Corpus et pathologie mentale : particularités dans la constitution et l'analyse d'une ressource

Maxime Amblard (LORIA), Michel Musiol (ATILF), Manuel Rebuschi (AHP-PReST)

Résumé

Le projet SLAM (Schizophrénie et Langage : Analyse et Modélisation) s'intéresse à l'analyse systématique de la capacité de communication des patients schizophrènes. Dans ce cadre, il est important de disposer de corpus le plus larges possible, qui pour notre part sont constitués de données issues de tests neuropsychologiques et de transcriptions d'enregistrements audio d'entretiens avec des psychologues. Ces corpus sont construits en interaction avec différents centres hospitaliers. L'explicitation d'un protocole de recueil de données ne résout pas les difficultés rencontrées dans la constitution des corpus spécifiquement liée à la pathologie mentale. Dans cet article, nous nous proposons de revenir sur ce type de corpus dont la nature est singulière, tant la communauté des psychologues participe peu à la constitution de corpus telle qu'elle est traditionnellement pratiquée en linguistique. Nous présentons les différentes phases de constitution de la ressource, ainsi que la méthodologie d'analyse choisie pour en identifier les spécificités.

1. Introduction

Une tradition de recherches en psycholinguistique qui s'étend tout au long de la seconde moitié du vingtième siècle s'appuie sur l'analyse des productions langagières de sujets atteints de pathologies mentales, avec pour objectif d'y déceler les indices de possibles dysfonctionnements cognitifs. Historiquement constitués au cours d'entretiens personnels, les corpus issus d'entretiens avec des patients constituent selon cette tradition la ressource privée de chercheurs, et pour diverses raisons, notamment éthiques, des données rarement partagées au sein de la communauté scientifique. Dans le contexte de travaux interdisciplinaires, ces pratiques traditionnelles entrent en conflit avec des exigences de publicité difficilement conciliables.

L'analyse systématique de la capacité de communication des patients schizophrènes, qui est l'objet du projet interdisciplinaire SLAM (Schizophrénie et Langage : Analyse et Modélisation), est naturellement confrontée à la nécessité de disposer de corpus homogènes. Ces corpus sont composés à la fois de données issues de tests neuropsychologiques, de différents enregistrements par des dispositifs spécifiques comme des *eye-trackers* ou des EEG, et de transcriptions d'enregistrements audio de conversations.

Dans le cadre de ce projet, nous avons conduit plusieurs campagnes pour constituer nos corpus en collaboration avec différentes institutions hospitalières. S'il est relativement aisé de définir un protocole permettant la constitution de ces corpus, force est de constater que la qualité des résultats est très variable, de campagnes réussies avec plus de 30 patients à des campagnes en rassemblant moins de 5. Cela est notamment dû à la spécificité de l'objet d'étude, la pathologie mentale impliquant des interactions sociales délicates et instables.

Dans cet article, nous nous proposons de revenir sur ce type de corpus dont la nature est singulière, tant la communauté des psychologues participe peu à la constitution de corpus telle qu'elle est usuellement pratiquée en linguistique. Nous présentons les différentes phases de constitution de la

ressource, ainsi que la méthodologie d'analyse choisie pour en identifier les spécificités. Dans la section 2, nous présentons le projet SLAM, ses objectifs et sa méthodologie. La section 3 est consacrée à la première étape dans la constitution du corpus, à savoir le recueil des données, et en aborde les obstacles matériels et institutionnels. La section 4 présente enfin les différents types d'annotation, automatiques et manuels, auxquels le corpus est soumis dans le cadre de nos analyses.

2. Le projet SLAM

Le projet SLAM est issu de la rencontre entre plusieurs chercheurs intéressés par la question du langage, en psychologie, en philosophie et en informatique linguistique.

Du côté de la psychologie et plus précisément de la psycholinguistique, une école nancéienne s'est développée depuis les années 1980 autour des travaux d'Alain Trognon sur la logique interlocutoire. L'idée est qu'une appréhension de la pathologie mentale peut être réalisée au travers de l'analyse de conversations semi-dirigées plus utilement que dans le cadre d'expériences contrôlées, la conversation constituant le lieu naturel d'interactions dont l'expression peut être saisie sans biais expérimentaux. Les analyses conversationnelles de cette école, conduites suivant une méthodologie pragmatique informelle (Roulet et al. 1985), visaient fondamentalement à mettre au jour la structure hiérarchique et fonctionnelle complexe de conversations avec des patients, avec un intérêt spécifique pour les lieux de surgissement de ruptures manifestes.

Pour sa part, la philosophie du langage a apporté de nombreuses contributions tout au long du vingtième siècle sur les rapports entre sémantique, entendue comme approche théorique de la signification, et logique, ainsi que sur les relations entre sémantique et pragmatique. Les normes de la rationalité sont ainsi mobilisées selon de nombreux philosophes du langage dans la constitution de la signification comme dans la compréhension linguistique. Dès lors, les conversations pathologiques constituent un terrain très particulier de mise à l'épreuve des théorisations : quand la signification conversationnelle paraît s'évanouir, quand tout au moins la compréhension n'est plus immédiate mais exige une interprétation, quand la cohérence logique est en question, l'analyste pourrait être tenté de jeter l'éponge. Pourtant, le fait que ces conversations « hors-normes » aient lieu constitue une donnée brute qui ne peut être simplement écartée.

La linguistique enfin, et particulièrement l'informatique linguistique, développe depuis près d'un demi-siècle des théories de sémantique et de pragmatique formelles, avec pour vocation ultime l'automatisation de la construction des significations, et notamment de l'interprétation des discours en langue naturelle. Aux difficultés habituelles de traitement posées par les multiples écarts des productions langagières d'avec les théories grammaticales, le contexte de la pathologie mentale vient ajouter une difficulté de taille en introduisant la contestation directe des règles encodées par ces théories. Là encore, on ne peut pas se contenter de rejeter purement et simplement les données hors du champ de l'investigation. A l'inverse, cette mise à l'épreuve des modèles théoriques suggère d'ouvrir une réflexion sur le statut des règles et normes grammaticales, et au-delà sur les relations subtiles entre sémantique et pragmatique dans le jeu de l'interaction conversationnelle.

Motivés par ces questions et convergeant sur le même objet, nous avons au fil des ans élaboré une méthodologie d'analyse des conversations pathologiques et un projet interdisciplinaire dont les résultats provisoires ont pu être questionnés par des chercheurs des différentes communautés.

Les travaux originels du groupe ont porté sur la représentation du discours schizophrénique. La question sous-jacente était double : d'une part identifier des dysfonctionnements psycholinguistiques chez des patients atteints de schizophrénie, d'autre part travailler à définir l'état de folie communément accepté pour cette pathologie. Du point de vue épistémologique la folie est un concept important en ce qu'il oblige à définir tout à la fois le monde réel, le processus communicationnel en œuvre dans le langage et le positionnement de chaque réalité dans ce processus.

La première partie de ces travaux a consisté à former un corpus d'entretiens avec des schizophrènes ainsi qu'avec un groupe témoin apparié et à chercher, comme l'ont montré (Trognon & Musiol, 1996), des discontinuités significatives dans ces discours. Ces dernières sont définies comme présentant une rupture manifeste sur plusieurs tours de parole pendant l'entretien. Il est apparu que seul le sous-groupe des schizophrènes paranoïdes faisait apparaître ce type de discontinuité (Musiol & Verhaegen, 2014). Ainsi, la notion de discontinuité dialogique est apparue comme un indice linguistique discriminant dans l'identification de la pathologie.

Plusieurs questions se sont posées à l'origine de notre travail. Tout d'abord, comment appréhender le type d'échange dans ces entretiens ? Il ne s'agit pas à proprement parler d'un discours monologique puisque l'échange met en jeu deux intervenants, mais pas non plus d'un dialogue ordinaire puisque la fonction discursive de l'un des intervenants (ici le psychologue) n'est pas naturelle. Il s'agit d'abord pour celui-ci de maintenir l'échange en faisant parler le patient. Le psychologue aide ainsi le schizophrène à verbaliser et à construire un discours. Les théories sémantico-pragmatiques sont alors tout à fait pertinentes pour la modélisation. Ensuite, dans un mouvement inverse, modéliser des phénomènes pathologiques permet d'interroger la validité cognitive de ces théories.

Les travaux du groupe SLAM s'appuient ainsi en premier lieu sur des corpus issus de conversations impliquant des personnes diagnostiquées schizophrènes. Dans ce qui suit, nous présentons les différentes étapes de la constitution et de l'analyse de ces corpus : recueil des données, transcriptions, et annotations automatiques et manuelles.

3. Le recueil des données

Dans la communauté des psychologues, les corpus sont habituellement recueillis en contexte institutionnel, donc à l'hôpital ou en clinique c'est-à-dire à l'endroit même où les patients bénéficient de soins et d'une prise en charge. En raison de leurs paradigmes de référence, la plupart du temps réductionnistes (psychanalyse, approche cognitivo-comportementale, approche neuropsychologique, approche psychopharmacologique), les psychiatres et les psychologues ont tendance à exploiter leurs données (dont les corpus) comme si elles véhiculaient des traces indélébiles que l'on peut rapporter « directement » à la psychopathologie du patient, à tout le moins à ses états mentaux. Il en résulte que le matériel verbal est *a priori* abordé indépendamment de facteurs linguistiques et

discursifs, de manière abstraite, donc selon une approche qui ne tient pas compte des conditions cotextuelles et situationnelles dans lesquelles les entretiens ont lieu.

Il arrive de plus en plus fréquemment que des chercheurs davantage rompus aux techniques de recueil et d'analyse des données en sciences du langage soient associés à des programmes de recherche dans le domaine de la clinique et de la psychopathologie. Mais dans le contexte psychiatrique par exemple, ces programmes sont obligatoirement placés sous l'égide ou la tutelle d'un médecin, qui seul, endosse la responsabilité de l'étude, et celle de la protection des données. Les protocoles de recherche et de recueil de données doivent en effet être ratifiés préalablement par une commission déontologique (i.e. CPP, Comité de Protection des Personnes) qui impose cette règle. Il en résulte que le médecin est au regard de la loi responsable des conditions dans lesquelles toutes les données qui concernent les patients, corpus compris, vont être recueillies et publiées. Ce qui a pour conséquence de limiter considérablement la marge de manœuvre des chercheurs et d'engendrer des collaborations de recherche compliquées étant entendu que les objectifs, les références et les habitudes de travail des chercheurs provenant du champ de la psychopathologie sont distinctes de ce qui se pratique habituellement en sciences du langage ou en psycholinguistique.

Dans le champ de la psychopathologie et de la clinique par exemple, les conceptions implicites des relations entre langage et pensée sont inspirées de la pensée de Bleuler au tout début du 20^{e} siècle selon laquelle le langage reflète « directement » la pensée. La psychiatrie et la psychanalyse s'intéressent aux troubles du langage et de la pensée depuis le début des années 1900. Il aura toutefois fallu attendre les années 1970 pour que les linguistes contribuent à la recherche et apportent des méthodologies d'analyse tant soit peu formalisées (Chaïka, 1974 ; Fromkin, 1975). Les débats engagés consistent alors à se poser la question de savoir si les troubles que les patients schizophrènes expriment en entretien ou dans la communication d'une manière générale, s'apparentent à des dysfonctionnements qui relèvent de la compétence langagière ou d'un désordre de la pensée. Aussi importants soient-ils pour la question diagnostique, pour la compréhension du trouble voire pour la prise en charge psychothérapeutique des patients, ces débats concernent aujourd'hui encore surtout les chercheurs et sont d'ailleurs modestement enseignés dans les formations professionnelles en psychologie, qui plus est en psychiatrie. Les psychiatres et les psychologues ne bénéficient que très rarement d'enseignements en sciences du langage. Les professionnels sont donc peu sensibilisés a priori au recueil de données systématique et à l'élaboration de corpus. Parallèlement, les querelles de paradigme au sein même de la communauté des cliniciens et plus généralement entre les défenseurs de la méthode clinique d'une part et les tenants des méthodes dites « scientifiques » d'autre part, se sont traduites par un repli sur soi et la rétention d'un maximum d'informations, conduisant à l'absence de transparence sur « ce que fait le clinicien avec son patient » et fermant la porte aux curiosités des personnes extérieures au champ de la psychiatrie, limitant donc les investigations des chercheurs.

Plus tard, à l'approche des années 2000, les méthodes des sciences naturelles, en l'occurrence l'expérimentation et la modélisation formelle pénètrent enfin la psychiatrie au gré de programmes de recherches pluridisciplinaires. Mais c'est aussi l'époque où les règles de déontologie deviennent de plus en plus strictes, alourdissant les procédures d'anticipation des projets de recherche et décourageant les chercheurs. Du côté du corps médical, on peine à s'engager dans un programme de recherche. L'accueil d'une équipe de recherche à l'hôpital prend beaucoup de temps et mobilise les

forces de travail alors que ces institutions manquent le plus souvent cruellement de ces deux ingrédients. Enfin les médecins responsables, qui de fait sont rarement au premier plan dans l'exécution de la recherche, craignent la fuite, la perte, ou la divulgation fortuite d'informations relatives à l'identité des patients et par conséquent d'éventuels procès ou plaintes à l'initiative des familles.

Deux hypothèses pour le moins expliquent les raisons pour lesquelles le recueil systématisé et contrôlé des corpus est aujourd'hui encore peu répandu : (1) la nécessité de protéger les patients en un sens juridique ; (2) leur singularité, leur souffrance et leur fragilité qui sont telles qu'ils ne s'expriment librement que dans des situations et des contextes interactionnels très singuliers.

Ainsi, pour ce qui est du second point, on rappellera qu'il n'est pas aisé d'apparier un corpus témoin à un corpus « pathologique » pour différentes raisons : par exemple, le QI des patients est la plupart du temps moins élevé que celui de la population générale adulte et leur niveau d'étude dans une même classe d'âge est plus bas. Les entretiens avec les patients ont donc nécessairement lieu (au moins indirectement) sous contrôle médical et en contexte psychiatrique. Ils sont par nature fortement asymétriques en raison de la relation de dépendance qui soumet le patient au dispositif clinique et psychiatrique et de fait, à l'interlocuteur. D'ailleurs, on n'imagine pas solliciter un patient dans une situation expérimentale avec un interlocuteur qui ne soit pas psychologue, psychiatre, ou au moins étudiant de fin de second cycle en psychologie. En outre, les patients sont nécessairement tous soumis à un protocole de prise en charge psychothérapeutique ou clinique varié, à un traitement médicamenteux complexe, distinct d'un patient à l'autre, susceptible d'effets secondaires difficiles à anticiper et qu'il est évidemment impossible d'induire chez les témoins pour des raisons déontologiques évidentes.

Le problème le plus sensible auquel nous sommes confrontés tient sans doute à l'appréhension de la volonté des patients. Etant donnée leur dépendance à l'institution psychiatrique et à ses représentants officiels, dont les psychologues dans leur ensemble font partie, les patients n'ont pas intérêt à, ou ne se sentent pas vraiment libres de s'opposer à une demande d'un médecin ou d'une personne « ayant soin ». Ils peuvent en effet penser, et beaucoup se comportent en tout cas comme si c'était le cas, qu'ils ont un intérêt personnel (hâter le processus de sortie, alléger les conditions de la prise en charge) à participer à un protocole de recherche et le cas échéant à accepter de rencontrer un expérimentateur. Quoi qu'il en soit, la notion de volonté chez un sujet atteint d'une pathologie mentale est extrêmement difficile à cerner.

La réflexion éthique en matière biomédicale a amplement contribué à la mise en place progressive d'un vaste dispositif de régulation, d'encadrement et de contrôle des pratiques (à la fois thérapeutiques et de recherche), avec pour préoccupation centrale affichée « la protection des personnes ». La notion de « consentement **»** occupe dans ce dispositif une place centrale. Dès 1947, le texte fondateur dit « code de Nuremberg » y consacrait son Article Premier, et tous les textes ultérieurs, dans leur foisonnement et leur diversité, se sont toujours efforcé d'en préciser les contours (modalités d'obtention, portée, conditions de validité...) sans jamais cesser d'en réaffirmer le caractère incontournable et la valeur pour ainsi dire **«** absolue **»**. Envisagé comme acte, le consentement engage la personne en tant que « sujet », pouvant être tenu pour l'auteur de ses actes, c'est-à-dire susceptible de se les voir imputer, capable d'en rendre raison, disposé à les

assumer (Grillo, 2017). A l'instar de cet auteur, nous considérons que le consentement constitue la « croix » de la recherche en psychologie et sciences du comportement : il impose au nom de l'éthique des critères d'acceptabilité que l'extrême vulnérabilité des populations concernées (les « malades mentaux ») rend parfois inaccessibles, enfermant du même coup le chercheur dans un dilemme permanent : soit, au nom du principe déontologique de non-malfaisance, s'abstenir d'associer à une recherche toute personne dont le consentement ne peut raisonnablement être tenu pour effectif, tout franchissement de cette limite pouvant être assimilée à une forme de violence. Mais cela revient à barrer le chemin de la recherche...

4. Les transcription et annotations

Le corpus est l'élément fondateur du projet, les différentes analyses étant articulées à partir de cette ressource. Les transcriptions de départ se font par des fichiers contenant des marques temporelles. Nous produisons d'abord une version lisible des entretiens, puis nous augmentons la ressource de nouvelles annotations. Le premier objet était d'identifier les ruptures décisives, ce que nous avons conduit par un travail de repérage systématique, et sur chacune des ruptures nous avons construit une annotation sémantique. Si la première phase du projet a été motivée par ces identifications, il est rapidement apparu intéressant de ne pas limiter l'analyse linguistique au seul niveau sémantico-pragmatique. Nous avons ajouté des annotations à l'aide d'outils automatiques. La figure 1 présente cette organisation.

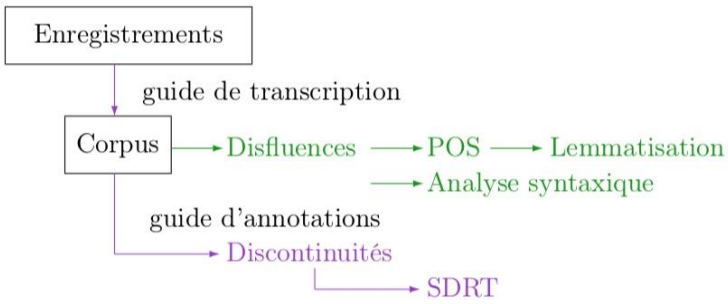

Figure 1. Organisation de l'annotation

Nous revenons dans la suite de cette section sur les différents aspects. D'abord nous nous intéressons à l'étape clé de la transcription qui convertit les enregistrements en une version écrite. Puis nous revenons sur les annotations produites automatiquement (partie en vert sur la figure). Enfin, nous nous intéresserons à la production d'annotation manuelle sur le plan de la sémantique et du discours (partie en violet).

4.1. La transcription

La transcription est la première composante pour le projet, sa qualité est donc primordiale. Nous avons préféré investir pour disposer d'une transcription manuelle, supposée de qualité. Avant cela nous avions testé des outils de transcription automatique de qualité. Bien qu'au niveau de l'état de l'art, ces outils produisent trop régulièrement des erreurs, de l'ordre d'un mot tous les dix mots. La correction s'avère avoir un coût très important rendant leur utilisation inefficace.

Le travail de transcription s'attache à rester traditionnel au sens où l'objectif est de produire une version écrite des entretiens. N'étant pas méthodologiquement outillés pour porter des analyses phonétiques et/ou phonologiques, ni à inscrire ce travail dans une communication multimodale, seule la production d'une version écrite a été retenue. Aucune marque représentant les mouvements ou les déplacements n'est incluse, et seules les marques de bruits de bouche ou de gorge nécessaires à l'interprétation sémantique sont retenues. Il s'agit donc de produire une transcription orthographique et non prosodique.

Nous avons défini spécifiquement un guide d'annotations, inspiré de (Blanche-Benveniste et Jeanjean, 1987) et (Blanche-Benveniste, 1997), qui décrit le plus exhaustivement possible comment la transcription doit être réalisée. Nous avons utilisé le logiciel Transcriber (Barras et al., 1998) pour ces opérations. Nous avons également procédé à des tests qualitatifs en faisant transcrire un même extrait par plusieurs transcripteurs pour calculer des accords inter-annotateurs (Cohen, 1960).

Une partie du guide explique aux annotateurs que nous cherchons à capter les réalités linguistiques et que l'objectif qualitatif du travail pour l'annotateur est de limiter les biais. Ainsi il n'est pas question de corriger ce qui est entendu pour le rendre plus acceptable ni même d'inventer de nouvelles formes écrites pour simuler la prononciation orale ou encore de chercher à donner une vision caricaturale des échanges. La nature des entretiens sur lesquels nous travaillons implique par ailleurs une certaine écoute entre les deux interlocuteurs, si bien que nous n'avons que très peu de phénomènes de recouvrement entre les locuteurs.

Les transcripteurs sont recrutés et payés pour leur travail. Du temps est pris au départ pour s'assurer de la bonne compréhension de la problématique et de la bonne maîtrise des outils. Le temps de sélection et de formation des transcripteurs s'élève à une vingtaine d'heures de travail par transcripteur. Pour cette phase du travail nous avons mobilisé 3 transcripteurs. À chaque patient inclus dans l'étude est associé l'enregistrement d'un entretien. Cet échange dure en moyenne 40 minutes, sachant que certains entretiens sont particulièrement courts. Nous avons évalué que le temps nécessaire à une transcription de bonne qualité de 5 minutes d'enregistrement était de 30 minutes. Nous avons choisi de considérer que le travail de transcription s'entendait pour la production d'une transcription complète, sa relecture et sa synchronisation.

Les annotateurs n'ayant pas une connaissance fine *a priori* des attendus de cette recherche, nous considérons qu'ils n'ont pas introduit de biais majeurs dans la constitution de la ressource. 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.

4.2. Les annotations automatiques

Les résultats de cette partie ont plus particulièrement fait l'objet de (Amblard, Fort, Demily et al., 2015). Une fois les entretiens transcrits, nous nous sommes intéressés à analyser la qualité de la production langagière pour différents niveaux d'analyse linguistique. Ne souhaitant pas porter une analyse exhaustive, nous avons utilisé des outils automatiques pour produire des annotations supplémentaires. Il s'agit des éléments en vert de la figure 1.

Un premier point a été de porter une analyse syntaxique automatique. Il n'existe pas d'analyseur pour le français parlé qui fasse consensus, en particulier parce que nous travaillons à partir de

transcriptions de l'oral où la notion de phrase avec structures syntaxiques bien définies n'est pas triviale. Par exemple (Deulofeu et al., 2010) argumentent pour une refonte en profondeur du schéma d'analyse de la syntaxe de l'oral. À l'inverse, (Abeillé et Crabbé, 2013) transposent les structures d'annotations de la syntaxe de l'écrit directement à celles de l'oral. La première étape de leur travail est l'identification des disfluences, i.e. des réalisations orales qui rompent la continuité syntaxique.

Pour repérer les disfluences, nous avons utilisé l'outil d'identification automatique Distagger (Constant et Dister, 2010), un outil libre qui présente de bonnes performances 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 %). Distagger permet d'identifier des réalisations de natures différentes, pour lesquelles quatre restent prédominantes dans les corpus oraux : les interjections d'hésitation ; la reprise à l'identique d'un mot ou d'un groupe de mots ; l'autocorrection immédiate ; l'interruption de morphème en cours d'énonciation. Distagger nous permet d'analyser la répartition des étiquettes de disfluences entre le groupe contrôle, les schizophrènes et les psychologues. Nous avons réalisé uniquement une analyse quantitative qui a révélé une très légère propension à la disfluence chez les schizophrènes par rapport aux autres groupes. Si cette différence reste faible, elle est validée par un test de significativité.

Une fois ces informations obtenues, nous nous sommes concentrés sur les étiquetages morphosyntaxiques (*part-of-speech* (POS)) et des lemmes. Un étiquetage morphosyntaxique est un processus qui associe à chaque mot d'un énoncé une information sur sa nature grammaticale. Il est aussi possible d'y associer des éléments morphologiques comme des propriétés de genre, nombre, etc. De manière simplifiée, le lemme associé à un mot d'un énoncé est la forme dont il est dérivé, par exemple le lemme de « relieront » est le verbe « relier ».

Cette analyse est réalisée avec l'outil MElt (Denis et Sagot, 2009), un tagger également librement disponible qui a été développé à partir de réseaux de neurones (perceptrons multicouches). Nous utilisons la version de l'outil entraînée sur le corpus TCOF-POS du français parlé (Benzitoun *et al*., 2012) et utilisons le lexique Le*fff* (Sagot, 2010). Ces ressources ont des caractéristiques suffisamment proches de notre corpus pour lui correspondre. Cet outil montre de bonnes performances, au niveau des outils de l'état de l'art avec une exactitude de 97,61% sur le *French Tree Bank*.

Nous appliquons l'outil sur chacune des transcriptions. Comme les modèles sont relativement lourds à mettre en mémoire, le temps de traitement reste significatif. L'annotation est appelée ligne par ligne sans contexte spécifique. L'outil découpe le texte en mots, et associe à chaque mot une catégorie morphosyntaxique et son lemme à partir de la phrase considérée. À partir de ces analyses nous avons constaté qu'aucun groupe ne se spécialisait du point de vue morphosyntaxique. Le nombre d'étiquettes lexicales reste très stable d'un entretien à l'autre et parfaitement cohérent sur l'ensemble du corpus. Les distributions des POS utilisées tant par les schizophrènes, que par le groupe témoin et le psychiatre semblent similaires.

Afin d'affiner nos résultats nous avons calculé la richesse lexicale de chaque groupe (ratio du nombre de lemmes par rapport au nombre total de formes) et la diversité lexicale (ratio du nombre de lemmes par rapport, cette fois, au nombre total de formes différentes (types)). Cette dernière réduit l'influence des mots très utilisés dans le calcul de la richesse lexicale. Ces calculs ne font pas non plus apparaître de comportement spécifique. Les schizophrènes ont une diversité lexicale et une richesse lexicale équivalentes aux autres catégories testées, ce qui est confirmé par le calcul de la

significativité que l'on retrouve dans (Amblard, Fort, Demily et al., 2015). La production langagière des schizophrènes apparaissant comme standard pour la morphosyntaxe, il serait intéressant d'analyser ce qu'il en est pour la maîtrise des structures syntaxiques complexes. On pourrait par exemple étudier les enchâssements multiples de relatives qui peuvent être considérés comme des indicateurs d'une capacité à planifier l'action discursive (compétence neuropsychologique et cognitive). Malgré l'attention portée à l'ensemble de l'étude, plusieurs biais restent présents. (Amblard, Fort, Demily et al., 2015) reviennent en détail sur ces aspects.

4.3. Les annotations manuelles

Nous avons également choisi de proposer des annotations pour le niveau sémantico-pragmatique. À la différence des précédentes, il n'est pas possible de produire des annotations couvrant entièrement la ressource, leur production étant très coûteuse. Nous avons donc tout d'abord cherché à identifier des extraits des transcriptions présentant des ruptures décisives au sens de Musiol (2009). Pour ces extraits, nous discutons de la production d'une représentation formelle de l'interaction. Ces deux étapes d'annotations sont réalisées par des experts. En effet, produire une représentation fine de l'interaction suppose d'avoir une connaissance préalable des propriétés des représentations. Par exemple, la définition des ruptures décisives implique de prendre en compte des interactions complexes sur plusieurs tours de parole. Cette analyse n'intervient pas strictement sur le plan sémantique, mais utilise des propriétés de la planification du dialogue.

Afin de rendre compte formellement de ces analyses, nous avons entrepris de modéliser les entretiens à l'aide de la SDRT (*Segmented Discourse Representation Theory*). Ce cadre logico-formel, créé par Asher & Lascarides (2003), permet une analyse des dimensions sémantiques et pragmatiques du discours, ce qui constitue une extension notable des théories formelles développées jusque-là, généralement cantonnées à l'une ou l'autre de ces dimensions. La SDRT est plus spécifiquement une extension de la DRT (*Discourse Representation Theory*) de Kamp (Kamp & Reyle 1993), qui vise à représenter la sémantique du discours en y incluant certains traits de la dynamique discursive, mais sans s'aventurer vers l'interprétation elle-même, ni sur le champ de la pragmatique.

Nous avons obtenu une première série de résultats à différents niveaux (Amblard *et al.* 2015a, 2015b). Nous avons tout d'abord restreint l'outillage théorique au strictement nécessaire : la modélisation des relations rhétoriques est prépondérante et en cela la partie plus classique issue de la DRT n'est pas directement utilisée. Il faut cependant nuancer ici car une unité thématique doit être vérifiée, qui peut aisément être validée au travers des formules logiques de la DRT. Nous avons donc étendu la SDRT à partir des arbres rhétoriques en leur ajoutant des boîtes thématiques qui représentent la portée des thèmes en jeu dans l'interaction. Ensuite, nous avons analysé chaque extrait révélant des discontinuités manifestes pour comprendre ce qui ne fonctionnait pas.

Il est apparu nécessaire de faire plusieurs postulats. Tout d'abord nous avons supposé que le patient schizophrène et le psychologue n'ont pas la même représentation mentale de l'interaction, au contraire de deux intervenants dans un dialogue ordinaire qui cherchent à construire une représentation commune. Ensuite, relevant que l'irrationalité apparente chez le schizophrène pouvait aisément être placée à différents niveaux, nous avons choisi d'adopter le principe de charité de Quine (1960), ce qui nous conduit à assumer que le patient est cohérent pour lui-même au

moment où il intervient dans l'interaction. Cela implique que les prises de paroles du patient sont alors logiques (au sens commun) au moment où elles sont réalisées, ceci expliquant la nécessité de construire deux représentations du discours différentes, une pour le patient, l'autre pour son interlocuteur.

Nous avons montré que les discontinuités significatives correspondaient à des usages non-usuels en SDRT. Deux phénomènes se présentent : la remontée au travers de l'arbre de représentations sans consistance de la partie précédemment explorée et la rupture de la frontière droite. Cette dernière propriété est fondamentale dans la définition de la SDRT et permet de limiter les lieux de la structure à partir desquels la construction peut se poursuivre. Elle permet par exemple de résoudre les problèmes de surgénération de la DRT pour la résolution d'anaphores. Un fait important, est que si la règle de la frontière droite permet d'améliorer le calcul, elle n'a pas par principe de réalité cognitive. Le fait qu'elle puisse être à l'œuvre dans ce phénomène pathologique permet d'en valider l'utilisation du point de vue cognitif également.

La représentation en SDRT de chaque extrait discontinu a été discutée au moins deux heures par trois experts pour obtenir un consensus. Cette indication de temps cherche à illustrer le type de complexité que les phénomènes recouvrent. Cependant, afin de vérifier la qualité des représentations et limiter la validation intersubjective, nous avons mis en place un protocole d'annotation par des non experts des extraits identifiés. Ce travail a consisté dans un premier temps à la définition d'un guide d'annotation qui présente de manière simple le type de relations utilisable. Nous avons défini une interface graphique fondée sur le logiciel Glozz (Mathet et Widlöcher, 2011). Les premières annotations ont montré que la tâche était particulièrement complexe. Le temps nécessaire à la production des annotations s'est révélé prohibitif. Une seconde phase d'annotation sur un schéma simplifié et des extraits resserrés donne des résultats plus encourageants. La figure 2 présente une annotation produite par un annotateur naïf pour la tâche. Nous disposons actuellement de l'annotation par 46 annotateurs de trois extraits. Les analyses de similarités sont en cours, mais les premiers résultats montrent que des formes de consensus apparaissent sans toutefois présenter une convergence entre tous les annotateurs. La définition d'une métrique d'accord ou de convergence reste particulièrement délicate pour ce type d'annotations.

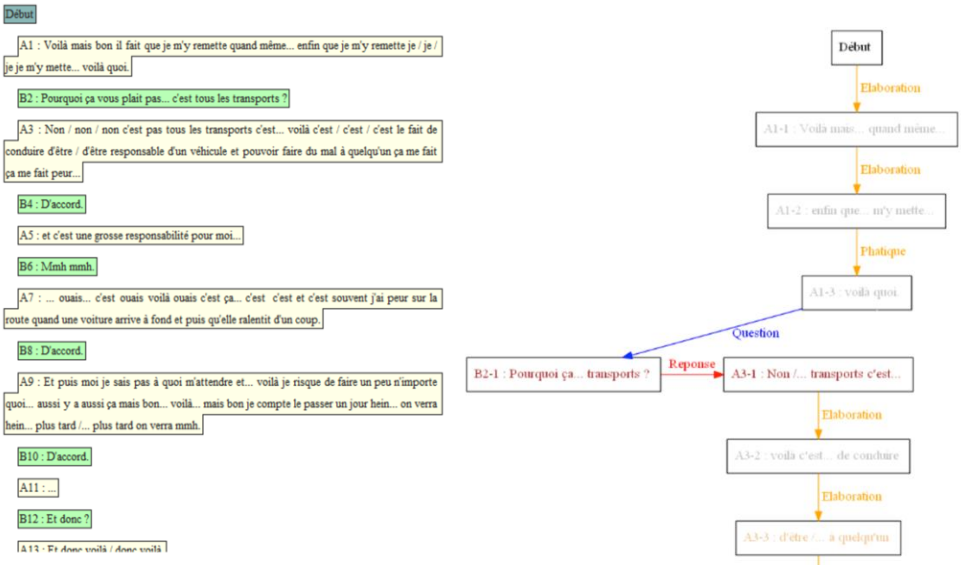

Figure 2. Annotation produite par un annotateur naïf avec Glozz

5. Conclusion

Les recherches succinctement présentées dans cet article rencontrent de multiples obstacles. Les obstacles matériels et institutionnels surgissent dès le recueil de données, alors que celles-ci constituent la matière première de l'ensemble des travaux. Les enjeux éthiques et juridiques affleurent dès lors que la santé mentale est en jeu et que le libre-arbitre des sujets est en question. On ne peut pas, d'un point de vue éthique ni juridique, faire fi du consentement des sujets et de l'avis de leur entourage familial et médical, lorsque l'on travaille sur un corpus impliquant des personnes schizophrènes. L'équilibre est difficile à tenir entre les exigences méthodologiques de la linguistique de corpus, pour laquelle la publicité des données est un prérequis, et l'exigence éthique de la protection des données et des personnes, quand bien même la recherche vise à une meilleure compréhension de la pathologie pour le bénéfice de ces personnes.

Nous avons présenté les différentes étapes de la recherche : recueil des données dans un environnement complexe et strictement réglementé ; transcription pour l'obtention d'une ressource primaire à soumettre à l'analyse ; traitement automatisé pour le repérage des disfluences et l'étiquetage morphosyntaxique ; traitement manuel pour l'analyse sémantico-pragmatique formelle. La dernière étape est certainement celle qui recèle le plus d'enjeux. S'intéressant aux compétences langagières de haut niveau des interlocuteurs, la conduire suppose la maîtrise de concepts théoriques généraux que l'on ne trouve pas communément chez les non linguistes. Les campagnes d'annotation butent sur cette difficulté qu'elles ne peuvent pas seulement en appeler aux compétences ordinaires de locuteurs-annotateurs natifs puisqu'elles doivent les informer minimalement des concepts utiles à l'annotation.

Cette tension n'est toutefois peut-être rien d'autre que la manifestation des prérequis les plus généraux de l'intercompréhension. En effet, à tenter d'appréhender la cohérence des conversations schizophrènes, les travaux du groupe ont ouvert une voie significativement différente de celles empruntées jusque-là. Le point de vue reconstruit est celui d'une appréhension en première personne, qui s'oppose clairement au neuro-réductionnisme qui inspire largement les recherches sur la pathologie mentale (Rebuschi et al. 2013). L'approche rejette ainsi *de facto* la réduction de l'explication du dysfonctionnement à l'expression d'un gène ou à l'activation d'une structure neuronale particulière chez le patient. Nous assumons ainsi le fait que d'autres aspects du langage et de la psychologie sont nécessaires pour expliquer et comprendre la pathologie.

Références

Abeillé, A. et Crabbé, B. (2013). Vers un treebank du français parlé. *Traitement Automatique des Langues Naturelles* (TALN). Les Sables d'Olonne, France, 174–187.

Amblard, M., Fort, K., Demily C., Franck, N. et Musiol, M. (2015). Analyse lexicale outillée de la parole transcrite de patients schizophrènes. *Traitement Automatique des Langues. Natural Language Processing and Cognition* 55.3, 91–115.

Amblard, M., Musiol, M., and M. & Rebuschi, M. (2015b). L'interaction conversationnelle à l'épreuve du handicap schizophrénique. *Recherches sur la Philosophie et le Langage*, 31, 67-89.

Asher, N. Lascarides, A., (2003). *Logics of Conversation*. Cambridge, Cambridge University Press.

Barras, C. et al. (1998). Transcriber: A Free Tool for Segmenting, Labeling and Transcribing Speech. *International Conference on Language Resources and Evaluation* (LREC), Granada, 1373–1376.

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

Blanche-Benveniste, C. (1997). *Approches de la langue parlée en français*. Collection L'Essentiel français. Gap, France : Ophrys

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

Chaïka, E. O. (1974) A linguist looks at 'schizophrenic' language. *Brain and Language*, 1, 257-276.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1), 37–46.

Constant, M. et Dister, A. (2010). Automatic detection of disfluencies in speech transcriptions. In Pettorino, M. *et al.* (eds.), *Spoken Communication*, T. 1. Cambridge Scholars Publishing, 259–272 .

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

Deulofeu, J. *et al*. (2010). Depends on What the French Say Spoken Corpus Annota- tion with and Beyond Syntactic Functions, *Proceedings of the Fourth Linguistic Annotation Workshop* (LAW IV '10). Uppsala, Sweden, Association for Computational Linguistics, 274–281.

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

Grillo, E. (2017). Quand y a-t-il "Personne"? L'éthique de la recherche au défi du consentement. Workshop *Ethique de la recherche en psychologie et sciences du comportement*. Société Française de Psychologie, Université Toulouse 2, 29 mars 2017.

Kamp, H., Reyle, U., (1993). *From Discourse to Logic*. Kluwer Academic Publishers.

Mathet, Y., Widlöcher, A. (2011). Stratégie d'exploration de corpus multi- annotés avec GlozzQL. *Actes de la 18e Conférence Traitement Automatique des Langues Naturelles* (TALN'11), Vol. 2, papiers courts. M. Lafourcade & V. Prince (eds.), Montpellier, France, 143–148.

Musiol, M. (2009). Incohérence et formes psychopathologiques dans l'interaction verbale schizophrénique. In J. Rozenberg, N. Franck & C. Hervé (eds.), *Des neurosciences à la psychopathologie : Action, Langage, Imaginaire*. Bruxelles : De Boeck, 219-238.

Musiol, M. & Verhaegen, F. (2014). Investigating discourse specificities in schizophrenic disorders. In M. Rebuschi et al. (eds.), *Interdisciplinary Works in Logic, Epistemology, Psychology and Linguistics*, Springer, Dordrecht, 317-344.

Quine, W. V. O., *Word and Object*, Cambridge, Mass., MIT, 1960.

Rebuschi, M., Amblard, M., Musiol, M. (2013). Schizophrénie, logicité et perspectives en première personne. *L'Evolution Psychiatrique*, 78/1, 127-141.

Roulet, E. *et al*. (1985) *L'articulation du discours en français contemporain*. Berne, Peter Lang.

Sagot, B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. *7th international conference on Language Resources and Evaluation* (LREC 2010). Valletta, Malta.

Trognon, A., Musiol M. (1996). L'accomplissement interactionnel du trouble schizophrénique. *Raisons Pratiques* 7, 179-209.