



HAL
open science

Vers un repérage automatique des discontinuités dans le discours pathologique du sujet schizophrène

Vincent-Thomas Barrouillet, Maxime Amblard, Michel Musiol

► To cite this version:

Vincent-Thomas Barrouillet, Maxime Amblard, Michel Musiol. Vers un repérage automatique des discontinuités dans le discours pathologique du sujet schizophrène. Journées scientifiques du GDR Lift - LIFT 2023, Nov 2023, Nancy, France. hal-04356448

HAL Id: hal-04356448

<https://inria.hal.science/hal-04356448>

Submitted on 20 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Vers un repérage automatique des discontinuités dans le discours pathologique du sujet schizophrène

Barrouillet Vincent-Thomas¹ Michel Musiol^{1,2} Maxime Amblard²

(1) Université de Lorraine, CNRS, ATILF, F-54000 Nancy, France

(2) Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
vincent-thomas.barrouillet@univ-lorraine.fr

RESUME

Le discours du sujet schizophrène en phase aiguë présente de multiples bizarreries. Un sous-ensemble de ces dernières a été modélisé en tant que « ruptures ». En étude systématique de corpus, on dégage 27 séquences comportant 47 ruptures dans des entretiens entre des patients et des psychologues. Nous définissons un modèle qui vise à retrouver de façon *in fine* automatique ces ruptures. À partir de la segmentation en actes des tours de parole, nous définissons les actes pertinents pour chaque interactant au moyen de règles de décision non ambiguës. Enfin, un calcul algorithmique permet de repérer les ruptures attendues. Nous obtenons une F-mesure de 88.9 % et une exactitude de 96.2 %. L'analyse qualitative des données montre que les 40 vrais positifs et les 3 faux positifs relèvent d'une même catégorie, les « ruptures thématiques », tandis que le système échoue à repérer les autres catégories de ruptures.

ABSTRACT

Toward an automatic identification of discontinuities in the pathological discourse of patient with schizophrenia

Pathological discourse of schizophrenics in acute phase presents multiple oddities. A subset of them was modeled as “discontinuities”. Here, 27 sequences with 47 discontinuities were manually identified in our corpus of interviews between a patient and a psychologist. We aim to find these discontinuities *in fine* automatically. From the segmentation of the turns into acts, we precise all the relevant acts for each interactant and each turn, thanks to unambiguous decision rules. Finally, an algorithm identifies the discontinuities. We achieve a F-measure of 88.9% and an accuracy of 96.2%. By analyzing the data, we point out that the 40 true positives and the 3 false positives fall into the same category, “thematic discontinuity”, while the system fails to identify all other categories of discontinuities.

MOTS-CLES : discours, discours pathologique, schizophrénie, corpus

KEYWORDS: discourse, pathological discourse, schizophrenia, corpus

1 Contextualisation

Le discours pathologique du sujet schizophrène est parsemé de bizarreries que l'on repère intuitivement. Un sous-ensemble de ces bizarreries a été formalisé sous la forme de ruptures, qui sont des discontinuités pragmatiques dans le discours. Un modèle général a ainsi émergé identifiant quatre types de ruptures (Musiol, 2009) qui sont le débrayage conversationnel et la déféctuosité de l'initiative conversationnelle d'une part, et les ruptures inter et intra d'autre part. Les deux premières sont prédites par un modèle adapté de la SDRT (Asher & Lascarides, 2003; Rebuschi et al., 2014). Ici, nous nous intéressons plus particulièrement à la partie du modèle prévoyant les deux dernières catégories : les ruptures inter, i.e. ayant lieu entre deux tours de parole, et les ruptures intra, i.e. ayant lieu à l'intérieur d'un tour de parole. Ces ruptures sont objectivement et manuellement mises à jour au moyen de la modélisation hiérarchique (Roulet et al., 2001), mais nous cherchons à les identifier par un système *in fine* automatique qui ne passe plus par la modélisation hiérarchique.

2 Présentation du corpus

Le corpus est constitué d'entretiens qui se sont tenus à l'hôpital psychiatrique de Rouen en 2005 et à l'hôpital de Tizi-Ouzou, en Algérie, en 2019. Nous avons obtenu un CPP (« Comité de Protection de la Personne ») pour le corpus de Rouen, mais l'absence de comité d'éthique à Tizi Ouzou ne nous a pas permis d'en présenter un. Aussi, afin de garantir la qualité de la méthodologie appliquée, nous avons réemployé un protocole pour lequel nous avons obtenu un accord la même année auprès du CHRU d'Aix-En-Provence. Le protocole consiste en la réalisation d'un ou deux entretiens non directifs, sans but explicite, permettant au sujet d'élaborer sur son environnement et sa vie quotidienne. Les entretiens sont conduits par plusieurs psychologues ayant reçu les formations nécessaires.

Le corpus, qui rassemble 26 heures d'entretiens, a ensuite été retranscrit manuellement par quatre transcripateurs entraînés. Un guide de transcription a été élaboré pour normaliser la production, et les deux transcripateurs ont été entraînés sur des données parallèles afin de contrôler leur production. Nous n'avons pas calculé le Kappa de Cohen car leur entraînement conjoint a été fait sur un volume réduit de données. Le matériel textuel obtenu est désidentifié des noms de personne, de ville, d'institution, etc., et augmenté de balises *ad hoc* telles que <rire> <chuchotement> <claquement de porte>, etc.

Une étude manuelle par deux spécialistes différents des transcripateurs a été effectuée, en application du modèle de (Roulet et al., 2001), pour identifier l'ensemble exhaustif des séquences où les ruptures apparaissent. L'étude a permis d'identifier 27 séquences de ruptures. Les séquences sont composées de plusieurs tours de parole que nous segmentons

en actes de langage (Searle & Vanderveken, 1985). Comme il n'existe pas de système automatique ayant les performances suffisantes, la segmentation est réalisée par un annotateur, et le résultat est corrigé et validé par un second annotateur.

3 Méthodologie

Notre méthode est basée sur la notion de « mise à jour ». Nous assumons qu'un acte est mis à jour, au sens large, par un acte d'un tour ultérieur du même interactant si le second a explicitement besoin du premier pour exister, comme dans un couple question / réponse, ou dans le cas d'une reformulation, d'une reprise, d'une élaboration, ou encore d'une répétition. Nous envisageons l'un après l'autre l'ensemble des actes d'un tour pour l'un puis pour l'autre des interactants et nous formalisons 5 règles explicitant les traitements à appliquer à chaque acte. Si un acte n'est pas mis à jour par la suite, il n'est pas retenu comme acte pertinent (règle A1). S'il est mis à jour par la suite pour le même interactant, il est retenu comme acte pertinent de tour en tour pour cet interactant (règle A3) jusqu'à sa mise à jour où il n'est plus pertinent (règle A4). L'ensemble des autres règles s'applique pour statuer sur la pertinence éventuelle de l'acte qui acte met à jour le précédent. Et il n'y a aucun acte pertinent dans le dernier tour de parole de la séquence (règle A5).

Toutes les règles font l'objet d'une description formelle, mais la règle A2 reste la plus complexe d'un point de vue conceptuel. Chaque tour de parole possède un et un seul acte dit « prévalent ». L'acte prévalent est l'acte « le plus important » du tour de parole qui se calcule comme celui dont les autres actes dépendent. Il est toujours possible de retirer soit un, soit un groupe d'actes secondaires, mais la suppression de l'acte prévalent dans le tour de parole lui retire la possibilité de faire sens.

Ce concept est une modélisation de l'« acte principal » défini dans le modèle hiérarchique de (Roulet et al., 2001) et des travaux en cours visent à automatiser le calcul de l'acte prévalent, afin de se libérer totalement de la modélisation hiérarchique dans la détection automatique des ruptures. La règle A2 stipule que l'acte prévalent d'un interlocuteur est forcément pertinent pour le tour en cours et pour l'interlocuteur concerné. Et s'il n'est pas mis à jour par la suite, il n'apparaîtra qu'une fois dans le tour de parole qu'il structure.

Nous utilisons un tableau à deux entrées contenant les tours de parole pour les deux interactants. Dans chaque case se trouvent les actes pertinents à échéance d'un tour de parole pour le patient ou le psychologue. Les actes pertinents en vertu de la règle A3 portent en indice le numéro de l'acte qui le mettra à jour, et les actes pertinents en vertu de la règle A2 ne portent pas d'indice, car ils n'apparaissent qu'une fois.

Par la suite, un algorithme détermine parmi les actes pertinents ceux qui ne doivent pas être pris en compte dans le calcul des ruptures, et qu'il faut retirer — ces actes sont appelés « actes fantômes ». Enfin, si les deux ensembles d'actes restants, patient et psychologue, ne

sont pas identiques pour un tour de parole donné, le système repère une rupture à échéance de ce tour.

4 Pseudo-code de l'algorithme de détection des ruptures

Dans cette section, nous proposons une description algorithmique de l'étape finale de la détection des ruptures. Nous commençons par les notations initiales.

$R := [R_0, R_1, \dots, R_k]$ # le tableau complet R se compose des tours de parole R_0 à R_k

$R_i := [R_{i0}, R_{i1}]$ # le tour R_i se compose des cases du patient (0) et du psychologue (1)

$R_{ij} := [A_0, A_1, \dots]$ # la case R_{ij} se compose d'une liste d'actes A_0, A_1, \dots

$A_n := [\text{corps}, \text{indice}]$ # l'acte A_n se définit comme un corps et un indice

corps := étiquette # le corps d'un acte est une étiquette qui en donne le numéro (A, B, C...)

indice := rien ou étiquette # l'indice d'un acte est l'étiquette de l'acte qui le mettra à jour. Il n'y a pas d'étiquette en cas d'application de la règle A2

Nous poursuivons par la définition d'une fonction supprimant les éléments non pertinents.

fonction $R_{ij_sans_fant\^ome}(i, j)$: # retirer les actes fantômes de $R(i, j)$

. liste_fantômes est initialisé comme liste vide

. pour tous les actes A de $R(i, j)$

. . pour tous les actes B de $R(i, 1 - j)$

. . . si $A = B$:

. . . . alors : passer au A suivant

. . . . sinon : pour tous les actes C de $R(i - 1, 1 - j)$

. si (corps de $A =$ corps de C) et (corps de $B =$ indice de C) alors :

. A est placé dans liste_fantômes

. pour tous les actes de $R(i, j)$, retirer les fantômes et renvoyer le résultat

fin de la fonction

Enfin, nous abordons la fonction qui identifie les ruptures.

fonction trouver_rupture(k): # k est le nombre de tours

. pour i allant de 0 à k # pour chaque tour i

. . pour les deux valeurs de j # les deux interactants

. . . R(i, j) prend la valeur de Rij_sans_fantome(i, j)
 . . . si R(i, 0) différent de R(i, 1), alors, il y a rupture à échéance du tour i
 fin de la fonction

5 Exemple : Voyage & Incontinence

Nous développons dans la suite un exemple complet d'analyse d'une rupture.

Actes de langage par tour de parole	Interactant S $R_{i,0}$	Interactant P $R_{i,1}$
S_0 A = J'ai toujours une belle-mère et un beau-père. <u>B</u> = Mon beau-père, depuis qu'il est à la retraite, s'absente neuf mois... C = Et quand il revient, il est violent.	$R_{0,0}$ $\{B_M ; C_O\}$	$R_{0,1}$ $\{B_D ; C_D\}$
P_0 <u>D</u> = Il va où pendant ces neuf mois ?	$R_{1,0}$ $\{B_M ; C_O ; D_E\}$	$R_{1,1}$ $\{D_F\}$
S_1 <u>E</u> = Oh, je ne sais pas où il va, puisque je ne l'ai jamais accompagné...	$R_{2,0}$ $\{B_M ; C_O ; E\}$	$R_{2,1}$ $\{E_F ; D_I\}$
P_1 <u>F</u> = Il vous dit pas ?	$R_{3,0}$ $\{B_M ; C_O\}$	$R_{3,1}$ $\{F_I\}$
S_2 <u>G</u> = Il est incontinent. H = Hachem. I = Il est incontinent mon beau-père. J = Mon beau-père est incontinent comme un bébé. K = Ou un petit garçon jusqu'à trois ans.	$R_{4,0}$ $\{B_M ; C_O ; G\}$	$R_{4,1}$ $\{F_I\}$
P_2 <u>L</u> = Et du coup, il s'absente 9 mois ?	$R_{5,0}$ $\{B_M ; C_O ; L_M\}$	$R_{5,1}$ $\{L\}$
S_3 <u>M</u> = Depuis qu'il est à la retraite, oui, il voyage.	$R_{6,0}$	$R_{6,1}$

N = Il s'absente 9 mois O = Et quand il revient, il est violent. ... U = Et je suis tombée. V = Et j'ai saigné.	\emptyset	\emptyset
---	-------------	-------------

Les $R_{i,j}$ du tableau sont l'ensemble des actes pertinents au regard des règles de décision. L'algorithme donné ci-dessus calcule les actes fantômes qui sont représentés ici en vert. Lorsque $R_{i,0}$ et $R_{i,1}$ ne sont pas identiques, après le retrait des actes fantômes, la rupture est repérée, ici pour $i = 3$ et pour $i = 4$, ce qui signifie qu'il y a une rupture, ici inter, à échéance de ces tours de paroles. Le tour de la patiente noté S2 est donc encadré par deux ruptures thématiques et se constitue donc une sorte d'ilot isolé du reste de l'interaction.

6 Résultats et discussion

Parmi les 27 séquences du corpus, le système repère 40 vrais positifs et 213 vrais négatifs, ainsi que 3 faux positifs et 7 faux négatifs. On obtient donc une exactitude de 96,2 % et une F-mesure de 88,9 %, ce qui est un bon résultat au vu de la complexité de la tâche. Plus finement, le système a une précision de 93.0 % et un rappel un peu inférieur, à 85.1 %.

L'analyse qualitative des résultats conduit à plusieurs remarques. Dans l'interaction, les ruptures peuvent être de nature différente : ruptures thématiques ou argumentatives. L'ensemble des positifs (vrais et faux) sont des ruptures thématiques, contrairement à l'ensemble des faux négatifs. Le système ne reconnaît aucune rupture argumentative, et se focalise sur les ruptures thématiques, y compris les 3 fausses. Nous disposons ainsi d'un système formel permettant d'établir la présence de ruptures thématiques avec un très bon niveau de performance. Mais si ces ruptures sont les plus nombreuses dans le corpus, le système nécessite une adaptation plus fine pour identifier les autres types de ruptures qui apparaissent au niveau de l'interaction.

7 Conclusion

Les « ruptures », dans le discours pathologique du sujet schizophrène, sont des points de discontinuité pragmatique que l'on peut mettre au jour objectivement et manuellement au moyen de la modélisation hiérarchique (Roulet et al., 2001). Dans le cadre d'un corpus d'entretiens patients et psychologue, et afin d'automatiser le repérage des ruptures dans le but de se passer de la modélisation hiérarchique, nous proposons un système fondé sur la notion de « mise à jour » : nous assumons qu'un acte de langage est mis à jour par un acte postérieur si le second a besoin du premier pour exister (e.g. un couple question / réponse).

Cinq règles formelles décident pour chaque acte de chaque tour de parole, pour le patient et le psychologue, si l'acte est « pertinent » ou pas. Si c'est le cas, l'acte se retrouve dans une case $R_{i,j}$ d'un tableau à deux entrées, i étant le numéro du tour et j valant 0 pour le patient, 1 pour le psychologue. Un algorithme, dont nous avons donné le pseudocode, calcule alors les ruptures dans la séquence, et nous obtenons sur les 27 séquences du corpus, comportant 47 ruptures, une exactitude de 96,2 % et une F-mesure de 88,9 %, soit de bons résultats au regard de la complexité de la tâche.

Références

- ASHER N. & LASCARIDES A. (2003) *Logics of conversation*. Cambridge University Press.
- MUSIOL M. (2009) Chapitre 13. Incohérence et formes psychopathologiques dans l'interaction verbale schizophrénique in *Psychose, langage et action*. Louvain-la-Neuve, pp. 217–238. De Boeck Supérieur (Neurosciences & cognition).
- REBUSCHI M., AMBLARD M. & Musiol, M. (2014) Using SDRT to analyze pathological conversations: Logicality, rationality, and pragmatic deviances in *Interdisciplinary works in logic, epistemology, psychology and linguistics*, pp. 343–368. Springer.
- Roulet E. et al (2001) Un modèle et un instrument d'analyse de l'organisation du discours, vol 62. Peter Lang GmbH, collection « *Sciences pour la communication* ». Berne.
- SEARLE J.R. & VANDERVEKEN D. (1985) *Foundations of illocutionary logic*. CUP Archive.