Étude quantitative des disfluences dans le discours de schizophrènes : automatiser pour limiter les biais

Maxime Amblard, Karën Fort
Université de Lorraine, LORIA, INRIA, CNRS
UMR 7503, Vandœuvre-lès-Nancy 54500 France
{maxime.amblard, karen.fort}@loria.fr

Résumé. Nous présentons dans cet article les résultats d'expériences que nous avons menées concernant les disfluences dans le discours de patients schizophrènes (en remédiation). Ces expériences ont eu lieu dans le cadre d'une étude plus large recouvrant d'autres niveaux d'analyse linguistique, qui devraient aider à l'identification d'indices linguistiques conduisant au diagnostic de schizophrénie. Cette étude fait la part belle aux outils de traitement automatique des langues qui permettent le traitement rapide de grandes masses de données textuelles (ici, plus de 375 000 mots). La première phase de l'étude, que nous présentons ici, a confirmé la corrélation entre l'état schizophrène et le nombre de disfluences présentes dans le discours.

Abstract. We present in this article the results of experiments we led concerning disfluencies in the discourse of schizophrenic patients (in remediation). These experiments are part of a larger study dealing with other levels of linguistic analysis, that could eventually help identifying clues leading to the diagnostic of the disease. This study largely relies on natural language processing tools, which allow for the rapid processing of massive textual data (here, more than 375,000 words). The first phase of the study, which we present here, confirmed the correlation between schizophrenia and the number of disfluences appearing in the discourse.

Mots-clés : discours pathologique, schizophrénie, disfluences.

Keywords: pathological discourse, schizophrenia, disfluencies.

1 Introduction

1.1 Contexte et motivations de l'étude

Cette étude participe d'un projet plus large portant sur les pratiques langagières chez les schizophrènes en situation d'entretiens semi-dirigés par un psychologue. Cette étude s'inscrit dans la continuité des premiers travaux de Chaika (1974) et Fromkin (1975), qui cherchaient à mettre en avant les particularités langagières chez les schizophrènes.

Plusieurs aspects sont ainsi étudiés, notamment les capacités neuro-cognitives par une série de tests, le comportement oculomoteur du patient par une série d'enregistrements par oculomètre (*eye-tracker*), l'activité encéphale par des enregistrements par électro-encéphalographe (EEG) et la pratique langagière par l'étude linguistique des entretiens. Dans cette partie du projet, nous nous concentrons sur le dernier aspect et laissons donc de côté les autres mesures. Cependant, il conviendra dans une phase ultérieure de revenir sur l'ensemble des analyses pour identifier des corrélations spécifiques.

Concernant la partie linguistique du projet, nous nous basons sur un résultat de psycho-linguistique mettant en avant des usages pathologiques de la langue chez les schizophrènes, au travers de la notion de discontinuités pragmatiques décisives (Musiol & Trognon, 1996; Verhaegen, 2007). Rebuschi *et al.* (2013) et Musiol *et al.* (2013) ont montré que dans la succession des focus thématiques de la conversation, les schizophrènes rejouent une ambiguïté linguistique précédemment introduite, rendant l'interprétation pragmatique et rhétorique impossible. Comme eux, nous souhaitons produire des analyses formelles des extraits discontinus, afin d'en donner une interprétation dans un modèle formel du type SDRT (Asher & Lascarides, 2003) (*Segmented Discourse Representation Theory*), extension à la rhétorique et à la pragmatique de la DRT (Kamp & Reyle, 1993) (*Discourse Representation Theory*). Par ailleurs, il apparaît nécessaire de discuter les

règles du cadre formel car les extraits de dialogue nécessitent l'usage de règles non conventionnelles pour rattraper la construction de telles structures.

Le projet général cherche à ré-interroger ces résultats sur un corpus plus large et sur un faisceaux d'indices diversifiés. D'où, en particulier, l'utilisation des oculomètres et des EEG dans les protocoles. Nous cherchons à interroger également d'autres niveaux linguistiques, en proposant une annotation multi-niveaux de la ressource.

Pour ce faire, la transcription, qui est la clé de voûte de l'ensemble du projet, doit être de qualité. Les outils de transcription automatique que nous avons pu tester ont donné des résultats insuffisants. Elle a donc été réalisé manuellement. Nous avons défini un guide d'annotation précis, dans la tradition de Blanche-Benveniste & Jeanjean (1987). Cependant, l'une des difficultés du projet réside dans la nature des sujets, qui implique une gestion stricte de l'anonymat (voir section 2.3), dont la conséquence est que nous devons minimiser le nombre de personnes ayant accès aux données non transformées. Nous ne disposons donc pour l'instant que d'une seule version des transcriptions, ce qui ne nous permet pas de les évaluer correctement en calculant un accord inter-annotateur.

À partir de ces transcriptions, plusieurs autres annotations vont être proposées, dont une partie va être produite par des outils de traitement automatique des langues (TAL), et une autre par des humains (autre que les transcripteurs). Nous présentons ici le premier niveau d'annotation du corpus, l'annotation en disfluences, réalisée grâce à l'outil `Distagger` (Constant & Dister, 2010). Outre son utilité intrinsèque, l'annotation en disfluences permet de normaliser les corpus avant d'y appliquer des analyseurs syntaxiques ou sémantiques.

1.2 Travaux précédents

Les travaux précédents menés sur le discours des schizophrènes ont donné peu de résultats concernant les disfluences et ces résultats ne sont souvent qu'un élément accessoire de l'expérience décrite. Ainsi, Feldstein (1962) a travaillé sur l'impact du type de contenu d'éléments à commenter (affectif ou non) et a, ce faisant, montré que les perturbations du discours (*speech disturbances*) étaient plus élevées chez les schizophrènes [1]. Ces résultats sont confirmés par la méta étude de Maher (Maher, 1972). Plus récemment, Kremen *et al.* (2003) ont montré, dans le cadre d'une étude concernant la comparaison entre fluence phonémique et fluence sémantique chez les schizophrènes, que ceux-ci ont une fluence verbale (quel que soit son type) légèrement dégradée par rapport aux témoins et aux patients bipolaires [2].

Si les résultats semblent concorder, il n'existe à notre connaissance aucune étude publiée à ce sujet concernant des patients francophones. Par ailleurs, la manière dont les données précédentes ont été annotées ou notées n'est jamais précisée, mais on peut aisément supposer qu'elles l'ont été par des humains, ce qui, en l'absence d'accord inter-annotateurs, est évidemment source de biais [3]. L'étude que nous proposons se singularise donc par l'utilisation d'outils de TAL.

Le présent article est organisé de la manière suivante : nous détaillons dans la section 2 la constitution du corpus et ses implications dans l'étude, tant sur la couverture qu'il propose que sur la délicatesse avec laquelle il est nécessaire de manipuler les données ; puis nous présentons l'outil utilisé pour produire les annotations en disfluences et le protocole d'expérimentation dans la section 3 ; à partir de ces expérimentations, nous analysons dans la section 4 les résultats obtenus, leur significativité et les biais potentiels de l'étude ; enfin, nous présentons les travaux à venir dans la conclusion.

2 Difficultés de constitution du corpus

2.1 Présentation du corpus

Le corpus utilisé pour cette étude est constitué de transcriptions d'entretiens. L'étude fait intervenir 79 sujets, 48 schizophrènes et 31 témoins. Les entretiens ont été réalisés par des psychologues, en milieu hospitalier. Deux recueils de données ont pu être réalisés, dans des unités médicales spécialisées : le premier à Ville1 [4], par deux psychologues, et le second à Ville2, par une seule psychologue.

Le sous-corpus Ville1 a été constitué au second semestre 2013. Il est composé de 18 patients diagnostiqués schizophrènes

1. Cette étude a impliqué 30 schizophrènes et 30 témoins.
2. Cette étude a impliqué 83 schizophrènes, 15 patients bipolaires et 83 témoins.
3. Ce même biais affecte nos transcriptions, mais nos annotateurs ne savaient pas à quoi celles-ci allaient servir, ce qui limite l'impact du biais.
4. Nous avons anonymisé les noms de villes par respect pour la confiance des patients 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 30 patients diagnostiqués schizophrènes en remédiation et sous traitement, à l'exception de sept d'entre eux (qui n'étaient pas sous traitement), et de huit témoins. Le tableau 1 présente la ventilation des sujets en fonction de leur type (schizophrène ou témoin) et de leur sexe.

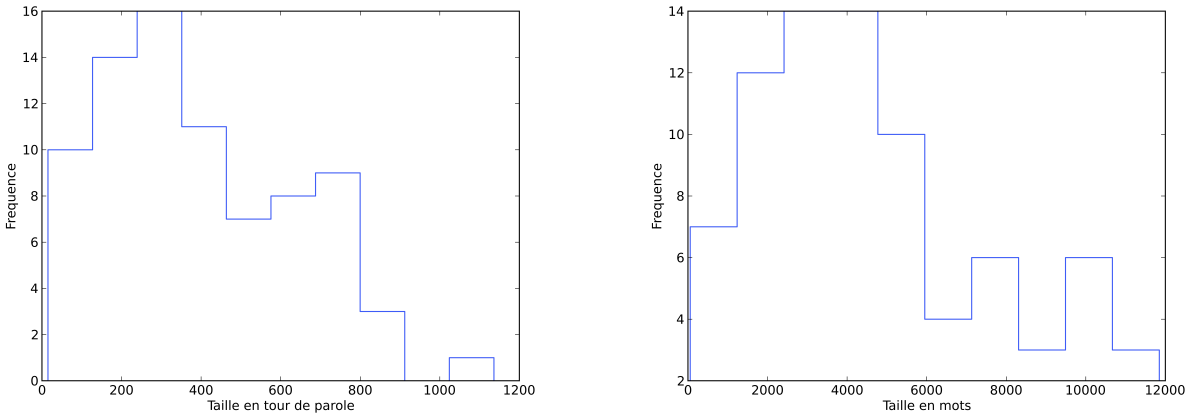
	corpus Ville1			corpus Ville2			total
	hommes	femmes	total	hommes	femmes	total	
schizophrènes	15	3	18	20	10	30	48
témoins	15	8	23	4	4	8	31
total	30	11	41	24	14	38	79

TABLE 1 – Répartition des sujets dans le corpus.

L'interaction mise en place pour cette étude est un entretien semi-directif conduit par un psychologue. Dans ce type d'entretien, le psychologue n'est pas personnellement engagé dans l'interaction. Il doit maintenir un échange dans lequel le patient revient sur son environnement et ses relations au sein de l'hôpital et avec l'extérieur. Il est clairement expliqué, tant à l'équipe médicale qu'au patient, que le contenu de l'entretien ne peut être utilisé comme base médicale.

Le protocole expérimental a consisté à identifier des patients intéressés, puis à leur faire passer des tests de mesure de capacité cognitive. Dans le sous-corpus de Ville2 aucune mesure supplémentaire n'a été faite, dans le sous-corpus de Ville1, les entretiens ont eu lieu en présence d'un double système d'oculomètre [5].

Le protocole a été défini de manière à être le moins invasif possible. Pour le sous-corpus Ville1, trois tests psychocognitifs mesurant les capacités de mémoire à court terme, d'attention, et la mémoire de travail ont été passés par les sujets : (i) le `Wechsler Adult Intelligence Scale-III` (mesure du quotient intellectuel, ou QI), (ii) le `California Verbal Learning Test` (capacité cognitive et de stratégie), et (iii) le `Trail Making Test` (dépréciation de la flexibilité cognitive et de l'inhibition). Nous n'utiliserons ici que les résultats du test de QI.



(a) (b)

FIGURE 1 – Distribution des entretiens par la taille : (a) en nombre de tours de parole ; (b) en nombre de mots.

Une fois les entretiens enregistrés, ils ont été transcrits manuellement par deux annotateurs (le ou la psychologue qui a mené tout ou partie des entretiens, et une autre personne) qui n'ont pas annoté en parallèle et qui ne savaient pas que nous allions compter les disfluences. 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 le découpage en tours de parole et en mots du corpus, et la figure 1 illustre la distribution des corpus en fonction de leur taille en nombre de tours de parole et de mots. Le caractère spécifique de l'entretien semi-directif transparaît ici clairement : le psychologue produit quasiment le même nombre de

5. Ce double système permet de capter les points de fixation du regard du sujet sur 'ce qu'il voit', mais également de capter ceux du psychologue. Ainsi, il est possible d'interpréter si un mouvement est déclenché par une interaction visuelle (regard de l'interlocuteur) ou non.

	corpus Ville1				corpus Ville2			
	nb tours de parole		nb de mots		nb tours de parole		nb de 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	

TABLE 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).

tours de parole que le sujet, pour un volume de mots très inférieur. Par exemple, dans le sous-corpus Ville1, le ratio entre le nombre de tours de parole des schizophrènes et des psychologues devant un schizophrène est de 1,003 (seulement 44 tours de parole), alors qu'il est de 1,54 en nombre de mots. Seuls les témoins du sous-corpus de Ville1 ne présentent pas cette caractéristique, mais une analyse plus fine des entretiens montre que pour six entretiens, les témoins sont restés réticents à prendre la parole.

2.2 Difficultés d'accès aux patients

Le nombre de 79 sujets peut sembler limité pour une étude de ce type, 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 un corpus significatif.

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

Mais ce qui rend la constitution d'une telle ressource complexe 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 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 à l'étude. Une première réticence vient du fait qu'il n'y a pas de conséquence positive, en terme médical, pour le patient à participer à l'étude. Il faut ajouter à cela des inquiétudes compréhensibles des patients schizophrènes concernant la possible publication de leur histoire, bien qu'une anonymisation totale soit garantie par le protocole. Bien entendu, la sous-catégorie des schizophrènes paranoïdes est encore plus difficile à rencontrer.

Par ailleurs, le protocole requérant de passer des tests psycho-cognitifs et un entretien, le temps nécessaire est relativement élevé, de l'ordre de deux heures. 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 également des défections. À titre d'exemple, lors de la phase de collecte des entretiens du sous-corpus Ville1, 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.

2.3 Anonymisation

La tâche d'anonymisation recouvre deux phases. La première, tout à fait classique, consiste à identifier les entités nommées et à les substituer par des marqueurs sémantiquement vides. Un outil automatique performant a été identifié pour ce faire, mais n'a pu être opérationnel à temps pour cette étude. Nous avons pour cela programmé une série de scripts en Python qui recherchent, grâce à des expressions régulières, les mots commençant par une majuscule qui ne sont pas

en début de phrase. Une intervention humaine a été ensuite nécessaire pour classer ces mots en 10 catégories : *prenomF*, *prenomM*, *nom*, *pays*, *département*, *ville*, *capitale*, *institution*, *montagne* et *non_pris_en_compte*. Les éléments de cette dernière catégorie ont été laissés tels quels dans le corpus, les autres ont été substitués par le nom de la catégorie suivie d'un identifiant unique. Ainsi, les références à Paris sont toutes identifiées par *capitale1*. Une fois ces substitutions réalisées, nous avons extrait l'ensemble des débuts de phrases et procédé à une vérification manuelle pour affiner les listes des catégories précédentes. Nous pouvons ainsi assurer une anonymisation fiable du corpus.

L'anonymisation du corpus ne s'arrête cependant pas là. En effet, les sujets relatant des événements s'inscrivant dans une temporalité et une géographie particulière, un certain nombre d'indices sont disséminés dans les entretiens. Il est donc relativement aisé d'identifier les personnes et il est difficile de trouver une solution à ce problème tout en conservant l'intégrité des entretiens. Cette particularité a des conséquences importantes sur notre projet.

Pour les traitements qui ne nécessitent qu'un faible contexte, en général celui de la phrase ou du tour de parole, nous avons créé une version de la ressource constituée de tous les tours de paroles randomisés. Les 31 575 tours de paroles sont donc mélangés et il devient impossible de reconstituer les historiques de chacun. Pour reconstruire les entretiens originaux, nous conservons une trace de la randomisation sous forme de table. Il est donc tout à fait possible de fournir la ressource pour des analyses du type morpho-syntaxe ou analyse syntaxique en dépendances, sans compromettre les données initiales.

Mais l'un des objectif du projet global reste l'analyse sémantico-pragmatique et, pour ces aspects, il est impossible de dissocier une prise de parole de son contexte sans perdre l'essence même de l'entretien. Seuls les membres engagés dans le projet et soumis à un devoir de confidentialité peuvent donc travailler sur cette partie. Un problème similaire se pose pour la partie transcription, puisque, bien que les bandes puissent être bippées, elles ne peuvent pas être randomisées en tours de parole. Cette contrainte explique que le nombre d'intervenants sur la transcription reste limité.

3 Protocole expérimental

3.1 Traitements automatiques

Étant donnée la taille importante des corpus et notre volonté de limiter les interventions humaines, nous avons utilisé l'outil `Distagger` (Constant & Dister, 2010) pour identifier automatiquement les disfluences dans les textes transcrits.

Pour cet outil, les disfluences regroupent plusieurs types de réalisations orales qui brisent la continuité syntaxique. Il est donc possible de produire une version reconstruite de la ressource pour obtenir des tours de parole plus cohérents du point de vue syntaxique, donc améliorer les résultats des annotations pour d'autres couches (en particulier l'analyse morphosyntaxique ou en dépendances).

L'outil 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. Nous revenons sur chacune d'elles en présentant un exemple extrait du corpus.

1. Les différentes réalisations de *euh* sont définies dans un fichier passé en argument de `Distagger`.

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

2. Les répétitions sont entendues comme la reprise explicite et identique d'un même mot ou d'un même groupe de mots dans le contexte immédiat d'apparition. La répétition peut malgré tout 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

3. 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

4. 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 auteurs ont évalué leur outil sur un corpus oral marqué en disfluences et validé manuellement. Le corpus de référence comprend au total 1 297 tours de parole, 22 476 mots, 5 817 méta-étiquettes et 1 280 disfluences. Ils obtiennent des f-scores significatifs, de 95,5 % (précision de 95,3 %, rappel 95,8 %) [6].

L'outil prend en entrée des données au format `Valibel` [7] (sans structure prédéfinie) ou au format `transcriber` (Barras *et al.*, 1998) (structuré et semi-annoté). Il fournit deux types de sortie, l'un correspondant au format `Valibel`, l'autre au format `transcriber`. `Distagger` est implémenté en Java, et peut être appelé en ligne de commande, ce qui nous a permis de l'intégrer facilement à notre chaîne de traitement. Par ailleurs, l'outil ajoute plusieurs annotations particulières structurant son résultat, qui ne sont pas informatives pour les disfluences.

Les annotations de `Distagger` sur le corpus font apparaître sept étiquettes : $\{IGN+EUH\}$, $\{IGN+REP\}$, $\{IGN+CORR\}$, $\{IGN+FRAG\}$, $\{IGN+short\_pause\}$, $\{IGN+slot\}$ et $\{IGN+speaker\}$. Les deux dernières sont des étiquettes spécifiques permettant de repérer les tours de parole et les interlocuteurs à qui sont associés ces tours de parole. Leur nombre n'apporte pas d'information caractéristique ici et elles seront écartées dans la suite. Par ailleurs, les premiers traitements ont fait apparaître les étiquettes $\{IGN+short\_pause\}$ et $\{IGN+FRAG\}$ dans des volumes très faibles (respectivement 5 et 1 étiquettes). Les $short\_pause$ correspondent à des reliquats de scories de la transcription qui ont été mal interprétés par l'outil.

Nous avons par ailleurs mis en place une série de programmes en Python pour pré-traiter les corpus, appliquer `Distagger` et post-traiter les résultats fournis.

3.2 Normalisation des corpus

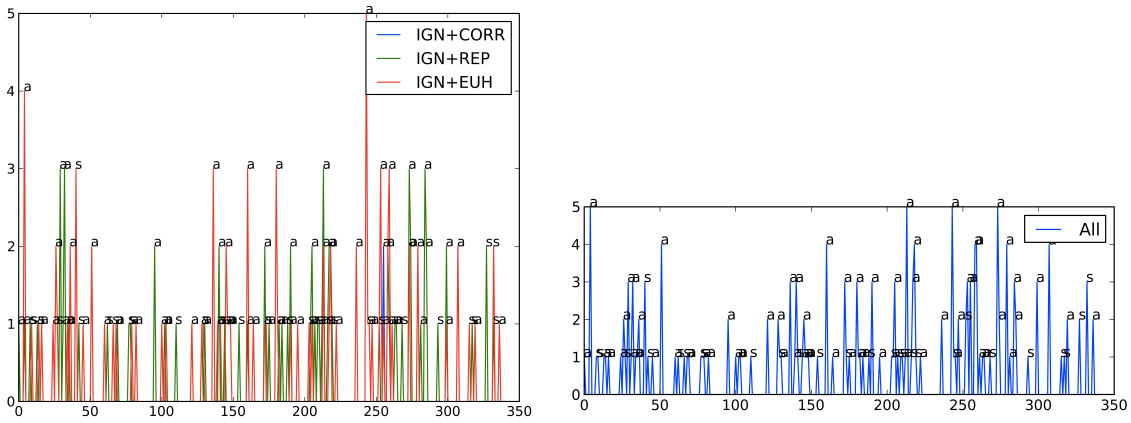
Le sous-corpus Ville2 initial n'ayant pas été prévu pour être traité par des outils de TAL, une première étape a donc consisté à extraire le contenu des documents en format `MS Word` et à normaliser le corpus à l'aide d'une trentaine d'expressions régulières. Il a fallu réinterpréter les marques spécifiques à la transcriptions originelle vers des marques explicites pour `Distagger` (↑ pour une intonation montante, ↓ pour une intonation descendante, etc.). Cinq traitements sont nécessaires pour ajouter aux fichiers les informations permettant à `Distagger` de fonctionner (utilisation de $spk1$ et $spk2$ pour les interlocuteurs [8], chemin explicite des fichiers, etc.). Ainsi, pour chaque fichier du sous-corpus de départ, nous obtenons sa version annotée par `Distagger`.

Puis, l'ensemble des résultats obtenus est fusionné pour produire une représentation générale pour ce sous-corpus, en associant à chaque tour de parole le numéro du corpus (ici 1), l'identifiant du sujet (deux chiffres et trois lettres), le numéro du tour de parole dans l'entretien, ainsi qu'une marque explicite de qui est l'interlocuteur (Pa pour le patient, et P pour le psychologue). Enfin, il manque une information discriminante qui est le statut du sujet : schizophrène ou témoin. Pour cela, une base qui distingue entre les deux, et dont la hiérarchie est fixée préalablement, est automatiquement produite à partir de la structure du corpus de départ. Nous construisons alors une représentation abstraite du corpus à partir de la fréquence d'apparition des différentes étiquettes de `Distagger`.

Le sous-corpus Ville1 s'inscrit dans le cadre du projet général. Les annotateurs ont utilisé l'outil `CLAN` pour réaliser la transcription, ce qui nous a permis d'extraire facilement le contenu textuel. Cependant, de nombreuses marques dépendantes du logiciel perdurent et il a été nécessaire de les supprimer. À nouveau, une trentaine d'expressions régulières a été utilisée. Comme pour Ville2, nous produisons une ressource contenant l'ensemble des tours de parole randomisé. C'est sur cette ressource que `Distagger` est utilisé. Puis, en utilisant la table de mémorisation, les entretiens sont reconstruits et envoyés à la même série de traitements que le corpus Ville2. Enfin, pour faciliter les études sur le contenu des entretiens, chacun est mémorisé sans aucune autre marque que l'interlocuteur (*P* ou *S* ou *T*) [9].

Il apparaît dans notre corpus que certaines annotations ne correspondent pas à des disfluences traditionnelles. En effet, le psychologue ayant une interaction particulière dans l'entretien, puisque son rôle est d'abord de maintenir l'échange, il utilise régulièrement des interventions de type *mmh mmh*, ou *oui oui*, ou *non non*. Nous avons donc mis en place un post-traitement qui redresse les résultats en supprimant les étiquettes correspondantes. De plus, la forme pronominale est très fréquemment utilisée, puisque les interlocuteurs se vouvoient et que le psychologue relance la conversation en posant des questions à caractère personnel. Nous avons donc inclus à ces post-traitements les formes *vous vous* qui apparaissent

6. Une évaluation de l'outil sur un échantillon de 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.
7. Voir : `https://www.uclouvain.be/cps/ucl/doc/valibel/documents/conventions_valibel_2004.PDF`.
8. $spk1$ est attribué au premier locuteur et ne correspond pas nécessairement au psychologue.
9. Il nous semble important de prendre en considération les disfluences du psychologue qui influencent le cours de l'interaction.



(a) (b)

FIGURE 2 – Distribution des étiquettes de disfluence au cours de l'entretien 011HAM du sous-corpus Ville1 (a est un tour de parole du patient et s, du psychologue).

en nombre très supérieur par rapport aux répétitions de *vous*.

Enfin, nous avons automatisé la production des graphiques pour chaque entretien, reprenant le nombre de disfluences par tour de parole, comme présenté dans la figure 2. Pour chaque sous-corpus nous produisons une surface où chaque entretien est ramené à l'échelle en pourcentage (le nombre de tours de parole fluctuant beaucoup) en fonction du nombre de disfluences présent dans la portion de texte (voir figure 3. Enfin, nous produisons plusieurs documents LaTeX reprenant les moyennes des disfluences par tour de parole et par nombre de mots, et nous calculons pour chaque sous-corpus leur indice de significativité. Nous produisons ensuite une représentation graphique de la position des disfluences dans l'entretien. L'ensemble de ces programmes de normalisation représente environ 1 500 lignes de code.

4 Résultats

4.1 Analyse quantitative

La figure 2 présente un exemple des résultats obtenus, pour le patient 011HAM (sous-corpus Ville1). Dans la première figure (a), une couleur de courbe est attribuée à chacune des trois étiquettes principales, l'axe des abscisses correspond aux tours de parole de 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 ajoutée, a pour les tours de paroles du patient, et s pour le psychologue. La seconde figure présente les mêmes données, en exhibant 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. Il se trouve que dans ce tour de parole `Distagger` identifie 4 *euh* et 1 *rep*, ce qui explique la différence observée.

Une lecture de l'ensemble des graphiques fait apparaître une régularité avec deux pics de disfluences, le premier en début d'entretien, le second au cours du dernier tiers. Le premier pic peut simplement s'entendre comme un pic de stress en début d'échange. Il est intéressant de constater que le second pic amène la fin de l'entretien. Du côté des témoins, on retrouve une entame d'entretien avec un pic, mais pas nécessairement un second.

La table 3 présente l'ensemble des résultats de `Distagger`. Pour chacun des deux sous-corpus nous calculons, pour les trois étiquettes principales ($IGN{+}CORR$, $IGN{+}REP$ et $IGN{+}EUH$), leur fréquence d'apparition dans les tours de parole, des schizophrènes (S), des témoins (T) ou du/de la psychologue (P). Nous normalisons d'une part par rapport au nombre de tours de parole (pour chaque catégorie d'interlocuteur), et d'autre part par rapport au nombre de mots (à nouveau pour chaque type d'interlocuteur). Nous calculons les mêmes valeurs pour les sujets (S + T), le ou la psychologue lorsqu'il ou elle est en face d'un schizophrènes (P+S) ou devant un témoin (P+T). Les résultats totaux reprennent la

	corpus Ville2						corpus Ville1					
	S	T	S+T	P+S	P+T	P	S	T	S+T	P+S	P+T	P
IGN+CORR												
par tour de parole	0,0087	0,0071	0,008	0,0015	0	0,0012	0,0180	0,0085	0,0127	0,0052	0,0109	0,0084
par / nb mots	0,0004	$9e-05$	0,0003	0,0001	0	0,0001	0,0013	0,0007	0,0010	0,0006	0,0007	0,0006
IGN+REP												
par tour de parole	0,2223	0,2519	0,2285	0,0646	0,0897	0,0699	0,2735	0,138	0,1978	0,1336	0,2608	0,205
par / nb mots	0,0125	0,0078	0,0115	0,0064	0,0079	0,0067	0,0211	0,0134	0,0168	0,0171	0,0177	0,0174
IGN+EUH												
par tour de parole	0,3107	0,2999	0,3084	0,0738	0,0616	0,0712	0,4201	0,3372	0,3736	0,1948	0,4651	0,3464
par / nb mots	0,0190	0,0089	0,0169	0,0077	0,0058	0,0073	0,0369	0,0326	0,0345	0,0244	0,0312	0,0282
Resultats totaux												
par tour de parole	0,5417	0,5589	0,545	0,1400	0,1513	0,1424	0,7117	0,484	0,5842	0,3338	0,7369	0,5599
par / nb mots	0,032	0,0168	0,0288	0,0144	0,0138	0,0142	0,0595	0,0468	0,0524	0,0421	0,0496	0,0463

TABLE 3 – Répartition quantitative des étiquettes de `Distagger` dans les sous-corpus.

somme des valeurs intermédiaires pour chaque catégorie d'interlocuteur. `Distagger` n'annote aucune correction pour le psychologue avec des témoins dans le sous-corpus Ville2, ce qui explique les deux valeurs à zéro.

La lecture des résultats totaux met en avant une variabilité importante des résultats normalisés par rapport au nombre de tours de parole. Les résultats normalisés par rapport au nombre de mots sont plus significatifs. En effet, si les disfluences produites par les témoins et le psychologue (quel que soit son interlocuteur) sont du même ordre : $1,68$ % et $1,42$ % pour le sous-corpus Ville2, et $4,68$ % et $4,63$ % pour le sous-corpus Ville1, les productions des schizophrènes sont bien supérieures : $3,2$ % et $5,95$ %. Il existe ainsi une différence entre le nombre de disfluences identifiées chez les schizophrènes et les non schizophrènes de $1,63$ % dans le sous-corpus Ville2 et de $1,29$ % dans le sous-corpus Ville1.

La variabilité des résultats peut s'expliquer par la différence des transcriptions entre les deux sous-corpus, ainsi que par le nombre de sujets dans chacun. 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.

Enfin, pour visualiser la répartition des résultats, nous produisons une surface où nous normalisons le nombre de tours de parole sur une échelle de 100. Pour chaque entretien seuls les tours de parole, soit du schizophrène, soit du témoin, sont utilisés et nous recalculons le nombre de disfluences sur un intervalle de 1 % de l'entretien. Nous obtenons ainsi des pics où, en valeur, le nombre est supérieur aux résultats précédents, mais qui correspondent à une combinaison linéaire de plusieurs tours de parole. La figure 3 présente les surfaces calculées pour les deux sous-corpus. Les deux figures de gauche correspondent au sous-corpus de Ville1, les deux figures de droite à celui de Ville2. Les figures en haut correspondent aux tours de parole des schizophrènes et celles en bas à ceux des témoins.

Dans le cas de Ville2, il apparaît de manière évidente que les témoins produisent beaucoup moins de disfluences que les schizophrènes. La surface en bas à gauche est en effet quasiment plane. L'interprétation des graphiques pour le sous-corpus Ville1 est plus délicate. Pour cela, nous présentons la projection de l'ensemble des entretiens sur l'axe *pourcentage* du graphique. La couleur bleu correspond à une densité importante sur la projection et la couleur rouge à une densité plus faible. En étudiant les valeurs ainsi trouvées pour le sous-corpus, on trouve une distribution régulière chez les témoins, mais toujours plus marginale que chez les schizophrènes.

4.2 Significativité

Afin de valider les résultats que nous avons pu mettre en avant, nous reprenons ici la mesure de significativité utilisée dans (de Mareüil *et al.*, 2013). Celle-ci permet de calculer un indice de distribution en fonction du nombre de mots entre deux catégories d'interlocuteurs. La valeur trouvée doit être supérieure à 1,96 pour être considérée comme significative.

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

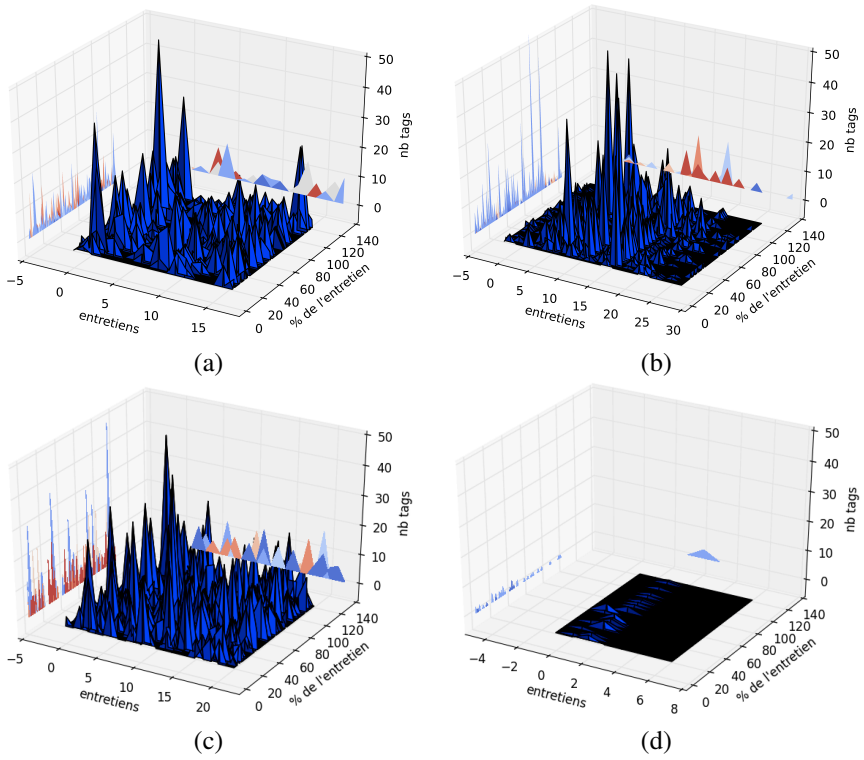


FIGURE 3 – Répartitions des étiquettes de disfluences : dans les entretiens de schizophrènes, sous-corpus Ville1 (a), Ville2 (b) ; dans les entretiens de témoins, sous-corpus Ville1 (c), Ville2 (d).

où :

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

– n_1 est le nombre de mots prononcés par la première catégorie d'interlocuteur,
– n_2 est le nombre de mots prononcés par la seconde catégorie d'interlocuteur,
– p_1 est la proportion de disfluences produites par la première catégorie d'interlocuteur,
– p_2 est la proportion de disfluences produites par la seconde catégorie d'interlocuteur.

Cette formule nous permet d'interpréter les résultats entre deux catégories d'interlocuteurs. Nous calculons donc l'indice pour les trois appariements que nous pouvons proposer. Les résultats obtenus sont présentés dans le tableau 4.

	corpus Ville1	corpus Ville2
S et Psy	$10,6806923083$	$19,4197596818$
T et Psy	$0,422898291704$	$3,23530253756$
S et T	$10,2827554261$	$16,0376100956$

TABLE 4 – Significativité des différences dans le nombre de disfluences entre interlocuteurs.

Il apparaît que les significativités entre les témoins et les psychologues sont faibles, voire non significatives, ce qui permet de rapprocher le comportement des témoins de celui des psychologues. Par contre, la significativité est importante (toujours supérieure à 10) dans les appariements qui comprennent des schizophrènes, ce qui nous permet de conclure que le nombre de disfluences produites par des schizophrènes est significativement différent de celui des non-schizophrènes de l'expérimentation (psychologue et témoins). Ce résultat est d'autant plus pertinent qu'il est mis en avant par des outils automatiques et ne fait pas intervenir de subjectivité humaine.

	nb étiquettes	/nb mots	/ nb tours de parole
$IGN{+}ENR$	2	$2e-05$	$0,00037$
$IGN{+}REP$	2 208	$0,01857$	$0,40314$
$IGN{+}Meta$	161	$0,00135$	$0,0294$
$IGN{+}EUH$	2 773	$0,02333$	$0,5063$
$IGN{+}CORR$	138	$0,00116$	$0,0252$

TABLE 5 – Analyse des disfluences dans le corpus TCOF-POS par `Distagger`.

4.3 Biais potentiels des expériences

Malgré la significativité des résultats que nous avons obtenus, il nous paraît important de revenir sur plusieurs biais potentiels de l'étude.

Les deux sous-corpus ayant été transcrits par des moyens différents et jamais en parallèle, comme nous l'avons discuté dans la section 2, il est difficile de proposer une évaluation qualitative des transcriptions. Afin de vérifier que nos résultats ne dévient pas de la réalité linguistique, nous avons appliqué `Distagger` sur le corpus de parole spontanée TCOF-POS (Benzitoun *et al.*, 2012). La table 5 reprend la ventilation des étiquettes trouvées par `Distagger`. Sur les trois étiquettes que nous retrouvons et avons analysé, le nombre de disfluences est de $4,3$ %, ce qui est comparable aux résultats précédents et nous conduit à considérer la transcription comme un faible biais.

Un autre biais réside dans la répartition entre témoins et patients à l'intérieur même de chaque sous-corpus. Ainsi, le sous-corpus Ville2 ne contient que 8 témoins, alors que le sous-corpus Ville1 en contient 23. Nous avons décidé de conserver tous les témoins dont nous disposions pour équilibrer le corpus général. Le projet s'attache à rééquilibrer la répartition. Les témoins du sous-corpus Ville1 produisent davantage de disfluences que ceux du sous-corpus Ville2. Une lecture des entretiens montre que la psychologue qui a recueilli les entretiens du sous-corpus Ville1 produit plus de disfluences, ce qui peut inciter les interlocuteurs à l'imiter, donc à produire davantage de disfluences.

Il existe par ailleurs une différence d'âge et de QI entre les participants schizophrènes et témoins (voir tableau 6). En effet, les schizophrènes sont significativement plus âgés que les témoins (près de 29 ans au lieu de 23, avec $p = 0,0058$ [10]) et leur QI est inférieur (à peu près 95 au lieu de 103, avec $p = 0,0203$), pour un nombre d'années d'école ou d'études très semblable (environ 13 pour les témoins et 12,4 pour les schizophrènes).

	QI	années d'études	âge
femmes			
témoins	$105,5$	13	$22,37$
schizophrènes	$98,33$	13	30
hommes			
témoins	$102,73$	$13,26$	$23,66$
schizophrènes	$94,53$	$12,28$	$28,66$
moyenne générale			
témoins	$103,70$	$13,17$	$23,22$
schizophrènes	$95,17$	$12,41$	$28,89$

TABLE 6 – Moyennes des QI, du nombre d'années d'études et des âges sur les participants au corpus Ville1.

Un autre biais important, mais inévitable, de l'étude est que les patients sont en remédiation, donc sous traitement (Chlorpromazine à Ville2 et neuroleptiques non spécifiés à Ville1). Levy (1968) a identifié des effets négatifs (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,

10. Les significativités ont ici été calculées à l'aide du test de Student.

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. Il est donc aujourd'hui extrêmement difficile d'évaluer l'influence exacte du traitement sur les productions des patients. Cette question est récurrente dans la littérature. Il apparaît néanmoins que les effets secondaires des médicaments sont moins prégnants aujourd'hui, les traitements ayant considérablement évolué depuis les années 60. Par ailleurs, nous disposons d'un sous-groupe de 7 patients schizophrènes sans traitement dans le sous-corpus Ville2. Les différences entre eux et les autres patients schizophrènes n'apparaissent pas significatives.

5 Conclusions et perspectives

Cette étude nous a permis de mettre en lumière un usage pathologique des disfluences chez les patients schizophrènes grâce à des outils et des méthodes issus du TAL. Pour cela, nous avons utilisé l'outil `Distagger` pour procéder à une annotation des disfluences. Il apparaît que les schizophrènes produisent, respectivement dans chaque corpus, 1,63 % et 1,29 % de disfluences de plus (par rapport au nombre de mots) que des sujets non diagnostiqués. Nous avons validé ce résultat par un calcul de significativité qui isole clairement les patients schizophrènes. Ce sont les outils de TAL qui ont permis d'aboutir à cette conclusion, en dehors de toute interprétation humaine.

Nous avons par ailleurs discuté des différents biais possibles de l'étude, tant sur la constitution du corpus de départ que sur la méthodologie utilisée. La suite du projet s'attachera à revenir sur ces derniers pour les corriger lorsque cela est possible, en particulier en calculant une mesure de qualité des transcriptions. Mais l'une des difficultés principales réside dans la nature de l'objet d'étude. D'une part, les patients schizophrènes doivent être suffisamment communiquant pour passer le protocole de tests et donc, généralement, être sous traitement. D'autre part, le caractère personnel de l'entretien pose plusieurs questions d'éthique. Par conséquent, si il est difficile d'accéder aux patients, il est tout aussi délicat de gérer l'accès à la ressource.

Dans la continuité du projet, nous souhaitons annoter la ressource en morpho-syntaxe et en syntaxe en dépendances. L'annotation en morpho-syntaxe nous permettra de réaliser automatiquement une lemmatisation du corpus, notamment pour évaluer la richesse du vocabulaire des sujets. Pour cela, nous allons utiliser l'outil `MElt` (Denis & Sagot, 2009) entraîné pour le français oral sur le corpus TCOF-POS (Benzitoun *et al.*, 2012) [11]. L'analyse en dépendances nous permettra de revenir sur les résultats anciens auxquels nous avons fait référence sur la complexité de la syntaxe utilisée par les patients. Pour cela, plusieurs outils sont disponibles, dont `FRMG` (De La Clergerie *et al.*, 2009), `Leopar` (Perrier & Guillaume, 2013) et `Talismane` (Urieli & Tanguy, 2013), plusieurs tests préliminaires ont d'ailleurs été d'ores et déjà réalisés. L'utilisation de plusieurs outils nous permettra de valider les analyses proposées.

Comme il a été fait mention dans la première partie de cet article, l'objectif est également de proposer une annotation en sémantique-pragmatique. Pour cela, nous conduirons deux campagnes d'annotation manuelle, l'une pour identifier les discontinuités décisives, l'autre, sur les extraits identifiés, pour annoter en SDRT. L'ensemble de ces indices seront corrélés avec les autres mesures dont nous disposons, dont les résultats aux tests psycho-cognitifs, les mesures oculométriques et les EEG.

La contribution apportée par cette étude au projet général montre l'importance d'utiliser des outils automatiques pour mettre en avant des indices objectifs. Il n'en reste pas moins qu'il est nécessaire d'affiner les résultats. Notre perspective principale est de proposer une ressource normalisée, riche en méta-données (dont le manque et l'importance sont mis en valeur dans (Ghio *et al.*, 2006)), malgré les nombreuses difficultés éthiques que posent ces travaux.

Références

ASHER N. & LASCARIDES A. (2003). *Logics of Conversation*. Studies in Natural Language Processing. Cambridge University Press.

BARRAS C., GEOFFROIS E., WU Z. & LIBERMAN M. (1998). Transcriber : a free tool for segmenting, labeling and transcribing speech. In *International Conference on Language Resources and Evaluation (LREC)*, p. 1373–1376.

BENZITOUN C., FORT K. & SAGOT B. (2012). TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe. In *Traitement Automatique des Langues Naturelles (TALN)*, p. 99–112, Grenoble, France.

11. Les performances de ce outil avec ce modèle atteignent 97,61 % d'exactitude.

BLANCHE-BENVENISTE C. & JEANJEAN C. (1987). *Le Francais parlé. Transcription et édition*. Paris, France : Didier Érudition.

CHAIKA E. (1974). A linguist looks at "schizophrenic" language. *Brain and Language*, **1**(3), 257–276.

CONSTANT M. & DISTER A. (2010). Automatic detection of disfluencies in speech transcriptions. In I. C. F. D. M. PETTORINO, A. GIANNINI, Ed., *Spoken Communication*, volume 1, p. 259–272. Cambridge Scholars Publishing.

DE LA CLERGERIE É., SAGOT B., NICOLAS L. & GUÉNOT M.-L. (2009). FRMG : évolutions d'un analyseur syntaxique TAG du français. In É. VILLEMONTE DE LA CLERGERIE & P. PAROUBEK, Eds., *Journée de l'ATALA sur : Quels analyseurs syntaxiques pour le français ?*, Paris, France : ATALA. Journée de l'ATALA organisée conjointement à la conférence IWPT 2009.

DE MAREÜIL P. B., ADDA G., ADDA-DECKER M., BARRAS C., HABERT B. & PAROUBEK P. (2013). Une étude quantitative des marqueurs discursifs, disfluences et chevauchements de parole dans des interviews politiques. *TIPA. Travaux interdisciplinaires sur la parole et le langage [En ligne]*, **29**. mis en ligne le 19 décembre 2013, consulté le 14 février 2014.

DENIS P. & SAGOT B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. In *Pacific Asia Conference on Language Information and Computing (PACLIC)*.

FELDSTEIN S. (1962). The relationship of interpersonal involvement and affectiveness of content to the verbal communication of schizophrenic patients. *Journal of Abnormal and Social Psychology*, **64**, 39–45.

FROMKIN V. A. (1975). A linguist looks at "a linguist looks at 'schizophrenic language"'. *Brain and Language*, **2**(0), 498 – 503.

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. (2006). Corpus de parole pathologique, état d'avancement et enjeux méthodologiques. *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence (TIPA)*, **25**, 109–126. Autorisation No.3015 : TIPA est la revue du Laboratoire Parole et Langage 3015 3015.

GOLDMAN-EISLER F., SKARBEK A. & HENDERSON A. (1965). The effect of chorpromazine on speech behaviour. *Psychopharmacologia*, **7**(3), 220–229.

KAMP H. & REYLE U. (1993). *From Discourse to Logic*. Kluwer Academic Publishers.

KREMEN W. S., SEIDMAN L. J., FARAONE S. V. & TSUANG M. T. (2003). Is there disproportionate impairment or phonemic fluency in schizophrenia ? *Journal of the International Neuropsychological Society*, **9**, 79–88.

LEVY R. (1968). The effect of chlorpromazine on sentence structure of schizophrenic patients. *Psychopharmacologia*, **13**(5), 426–432.

MAHER B. (1972). The language of schizophrenia : A review and interpretation. *The British Journal of Psychiatry*, **120**, 3–17.

MUSIOL M., AMBLARD M. & REBUSCHI M. (2013). Approche sémantico-formelle des troubles du discours : les conditions de la saisie de leurs aspects pyscholinguistiques. In *27ème Congrès International de Linguistique et de Philologie Romanes*, Nancy, France.

MUSIOL M. & TROGNON A. (1996). L'accomplissement interactionnel du trouble schizophrénique. *Raisons Pratiques 7*, p. 179–209.

PERRIER G. & GUILLAUME B. (2013). Leopar : an Interaction Grammar Parser. In *Workshop on High-level Methodologies for Grammar Engineering, ESSLLI*, p. 121–122, Dusseldorf.

REBUSCHI M., AMBLARD M. & MUSIOL M. (2013). 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, p. 1–24. Springer.

URIELI A. & TANGUY L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, p. 188–201, Les Sables d'Olonne, France.

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

WARD D. (2008). *Stuttering and Cluttering : Frameworks for Understanding and Treatment*. Taylor & Francis.