L'impossibilité de l'anonymat dans le cadre de l'analyse du discours

Maxime Amblard[1,2], **Karën Fort**[3], **Michel Musiol**[1,4], **Manuel Rebuschi**[1,5]
[1]Université de Lorraine, MSH-Lorraine USR 3261 [2]LORIA, UMR 7503 (INRIA/CNRS/UL)
[3] STIH, EA 4089 (Université Paris-Sorbonne) [4]ATILF, UMR 7118 (CNRS /UL) [5]LHSP-AHP, UMR 7117 (CNRS/UL)
maxime.amblard, manuel.rebuschi, michel.musiol@univ-lorraine.fr, karen.fort@paris-sorbonne.fr

Abstract

Cette proposition revient sur les questions soulevées par le projet SLAM qui s'intéresse au discours des schizophrènes. Ces derniers manifestent un dysfonctionnement du maintien de la cohérence de l'interaction dialogique. Cette étude nous a confronté à plusieurs niveaux de difficultés dans la constitution du corpus sur lesquelles nous revenons ici. Par ailleurs, il est évidemment délicat de proposer des modèles de représentation de l'action de pensée par le biais de la pathologie sans avoir une réflexion épistémologique et éthique sur l'implication sociétale pour les patients des conclusions avancées.

Keywords: anonymisation, annotation, dialogue, discours, schizophrène, troubles mentaux

1. Introduction

Nos travaux s'inscrivent dans le cadre du projet SLAM[1] (Schizophrénie et Langage : Analyse et Modélisation). L'un des objets de ce projet est de proposer une analyse formelle du discours schizophrénique. L'étude linguistique de la production langagières des troubles de la pensée a émergé dans les année 70, notamment dans (Chaika, 1974). Par ailleurs, les avancées en traitement automatique des langues (TAL) sont nombreuses depuis cette époque, et ce type de données n'a pas été investigué par le prisme de ces outils. C'est là l'un des objectifs du projet SLAM qui est décrit dans la section 2. La constitution du corpus de travail a soulevé plusieurs points méthodologiques relevant de questions éthiques sur lesquels nous revenons dans la section 3. Quand bien même les outils actuels permettent d'assurer un anonymat satisfaisant, la nécessité de travailler sur de longs extraits d'entretiens conduit à une impossibilité de garantir l'anonymat, comme discuté dans la section 4. Enfin, étudier le langage par la pathologie, en particulier des troubles mentaux, induit de respecter une forme d'éthique quant aux répercussions sur les patients étudiés. Nous revenons sur ces aspects dans la section 5.

2. Le projet SLAM

Le projet SLAM s'intéresse aux pratiques langagières chez les schizophrènes en situation d'entretiens avec un psychologue. Outre le contenu explicite de l'entretien, d'autres aspects sont analysés dont 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 de l'encéphale par des enregistrements par électro-encéphalogramme (EEG).

La motivation originelle du projet provient des résultats de (Trognon and Musiol, 1996) qui mettent en avant des *discontinuités pragmatiques* dans l'accomplissement de l'interaction verbale. Ces dysfonctionnements ont été interprétés comme des manifestations pathologiques de la planification du discours. Certaines de ces discontinuités, dites décisives, seront analysées ensuite comme véhiculant la trace de troubles cognitifs fonctionnels affectant le langage et/ou les processus psychologiques (Musiol, 2009). En cela, les patients schizophrènes rejouent des ambiguïtés linguistiques précédemment résolues dans l'interaction. L'interprétation pragmatique et rhétorique devient alors impossible. Dans le projet SLAM, nous faisons appel aux théories formelles du discours développées ces dernières années, en particulier la SDRT, (Asher and Lascarides, 2003), et nous les confrontons à ces manifestations. (Rebuschi et al., 2014) ont ainsi mis en avant des corrélations explicites.

Le projet se décompose en trois thématiques, chacune confrontée à des problèmes d'éthique :

- aspects épistémologiques (norme, folie, rationalité)
- aspects formels de la modélisation du dialogue
- constitution du corpus, objet de la section suivante.

3. Constitution de la ressource

Le corpus utilisé est constitué de transcriptions d'entretiens. L'étude fait intervenir 79 sujets, 48 patients schizophrènes en remédiation et sous traitement (à l'exception de 7), et 31 témoins. Les entretiens sont réalisés par des psychologues, en milieu hospitalier.

L'interaction est un entretien semi-directif conduit par un psychologue : ce dernier 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 l'entretien ne peut être utilisé comme base médicale.

3.1. L'accès aux patients

Le nombre de 79 sujets peut sembler limité, mais la constitution d'une telle ressource implique de surmonter de nombreuses difficultés, en particulier pour accéder aux patients. Dès lors, 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 nécessaire d'obtenir une autorisation du CPP (Comité de Protection de la Personne) de la région de l'établissement. Les demandes contiennent explicitement et exactement le protocole de test. L'instruction

[1]Le projet SLAM est soutenu par la MSH-Lorraine et le CNRS au travers d'un PEPS HuMaIn.

du dossier requiert plusieurs mois et elle exige la contraction d'une assurance (pour couvrir les possibles dommages). De ce fait, les budgets nécessaires augmentent considérablement. En outre, une fois les accords obtenus, il n'est plus possible de modifier les protocoles.

Assurer la participation des patients est aussi difficile : il faut identifier, au sein d'un service, ceux répondant aux critères de l'étude et en capacité d'interagir avec une personne tierce à ce service ; au sein de cette population, trouver ceux qui acceptent de participer. Une première réticence vient du fait qu'il n'y a pas, en terme médical, de conséquence positive pour le patient à participer à l'étude. Il faut ajouter à cela des inquiétudes compréhensibles concernant la possible publication de leur histoire, bien qu'une anonymisation totale leur soit garantie.

Par ailleurs, le temps nécessaire pour passer le protocole est significatif (environ deux heures). Ce n'est pas tant la disponibilité des patients qui est alors en jeu, que leur concentration. Lorsque le patient présente soudainement des difficultés, il faut convenir d'un second rendez-vous, au risque de générer des défections. À titre d'exemple, lors d'une phase récente de collecte, 45 % (18 sujets) des patients ont refusé de participer, 10 % (4) ont accepté un premier rendez-vous mais ne sont pas présentés au second, et 45 % (18) ont participé à toute l'étude.

3.2. Anonymisation classique

L'anonymisation est la première tâche après la transcription. Nous suivons une approche tout à fait classique qui consiste à identifier les entités nommées et à les substituer par des marqueurs sémantiquement vides. Un outil automatique performant a été identifié, mais n'a pu être opérationnel à temps. Nous avons donc programmé une série de scripts en Python basés sur des expressions régulières et mis en place un post traitement humain. Nous pouvons ainsi assurer une anonymisation fiable de la transcription et de la bande son par ajout de *bips* sonores.

4. De l'impossibilité de l'anonymat

L'anonymisation du corpus ne s'arrête cependant pas là. Les sujets relatant des événements s'inscrivant dans une temporalité et une géographie particulières, des indices sont disséminés dans les entretiens. Il est donc possible d'identifier des élément biographiques. Il est difficile de trouver une solution à ce problème tout en conservant l'intégrité des entretiens. Cette particularité a des conséquences importantes sur le projet.

Pour les traitements ne nécessitant qu'un faible contexte (phrase ou tour de parole), nous avons créé une version randomisée de la ressource. Les tours de paroles étant mélangés, il devient impossible de reconstituer les historiques. Il est donc possible de fournir la ressource pour des analyses morpho-syntaxiques ou syntaxiques, sans compromettre les données initiales. Une trace de la randomisation est cependant conservée sous forme de table pour reconstruire les entretiens originaux.

Mais l'un des objectifs du projet 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. Ce problème est intrinsèque au niveau linguistique choisi pour l'analyse. 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 implique une restriction d'accès à ces données.

5. De la réalité des patients

Nous considérons important de revenir sur les aspects épistémologiques et éthiques de l'étude en elle-même. En effet, analyser formellement le langage relève de la volonté de définir une norme dont la déviance serait manifeste d'un dysfonctionnement, en l'occurrence ici d'un trouble de la pensée. Or, tout locuteur est confronté quotidiennement à des troubles du langage provenant de personnes saines. Si le projet permet la définition d'indices qui participent à la pose d'un diagnostic et à la mesure de l'efficacité des processus de remédiation, ils ne doivent pas en être l'argument. En effet, il apparaît clairement dans la littérature que la vision et la prise en charge des troubles mentaux est une question qui évolue grandement, et qui doit être abordée avec délicatesse, en particulier lorsqu'il s'agit de schizophrénie. Les traitements sont lourds et le diagnostic ne peut souffrir de l'approximation récurrente dans les outils du TAL.

6. Conclusion

La matière de ce type de projet est confrontée, à tous les niveaux, à des questionnements et une gestion des aspects éthiques qu'on ne retrouve pas habituellement dans le TAL. Il nous faut à la fois surmonter les difficultés dans l'accès aux patients et dans la gestion du groupe de patients, et identifier les conséquences sur la vie des patients que peuvent avoir les conclusions de l'étude. Par ailleurs, il apparaît impossible de ne pas rompre l'anonymat des personnes sans briser la notion de discours qui est au cœur même du projet. En conséquence seuls les membres engagés dans le projet et soumis à un devoir de confidentialité peuvent travailler sur son intégralité.

7. References

Nicholas Asher et Alex Lascarides. 2003. *Logics of Conversation*. Studies in Natural Language Processing. Cambridge University Press.

Elaine Chaika. 1974. A linguist looks at "schizophrenic" language. *Brain and Language*, 1(3):257–276, July.

Michel Musiol. 2009. Incoherence et formes psychopathologique dans l'interaction verbale schizophrenique. In *Psychose, langage et action (approches neuro-cognitives)*, pages 219–238. De Boeck, Bruxelles.

Manuel Rebuschi, Maxime Amblard, et Michel Musiol. 2014. Using SDRT to analyze pathological conversations. Logicality, rationality and pragmatic deviances. In *Interdisciplinary Works in Logic, Epistemology, Psychology and Linguistics: Dialogue, Rationality, and Formalism*, Logic, Argumentation & Reasoning, pages 1–24. Springer.

Alain Trognon et Michel Musiol. 1996. L'accomplissement interactionnel du trouble schizophrénique. *Raisons Pratiques 7*, pages 179–209.