Analyse lexicale outillée de la parole transcrite de patients schizophrènes

Maxime Amblard, Karën Fort, Caroline Demily, Nicolas Franck, Michel Musiol

Analyse lexicale outillée de la parole transcrite de patients schizophrènes

Maxime Amblard* — **Karën Fort****
Caroline Demily*** — **Nicolas Franck***** — **Michel Musiol******

* *LORIA, UMR 7503, Université de Lorraine, CNRS, Inria*
** *Université Paris-Sorbonne / EA STIH*
*** *UMR 5229 CH le Vinatier CNRS et Université Lyon 1*
**** *ATILF, UMR 7118, Université de Lorraine, CNRS*

RÉSUMÉ. *Cet article détaille les résultats d'analyses réalisées sur la transcription d'entretiens avec des patients schizophrènes, aux niveaux de la production orale (disfluences) et du lexique (morpho-syntaxe et lemmes). L'étude s'inscrit dans le cadre d'un projet plus large qui prévoit d'autres niveaux d'analyse (syntaxique et du discours), les résultats obtenus devant nous permettre de réfuter ou d'identifier de nouveaux indices linguistiques présents dans la manifestation d'un dysfonctionnement à ces différents niveaux. Le corpus traité contient plus de 375 000 mots, son analyse a donc nécessité l'utilisation d'outils de traitement automatique des langues (TAL) et de textométrie. Nous avons en particulier séparé le traitement des disfluences du traitement lexical, ce qui nous a permis de montrer que si les schizophrènes produisent davantage d'achoppements et de répétitions (disfluences) que les témoins, la richesse de leur lexique n'est pas significativement différente.*

ABSTRACT. *This article details the results of analyses we conducted on the discourse of schizophrenic patients, at the oral production (disfluences) and lexical (part-of-speech and lemmas) levels. This study is part of a larger project, which includes other levels of analyses (syntax and discourse). The obtained results should help us rebut or identify new linguistic evidence participating in the manifestation of a dysfunction at these different levels. The corpus contains more than 375,000 words, its analysis therefore required that we use Natural Language Processing (NLP) and lexicometric tools. In particular, we processed disfluencies and parts-of-speech separately, which allowed us to demonstrate that if schizophrenic patients do produce more disfluencies than control, their lexical richness is not significatively different.*

MOTS-CLÉS : *schizophrénie, disfluences, morpho-syntaxe, lemmes, discours pathologique*

KEYWORDS: *schizophrenia, disfluencies, POS, lemmas, pathological discourse*

3e soumission à *TAL 55-03*, le 01/09/2015

1. Introduction

De nombreuses études ont porté sur la définition, l'implémentation et l'évaluation d'outils pour analyser les pratiques langagières. Leurs motivations s'inscrivent dans des tâches bien définies, mais la question de la validité cognitive des théories et modèles est souvent reléguée à des ouvertures plus qu'à de véritables arguments. S'il apparaît évident qu'il s'agit d'une question complexe, il n'en est pas moins nécessaire d'interroger ces propositions sous l'angle du fonctionnement cognitif.

Une manière d'appréhender cette problématique est de s'intéresser à des manifestations explicites du fonctionnement cognitif, par exemple des dysfonctionnements plutôt qu'à des usages supposés normaux. Ici, nous nous intéressons à l'étude de la réalisation de phénomènes spécifiques chez les schizophrènes au travers de leur pratique langagière. Ces phénomènes sont analysés comme des dysfonctionnements dans la planification du discours, symptôme d'un dysfonctionnement cognitif (Rebuschi *et al.*, 2013). Nous ne tentons pas de définir comment le cerveau produit ce dysfonctionnement mais comment et où ce dysfonctionnement apparaît du point de vue linguistique. Une fois ce dysfonctionnement circonscrit, il deviendra possible de travailler à un modèle pour en rendre compte précisément. Dans un mouvement inverse, étudier ces dysfonctionnements apparaissant dans la langue permet de donner une validité cognitive aux outils capables de les identifier.

Ces travaux s'inscrivent dans le cadre d'une étude large portant sur les pratiques langagières de patients schizophrènes. Le matériel de cette étude provient principalement d'entretiens semi-dirigés par des psychologues. Ces entretiens sont définis pour minimiser l'apport du psychologue, laissant une place importante à la parole du patient. Il s'agit de recueillir l'expression de sa pensée, pour procéder à une analyse, en général psychologique ou psychiatrique. Mais ce n'est pas notre sujet. Nous appréhendons ce matériau comme l'expression d'une pensée en action, au sens cognitif, et l'analysons du point de vue de la pratique langagière.

Dans la continuité des travaux de (Chaika, 1974) et (Fromkin, 1975) qui, les premiers, ont cherché à mettre en avant des indices spécifiques à la capacité langagière des schizophrènes. Ils se sont appuyés sur l'hypothèse forte que la forme de l'expression de leur pensée véhiculait des informations sur les processus cognitifs en œuvre. En même temps, les schizophrènes manifestent des particularités. Si Chaïka s'intéresse à la capacité d'appliquer des règles syntaxiques, (Landre *et al.*, 1992) rapporte que les schizophrènes font le même type d'erreurs que les aphasiques, ce qui les conduit à donner une origine extralangagière au dysfonctionnement, le positionnant à un plus haut niveau cognitif. (Besche *et al.*, 1996) ont étudié la pratique lexicale des patients schizophrènes pour également réfuter l'idée qu'ils auraient un trouble généralisé de traitement du contexte, à nouveau inscrivant les dysfonctionnements à un niveau cognitif plus élevé.

Cependant, ces études restent très limitées, tant dans le nombre de patients pris en considération que dans l'ampleur des phénomènes analysés. En général, et au vue de la difficulté de rencontrer de tels patients, ces études incluent seulement une vingtaine de

participants. Par ailleurs, les moyens tant matériels que théoriques à la disposition des auteurs les contraignent à réaliser à la main des tests relativement peu avancés. Nous ne souhaitons aucunement remettre en cause leurs méthodes, mais utiliser des outils et méthodologies développés dans le cadre du traitement automatique des langues (TAL) sur ces données particulières.

Dans une première partie, nous revenons sur le contexte de cette étude, tant du point de vue de son organisation que de son contexte scientifique. Nous en précisons également le cadre et les limites. Puis nous présentons le corpus en revenant sur sa constitution et les difficultés de la création d'une telle ressource. Enfin, nous détaillons les outils utilisés et les résultats obtenus sur le corpus en analysant les achoppements et les répétitions (disfluences), les catégories morpho-syntaxiques, et les lemmes produits. Nous proposerons ensuite une brève analyse textométrique avant de conclure.

2. Contexte de l'étude

Si nous disposons aujourd'hui de nombreuses références d'articles traitant du sujet de la production langagière des schizophrènes, il n'est pas aussi simple d'en tirer des conclusions. Outre que ces articles proviennent de domaines variés (psychologie, médecine, linguistique, etc.) et qu'ils sont plus ou moins récents et plus ou moins facilement disponibles selon les traditions de chaque domaine, les conditions des expériences décrites sont d'une telle variabilité qu'il est difficile d'en mettre les résultats en cohérence. En effet, les tailles de corpus et les protocoles varient énormément, la langue diffère, les patients sont pour certains en remédiation (et sous-traitement), d'autres non. Enfin, les résultats sont comparés dans certains cas à des témoins et dans d'autres à des patients souffrant d'autres désordres ou pathologies.

La méta-étude de Brendan Maher (Maher, 1972) est très intéressante de ce point de vue, car l'auteur signale les biais de telle ou telle étude ou leurs différences. S'il présente ensemble les résultats concernant les répétitions et ceux concernant la richesse lexicale, il est l'un des rares à les distinguer. Ses conclusions sur les répétitions, déduites des TTR (*Type-Token Ratio*), sont relativement claires : les patients schizophrènes ont un TTR inférieur, ce qui signifierait qu'ils se répètent davantage. Des perturbations du discours des schizophrènes (achoppements, répétitions) ont également été observées par d'autres, notamment dans (Feldstein, 1962) et (Kremen *et al.*, 2003).

Par ailleurs, (Maher, 1972) cite, tout en émettant certaines réserves (les données étant trop limitées), des résultats qui montreraient que les patients schizophrènes utiliseraient un vocabulaire plus restreint. Une étude portant sur les familles de schizophrènes et détaillée dans (DeLisi, 2001) montre elle-aussi que les schizophrènes chroniques utilisent significativement moins de mots que les témoins.

L'analyse que nous présentons vise à vérifier ces résultats, sur une cohorte relativement large. Elle est à notre connaissance la seule portant sur des patients francophones, surtout, elle a été réalisée à l'aide d'outils de TAL au niveau état de l'art. Il est à noter que, comme les travaux que nous présentons s'inscrivent dans un projet plus

général, le corpus sur lequel nous travaillons est partagé avec d'autres recherches. Le protocole utilisé couvre lui l'ensemble du projet. En particulier, nous mesurons les capacités neuro-cognitives par une série de tests avant l'entretien et au cours de certains, nous enregistrons le comportement oculomoteur du patient avec un oculomètre (*eye-tracker*) et/ou l'activité encéphale par électro-encéphalographe (EEG). Dans cet article, nous n'utilisons que les enregistrements sonores transcrits des entretiens.

3. Constitution du corpus

Notre étude, comme la plupart de celles sur les pratiques langagières des patients schizophrènes, est confrontée à de nombreux obstacles pour la constitution du corpus.

3.1. *Répartition des sujets*

Aux vues des difficultés pour identifier les patients et les faire intervenir dans l'étude, notre corpus a été constitué en plusieurs phases, dans différents centres hospitaliers. Dans l'analyse présentée ici, nous considérons les résultats de deux cohortes comportant en tout 80 sujets qui se répartissent en 49 schizophrènes et 31 témoins. Le tableau 1 présente la ventilation des sujets en fonction de leur type (schizophrène ou témoin) et de leur sexe.

Le corpus est divisé en deux cohortes, correspondant aux villes des unités médicales spécialisées des recueils. Par respect pour la confiance accordée par les patients, nous anonymisons ces noms de villes en Ville1 et Ville2. Le recueil de Ville1 a été réalisé par une psychologue pour les patients et trois psychologues pour les témoins, et celui de Ville2 par les deux mêmes psychologues pour les patients et les témoins.

Le sous-corpus Ville1 a été constitué au second semestre 2013. Il est composé de 18 patients diagnostiqués schizophrènes, en remédiation et sous traitement, ainsi que de 23 témoins. Le sous-corpus Ville2 a été constitué au printemps 2002. Il est composé de 31 patients diagnostiqués schizophrènes en remédiation et sous traitement, à l'exception de sept d'entre eux (qui n'étaient pas sous traitement), et de 8 témoins.

	corpus Ville1			corpus Ville2			total
	hommes	femmes	total	hommes	femmes	total	
schizophrènes							
sous traitement	15	3	18	21	3	24	
sans traitement	0	0	0	1	6	7	
total			18			31	49
témoins	15	8	23	4	4	8	31
total	30	11	41	26	13	39	80

Tableau 1 : Répartition des sujets dans le corpus en fonction des cohortes et du sexe.

3.2. *Protocole de collecte*

L'interaction choisie pour cette étude s'organise autour d'un entretien semi-dirigé conduit par un psychologue. Ce type d'entretien est bien défini dans la communauté psychologique et psychanalytique (bien que la terminologie puisse varier) : il s'agit pour le psychologue de maintenir une interaction dans laquelle l'interlocuteur parle librement de lui-même. Pour cela il revient sur son environnement matériel direct, ses relations humaines dans son cadre, ainsi qu'à l'extérieur de son cadre. Le psychologue n'est en aucun cas personnellement engagé dans l'interaction, et sa contribution principale est de relancer l'échange ou de préciser certains éléments.

Par ailleurs, lors de la constitution du sous-corpus Ville1, les sujets ont passé une série de tests permettant de mesurer certaines compétences cognitives. Les tests choisis sont classiques, au sens où ils sont régulièrement utilisés dans la littérature pour des analyses similaires. Les trois tests psychocognitifs choisis mesurent les capacités de mémoire à court terme, d'attention, et la mémoire de travail :

1) le *Wechsler Adult Intelligence Scale-III* (mesure du quotient intellectuel, ou QI),

2) le *California Verbal Learning Test* (capacité cognitive et de stratégie),

3) le *Trail Making Test* (dépréciation de la flexibilité cognitive et de l'inhibition, déficit qui peut affecter la vitesse du système perceptif-moteur, la flexibilité spontanée ou la flexibilité de réaction).

Dans le présent article, nous n'utiliserons que les résultats du test de QI. Il nous semble important d'insister sur le fait que le protocole stipule explicitement que le contenu de l'entretien ne peut et ne doit pas être utilisé ni pour, ni contre le patient. Le fait de ne pas utiliser le contenu contre le patient leur permet une certaine liberté d'expression, et dans un mouvement inverse ne pas l'utiliser pour eux limite la tentation de renvoyer une image trop positive d'eux-mêmes dans le contexte hospitalier.

3.3. *Transcription de la parole*

Nous récupérons les enregistrements des entretiens sous forme de fichier sonore mp3. Ils sont alors transcrits. Nous considérons la transcription comme le premier niveau d'annotation de la ressource. Les deux sous-corpus ayant été constitués à des moments très écartés dans le temps (plus de 10 ans les séparent), les processus de transcription n'ont pas pu être les mêmes. Cependant, dans les deux cas, les transcriptions ont été réalisées par plusieurs annotateurs. Il s'est agit du ou de la psychologue qui a mené tout ou partie des entretiens, ainsi que d'une seconde personne. L'investissement en temps sur cette tâche étant limité, les transcriptions n'ont malheureusement pas pu être réalisées en parallèle.

Il est important de noter que les transcripteurs n'ayant pas connaissance de l'utilisation de leur travail pour des tâches de TAL, n'ont probablement pas pu influencer les résultats dans un sens ou un autre. Les annotateurs ont suivi les recommandations

de base fournies avec `Transcriber` pour une transcription fine et la transcription a été post-traitée suivant les préconisations de (Blanche-Benveniste et Jeanjean, 1987). Nous avons réalisé une relecture partielle *a posteriori* pour identifier les unifications d'annotations minimales à apporter à l'ensemble de la ressource par une série de scripts de normalisation tant sur le codage du texte, le format des fichiers que les annotations elles-mêmes.

En moyenne, les entretiens du sous-corpus de Ville1 sont constitués de 552,73 tours de parole, alors que les entretiens du sous-corpus Ville2 en contiennent 234,5. L'ensemble du corpus comprend 31 575 tours de parole, soit environ 375 000 mots. Le tableau 2 présente la répartition en tours de parole et en mots de l'ensemble du corpus. Il faut noter que, du point de vue du TAL, ce corpus reste de taille modeste. Cependant, nous considérons qu'il atteint une taille raisonnable pour l'utiliser, au vue de sa spécificité, aspect sur lequel nous revenons dans la section suivante.

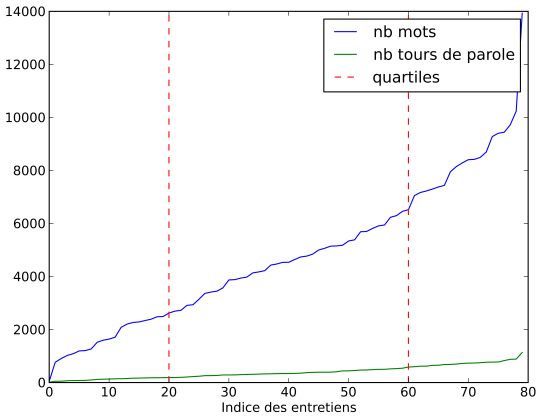


Figure 1 : Distribution des tailles des entretiens en nombre de mots (en bleu) et en nombre de tours de parole (en vert), ainsi que les quartiles (en rouge).

La figure 1 illustre la distribution des tailles des entretiens en nombre de mots et de tours de parole, pour les deux cohortes. Les quartiles apparaissent en rouge. On constate que les entretiens contiennent pour la moitié d'entre eux entre 2 500 et 6 000 mots. Par ailleurs, le nombre de tours de parole est relativement constant dans le corpus. On calcule qu'en moyenne un entretien est composé de 393 tours de parole pour 4 792 mots. Le caractère spécifique de l'entretien semi-directif apparaît dans le corpus : le psychologue produit le même nombre de tours de parole que le sujet, pour un volume de mots très inférieur. Par exemple, dans le sous-corpus Ville1, le nombre de tours de parole des schizophrènes et des psychologues devant un schizophrène est le même, alors que le volume de mots des schizophrènes est 1,54 fois plus important que celui des psychologues. Les témoins du sous-corpus de Ville1 ne présentent pas cette caractéristique, mais une analyse plus fine des entretiens montre que pour six d'entre eux, les témoins ont été réticents à prendre la parole.

	corpus Ville1				corpus Ville2			
	nb tours		nb mots		nb tours		nb mots	
S	3 863	11 145	46 859	119 762	4 062	4 433	66 725	79 081
T	7 282		72 903		371		12 356	
$P+S$	3 819	11 517	30 293	138 571	4 098	4 480	33 686	37 842
$P+T$	7 698		108 278		382		4 156	
$total$		22 662		258 333		8 913		116 923

Tableau 2 : Décomposition du corpus en sous-corpus, en nombre de tours de parole et nombre de mots, en fonction du type d'interlocuteur : S (schizophrènes), T (témoins), P + S (psychologue avec un schizophrène), P + T (psychologue avec un témoin).

Notre étude se focalise sur des aspects relevant du lexique et de la morpho-syntaxe. Aussi nous n'avons pas exploité les aspects phonétiques, comme le temps de parole ou la vitesse d'élocution des locuteurs, ni le recouvrement des tours de parole. Ces données restent cependant disponibles dans le corpus pour une étude ultérieure.

3.4. *Difficultés d'accès aux patients*

Le nombre de 80 sujets peut sembler limité, mais la constitution d'une telle ressource implique de surmonter de nombreuses difficultés, en particulier pour accéder aux patients. De ce fait, disposer d'une cinquantaine de transcriptions d'entretiens avec des schizophrènes représente déjà un corpus significatif.

Pour s'entretenir avec une personne prise en charge en milieu hospitalier, il est nécessaire d'obtenir une autorisation du CPP (Comité de Protection de la Personne) de la région de l'établissement. Les demandes déposées contiennent explicitement et exactement le protocole. L'instruction du dossier requiert plusieurs mois et demande la contraction d'une assurance (pour prendre en charge les possibles dommages). Ces assurances augmentent considérablement les budgets nécessaires à ce type d'expérience. Une fois les accords obtenus, il n'est alors plus possible de modifier les protocoles.

Mais ce qui rend complexe la constitution d'une telle ressource est principalement la difficulté de faire participer les patients. Plusieurs problèmes se posent. Il faut d'abord identifier, au sein d'un service, les patients répondant aux critères de l'étude et en capacité d'interagir avec une personne tierce au service. Puis il faut, au sein de cette population, trouver les patients qui acceptent de participer. Une première réticence vient du fait qu'il n'y a pas de conséquence positive, en terme médical, à participer à l'étude. Il faut ajouter à cela des inquiétudes compréhensibles des patients concernant la possible publication de leur histoire, bien qu'une anonymisation soit garantie.

Par ailleurs, le protocole requérant de passer des tests psycho-cognitifs et un entretien, le temps nécessaire est de l'ordre de deux heures, ce qui est relativement élevé. Ce n'est pas tant la disponibilité des patients qui est alors en jeu, que leur aptitude à rester concentrés. Lorsque le patient présente soudainement des difficultés, il faut convenir d'un second rendez-vous pour finaliser le protocole. La multiplication des rendez-vous génère des défections. À titre d'exemple, lors de la phase de collecte des entretiens du sous-corpus Ville1 qui s'est déroulée dans d'excellente conditions matérielles et administratives, 45 % (18) des patients contactés ont refusé de participer, 10 % ont accepté un premier rendez-vous mais ne sont pas présentés au second, et 45 % (18 sujets) ont participé à toute l'étude.

3.5. *Anonymisation*

L'anonymisation d'un corpus est une tâche complexe qui recouvre plusieurs dimensions. Nous avons dans un premier temps cherché à désidentifier le corpus (Meystre *et al.*, 2010), c'est-à-dire à identifier les entités nommées et à leur substituer des marqueurs neutres.

Afin de conserver la lisibilité des documents, nous avons cherché par lectures successives les catégories à masquer. Nous en avons identifié 10. Concernant les personnes, nous identifions les noms et nous avons choisi de conserver une marque explicite du sexe pour le prénom, ce qui permet de lever des ambiguïtés dans plusieurs entretiens. De manière tout à fait classique (Grouin, 2013), nous avons choisi d'utiliser une catégorie pour les institutions, qui ici sont nombreuses, car il est souvent fait mention d'hôpitaux ou de différents services d'un même hôpital. Cette analyse nous impose de conserver un identifiant unique pour chaque entité. Ainsi, si la première référence faite à l'hôpital X est substituée par *institution3*, ce même hôpital restera *institution3* dans tout l'entretien. Il ne serait plus possible de comprendre les échanges sans cette contrainte. Cela étant, nous ne conservons cette cohérence de dénotation qu'à l'intérieur d'un même entretien et non sur l'ensemble du corpus. Par extension, cette propriété est conservée pour toutes les catégories. Les entretiens étant situés dans une géographie particulière, nous utilisons également les catégories *pays*, *département*, *ville*, *capitale*, *montagne*. Ces catégories pourraient être amenées à évoluer, en particulier autours des relations ontologiques qu'elles entretiennent. Le choix a été fait de les fixer à cela pour cette phase du projet. Enfin, nous avons ajouté une dernière catégorie rassemblant tous les autres cas : *non_pris_en_compte*. Nous ne présentons pas ici de répartition de ces annotations d'anonymisation, mais elles apparaissent en très faible quantité.

Nous avons identifié un outil d'anonymisation performant, MEDINA (Grouin, 2013), mais nous n'avons pu l'obtenir à temps pour des raisons administratives[1]. Nous avons donc implémenté une série de scripts en `Python` basés sur les expressions régulières. Une intervention humaine reste nécessaire pour superviser l'application des

1. Il faut, pour utiliser cet outil, faire signer sa licence par les laboratoires.

scripts qui peuvent lever des exceptions en cas d'ambiguïté. Cette tâche étant réalisée une fois pour toute, nous avons choisi d'en privilégier la qualité, quitte à perdre en efficacité. À partir du résultat, nous avons donc procédé à une vérification par extraction automatique de toutes les positions potentiellement ambiguës.

Nous nous sommes rendus à l'évidence, comme d'autres avant nous (notamment (Eshkol-Taravella *et al.*, 2014)), que s'il nous était possible de cacher le prénom et le nom des personnes, les sujets exprimant de nombreuses informations tant sur leur histoire que sur leur famille, voire leur localisation, il nous est impossible de garantir un véritable anonymat. Ainsi, dans l'un des entretiens, le patient explique qu'il a intégré une classe préparatoire dans une ville du Nord, avant de retourner s'inscrire à l'université dans une autre ville, suite à une dépression. Ces deux éléments peuvent paraître peu, mais mis ensemble, qui plus est en ajoutant d'autres informations sur sa famille disséminées dans l'entretien, il devient, sinon possible de le désigner nommément, au moins d'identifier un groupe restreint de personnes correspondant. Nos craintes peuvent paraître excessives, mais les conséquences, tant pour les patients que pour leurs proches, pouvant être lourdes, nous ne pouvons accepter de considérer la tâche d'anonymisation comme pleinement réalisée.

Nous travaillons donc à partir de la ressource transcrite et désidentifiée. Nous avons, dans un deuxième temps, construit une version du corpus où les tours de paroles ont été mélangés, en ne conservant que la catégorie du locuteur (psychologue - témoin - patient). Il devient alors très difficile (impossible) de reconstruire les histoires, les temporalités ou encore les géographies, et nous pouvons raisonnablement considérer que l'anonymat est garanti.

4. Protocole expérimental

Nous avons annoté le corpus automatiquement en disfluences, en morpho-syntaxe et lemmes, puis nous avons réalisé une analyse textométrique outillée. Cette section présente les outils utilisés.

4.1. *Annotation automatique des disfluences :* `Distagger`

L'annotation en disfluences a été réalisée pour en étudier la pratique chez les schizophrènes (section 5.2).

`Distagger` (Constant et Dister, 2010) est un outil d'annotation automatique des disfluences librement disponible dont les performances ont été évaluées, sur un corpus de référence de 22 476 mots et 1 280 disfluences, à 95,5 % de F-score (précision de 95,3 %, rappel 95,8 %)[2].

2. Une évaluation de l'outil sur un échantillon de nos données (4 entretiens) a mis au jour un taux d'erreur compris entre 5 et 10 %. Une analyse de ces erreurs a montré qu'elles étaient majoritairement dues à des interruptions mal identifiées, problème que nous avons corrigé depuis.

De manière tout à fait classique pour la question des disfluences, `Distagger` les définit comme des réalisations orales qui rompent la continuité syntaxique. Il permet d'identifier des réalisations de natures différentes, pour lesquelles quatre restent prédominantes dans les corpus oraux : les *euh*, les répétitions, les autocorrections immédiates et les amorces de morphèmes. Ces différentes réalisations sont définies comme suit (les exemples proviennent de notre corpus) :

– Les *euh* :

(1) moi ça m'est presque plus euh difficile et euh anti-naturel de parler

– Les répétitions sont entendues comme la reprise explicite et identique d'un mot ou d'un groupe de mots dans le contexte immédiat d'apparition. La répétition peut contenir ou être précédée d'un mot creux comme *oui*, *non*, ou un *euh* :

(2) j' arrive à être à être concentrée quand il faut faire quelque chose

– L'autocorrection immédiate est une variante de la répétition dans laquelle un trait morphologique peut varier (ce qui apparaît régulièrement avec les déterminants) :

(3) enfin je sais pas trop le les termes

– L'amorce est une interruption de morphème en cours d'énonciation. La fin du mot est marquée par un -.

(4) pis progressivement vous av- pouvez travailler sur votre concentration

Les annotations de `Distagger` sur le corpus font apparaître sept étiquettes dont deux étiquettes spécifiques permettant de repérer les tours de parole et les interlocuteurs à qui ils sont associés. Par ailleurs, deux autres apparaissent dans des volumes trop restreints pour être significatifs (respectivement 5 et 1 étiquettes). Dans la suite nous utiliserons les trois étiquettes : $\{EUH\}$, $\{REP\}$ et $\{CORR\}$. Une remarque importante est que les amorces sont soit reconnues comme des répétitions, soit comme des corrections. Comme nous nous intéressons uniquement au volume de disfluences et non à leur distribution en catégories, nous conservons cette version de l'annotation. Il va de soit qu'un prolongement de notre proposition nécessiterait de revenir plus en détails sur cette distribution et sur les disfluences combinées, sans se contenter du *reparandum*[3].

4.2. *Annotation automatique en morpho-syntaxe :* `MElt`

`MElt` (Denis et Sagot, 2009) est un analyseur morpho-syntaxique (*tagger*) librement disponible reposant sur des perceptrons multiclasses. Il est distribué avec un modèle pour le français parlé entraîné sur le corpus TCOF-POS (Benzitoun *et al.*, 2012) et utilisant le lexique Le*fff*. Les performances de cet outil avec ce modèle atteignent 97,61 % d'exactitude, elles sont donc au niveau de l'état de l'art.

3. Le reparandum est la partie de l'énoncé précédent la disfluence et qui doit être corrigée.

MElt est appelé en ligne de commande, directement sur les documents textuels auxquels nous apportons des méta-données. Nous avons donc implémenté une série de scripts en Python pour pré-traiter le corpus et lui appliquer MElt.

Afin de conserver l'intégrité des données, nous avons fait le choix d'appliquer MElt sur chacun des entretiens, individuellement. Du point de vue opérationnel ce choix n'apporte pas la meilleure efficacité (le temps de chargement des ressources de MElt étant important), mais nous pouvons ainsi post-traiter chacun des entretiens.

Techniquement, nous divisons le corpus en deux fichiers, l'un contenant l'identification du locuteur, l'autre le contenu du tour de parole. Sur ce second fichier nous appelons MElt, puis nous reconstruisons la ressource originale augmentée des annotations en fusionnant ces deux fichiers.

Les annotations ont la forme suivante, le caractère $*$ étant utilisé pour annoter les lemmes des mots inconnus du lexique :

(5) Voilà alors peut-être vous pouvez m'e/ m'expliquer

Voilà/FNO/voilà alors/ADV/alors peut-être/ADV/peut-être vous/PRO :cls/vous pouvez/VER :pres/pouvoir m'/PRO :clo/me e/ADV/*e //MLT/*/ m'/PRO :clo/me expliquer/VER :infi/expliquer

4.3. *Analyse textométrique : TXM*

TXM (Heiden *et al.*, 2010) est un outil d'analyse textométrique, librement disponible, de corpus textuels, incluant des fonctionnalités d'analyse statistique (*via* le logiciel R). Outre sa facilité d'utilisation, TXM présente un avantage décisif par rapport à des logiciels d'analyse statistique générique : il offre un accès direct au contexte, ce qui permet d'affiner les résultats quantitatifs par une analyse qualitative manuelle.

Par défaut et pour des raisons de compatibilité logicielle, TXM étiquette les corpus avec TreeTagger (Schmid, 1994), un étiqueteur morpho-syntaxique dont les performances ne sont pas au niveau état de l'art[4] et qui ne propose pas de modèle spécifique pour le français parlé. Nous avons donc annoté et lemmatisé le corpus avec MElt, et l'avons importé ainsi enrichi dans TXM, en prenant soin de neutraliser TreeTagger.

Nous avons utilisé TXM pour évaluer la sur-représentation ou la sous-représentation de certains mots dans les sous-corpus de schizophrènes par rapport aux sous-corpus de témoins. Nous avons pour cela créé une partition du corpus, puis nous avons calculé les spécificités (Lafon, 1980) de chaque lemme, ce qui permet de prendre en compte les déséquilibres entre sous-corpus.

Nous en avons profité pour réaliser un calcul de la richesse lexicale de chaque sous-groupe (schizophrènes, témoins et psychologues) qui est le ratio du nombre de

4. TreeTagger atteint 95,7 % d'exactitude sur le français (Allauzen et Bonneau-Maynard, 2008), ce qui correspond à peu près à deux fois plus d'erreurs que MElt.

lemmes par rapport au nombre total de forme et un indice de diversité lexicale, qui est le ratio du nombre de lemmes par rapport, cette fois, au nombre total de formes différentes (types). Il est à noter que la richesse lexicale se différencie du TTR par le fait que nous prenons en compte les lemmes et non les types (qui sont des formes). La diversité lexicale permet elle de minimiser l'impact des mots très utilisés dans le calcul de la richesse lexicale. Plus la valeur est proche de 1, plus l'interlocuteur utilise de termes différents, indépendamment de leurs dérivations morphologiques.

Enfin, nous avons manuellement examiné les contextes des lemmes présentant des spécificités élevées, afin de vérifier à quoi ceux-ci correspondent dans le corpus.

5. Résultats sur le corpus

5.1. *Significativité*

Dans la suite de notre présentation, nous allons revenir sur plusieurs résultats calculés sur différents volumes de données. Si une lecture directe des résultats peut sembler apporter des éléments d'interprétation, nous avons choisi de valider ces interprétations en faisant appel à une mesure de significativité. Pour cela nous avons repris celle utilisée dans (de Mareüil *et al.*, 2013) pour des contextes similaires, c'est-à-dire l'analyse de disfluences dans des entretiens journalistiques.

Cette mesure permet de calculer un indice de distribution en fonction du nombre de mots entre deux catégories d'interlocuteurs. Nous le calculons pour les trois appariements possibles (psychologue/schizophrène, psychologue/témoin, témoin/schizophrène) :

$$s = \frac{(p_1 - p_2)}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

où :

– $p = (n_1 p_1 + n_2 p_2)/(n_1 + n_2)$

– n_1 est le nombre de mots[5] prononcés par la première catégorie d'interlocuteurs,

– n_2 est le nombre de mots prononcés par la seconde catégorie d'interlocuteurs,

– p_1 est la proportion du phénomène attribuée à la première catégorie d'interlocuteurs,

– p_2 est la proportion du phénomène attribuée à la seconde catégorie d'interlocuteurs.

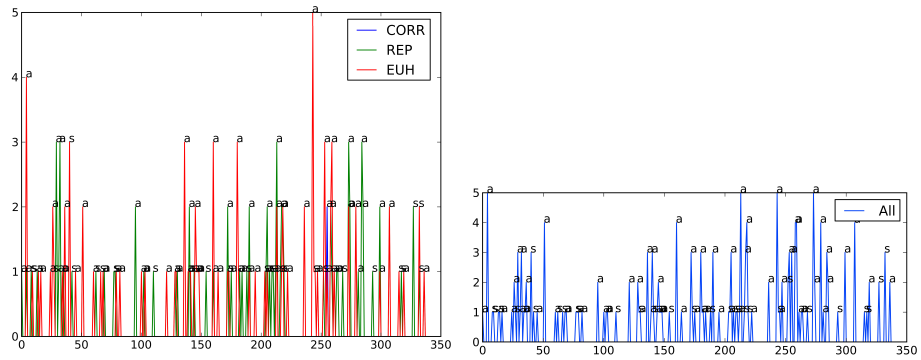
Cette mesure présente l'avantage de comparer des travaux sur les disfluences du français. Elle cherche à déterminer si une production p_α est significativement différente d'une autre p_β. Le résultat s suit une loi normale. L'hypothèse la plus simple est

5. Chaque token compte pour un mot (y compris dans les disfluences).

l'égalité entre les productions, hypothèse dite H_0. Nous tentons de rejeter H_0 si s est inférieur au quantile 2,5 ou supérieur au quantile 97,5 d'une loi normale, c'est-à-dire 1,96. La valeur trouvée doit donc être supérieure à 1,96 pour être considérée comme significative, avec un risque d'erreur de 5 %. Nous ne considérons pas les comparaisons multiples car nos résultats sont répartis sur trop peu de données.

5.2. *Analyse des disfluences*

Afin d'analyser les pratiques en disfluences, nous avons étudié d'une part la quantité de disfluences produites par chacun des groupes, et d'autre part la position des disfluences dans l'entretien.



(a) (b)

Figure 2 : Nombre d'étiquettes de disfluence par tour de parole pour un entretien du sous-corpus Ville 1. L'abscisse est la position du tour de parole dans l'entretien. Les tours de parole du psychologue sont notés par un s et ceux du schizophrène par un a.

La figure 2 présente un exemple des résultats obtenus, pour un patient (sous-corpus Ville1). Dans la première figure (a), une couleur est attribuée à chacune des trois étiquettes principales, l'axe des abscisses correspond à la position du tour de parole dans l'entretien et celui des ordonnées au nombre de disfluences dans ce tour de parole. Pour les points où l'ordonnée est différente de 0, une étiquette est apportée, a pour les tours de paroles du patient, et s pour le psychologue. La seconde figure présente les mêmes données, en affichant la somme du nombre d'étiquettes pour le même tour de parole. Ainsi, la première valeur significative est à 4 dans la figure a et à 5 dans la figure b. Dans ce tour de parole `Distagger` identifie 4 EUH et 1 REP.

Ces figures, même si elles n'apportent pas directement de résultats sur les disfluences des schizophrènes, nous ont permis d'identifier un schéma récurrent sur l'ensemble des entretiens. En effet, nous pouvons visualiser deux moments de stress où les disfluences augmentent chez le schizophrène. Le premier est en début d'entretien et

peut simplement s'entendre comme la tension dû à la rencontre avec le psychologue. Le second moment est plus difficile à interpréter, il intervient au bout de deux tiers de l'entretien, et ce, quelle qu'en soit la longueur. Nous pensons qu'il s'agit en fait d'un indice de fatigue que le psychologue intègre de manière implicite. Comme il ne s'agit pas de brutalement mettre fin à l'entretien, il accompagne le patient vers la fin de la rencontre dans le dernier tiers. Si on retrouve également des moments de stress dans les entretiens du groupe des témoins, cette régularité ne semble pas se confirmer. Elle appartiendrait bien au groupe des schizophrènes.

Au delà de cette première analyse, nous avons procédé à une évaluation quantitative systématique. Nous calculons sur l'ensemble des trois étiquettes $CORR$, REP et EUH, les fréquences d'apparition. Nous avons choisi de normaliser ces fréquences par rapport au nombre de mots prononcés en fonction de la catégorie de l'interlocuteur. Le tableau 3 rassemble toutes les données issues de l'analyse produite par `Distagger`. Nous présentons les résultats pour chacune des catégories (en sommant tous les résultats des psychologues, qu'ils soient devant un témoin ou un patient). Le total reprend la somme des valeurs de la catégorie d'interlocuteur.

	corpus Ville2			corpus Ville1		
	S	T	P	S	T	P
$CORR$	$0,0004$	$9e-05$	$0,0001$	$0,0013$	$0,0007$	$0,0006$
REP	$0,0125$	$0,0078$	$0,0067$	$0,0211$	$0,0134$	$0,0174$
EUH	$0,0190$	$0,0089$	$0,0073$	$0,0369$	$0,0326$	$0,0282$
$total$	$0,032$	$0,0168$	$0,0142$	$0,0595$	$0,0468$	$0,0463$

Tableau 3 : Répartition des étiquettes de `Distagger` dans les sous-corpus, normalisée par rapport au nombre de mots (T = témoins, S = schizophrène, P = psychologue).

La lecture du tableau montre que la production de disfluences des témoins et des psychologues sont du même ordre de grandeur : $1,68$ % et $1,42$ % pour le sous-corpus Ville2, et $4,68$ % et $4,63$ % pour le sous-corpus Ville1. Dans le même temps, les productions des schizophrènes sont supérieures : $3,2$ % et $5,95$ %. L'observation de la différence du pourcentage de disfluences entre les non-schizophrènes et les schizophrènes est alors relativement stable : $1,63$ % dans le sous-corpus Ville2 et $1,29$ % dans le sous-corpus Ville1.

La variabilité des mesures obtenues entre les deux sous-corpus peut paraître importante, mais elle nous semble venir de la qualité de la transcription. De plus, la répartition des sujets dans les deux sous-corpus est différente, ce qui peut aussi être une explication. Néanmoins, s'il n'est pas raisonnable de proposer le calcul d'un résultat pour l'ensemble du corpus, la constance de la différence de résultats conduit à notre conclusion.

Comme nous l'avons annoncé, nous avons calculé la significativité entre ces valeurs. Les résultats obtenus sont présentés dans le tableau 4.

	corpus Ville1	corpus Ville2
T - P	$0,42$	$3,23$
S - P	$10,68$	$19,42$
S - T	$10,28$	$16,04$

Tableau 4 : Significativité des disfluences entre les groupes d'interlocuteurs (T = témoins, S = schizophrène, P = psychologue).

Il apparaît que les différences entre les témoins et les psychologues sont faibles, voire non significatives, ce qui permet de rapprocher leurs comportements. Par contre, la significativité est importante (toujours supérieure à 10) dans les appariements qui comprennent des schizophrènes, ce qui nous conduit à conclure que le nombre de disfluences produites par des schizophrènes est significativement différent de celui des non-schizophrènes de l'expérimentation (psychologues et témoins).

5.3. *Analyse des catégories morpho-syntaxiques et lemmes*

L'objectif étant d'étudier la production langagière des schizophrènes, nous avons poursuivi les analyses en nous focalisant sur les catégories morpho-syntaxiques (*Part-of-Speech* - POS), ainsi que sur les lemmes. Ces objets nous intéressent en ce qu'ils devraient nous apporter des indices *a priori* sur la complexité de la production. Comme nous l'avons introduit dans la section 4.2, nous utilisons l'outil `MElt` avec un modèle adapté à l'oral pour annoter le corpus en catégories morpho-syntaxiques et en lemmes.

Une première analyse nous a conduit à calculer la distribution moyenne des étiquettes morpho-syntaxiques dans le corpus. Le tableau 5 rassemble les données relatives au nombre moyen d'étiquettes morpho-syntaxiques et au nombre moyen de tours de paroles dans le corpus. Nous avons calculé le ratio entre ces deux valeurs, ainsi que le nombre moyen d'étiquettes morpho-syntaxiques différentes dans chaque entretien. Nous pouvons observer que le nombre moyen de catégories par tour de parole est homogène dans le sous-corpus Ville1 contrairement au sous-corpus Ville2. Une analyse qualitative des données pour les témoins du sous-corpus Ville2 met en avant deux phénomènes : le fait que le psychologue est beaucoup moins intervenu que dans le cas des schizophrènes, et d'autre part, le fait que les échanges restent très courts par rapport à ceux réalisés avec les schizophrènes. Cela explique le chiffre de 31,78 pour les témoins, ainsi que l'augmentation observable chez les schizophrènes et la baisse chez les psychologues. Il est important de noter que le nombre moyen d'étiquettes morpho-syntaxiques différentes utilisées par chaque type d'interlocuteur est quant à lui très stable (variation de 35 à 40, mais avec une moyenne globale cohérente dans chaque sous-corpus). Il semble que la quantité d'étiquettes morpho-syntaxiques ne soit pas ici discriminante.

		VER	ADJ	ADV	NOM	DET	PRP	PRO	AUT	Ratio	Diff
Ville1	T	10,98	40	524	83	711	297	234	218	617	785
	S	13,12	38	416	61	537	238	183	190	497	632
	P	13,02	39	575	91	795	318	247	247	634	738
Ville2	T	31,78	35	247	45	173	199	146	141	265	243
	S	17,32	37	378	58	243	266	201	182	440	498
	P	9,42	35	192	27	135	117	86	80	231	194

Tableau 5 : Ratio moyen du nombre de catégories par rapport au nombre de tours de parole par entretien et nombre moyen d'étiquettes différentes par entretien, et répartition moyenne des catégories morpho-syntaxiques en grandes catégories : VERbe, ADVerbe, NOM, DÉTerminant, PRÉposition, PROnoms et AUTres (T = témoins, S = schizophrène, P = psychologue).

Nous nous sommes ensuite intéressés à une première forme d'étude qualitative des étiquettes morpho-syntaxiques en les classant en grandes catégories : verbe, adverbe, nom, déterminant, préposition, pronoms et autres. Le tableau 5 présente les valeurs moyennes par type d'interlocuteurs sur ces grandes catégories. Par exemple, la première valeur indique que les témoins ont utilisé 524 verbes en moyenne dans leur entretien dans le sous-corpus Ville1. Il convient d'utiliser des catégories plus fines, ce que nous laissons à une étude ultérieure.

La lecture de ces seules valeurs reste difficile, aussi avons nous produit une représentation graphique de cette répartition. Il apparaît que la répartition est homogène sur l'ensemble des catégories, quel que soit le type d'interlocuteur. Plus précisément, si certains entretiens contiennent une répartition divergente à celle de leur groupe, elle ne peut pas être rapprochée d'un autre comportement. La figure 3 présente la répartition en grandes catégories des POS pour les témoins et les shyzophrènes du sous-corpus Ville 1.

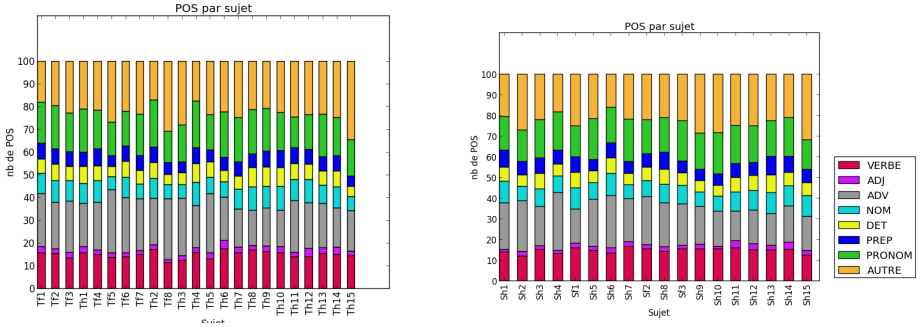


Figure 3 : Répartition des catégories morpho-syntaxiques dans le sous-corpus Ville1 pour les témoins (gauche) et les schizophrènes (droite).

	Ville1		Ville2									
					H		F		avec trait.		sans trait.	
	RL	DL	RL	DL	RL	DL	RL	DL	RL	DL	RL	DL
T	0,04	0.68	0,11	0.73	0,15	0.76	0,14	0.74				
S	0,05	0.69	0,06	0.70	0,07	0.72	0,08	0.71	0,06	0.71	0,10	0.72
P	0,02	0.64	0,06	0.68								

Tableau 6 : Richesse lexicale (RL) et diversité lexicale (DL) selon les sous-corpus, avec données en sexe pour le sous-corpus ville2 et en *avec traitement* ou *sans* pour les schizophrènes du même sous-corpus (T = témoins, S = schizophrène, P = psychologue).

À partir de l'information extraite sur les lemmes, les formes et le nombre de mots, nous avons calculé la richesse lexicale et la diversité lexicale, comme introduites dans la section 4.3. Le tableau 6 montre une richesse lexicale équivalente chez les témoins et les schizophrènes dans le sous-corpus Ville1 (0,04 et 0,05, respectivement), ce qui est confirmé par la DL (0,68 et 0,69, respectivement). En revanche, dans le sous-corpus Ville2, les témoins semblent avoir une richesse lexicale supérieure (0,11 *vs* 0,06 pour les schizophrènes). Or, le sous-corpus Ville2 comprend des paroles de schizophrènes qui ne sont pas sous traitement. La comparaison de leur richesse lexicale respective montre que celle des schizophrènes sans traitement est proche de celle des témoins (0,10 *vs* 0,11), alors que celle des schizophrènes sous traitement est bien inférieure (0,06). Mais cela n'est pas confirmé par la DL (0,71 et 0,73).

Parmi les schizophrènes sans traitement, une très large majorité (6 sur 7) sont des femmes. Nous avons donc voulu vérifier si la différence était liée au sexe, mais comme le montrent les résultats, ce n'est pas le cas (0,15 *vs* 0,14 chez les témoins et 0,07 *vs* 0,08 chez les schizophrènes). Une explication possible des différences observées pourrait résider dans le déséquilibre important du sous-corpus Ville2, qui ne comprend que 8 témoins et 7 schizophrènes sans traitement pour 24 schizophrènes sous traitement. La mesure choisie est en effet sensible à la taille du corpus (au même titre que le TTR, voir, entre autres (Richards, 1987)).

La richesse lexicale des psychologues est limitée dans le sous-corpus Ville1 (RL de 0,02 et DL de 0,64, la plus faible valeur calculée pour DL) ce qui s'explique facilement par le type d'entretien réalisé. Le psychologue ne fait que maintenir l'interaction et n'emploie donc qu'un vocabulaire limité et répétitif. Dans le sous-corpus Ville2 la richesse lexicale des psychologues est équivalente à celle des schizophrènes (RL de 0,06 et DL de 0,68) et bien inférieure à celle des témoins (RL de 0,11 et DL de 0,73), ce qui semble être cohérent avec le sous-corpus Ville1 si l'on prend en compte le déséquilibre entre témoins.

Au delà de cette analyse directe sur les données, nous avons calculé la significativité pour tous ces groupes. Le tableau 7 présente l'ensemble de ces résultats. Les va-

leurs non significatives, inférieures à 1,96, apparaissent en rouge et les valeurs moins significatives, inférieures à 10, apparaissent en bleu.

	Ville1		Ville2									
					H		F		avec trait.		sans trait.	
	RL	DL	RL	DL	RL	DL	RL	DL	RL	DL	RL	DL
T - P	21.38	18.24	21.14	11.17	21.70	7.67	8.44	7.08	17.57	8.01	4.76	5.92
S - P	24.27	20.08	0.14	7.85	1.42	6.97	4.79	2.53	0.61	5.52	0.46	3.90
T - S	4.72	4.28	22.70	6.95	22.19	4.12	13.79	5.51	18.56	4.77	4.47	2.29

Tableau 7 : Significativité de RL et DL pour les appariement T - S - P du tableau 6 (T = témoins, S = schizophrène, P = psychologue).

La première hypothèse d'une richesse et d'une diversité lexicales équivalentes entre les témoins et les schizophrènes se confirme avec des significativités de 4,72 et 4,28, à comparer avec 21,38 et 21, 27 d'une part et 18,24 et 20,08 d'autre part. Les témoins de Ville2 ont bien une richesse lexicale supérieure. La significativité entre les schizophrènes et les psychologues est très basse, 0,14, alors qu'en comparaison des témoins elle est très élevée (21,14 avec les psychologues et 22,70 avec les schizophrènes).

En ce qui concerne l'hypothèse de l'impact du traitement, la significativité poursuit sur cette interprétation. En effet, la différence avec les psychologues de ce groupe est très faible (significativité de 0,46) et constante dans la comparaison avec le groupe des témoins (4,76 entre les témoins et les psychologues et 4,47 entre les témoins et les schizophrènes sans traitement). Pour être plus précis il conviendrait de conclure que ce groupe n'a pas un comportement langagier commun (au sens où il reste significativement différent de celui du groupe témoin), mais proche du comportement attendu pour un entretien de type semi-dirigé.

Enfin, l'analyse sur le sexe est très clairement confirmée pour les hommes (très grande proximité entre les schizophrènes et les psychologues avec une significativité de 1,42, et significativité de 21,70 et 22,19 avec les psychologues et les témoins, respectivement). Les valeurs sont moins explicitent pour les femmes. Bien que les différences soient moindres, leur ordre reste le même.

L'ensemble des analyses proposées est confirmé par le calcul de la significativité. Il apparaît que les schizophrènes n'ont pas de comportement spécifique identifié autour des catégories morpho-syntaxiques et des lemmes.

5.4. *Analyse textométrique*

Nous avons souhaité approfondir l'analyse en allant au-delà de l'aspect purement quantitatif et avons pour cela utilisé l'outil de textométrie `TXM` (Heiden *et al.*, 2010) (voir section 4.3).

Nous avons considéré les deux sous-corpus Ville1 et Ville2 séparément, reflétant ainsi leurs différences, et avons créé pour chacun une partition selon le type de locuteur (psychologues, témoins, schizophrènes). Nous avons ensuite calculé les spécificités (Lafon, 1980) de chaque lemme prononcé et extrait ceux ayant une spécificité supérieure à 4, ce qui correspond à une spécificité à la fois supérieure à la zone de banalité et affichable. Les résultats de ces calculs pour les schizophrènes du sous-corpus Ville1 sont présentés dans la figure 4.

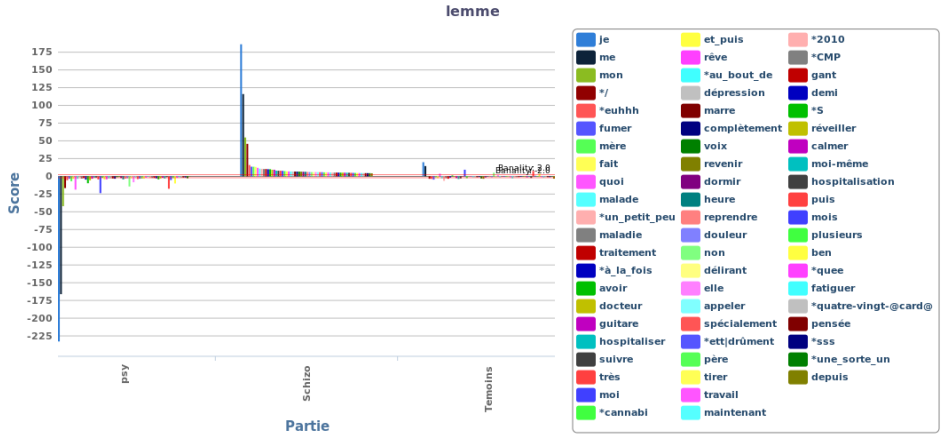


Figure 4 : Lemmes ayant une spécificité supérieure à 4 pour les schizophrènes du sous-corpus Ville1.

Au-delà des pronoms *je, me, mon*, dont l'usage extensif est dû au type d'entretien, on peut voir apparaître très rapidement (en position 5 et 3) les *euh* (ou *euhhh*), qui participent aux disfluences[6]. Cet indice semble confirmer ce que nous avons identifié par ailleurs : les patients schizophrènes tendent à produire davantage de disfluences que les témoins.

Si les données du sous-corpus Ville2 ne sont guère plus informatives, celles du sous-corpus Ville1 montrent une fréquence élevée du verbe *fumer* (position 6) et, encore plus intéressant, du mot *cannabis* (position 22). Nous avons accédé aux contextes des occurrences de ces mots et avons fait plusieurs constatations : d'une part, ce n'est pas le psychologue qui incite les patients à les prononcer (ils le font spontanément), d'autre part ils sont prononcés par 6 patients différents (sur un total de 18), auxquels il faut ajouter un patient qui parle de *shite*. Une explication pourrait être que, ces patients étant tous en remédiation, le personnel médical leur a probablement parlé des risques liés à la consommation de cannabis et qu'ils ont fait le lien avec leur pathologie. En revanche, *mère*, qui apparaît en position 7 dans la liste n'est pas significatif,

6. Rappelons que les lemmes commençant par une étoile correspondent à des mots inconnus du tagger `MElt`.

puisqu'il est en fait employé par un seul patient (SH10), comme le montre un calcul de spécificités localisé aux patients schizophrènes (voir figure 5).

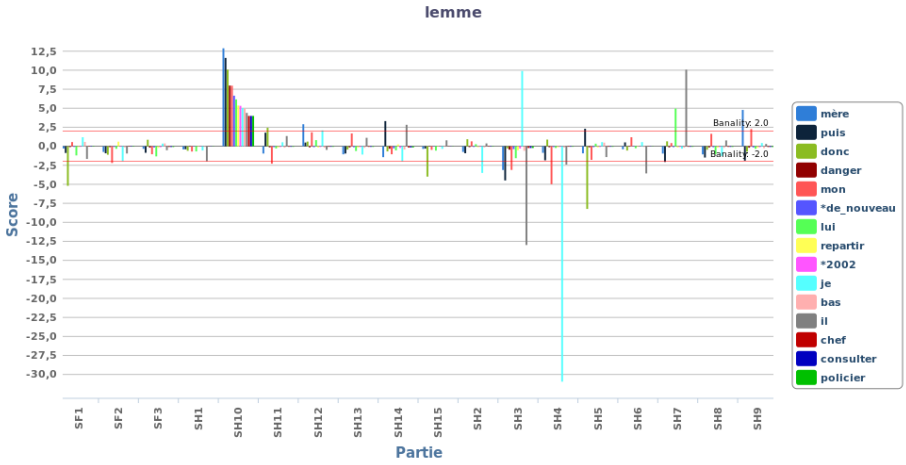


Figure 5 : Spécificités du lemme *mère* parmi les patients schizophrènes.

Les termes liés à la maladie sont évidemment très présents et sont prononcés par différents patients. On trouve ainsi *malade*, *maladie*, *traitement* (qu'il faut *suivre*), *douleur* (qu'il faut *calmer*), *docteur*, *hospitaliser*, *hospitalisation*. D'autres semblent plus spécifiques à la schizophrénie : *dépression*, *voix*, *délirant*.

Cette analyse montre l'intérêt d'utiliser un outil comme `TXM`, afin, non seulement d'obtenir des résultats statistiquement fiables, mais également de pouvoir accéder au contexte d'énonciation pour valider ou invalider les hypothèses émises.

6. Biais potentiels des expériences

Il nous apparaît nécessaire de revenir sur différents aspects de l'étude qui participent à en biaiser les résultats. Nous pensons que ces biais ne remettent pas cause l'ensemble de la méthodologie, ni les résultats, mais ils peuvent influer sur eux. Les biais potentiels concernant l'analyse des disfluences ont été présentés de manière extensive dans (Amblard et Fort, 2014). Nous ne reviendrons que sur les biais communs à tout le corpus.

Une première série de biais réside dans la constitution des sous-corpus. La méthodologie ayant été largement éprouvée et améliorée pour le sous-corpus Ville1 par rapport au sous-corpus Ville2, les résultats sont plus précis sur ces données, en particulier en ce qui concerne la transcription, qui a été largement normalisée à la suite de la première collecte.

Par ailleurs, un élément important est la répartition des sujets dans l'étude. Ainsi, le sous-corpus Ville2 ne contient que 8 témoins pour 31 schizophrènes, ce qui le rend

très déséquilibré. Nous disposons d'une autre série d'entretiens avec des témoins, réalisée à la même époque, mais dans des conditions différentes (conversations sans psychologue). Nous avons choisi d'écarter ces données qui apparaissent comme trop déviantes par rapport au reste du corpus.

	Schizophrènes			Témoins		
		femme	homme		femme	homme
âge	28,89	30	28,66	23,22	22,37	23,66
QI	95,17	98,33	94,53	103,70	105,5	102,73
années d'études	12,41	13	12,28	13,17	13	13,26

Tableau 8 : Moyennes des QI, du nombre d'années d'études et de l'âge des participants au corpus Ville1.

Une analyse plus fine des sujets intégrés dans l'étude montre aussi des différences sur les capacités neuro-cognitives et l'âge des sujets. Le tableau 8 rassemble les données moyennes relatives au QI, au nombre d'années d'études et à l'âge. On constate que les sujets schizophrènes sont significativement plus âgés que les sujets du groupe témoin avec un âge moyen de 28,89 contre 23,22. Le test de Student confirme que cette différence est significative (p=0,0058). Par ailleurs, ils ont un QI moyen de 95,17, donc inférieur au QI moyen des témoins qui est de 103,7. La significativité issue du test de Student nous indique que cette différence fait sens (p=0,0203). Il est intéressant de noter que le nombre d'années d'études est quant à lui plutôt constant sur l'ensemble du corpus (12,41 pour les schizophrènes et 13,17 pour les témoins).

Enfin, nous avons identifié un biais important, mais sur lequel nous ne pouvons que très difficilement influer. Les patients schizophrènes sont hospitalisés, donc sous traitement. Comme nous l'avons explicité précédemment, une très faible partie du corpus de Ville2 est constituée de patients sans traitement médicamenteux, mais, d'une part, il est très compliqué d'accéder à ces patients, et d'autre part, il n'est pas toujours évident d'interagir avec eux. Les entretiens sont généralement plus courts. Les données qui en sont issues restent pertinentes pour notre étude. La constitution d'un sous-groupe de sujets sans traitement médicamenteux, de taille équivalente au sous-groupe sous traitement n'est cependant pas réaliste.

Par ailleurs, nous n'avons pas mis en perspective la répartition de l'avancé dans la maladie des patients. Nous n'avons pas pris en compte la durée des hospitalisations, ni le fait que certains patients sont en remédiation. Pour cette dernière catégorie, cela influe directement sur leur relation à leur traitement médicamenteux (Chlorpromazine à Ville1 et neuroleptiques non spécifiés à Ville2). Les patients en remédiation et remédiation avancée sont stabilisés et prêts à être autonomisés en dehors de l'institution hospitalière. Les doses de neuroleptiques administrées peuvent être considérées comme minimales, contrairement aux patients récemment admis après une urgence qui, eux, reçoivent des doses beaucoup plus élevées.

L'influence des médicaments sur l'étude des patients schizophrènes reste une question commune à toute étude sur cette pathologie. (Levy, 1968) a identifié des effets né-

gatifs (en l'occurrence, une baisse des performances) de la Chlorpromazine sur la syntaxe de quatre patients schizophrènes en calculant le ratio du nombre de propositions subordonnées produites sur la totalité des propositions produites. En outre, cet antipsychotique semble provoquer des bégaiements (Ward, 2008). Cependant, (Goldman-Eisler *et al.*, 1965) a montré (sur des sujets non schizophrènes) que les effets de cette même molécule sur les temps de pause du locuteur sont très variables selon les individus et qu'un temps de pause supérieur permet au groupe testé de générer des structures verbales complexes, comme chez les témoins. Pour ajouter à ces incertitudes, (Kremen *et al.*, 2003) ont montré que des patients bipolaires sous antipsychotiques (dont fait partie la Chlorpromazine) présentent une meilleure fluence sémantique que les témoins. Une réserve sur ce contre-argument doit être apportée sur la capacité des médicaments qui ne cessent de faire des progrès. Les effets secondaires ont été considérablement réduits dans les 50 dernières années. Il reste particulièrement délicat de séparer ce qui appartient à l'influence du traitement médicamenteux de ce qui est propre à celle de la pathologie.

7. Conclusions et perspectives

Cet article s'inscrit dans un projet plus large sur l'étude des pratiques langagières des schizophrènes. Il a été montré que ces derniers présentaient un dysfonctionnement dans la gestion de la planification du discours (Musiol et Trognon, 1996 ; Verhaegen, 2007). L'interprétation de ce trouble se situe au niveau sémantico-pragmatique (Rebuschi *et al.*, 2013) et (Musiol *et al.*, 2013), ce qui a conduit à proposer des modélisation inspirée de la SDRT (*Segmented Discourse Representation Theory*) (Asher et Lascarides, 2003). Cette manifestation nous renseigne sur le fonctionnement cognitif, en particulier dans son rapport à l'expression de la pensée par le langage. Il nous est apparu nécessaire d'interroger d'autres niveaux d'analyse linguistique. Une étude manuelle n'étant pas réaliste, nous avons choisi de travailler à partir des résultats fournis par des outils de TAL existants.

Nous avons, dans un premier temps, mis en avant un usage spécifique des disfluences chez les sujets schizophrènes grâce à l'outil `Distagger`. En effet, les schizophrènes produisent, respectivement dans chaque corpus, 1,63 % et 1,29 % plus de disfluences (par rapport au nombre de mots) que des sujets témoins ou les psychologues. Ce résultat est confirmé par un calcul de significativité.

Dans un deuxième temps, nous nous sommes intéressés aux productions en catégories morpho-syntaxiques et en lemmes. Il apparaît en effet régulièrement dans la littérature que les patients schizophrènes auraient une capacité morpho-syntaxique et une diversité lexicale réduites par rapport aux sujets témoins. Cependant, notre étude tend à montrer le contraire. Les sujets schizophrènes produisent autant de catégories morpho-syntaxiques que les autres, avec une diversité similaire. Par ailleurs, leur richesse lexicale est également similaire. Plus exactement, à partir de nos données, nous ne pouvons pas identifier de sous-classe particulière, certains sujets ayant des comportements singuliers pour chaque type d'interlocuteurs. Ces indices nous permettent de

conclure que ces niveaux n'interviennent pas dans la défaillance cognitive que nous cherchons à circonscrire. Il ne s'agit alors pas ici d'une défaillance de la capacité, mais bien une défaillance de la gestion de l'expression. Le résultat précédent sur les disfluences s'inscrit tout à fait dans cette perspective.

Nous nous inscrivons donc en faux contre les précédents résultats. Il faut noter que notre corpus est significativement plus important que tous ceux utilisés dans les études auxquelles il est fait mention, qui ne dépassent jamais plus de 20 patients, alors que nous en étudions 49.

Nous avons poursuivi par une étude plus qualitative de la diversité lexicale avec l'outil `TXM`. Nous avons pu mettre en avant certaines thématiques particulières qui peuvent s'interpréter relativement facilement en fonction du contexte des entretiens. En effet, les patients schizophrènes sont en milieu hospitalier et ont une habitude des entretiens semi-dirigés. Ils ont tendance à aborder des thématiques plus classiques pour la psychologie et la psychiatrie comme leur rapport aux médicaments, à leurs dépendances, ou à leur famille.

Les prolongements de nos travaux sont de deux ordres. D'un côté nous souhaitons revenir sur le niveau de granularité de notre étude et de l'autre nous souhaitons inscrire ces résultats dans une perspective plus globale. En effet, pour les disfluences, il conviendrait d'étudier leurs lieux d'apparition à l'intérieur des tours de parole, leur dynamique dans l'interaction et de proposer des catégories plus précises (disfluences combinées, notamment). De manière similaire, pour les POS et les lemmes, il serait intéressant de regarder des catégories plus fines, en particulier concernant les verbes.

Un autre volet de l'analyse s'intéressera plus particulièrement à corréler les défaillances identifiées à la manifestation d'indices cognitifs. Bien évidemment il conviendra de les associer aux différents tests neuro-cognitifs, mais plus certainement aux comportements enregistrés par le double système oculométrique (*Eye-tracking*), ainsi qu'aux enregistrements de l'activité de l'encéphale (EEG).

Les outils de TAL sont pour nous les seuls à pouvoir proposer une cartographie précise de la manifestation de ces troubles. Ils nous permettent d'une part de produire une ressource normalisée riche en méta-données (dont le manque et l'importance sont mis en valeur dans (Ghio *et al.*, 2006)). Cependant, il apparaît très complexe de parvenir à constituer et gérer une telle ressource qui pose de nombreux problèmes éthiques.

Nous souhaiterions mettre ce corpus à disposition des chercheur(e)s, au moins une version manuellement anonymisée et randomisée par tour de parole, mais la non planification de cette étape dans les protocoles validés par la CPP rend cela impossible.

Remerciements

Nous tenons à remercier les relecteurs de la revue, qui, grâce à leurs remarques constructives et détaillées, nous ont permis d'améliorer de manière significative (non calculée ici) la clarté de cet article.

8. Bibliographie

Allauzen A., Bonneau-Maynard H., « Training and Evaluation of POS Taggers on the French MULTITAG Corpus », *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Maroc, mai, 2008.

Amblard M., Fort K., « Étude quantitative des disfluences dans le discours de schizophrènes : automatiser pour limiter les biais », *TALN - Traitement Automatique des Langues Naturelles*, Marseille, France, p. 292-303, July, 2014.

Asher N., Lascarides A., *Logics of Conversation*, Studies in Natural Language Processing, Cambridge University Press, 2003.

Benzitoun C., Fort K., Sagot B., « TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe », *Traitement Automatique des Langues Naturelles (TALN)*, Grenoble, France, p. 99-112, juin, 2012.

Besche C., Passerieux C., Segui J., Mesure G., Hardy-Baylé M.-C., « Étude du traitement des informations contextuelles syntaxiques lors d'une tâche de décision lexicale chez des sujets schizophrènes », *Revue Canadienne de Psychiatrie*, vol. 41, n° 9, p. 587-594, 1996.

Blanche-Benveniste C., Jeanjean C., *Le Francais parlé. Transcription et édition*, Didier Érudition, Paris, France, 1987.

Chaika E., « A linguist looks at "schizophrenic" language », *Brain and Language*, vol. 1, n° 3, p. 257-276, juillet, 1974.

Constant M., Dister A., « Automatic detection of disfluencies in speech transcriptions », *in* M. Pettorino, A. Giannini, I. Chiari, F. Dovetto (eds), *Spoken Communication*, vol. 1, Cambridge Scholars Publishing, p. 259-272, 2010.

de Mareüil P. B., Adda G., Adda-Decker M., Barras C., Habert B., Paroubek P., « Une étude quantitative des marqueurs discursifs, disfluences et chevauchements de parole dans des interviews politiques », *TIPA. Travaux interdisciplinaires sur la parole et le langage*, 2013.

DeLisi L. E., « Speech disorder in schizophrenia : Review of the literature and exploration of its relation to the uniquely human capacity for language. », *Schizophrenia Bulletin*, vol. 27, n° 3, p. 481-496, 2001.

Denis P., Sagot B., « Coupling an Annotated Corpus and a Morphosyntactic Lexicon for State-of-the-Art POS Tagging with Less Human Effort », *Pacific Asia Conference on Language Information and Computing (PACLIC)*, Hong-Kong, 2009.

Eshkol-Taravella I., Kanaan-Caillol L., Baude O., Dugua C., Maurel D., « Procédure d'anonymisation et traitement automatique : l'expérience d'ESLO », , Journée d'études ATALA, Ethique et TAL, 2014.

Feldstein S., « The relationship of interpersonal involvement and affectiveness of content to the verbal communication of schizophrenic patients », *Journal of Abnormal and Social Psychology*, vol. 64, p. 39-45, 1962.

Fromkin V. A., « A linguist looks at "a linguist looks at 'schizophrenic language'" », *Brain and Language*, vol. 2, n° 0, p. 498 - 503, 1975.

Ghio A., Teston B., Viallet F., Jankowski L., Purson A., Duez D., Locco J., Legou T., Pinto S., Marchal A., Giovanni A., Robert D., Révis J., Fredouille C., Bonastre J.-F., Pouchoulin G., Nguyen N., « Corpus de parole pathologique, état d'avancement et enjeux méthodologiques », *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence (TIPA)*, vol. 25, p. 109-126, 2006.

Goldman-Eisler F., Skarbek A., Henderson A., « The effect of chorpromazine on speech behaviour », *Psychopharmacologia*, vol. 7, n° 3, p. 220-229, 1965.

Grouin C., Anonymisation de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique, Thèse de doctorat, Université Pierre et Marie Curie, Paris, France, Juin, 2013.

Heiden S., Magué J.-P., Pincemin B., « TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement », *in* L. G. Sergio Bolasco, Isabella Chiari (ed.), *10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*, vol. 2-3, Edizioni Universitarie di Lettere Economia Diritto, Rome, Italie, p. 1021-1032, 2010.

Kremen W. S., Seidman L. J., Faraone S. V., Tsuang M. T., « Is there disproportionate impairment or phonemic fluency in schizophrenia ? », *Journal of the International Neuropsychological Society*, vol. 9, p. 79-88, 2003.

Lafon P., « Sur la variabilité de la fréquence des formes dans un corpus », *Mots : Saussure, Zipf, Lagado, des méthodes, des calculs, des doutes et le vocabulaire de quelques textes politiques*, n° 1, p. 127-165, octobre, 1980.

Landre N., Taylor M., Kearns K., « Language functioning in schizophrenia and aphasic patients », *Neuropsychiatry, Neuropsychology, and Behavioral Neurology*, 1992.

Levy R., « The effect of chlorpromazine on sentence structure of schizophrenic patients », *Psychopharmacologia*, vol. 13, n° 5, p. 426-432, 1968.

Maher B., « The Language of Schizophrenia : A Review and Interpretation », *The British Journal of Psychiatry*, vol. 120, p. 3-17, 1972.

Meystre S. M., Friedlin F. J., South B. R., Shen S., Samore M. H., « Automatic de-identification of textual documents in the electronic health record : a review of recent research », *BMC Med Res Methodol*, 2010.

Musiol M., Amblard M., Rebuschi M., « Approche sémantico-formelle des troubles du discours : les conditions de la saisie de leurs aspects pyscholinguistiques. », *27ème Congrès International de Linguistique et de Philologie Romanes*, Nancy, France, juillet, 2013.

Musiol M., Trognon A., « L'accomplissement interactionnel du trouble schizophrénique », *Raisons Pratiques*, vol. 7, p. 179-209, 1996.

Rebuschi M., Amblard M., Musiol M., « Using SDRT to analyze pathological conversations. Logicality, rationality and pragmatic deviances », *in* M. Rebuschi, M. Batt, G. Heinzmann, F. Lihoreau, M. Musiol, A. Trognon (eds), *Interdisciplinary Works in Logic, Epistemology, Psychology and Linguistics : Dialogue, Rationality, and Formalism*, Logic, Argumentation & Reasoning, Springer, p. 1-24, 2013.

Richards B., « Type/Token Ratios : what do they really tell us ? », *Journal of Child Language*, vol. 14, p. 201-209, 6, 1987.

Schmid H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », *Proceedings of International Conference on New Methods in Language Processing*, 1994.

Verhaegen F., Psychopathologie cognitive des processus intentionnels schizophréniques dans l'interaction verbale, PhD thesis, Université Nancy 2, France, 2007.

Ward D., *Stuttering and Cluttering : Frameworks for Understanding and Treatment*, Taylor & Francis, 2008.