de la langue, etc. ont été menés en sociolinguistique. Déjà en 1976, David Sankoff insistait sur « l'emploi maximum de techniques électroniques et automatiques dans le traitement des données. » Il ajoutait que « de telles méthodes sont indispensables à la compréhension des rapports subtils et complexes entre individus et à l'explication de la façon dont ils se servent de la langue. Souvent il est impossible de cerner ces rapports autrement qu'en termes quantitatifs. » (Sankoff et al., 1976 : 124)

Le corpus ESLO sur lequel nous avons travaillé n'a pas non plus échappé à ce type d'études. (Bergounioux, 2008), (Serpollet, 2007) ou encore (Abouda, 2009) ont étudié les corrélations entre les différents phénomènes linguistiques et les variables sociologiques. Les résultats ont également été démontrés de manière quantitative. La particularité de notre étude est l'analyse des valeurs numériques contenues dans les annotations des transcriptions. Nous essayons d'expliquer comment les annotations faites sur les transcriptions peuvent être analysées avec des outils statistiques. Les annotations peuvent et doivent être prises en compte dans les analyses car elles donnent des informations complémentaires et présentent donc elles-mêmes des variations intéressantes. Notre étude porte sur le lien que peut avoir la variation diastratique par rapport à celle des annotations numériques.

2. ESLO 1

La première Enquête SocioLinguistique à Orléans, ESLO1, a été conçue il y a quarante ans dans un cadre d'étude en Didactique du Français Langue Etrangère. Il s'agit d'enregistrements faits par des chercheurs britanniques des différentes couches de la population orléanaise dans les années 68-9. ESLO1 « comprend environ 200 interviews, toutes référencées (caractérisation sociologique des témoins, identification de l'enquêteur, date et lieu de passation de l'entretien), mais aussi une gamme d'enregistrements variés (conversations téléphoniques, réunions publiques, transactions commerciales, repas de famille, entretiens médicopédagogiques, etc.). » (Abouda et Baude, 2007 :164). Le corpus contient 300 heures (environ 4 500 000 mots). Dans les années 1980-90, une partie du corpus a été transcrite et étiquetée puis mise à disposition sur la toile dans le cadre du projet ELILAP/LANCOM1. Dans les années 1993-2001, le corpus a été repris par des chercheurs de l'Université de Louvain (Debrock et al., 2000). Dans le cadre du projet Variling, la totalité d'ESLO1 a été transcrite et la nouvelle enquête ESLO2 a été commencée 2008.

ESLO1 est un témoignage très intéressant sur le français parlé de cette époque mais aussi sur la vie d'une ville de province comme Orléans dans les années 70. Le sous-corpus sur lequel nous avons travaillé est constitué d'entretiens en face à face au cours desquels un certain nombre de questions a été posé aux Orléanais. Le questionnaire fermé comprend quatre questions préliminaires et cent dix-huit questions réparties entre quatre grandes rubriques : travail, loisirs, enseignement et politique auxquelles s'ajoute une branche «langue et culture». Certaines questions sont posées à tous et font partie du « tronc commun », d'autres sont

[1] ELILAP 1980-83 puis LANCOM 1993-2001, voir (Mertens, 2002).

facultatives et sont introduites sur décision de l'enquêteur. Parmi celles-ci, l'une d'entre elles nous a semblé intéressante à étudier : la question sur les événements de mai 68.

L'analyse du corpus composé des réponses de différents locuteurs à une question posée présente l'avantage d'avoir des données comparables et structurées. Comme le note (Abouda, 2009) « Le fait en effet de disposer d'un long paradigme de réponses, […], à une question simple et identique, produite dans les mêmes conditions discursives » constitue à ses yeux « un terrain privilégie pour une étude linguistique ».

3. Données à traiter

3.1. Transcription, segmentation et annotation sous Transcriber

Le corpus ESLO1 a été transcrit à l'aide du logiciel Transcriber[2]. La transcription est orthographique, sans signes de ponctuation à l'exception des points d'interrogation et des majuscules pour les entités nommées. Les fichiers Transcriber sont des fichiers XML où les balises contiennent :

- différentes métadonnées : des informations sur les locuteurs, les circonstances de l'enregistrement, la date et éventuellement la transcription-même ;
- l'annotation d'événements comme « toux », « rire », etc. qui semblent influer l'analyse des durées, ainsi que des phénomènes de prononciation ;
- des indications de temps permettant la synchronisation avec les fichiers son, comme le début et la fin des pauses et des énoncés, des tours de parole et des sections (voir Figure 1).

L'annotation sous Transcriber est manuelle. Leech en 1997 remarquait que tout commentaire (balisage des bruits, notes du transcripteur) appartient également au domaine de l'annotation et peut donc être considéré de l'ordre de l'interprétation. Dans les transcriptions ESLO, la segmentation est réalisée intuitivement par le transcripteur, sauf dans les cas de changement de locuteur(s). L'interruption du signal sonore est jugée intuitivement et non mesurée à l'aide d'appareils acoustiques. L'annotation est donc subjective.

Comme les transcriptions ne sont pas ponctuées, les balises Transcriber servent de repères de segmentation et permettent entre autre un traitement plus facile par les outils du TAL.

Le premier niveau de segmentation porte sur les énoncés, qui représentent les plus petits segments de parole. Ils sont inclus dans les tours et correspondent à la balise <Sync> de l'annotation Transcriber. Les énoncés sont caractérisés par un locuteur unique, mais il peut y avoir plusieurs interventions (énoncés) par tour de parole. Les tours de parole correspondent aux <Turn>. A la différence des énoncés, ils commencent systématiquement par un changement de locuteur, car ils sont utilisés pour délimiter la parole ; un transfert de l'énonciation d'un individu à un autre se traduit obligatoirement par un nouveau tour de parole. Sur la Figure 1, qui montre l'interface du logiciel Transcriber, les « rectangles bleus » désignent les tours de parole, et les « points bleus » sont des énoncés.

Enfin, le « rectangle rouge » délimite des sections correspondant aux questions. Elles sont codées QP3, T1, T2, etc. et se rapportent à des thématiques plus vastes comme le travail, les loisirs, la culture, la langue, la politique ou la cuisine.

[2] http://trans.sourceforge.net/en/presentation.php. Une nouvelle version de ce dernier est disponible depuis juillet 2011.

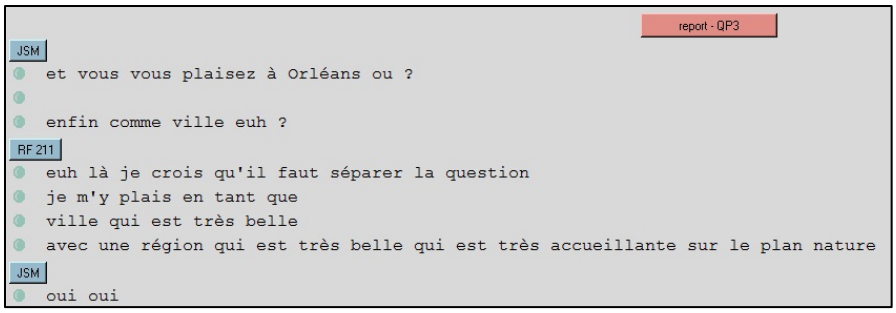


Figure 1 : Interface de Transcriber

C'est à ce niveau que nous sommes lorsque nous traitons de « Mai 68 », question qui appartient à l'ensemble supra-segmental qu'est la rubrique « Politique » du questionnaire.

3.2. Constitution du sous-corpus de mai 68

Le grand problème auquel nous avons été confrontés dans la constitution du corpus est la non homogénéité des données à traiter. Les questions sont inégalement réparties dans les diverses thématiques, en fonction des domaines d'investigations des chercheurs didacticiens. Les thématiques n'apparaissent pas systématiquement dans chaque interaction. Toutes ces restrictions et cette hétérogénéité des données ont rendu difficile l'automatisation du processus d'extraction. C'est pourquoi nous n'avons pris qu'un échantillon de 40 enregistrements.

Nous avons prélevé chacune des sections contenant la thématique de mai 68 de manière à constituer un paradigme rendant possible la comparaison entre locuteurs. La sélection de la question sur les événements de mai 68 s'est faite en deux étapes. En premier lieu, nous nous sommes servis des mots clé « soixante-huit » ou « mai » pour extraire les sections entières contenant ce mot. Nous avons utilisé ensuite les feuilles de style XSLT pour extraire sous forme de tableau les éléments nécessaires à notre analyse statistique.

3.3. Le sous-corpus de mai 68

Le sous-corpus extrait contenant la réponse sur les événements de mai 68 donnée par les orléanais de l'époque présente quelques particularités relevées par (Bergounioux, 2010) : «Tout d'abord, elle n'a été préparée par rien […] Elle ne sera suivie par aucune demande d'éclaircissement, par aucun prolongement […] Ces difficultés, accrues par la réticence commune à faire état d'opinions politiques, expliqueraient non seulement l'hésitation des enquêteurs à poser la question mais aussi le nombre de réponses souvent dilatoires, quand ce ne sont pas carrément des refus de répondre […]. Les personnes sollicitées préfèrent, le plus souvent, ne pas afficher de jugements trop tranchés. Les non réponses sont nombreuses, que les personnes se déclarent apolitiques, ou se réfugient derrière leur incompétence.[…] on note la grande prudence de nombreux témoins qui rechignent à fournir des explications, préférant décrire ce qu'ils ont vu dans des registres très détachés où ils se situent en tant que spectateurs plutôt qu'acteurs d'agissements qu'ils désapprouvent plus souvent qu'ils ne s'y reconnaissent. » (p.7-9)

Ces différentes observations de l'auteur présentent toutes un point commun : un certain malaise est provoqué par cette question. Sans reprendre le travail de Bergounioux nous le complétons par l'étude du non-dit ou plus exactement des pauses dans la réponse à cette question délicate. Notre objectif est d'analyser une autre information d'ordre implicite contenue dans les annotations du temps de parole. Il s'agit des valeurs numériques marquées par les balises XML. Si la question est inattendue et gênante pour le locuteur, la durée du

silence entre la question posée et la réponse doit être importante. Nous tenons compte aussi dans notre analyse d'autres informations d'ordre sociologique contenues dans la base de données du corpus ESLO. Nous nous demandons si la pause après la question posée ou encore si la durée totale de la réponse varient en fonction du profil sociologique du locuteur.

4. Valeurs à prendre en compte

L'objectif de ce travail est d'utiliser des méthodes statistiques pour analyser les données numériques contenues dans la transcription des réponses sur les événements de mai 68 par rapport aux critères sociologiques des locuteurs. Nous développerons dans cette partie, chaque valeur prise en compte par notre étude.

4.1. Valeurs numériques de transcription

4.1.1. Durée de pause après la question (DPAQ)

Le point de départ de notre travail est l'analyse des valeurs numériques contenues dans les annotations des transcriptions. Nous présupposons en nous basant sur l'étude de (Bergounioux, 2010) que cette question pourrait mettre en difficulté le locuteur et engager un silence important avant la réponse, que nous aimerions calculer et étudier.

Certains linguistes distinguent deux phénomènes : silence et pause silencieuse. Caroline Pount-Biset en 1994 dans l'article où elle étudie le silence et la pause dans un enregistrement d'ESLO définit le silence comme « l'absence de son, c'est-à-dire la réalité physique correspondant à l'absence de variation de la pression d'air » et la pause comme « toute interruption dans le signal sonore égale ou supérieure à la durée moyenne nécessaire à l'articulation d'un phonème ». L'auteure ajoute qu'« on peut distinguer le silence que le locuteur subit comme une contrainte biologique, et la pause que celui-ci réalise délibérément dans un but stylistique ou grammatical. » (p.2) Anne Dister dans sa thèse définit le silence comme « une pause perçue par le transcripteur comme particulièrement longue […] Ces pauses en fin de tour de parole, dans notre corpus, sont souvent liées au fait qu'après une question, le locuteur qui va prendre la parole a besoin d'un temps de réflexion (ou de réaction) avant de répondre ».(p. 260) Si l'on suit le même raisonnement, le temps étudié concerne plutôt le silence que la pause. Cependant, comme l'annotation de Transcriber ne fait pas cette distinction (le temps de non parole est marqué par la balise vide <Sync/>) :

<Sync time= « 643.462 »> on a beaucoup parlé des événements de mai soixante-huit</Sync>

<Sync time= « 646.48 »></Sync>

<Sync time= « 647.402 »>euh</Sync>

nous gardons le terme pause même s'il s'agit du temps après la question posée.

4.1.2. Durée de pause par section (DPS)

Si l'on étudie le temps de pause après la question, il est intéressant d'y ajouter une autre valeur, celle de la distribution des pauses au long de la section. C'est ce que signifie la valeur DPS calculée par la somme des pauses divisée par la durée de section. Il est intéressant de vérifier aussi s'il existe une relation entre la durée de pause après la question et la répartition globale des pauses pendant la réponse.

4.1.3. Durée de section (DS)

Comme nous l'avons décrit dans la partie sur l'annotation, chaque transcription est segmentée en sections qui décrivent une thématique de l'interview. Ainsi, la réponse sur les événements de mai 68 fait partie d'une section extraite. La durée de la section est indiquée par les attributs «startTime » et « endTime » :

<Section type="report" topic="to1" startTime="2.814" endTime="10.88">

Cette valeur montre le temps global de la réponse et varie de 1 (le temps minimal) à 4 (le temps maximal).

On pourrait donc se demander si la durée de la réponse ne varie pas en fonction du profil sociologique du locuteur. Dans l'analyse du sous-corpus des omelettes (Abouda, 2009) a constaté un lien entre la durée de la réponse et le sexe du témoin. Selon l'auteur « les hommes parlent en moyenne 12 secondes de moins que les femmes sur une durée moyenne de 69 secondes ». Nous vérifierons donc le lien de la variable de la durée de réponse avec des critères sociologiques multiples en considérant cependant qu'un critère de sexe ne suffit pas de faire des constatations objectives.

4.1.4. Autres variables (Nombre d'énoncés NES et Taux d'énoncés par section TXEPS)

Comme nous l'avons mentionné précédemment, chaque section est composée de tours de parole définis par le changement du locuteur. Chaque tour est divisé en énoncés. Un locuteur peut produire plusieurs énoncés avant de laisser la parole à son interlocuteur comme dans le cas suivant :

<Turn speaker="spk1" startTime="1104.46" endTime="1113.212">
<Sync time="1104.46"> on a fait semblant de faire des réformes</Sync>
<Sync time="1106.535"></Sync>
<Sync time="1108.07">
et rien y changer on a mis un peu de désordre partout et c'est tout mais il n'y a rien eu de positif </Sync>
</Turn>

Dans l'exemple ci-dessus, le tour de parole contient deux énoncés « on a fait semblant de faire des réformes » et « et rien y changer on a mis un peu de désordre partout et c'est tout mais il n'y a rien eu de positif » et une pause lorsque la balise <Sync> est vide.

L'autre variable, TXEPS représente en quelque sorte la fréquence relative des énoncés. Nous avons ajouté au début ces deux variables mais l'analyse menée a montré leur non pertinence par rapport à l'étude. Nous les avons donc mis de côté.

4.2. Valeurs sociologiques

En étudiant le corpus de mai 68, (Bergounioux, 2010) constate l'influence de variables sociologiques dans le refus de certains locuteurs de répondre à la question sur mai 68 « Le cas est particulièrement flagrant pour les femmes qui se sentent souvent moins qualifiées, à cette époque, pour se prévaloir d'une opinion » (p.7) ou « l'importance des refus de réponse, en particulier dans les classes populaires » (p.2). Nous décidons donc d'ajouter des variables sociologiques et, de cette manière, d'observer les liens entre les deux types de données : d'une part, les valeurs numériques de temps annotées par les transcripteurs, et, d'autre part, les métadonnées sur le profil sociologique de chaque locuteur.

Chaque transcription est liée à la base de données où sont répertoriées les informations sociologiques sur le locuteur principal de chaque entretien : date et lieu de naissance, sexe,

questions relatives à la famille et à l'éducation, profession et appartenance sociale. Ces critères sont renseignés autant que possible au moment de l'enquête.

La prise en compte des critères sociologiques présuppose une certaine catégorisation préétablie des locuteurs, d'où le danger constaté par (Cappeau et Gadet, 2007) de privilégier « une conception du social reposant sur des catégories pré-construites, dont les locuteurs sont censés être porteurs. » (p.106) Nous sommes donc très prudents avec les résultats présentés dans la partie suivante. Plus on considère de variables, plus l'analyse est globale et plus elle tente d'être objective. Ainsi, ne prendre en compte que le critère de sexe, présuppose qu'il y a un langage des femmes opposé à celui des hommes. « Il ne suffit pas de dégager des différences homme/femme dans des échantillons homogènes du point de vue des autres variables, où l'on recherche l'influence de la variable sexe, il faut que des femmes aient des productions linguistiques identiques, ou au minimum présentent des tendances analogues, convergentes, quels que soient les échantillons. Ce qui signifie différence d'âge, de milieu socio-culturel, éducatif, ou professionnel équivalents. » (Houdebine-Gravaud, 2003). D'ailleurs, notre analyse présentée dans la partie suivante montre que la variable sexe n'est pas du tout pertinente pour notre étude.

Essayant de tenir compte de toutes ces remarques et étant très prudents sur les résultats obtenus, nous utilisons dans notre étude six variables sociologiques :

- sexe ;

- âge classé en trois catégories de 1 à 3 où la classe 1 représente les personnes les plus âgées ;

- classe sociale mesurée avec l'échelle Alix Mullineaux en utilisant cinq degrés de A à E (A correspondant aux classes supérieures : patrons, professions libérales et enseignants du second degré et E aux classes populaires) ;

- âge fin d'études qui varie de 0 (études en cours : cas des étudiants) jusqu'à 4 (études avancées) ;

- origine (Loiret, Paris, France, Autre) ;

- politique (Gauche, Droite, Centre, Ne se classe pas, Non renseignée).

Ce dernier nous semble être intéressant à prendre en compte dans une réponse éventuellement connotée politiquement comme les événements de mai 68.

5. Analyse statistique

Pour traiter les données, nous utilisons trois méthodes différentes mais complémentaires de statistiques descriptives pour observer des tendances de rapprochement entre des variables différentes.[3]

5.1. Boîtes à moustaches

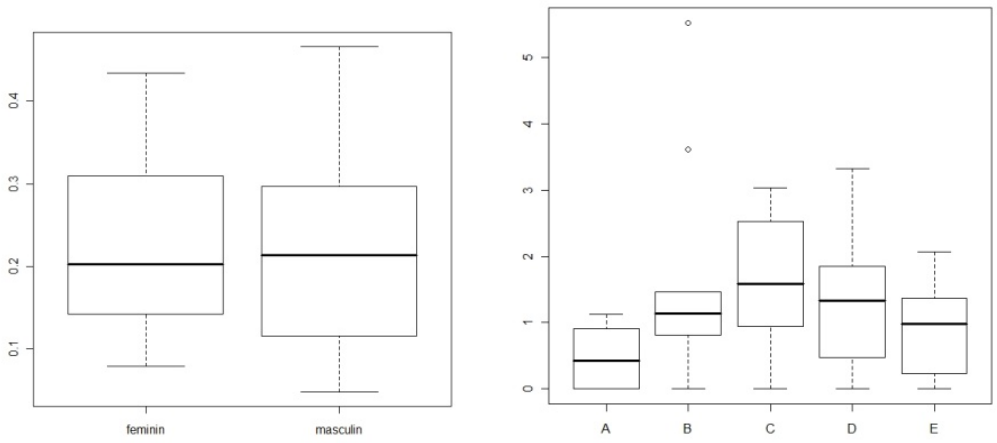
Les boîtes à moustaches permettent de représenter de manière schématique la répartition des valeurs prises par une variable numérique. Elle décrit les observations en représentant de bas en haut les valeurs du minimum, des quartiles, et du maximum[4]. On utilise des boîtes à

[3] On peut notamment trouver plus de détails concernant les méthodes d'analyse des composantes principales (ACP) et d'analyse des correspondances multiplies (ACM) dans (Lebart et al, 2006).

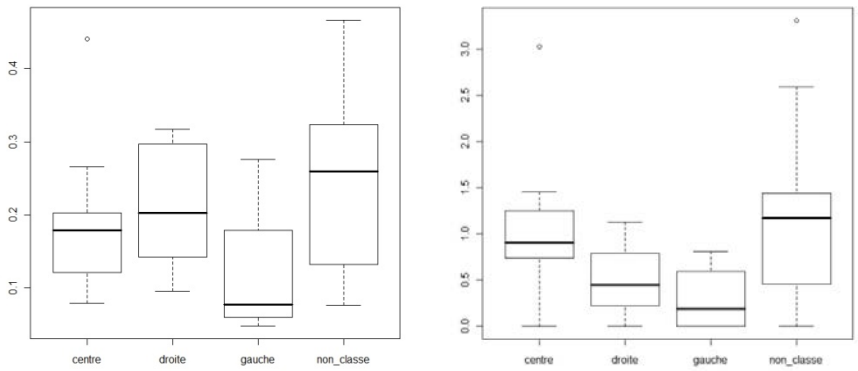
[4] Certaines valeurs « extrêmes » ne sont pas prises en compte pour calculer le minimum et le maximum. Elles sont représentées par des points hors de la boîte à moustache (cf. DPAQ/AM, catégorie B).

moustaches pour les populations correspondant aux différentes valeurs d'un critère qualitatif (par exemple le critère de position politique : gauche, centre, droite, ne se classe pas) afin de comparer les valeurs prises par une variable numérique en fonction d'un critère qualitatif sélectionné.

Observons des boîtes à moustaches représentant des sous-populations définies par trois critères sociologiques : sexe, échelle AM et politique. Le trait noir au milieu représente la médiane. C'est la valeur centrale des données (50% des observations lui sont supérieures, 50% lui sont inférieures). Le trait en dessous [respectivement en dessus] est le premier quartile (25% en dessous, 75% au dessus) [respectivement le troisième quartile (75% en dessous, 25% au dessus)]. Cette représentation des données permet de résumer la manière dont elles se répartissent. On sait que 50% des observations se situent autour de la médiane entre le premier et le troisième quartile. D'autre part, plus l'étirement de la boîte à moustache est important, plus forte est la variabilité de nos observations autour de la médiane.



DPS/sexe *DPAQ/AM*



DPS/P *DPAQ/P*

Figure 2 : Boîtes à moustaches

La première boîte montre la répartition de la variable DPS par rapport au critère sociologique du sexe. On voit que la médiane est semblable dans les deux cas et se rapproche de 0,2 secondes. On constate que 20% de la section chez les hommes comme chez les femmes est

silencieuse. Cette similitude dans le comportement entre les deux sexes se vérifie avec des boîtes à moustaches pour d'autres valeurs[5] comme DPAQ où 50% des femmes et des hommes environ ont environ une durée de pause inférieure à 1 seconde et 50% supérieure à 1 seconde, DS proche de 200 secondes dans les deux cas ou encore avec d'autres variables NES, TXEPS.

La deuxième boîte montre la répartition de la valeur numérique DPAQ par rapport à l'échelle AM. On constate que les locuteurs appartenant à la catégorie A (classe supérieure) répondent plus rapidement à la question posée (la durée de pause après la question est plus courte). Les gens qui semblent être plus gênés (la durée de pause est plus importante) sont ceux provenant de la catégorie C (classe moyenne). Globalement, la boîte montre que le temps de pause augmente jusqu'à la classe C et descend ensuite.

Les deux dernières boîtes répartissent les valeurs numériques DPAQ et DPS par rapport au critère politique. D'une manière générale, on observe un faible pourcentage de pauses le long de la section chez les gens de gauche, plus de variabilité chez les locuteurs de droite (leur valeur centrale, la médiane est plus haute que chez les personnes de gauche) et les personnes qui ne se classent pas politiquement montrent plus d'hésitation. En ce qui concerne la valeur de DPAQ, elle est plus faible chez les gens de gauche qui semblent répondre plus rapidement, ensuite viennent les locuteurs de droite, de centre et enfin ceux qui ne se classent pas. On observe de nouveau plus de variabilité dans le temps de réponse chez ces derniers.

5.2. Analyse des composantes principales (ACP)

Pour chaque entretien nous considérons quatre variables numériques (DPAQ, DPS, DS, et TXEPS). L'objectif de l'analyse en composantes principales est de chercher des combinaisons de ces variables numériques que l'on va appeler « axe factoriels » par rapport auxquels s'expriment principalement la variabilité entre les enregistrements de manière à pouvoir résumer les caractéristiques de chaque entretien en seulement deux ou trois dimensions (tout en exprimant le plus possible cette variabilité).

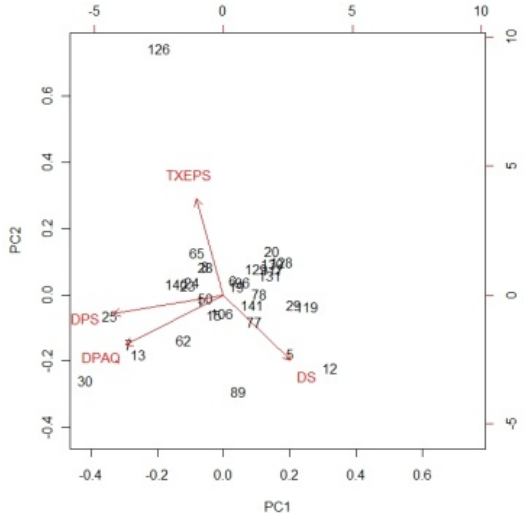


Figure 3 : ACP générale

Le graphique explique 77% de la variabilité entre les enregistrements. Il est intéressant d'observer un rapprochement important entre les deux variables numériques DPS et DPAQ (la

[5] La taille de cet article ne permet pas de mettre toutes les boîtes à moustaches que nous avons étudiées.

corrélation est 0,75) ce qui illustre que les gens prenant du temps à répondre à la question font des pauses généralement aussi le long de la section. Il n'y a donc pas de différence importante entre la pause après la question et les autres. On peut également observer des enregistrements qui ressortent du graphique. L'enregistrement 12, par exemple, a les valeurs DPS et DPAQ très faibles ce qui s'explique par l'absence de la pause après la question et le nombre peu élevé des pauses pendant le discours. Le locuteur donne des réponses longues sans beaucoup de pauses et prend du temps à répondre. On remarque très peu d'hésitation dans son discours :

> PB: alors on aimerait bien savoir quelles étaient vos impressions de ce qui c'est passé à ce moment-là ?
>
> LD 386: très volontiers très volontiers parce que c'est une période qui m'a beaucoup touchée et qui se prolonge d'ailleurs encore aujourd'hui les évènements de mai ne sont pas terminés et de loin contrairement aux apparences mes impressions de l'époque on- pour moi j'ai considéré que les évènements de mai étaient de d'heureux évènements pour la France parce que pour la première fois il y a eu une certaine prise de conscience collective que après tout les choses n'allaient peut-être pas si bien qu'elles en avaient l'air ou qu'on avait l'air de nous le dire

Il est intéressant de noter qu'il s'agit du locuteur qui a fait le plus d'années d'études et qui appartient à la catégorie sociale la plus élevé (A).

5.3. Analyse des correspondances multiples (ACM)

L'analyse des correspondances multiples est un outil très courant de statistique descriptive. L'objectif est de mettre en évidence les relations qui peuvent exister entre les modalités de plusieurs variables qualitatives (et éventuellement quantitatives discrètes, ou continues et classées). On recherche des axes factoriels permettant de résumer le mieux possible la variabilité entre les entretiens par rapport à la répartitions des modalités des variables que l'on considère. Dans ce cas cela n'a pas de sens de considérer directement des combinaisons de nos variables comme en ACP car nos données ne sont pas numériques. La démarche de l'ACM permet la construction d'axes factoriels en utilisant une distance particulière entre les profils des individus. Nous avons ici considéré des variables qualitatives correspondant à des critères sociologiques (P, Origine, AM, Sexe) ainsi que des variables numériques classées (DOBC, DPSC, DPAQC, DSC, AFEC)[6]. L'origine correspond au centre de gravité, c'est à dire à la tendance globale. La distance entre une modalité et l'origine quantifie le côté « atypique » des enregistrements où cette valeur est observée en ce qui concerne les valeurs prises par les autres variables. Si l'on a tendance à retrouver des valeurs de deux variables différentes pour les mêmes enregistrements, cela se traduit sur le graphique (coordonnées suivant les deux premiers axes) par une proximité de leurs représentations. Au contraire, on peut s'attendre à ce que des modalités qui ne se retrouvent que très rarement pour les mêmes enregistrements soient éloignées l'une de l'autre. Cet effet d'attraction (ou de répulsion) entre les modalités des variables est d'autant plus à prendre en compte qu'elles sont représentées loin de l'origine. Il est toutefois important de prendre garde que cette représentation n'explique qu'une faible part de la variabilité entre les enregistrements et que certaines proximités observées peuvent être trompeuses.

[6] C signifie classée.

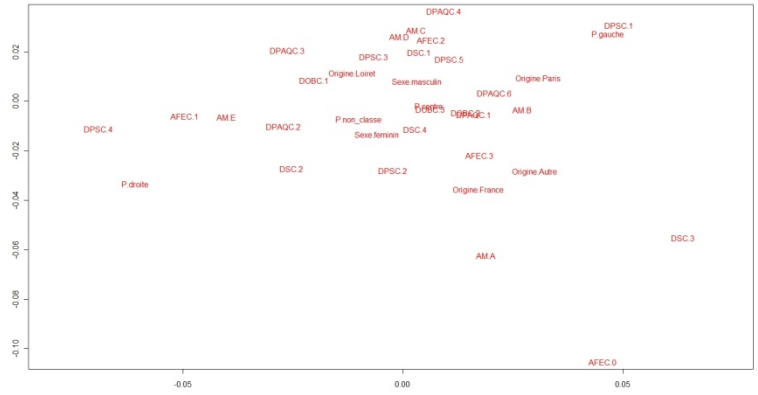


Figure 4 : ACM général

En observant le graphique de l'ACM général, on peut déjà affirmer que le sexe masculin et féminin se trouvent proches de l'origine (zéro) ce qui signifie que la répartition des autres variables pour les hommes et femmes est semblable. On remarque un regroupement entre Pgauche et DPS1 ce qui confirme l'analyse par des boîtes à moustaches. Les gens de gauche semblent avoir peu de pauses pendant le discours. La variable Pdroite est proche de DPS4 ce qui confirme de nouveau nos observations précédentes : les gens de droite prennent plus de pauses pendant la réponse.

Pour affiner les résultats, nous proposons un graphique similaire de la répartition de la valeur numérique DSC par rapport aux critères sociologiques (Figure 5).

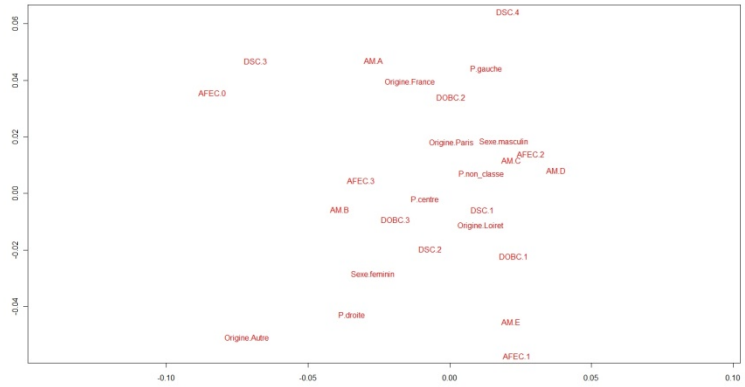


Figure 5 : ACM/DSC

Si l'on regarde la DS d'une façon générale chez les différents individus, on observe que la réponse des gens de gauche est plus longue (DSC4). Les étudiants (AFEC0) semblent aussi être « bavards » en répondant à la question. Toutes ces remarques doivent être considérées avec prudence car la méthode ACM tient compte de beaucoup de variables et la représentation ici des seuls deux premiers axes n'est pas toujours fiable. Ces hypothèses demandent donc plus d'analyse mais permettent cependant d'émettre les différents pronostics sur le comportement des différents locuteurs.

6. Conclusion

Ce travail présente un premier essai d'analyse statistique de données quantitatives d'ESLO 1 par rapport aux métadonnées sur le profil sociologique du locuteur contenues dans la base de

données. Il s'agit d'étudier des variations entre d'une part, des annotations de temps faites par les transcripteurs, d'autre part, des valeurs sociologiques. Nous sommes conscients des limites et contraintes de cette expérience qui ne porte que sur les valeurs numériques des enregistrements et ne prend pas en compte d'autres critères comme le lexique, la prosodie ou les catégories syntaxiques, par exemple.

Cette première ébauche a montré la pertinence de certains critères sociologiques pour notre étude comme l'âge de fin d'études, l'échelle AM, le positionnement politique ou encore l'âge du locuteur ainsi que la non pertinence des critères du sexe ou de l'origine. Une fois encore, toutes les tendances observées sont à pondérer en raison du nombre d'individus étudiés.

Les futurs recherches pourraient porter sur tout le corpus ESLO en analysant la variation pas seulement entre les locuteurs mais aussi au sein du même enregistrement entre les différentes questions posées. Il serait intéressant de tenir compte d'autres annotations faites sur ESLO : étiquetage morpho-syntaxique (Eshkol et al., 2010a), balisage en entités nommées et dénommantes (Eshkol et al., 2010b).

Ce travail a permis d'observer des premières tendances pour émettre certaines hypothèses sur le comportement des différents individus. Le travail entamé sera poursuivi sur un nombre plus important de données et en tenant compte de plus de critères.

Références

Abouda, L. et Baude, O. (2007). Constituer et exploiter un grand corpus oral : choix et enjeux théoriques. Le cas des Eslo. *Corpus en Lettres et Sciences sociales. Des documents numériques à l'interprétation*, Rastier, F. & Ballabriga M. dir, Actes du XXVII^{e} Colloque d'Albi, Toulouse, Presses Universitaires du Mirail : 161-168.

Abouda, L. (2009). Le temps des omelettes. Une nouvelle valeur pour le présent ? *5^{e} Rencontres de Sémantique et Pragmatique*, Université de Gabès (Tunisie), 22-24 avril 2009.

Bergounioux, G. (2010). Mai 68 vu d'Orléans. Actes du 2^{e} *Congrès Mondial de Linguistique Française*, Nouvelle-Orléans, 12-16 juillet 2010.

Debrock M., Mertens P., Truyen F. and Brosens V. (2000). ELICOP, Etude LInguistique de la COmmunication Parlée: Constitution et exploitation d'un corpus de français parlé automatisé, K.U.Leuven: Departement Linguïstiek.

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

Eshkol I., Maurel D., Friburger N. (2010a). Eslo : from transcription to speakers' personal information annotation », *Seventh language resources and evaluation conference (LREC 2010).*

Eshkol I., Tellier I, Taalab S. & Billot S.(2010b). Étiqueter un corpus oral par apprentissage automatique à l'aide de connaissances linguistiques, *10th International Conference on statistical analysis of textual data (JADT 2010).*

Houdebine-Gravaud, A.-M. (2003). Trente ans de recherche sur la différence sexuelle, ou Le langage des femmes et la sexuation dans la langue, les discours, les images. *Langage et société, 106,* 33.

Leech G. (1997). Introduction corpus annotation. In Garside R., Leech G., McEnery A., (Eds.), *Corpus annotation: Linguistic information from computer text corpora*. London: Longman, 1 :18.

Labov W. (1976). *Sociolinguistique*, Paris, Minuit, présentation de P. Encrevé.

Lebart, L., Morineau, A., Piron, M. *Statistique exploratoire multidimensionnelle*, Dunod, 2006.

Mertens, P. (2002). Les corpus de français parlé ELICOP : consultation et exploitation, in Binon, J., *et al.* (*éd*.) *Tableaux Vivants. Opstellen over taal-en-onderwijs aangeboden aan Mark Debrock*. Leuven: Universitaire Pers., 2002.

Pount-Biset, C. (1994). Le silence et les pauses dans le discours. Etude d'un corpus. *Travaux et Document* n3, pp.1-23.

Serpollet, N. (2007). Tell me how you cook and I will tell you who you are – How can a question such as "How do you make an omelette?" illustrate the morpho-syntactic and sociological variations found in the ESLO oral corpus?, in Davies, M., Rayson, P., Hunston, S. & Danielsson, P. (eds), *Proceedings of the Corpus Linguistics Conference, CL2007*, University of Birmingham, 27-30 juillet 2007 : 23-40.

Sankoff, D., Sankoff,G., Laberge, S., Topham, M. (1976). Méthodes d'échantillonnage et utilisation de l'ordinateur dans l'étude de la variation grammaticale. *Cahier de linguistique*, n° 6, pp. 85-125.