HAL open science

Analyse automatique du discours de patients pour la détection de comorbidités psychiatriques

Christophe Lemey, Yannis Haralambous, Philippe Lenca, Romain Billot, Deok-Hee Kim-Dufor

To cite this version:

Christophe Lemey, Yannis Haralambous, Philippe Lenca, Romain Billot, Deok-Hee Kim-Dufor. Analyse automatique du discours de patients pour la détection de comorbidités psychiatriques. Conférence Internationale Francophone sur la Science des Données, Aix-Marseille Université - LIS UMR 7020, Jun 2021, Marseille, France. pp.261-272. hal-03258036

HAL Id: hal-03258036

https://hal.science/hal-03258036v1

Submitted on 11 Jun 2021

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.

Analyse automatique du discours de patients pour la détection de comorbidités psychiatriques

Christophe Lemey*,**, Yannis Haralambous**, Philippe Lenca**,
Romain Billot**, Deok-Hee Kim-Dufor***

*Service hospitalo-universitaire de psychiatrie adulte, CHRU de Brest
christophe.lemey@chu-brest.fr,
**IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238 Brest, France
prenom.nom@imt-atlantique.fr
***EA 7479 SPURBO, Université de Bretagne Occidentale, Brest, France

Résumé. Les comorbidités sont très fréquentes en santé mentale et représentent un enjeu thérapeutique majeur ainsi qu'un levier pour une meilleure compréhension des mécanismes physiopathologiques des pathologies. Certains maladies, comme la schizophrénie, s'installent progressivement et de façon variable d'un individu à l'autre, les symptômes augmentant progressivement en intensité et en spécificité. Dans ces cas, les cliniciens cherchent à identifier au plus tôt les symptômes annonciateurs et les comorbidités associés, afin de proposer des interventions maximisant les effets thérapeutiques. La parole, et donc le langage, est un élément-clé sur lequel ils s'appuient lors des consultations pour comprendre l'état psychique des patients, et les systèmes d'analyse automatique de la langue peuvent fournir une aide à l'évaluation. Nous proposons une telle aide, ciblant la détection de comorbidités et fondée sur les grammaires de dépendances et des indicateurs paralinguistiques comme les pauses et les interjections, qui s'avèrent être des choix pertinents.

Mots-clés : comorbidités psychiatriques, schizophrénie, traitement automatique du langage, grammaires de dépendances, paralinguistique

1 Introduction

Une comorbidité est l'association de deux ou plusieurs maladies ou troubles dans le même temps. En santé mentale, une étude ((Roca et al., 2009)) a revelé que jusqu'à 30% des patients dans une cohorte nationale présentent des comorbidités, L'exploration des comorbidités représente un double enjeu, à la fois thérapeutique et étiologique. En effet, les comorbidités compliquent la mise en œuvre des traitements et leur étude permet de mieux comprendre les mécanismes physiopathologiques mis en jeu dans la génèse des troubles.

Parmi les principales maladies psychiatriques, la schizophrénie touche environ 1% de la population et est l'une des principales maladies entraînant un nombre important d'années vécues avec un handicap (Rössler et al., 2005; Anderson, 2019), en grande partie en raison du jeune âge auquel elle se déclare (souvent à l'adolescence), du poids élevé des handicaps (fonctionnels et sociaux) et de l'évolution chronique fréquente de la pathologie. Début 2020, en

France, environ 600 000 personnes souffraient de cette maladie [1], et il est estimé que parmi elles, une sur deux fera au moins une tentative de suicide, à court ou moyen terme.

La schizophrénie s'installe progressivement. Le cours évolutif de la pathologie est caractérisé par les phases suivantes : phase prémorbide, de la naissance du patient jusqu'à l'apparition des premiers signes ; phase prodromique au cours de laquelle apparaissent les premiers symptômes peu spécifiques, ces symptômes augmentant progressivement en intensité et en spécificité au cours de la phase qui précède les symptômes psychotiques francs ; phase psychotique avec les premiers signes psychotiques avérés qui déterminent le premier épisode de psychose. Lors de la phase active de la schizophrénie on constate une multitude de symptômes très variables (syndromes positifs : idées délirantes et hallucinations ; syndromes négatifs : retrait social et déficits cognitifs ; syndrome de désorganisation : trouble du contact). Il s'agit d'une maladie complexe dont la physiopathologie reste peu connue. Le modèle explicatif dominant actuellement, est le modèle de diathèse-stress qui combine deux facteurs : la vulnérabilité intrinsèque et le stress provenant d'expériences vécues (Howes et McCutcheon, 2017; Pruessner et al., 2017). Néanmoins, les mécanismes sous-jacents doivent encore être explorés. Lors de ces phases précoces d'évolution de la maladie, la présence de comorbidités est très fréquente (notamment les comorbidités anxieuses, dépressives et addictives). Jusqu'à 50% des patients en phase prodromique peuvent présenter une comorbidité (Lim et al., 2015).

La durée entre l'apparition des premiers symptômes psychotiques francs et le premier accès aux soins est en moyenne de deux à cinq ans (avec d'importantes différences entre régions du monde). Cette période est communément appelée «durée de la psychose non traitée» (Fusar-Poli et al., 2013). Les efforts vont dans le sens d'un traitement précoce et d'une réduction de cette durée. En effet, l'identification précoce et les interventions rapides au cours de l'évolution d'un trouble psychotique semblent maximiser les effets thérapeutiques et améliorer la qualité de vie des patients (McGlashan et Johannessen, 1996). Durant cette phase, des signes d'alerte avant la phase active de la maladie peuvent être détectés, ce qui permet d'optimiser les soins et de réduire la durée de la psychose non traitée (Olsen et Rosenbaum, 2006) en orientant les patients vers des centres de détection précoce des troubles psychotiques utilisant des outils spécifiques d'évaluation (Olsen et Rosenbaum, 2006; Yung et al., 2005).

Parmi ces outils d'aide à l'évaluation, et de façon générale en psychiatrie, se trouvent les systèmes d'analyse automatique de la parole et, plus particulièrement, du discours des patients (Le Glaz et al., 2021). En effet, le langage est l'un des éléments clés sur lequel les cliniciens peuvent s'appuyer lors des consultations pour mieux comprendre l'état psychique des patients (Mota et al., 2012). Ainsi, les psychiatres étudient tout le spectre des caractéristiques linguistiques du discours des patients, en tant que reflet des pathologies qu'ils présentent. Les techniques d'analyse informatisée du langage telles que l'analyse sémantique latente et l'analyse structurelle du discours indiquent une diminution de la cohérence chez les patients atteints de schizophrénie en corrélation avec les évaluations cliniques et une précision identique ou supérieure dans l'évaluation diagnostique (Hoffman et al., 1986; Elvevåg et al., 2007; Mota et al., 2012). Une combinaison d'analyses sémantiques et syntaxiques peut prédire avec une précision raisonnable la transition vers la schizophrénie et semble être plus efficace que l'évaluation clinique utilisant des outils standardisés (Bedi et al., 2015; Corcoran et al., 2018).

1. https://www.inserm.fr/information-en-sante/dossiers-information/schizophrenie, consulté le 23/03/2021

La psychose est accompagnée de comorbidités, en particulier les troubles de l'humeur, l'anxiété et la dépendance (Bazziconi et al., 2017). Il est donc important de les identifier afin de proposer des soins adaptés. Les analyses prosodiques des comorbidités psychiatriques se sont principalement concentrées sur la fréquence fondamentale (F0) et le débit de parole (Scherer et Bänziger, 2004; Audibert et al., 2005; van den Broek, 2004; Moore et al., 2003). Silber-Varod et al. (2016) considèrent les pauses et les disfluences dans les comorbidités anxieuses en étudiant principalement les caractéristiques prosodiques. Notre étude considère les mêmes facteurs en se concentrant sur la syntaxe.

Nous présentons, section 2, les dépendances syntaxiques, au cœur de notre approche, et introduisons un nouveau concept, le croisement interstitiel de dépendances. Dans la section 3 nous décrivons la façon dont les comorbidités sont évaluées, la constitution du corpus, notre méthodologie et nos résultats. Une conclusion et des perspectives sont proposées section 4.

2 Grammaires de dépendances et croisements interstitiels

À la fin des années trente, le linguiste français Lucien Tesnière a commencé à travailler sur une nouvelle théorie syntaxique fondée sur les relations entre les mots, théorie qui n'a été publiée qu'à titre posthume (Tesnière, 1959). Ses travaux resteraient méconnus au niveau international si un chercheur de la Rand Corporation, David Hays, n'avait pas présenté les idées de Tesnière à la communauté encore jeune des «linguistes computationnels» par une présentation au célèbre symposium de l'UCLA sur la traduction automatique (Hays, 1960), suivie d'un article dans la revue *Language* (Hays, 1964) et, enfin, d'un livre qui se trouve être *le premier livre consacré à la linguistique computationnelle* (Hays, 1967). C'est Hayes qui a introduit les termes de *grammaire de dépendances* et de *relation de dépendance*. Par la suite, l'utilisation des grammaires de dépendances a continué de se répandre et elles semblent avoir aujourd'hui supplanté les méthodes basées sur les constituants (Osborne, 2019). Les grammaires de dépendances ont déjà été utilisées dans le domaine psychiatrique, par exemple dans Tanana et al. (2016) où les séances d'entretiens de motivation ont été codées par ordinateur.

Dans une grammaire de dépendances, chaque phrase a une *tête* (généralement le verbe) qui est la racine d'un arbre orienté de *relations de dépendance*. Les arêtes sont orientées de manière que l'on puisse tracer des chemins (orientés) de chaque feuille à la racine. Chaque arête possède une étiquette, appelée *nature de dépendance*, qui décrit la relation entre la *dépendance* (source de l'arête) et le *gouverneur* (cible de l'arête).

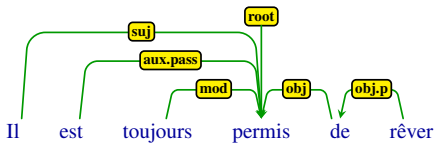

FIG. 1 – *Arbre de dépendance de «Il est toujours permis de rêver», tiré du* French Treebank Corpus *(Abeillé et al., 2003)*

Nous remarquons dans l'arbre de la figure 1 que le participe «permis» est la tête de l'arbre, et qu'il gouverne : le pronom «il» en tant que sujet (suj) ; le verbe «est» en tant qu'auxiliaire

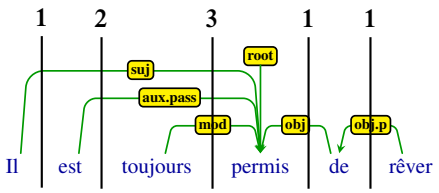

FIG. 2 – *Croisements interstitiels de dépendances*

(aux.pass) ; l'adverbe «toujours» en tant que modificateur (mod) ; et la préposition «de» en tant qu'objet (obj). De plus, nous remarquons que «rêver» est gouverné par «de» à travers une relation de dépendance de complément d'objet indirect prépositionnel (obj.p).

Liu (2008) explore les dépendances d'un point de vue cognitif et définit une mesure de complexité du langage, la *distance de dépendance moyenne* (DDM), qui quantifie le fait qu'une phrase anglaise telle que «The man the boy the woman saw heard left», bien que grammaticale, est bien plus difficile à comprendre que la phrase sémantiquement équivalente «The woman saw the boy that heard the man that left» (leurs valeurs DDM sont resp. de 3 et de 1,4). Si l'on définit DDM comme la distance moyenne entre gouverneur et gouverné, plus les dépendances sont «à longue distance», plus le DDM est élevé.

Par ailleurs, les dépendances ne se chevauchent pas, de sorte que nous avons une relation binaire irréflexive, asymétrique et transitive $\prec$ entre elles : $(a \rightarrow b) \prec (c \rightarrow d)$ lorsque $(\mathrm{pos}(a) < \mathrm{pos}(c)$ et $\mathrm{pos}(d) \geq \mathrm{pos}(b))$ ou $(\mathrm{pos}(a) \leq \mathrm{pos}(c)$ et $\mathrm{pos}(d) > \mathrm{pos}(b))$, où pos représente l'ordre linéaire des mots de la phrase.

Dans l'exemple de la figure 2, nous avons (toujours $\rightarrow$ permis) $\prec$ (est $\rightarrow$ permis) $\prec$ (Il $\rightarrow$ permis). La relation $x \prec y$ implique également que $\mathrm{longueur}(x) < \mathrm{longueur}(y)$, et est un ordre partiel de sorte que nous pouvons construire un treillis dont les nœuds sont des dépendances et les arêtes représentent $\prec$. Les chemins dans ce treillis peuvent être visualisés dans l'arbre de dépendance en traçant des lignes verticales entre les mots. Le fait que (toujours $\rightarrow$ permis) $\prec$ (est $\rightarrow$ permis), qui est un chemin de longueur 2 dans le treillis, est représenté par le fait que la deuxième ligne verticale traverse deux dépendances. De même, le chemin (toujours $\rightarrow$ permis) $\prec$ (est $\rightarrow$ permis) $\prec$ (Il $\rightarrow$ permis), qui est d'ordre 3, est représenté par le fait que la troisième ligne verticale croise trois dépendances. Comme nous le voyons, le nombre de croisements augmente lorsque nous nous approchons de la racine par la gauche puisque de nombreuses dépendances ciblant la racine s'accumulent, tandis qu'à droite, en raison de l'adjacence entre les nœuds, le nombre de croisements reste faible.

Outre les mots, notre corpus contient également des *éléments paralinguistiques*, tels que les *interjections* et les *pauses* qui, par définition, ont lieu dans des positions interstitielles. Pour les traiter de manière appropriée, nous avons introduit (Haralambous et al. (2020)) une nouvelle notion : le ***croisement interstitiel de dépendances***. Notre hypothèse est que les positions interstitielles ayant une valeur de croisement élevée sont stratégiques et que le fait d'y placer des «intrus» (interjections, pauses) peut être indicateur de trouble. Nous verrons qu'en combinant le nombre et la nature des dépendances croisées sur une pause ou une interjection nous obtenons des indicateurs de comorbidité.

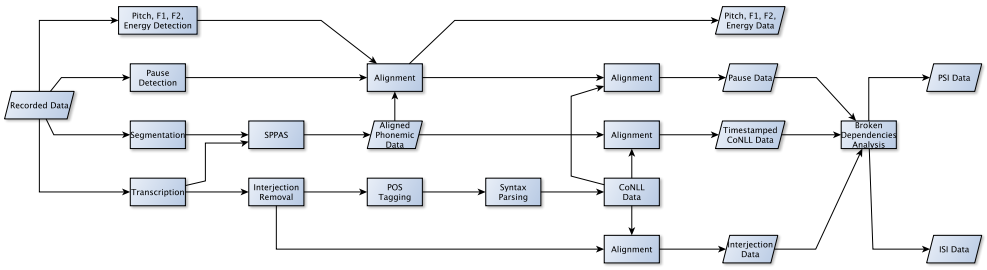

FIG. 3 – *Le processus d'extraction de données*

3 Vers une aide au diagnostic de comorbidités chez des patients courant le risque de développer une psychose

Par la suite nous allons décrire notre processus de traitement de données extraites d'entretiens psychiatriques visant à détecter, de manière précoce, un patient présentant un trouble psychotique (en général) et notamment la schizophrénie (en particulier). Notre corpus est relativement petit (des *small data*), comme c'est souvent le cas pour les données médicales – dans notre cas, une patienthèque, avec plusieurs centres de détection précoce, est en cours de constitution et plusieurs centaines d'entretiens seront disponibles d'ici deux ans. Nous avons fait une recherche d'indicateurs linguistiques et paralinguistiques à large spectre, qui nous a permis de conclure que certaines propriétés syntaxiques des pauses et interjections peuvent être corrélées avec certains groupes de comorbidités. Si nous nous limitons, pour le moment, aux comorbidités, c'est parce que ce n'est qu'après deux années de suivi, lorsque l'état clinique des interviewés aura évolué, que nous pourrons véritablement évaluer des prédictions de transition vers la psychose. En effet, ces patients bénéficient d'un suivi rapproché pendant deux ans afin de surveiller leur évolution clinique et de leur apporter les soins appropriés en cas d'aggravation de leur état.

3.1 Évaluation du risque de psychose et constitution de corpus

Les patients reçus au sein de la consultation de détection et d'intervention précoce du CHU de Brest (programme CEVUP = consultation d'évaluation de la vulnérabilité psychologique, Bazziconi et al. 2017) sont évalués par une équipe pluridisciplinaire comprenant un psychiatre, un psychologue, un infirmier et un neuropsychologue. L'évaluation initiale permet d'identifier leur niveau de risque et d'établir un protocole de soins personnalisé. Une réévaluation semestrielle est proposée pendant deux ans afin d'identifier les aggravations potentielles des troubles et l'apparition éventuelle d'une psychose. Cette transition du statut de patient «à risque de développer un trouble psychotique» à l'apparition d'une pathologie psychotique confirmée est appelée «transition vers la psychose» (24% des patients à risque développent un trouble psychotique dans les deux années qui suivent et 33% dans les trois années suivantes, cf. *loc. cit.*). Les résultats présentés dans cet article s'inscrivent dans le cadre d'un projet de recherche sur l'analyse informelle de la parole impliquant tous les patients orientés vers le centre de détection et d'intervention précoce (protocole de recherche NCT03525054 validé par le Comité de Pro-

ID	Genre	Durée	A_1	A_2	B	C	D_1	D_2	E	F	G	H	I	J	K_1	K_2	K_3	K_4	L	M	N	O	P	**THY**	**ANX**	**ADD**
15	F	$47'29''$	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
21	M	$47'45''$	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	3	0
23	F	$43'50''$	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	3	1	0
25	M	$30'59''$	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	3	1	0
27	M	$25'05''$	1	0	0	2	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	3	2	1
28	M	$27'26''$	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	2	2	0
30	M	$63'08''$	0	1	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2	1	1
44	M	$43'13''$	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	0	1	1	2

TAB. 1 – *Valeurs de comorbidités de notre corpus de patients*

tection des Personnes Est-III –N CPP : 18.04.03–). Il prévoit un enregistrement de l'entretien médical clinique initial et un suivi de deux ans.

Les comorbidités suivantes, représentées par les lettres A à P (variables oui (1) / non (0) sauf pour la suicidalité, qui est graduée selon trois intensités 0/1/2), sont évaluées lors de l'entretien selon le standard *Mini International Neuropsychiatric Interview* (Sheehan et al., 1998) :

A. Trouble dépressif majeur (A_1 : trouble dépressif majeur sans désordres psychotiques ; A_2 : trouble dépressif majeur avec caractéristiques psychotiques)
B. Dysthymie
C. Suicidalité
D. Épisode (hypo)maniaque (D_1 : hypomaniaque ; D_2 : maniaque)
E. Trouble de panique
F. Agoraphobie
G. Phobie sociale
H. Trouble obsessionnel compulsif
I. Syndrome de stress post-traumatique
J. Dépendance/abus d'alcool
K. Dépendance aux drogues (K_1 : opioïdes ; K_2 : cocaïne ; K_3 : cannabis ; K_4 : sédatifs)
L. Troubles psychotiques
M. *Anorexia Nervosa*
N. *Bulimia Nervosa*
O. Trouble d'anxiété généralisée
P. Trouble de la personnalité antisociale

Nous avons regroupé ces comorbidités en trois groupes selon la nature des troubles, afin de permettre des analyses statistiques sur un nombre limité de patients : troubles thymiques (**THY** : A, B, C, D) ; troubles anxieux (**ANX** : E, F, G, H, I, O) ; et troubles de dépendance/addiction (**ADD** : J, K, M, N). Les comorbidités L et P, ne concernant presque aucun patient de notre corpus, ont été omises par notre étude.

Les enregistrements sont retranscrits par un personnel médical, en respectant les conventions d'interjections et de respiration paralinguistique établies par Bigi (2015). Ces transcriptions sont ensuite relues et éditées par un correcteur indépendant (pour une deuxième vérification, garantissant, entre autres, l'anonymat au sein des retranscriptions). Le processus de constitution du corpus nécessite donc des ressources conséquentes (personnels, temps). Par ailleurs, la schizophrénie touche une très petite part de la population (environ 1%). De ce fait nous ne disposons, pour le moment, que de corpus de petite taille. Le corpus utilisé dans cette étude se compose uniquement de huit entretiens – nous verrons néanmoins qu'il permet d'exhiber des indicateurs concluants. Le tableau 1 présente les principales caractéristiques du corpus «brut» (genre des patients ; durée de l'entretien ; valeurs des comorbidités).

Chaque entretien enregistré est segmenté en répliques entre le soignant et le patient. Une fois la transcription soigneusement vérifiée, les deux flux de données (son et texte) sont fournis

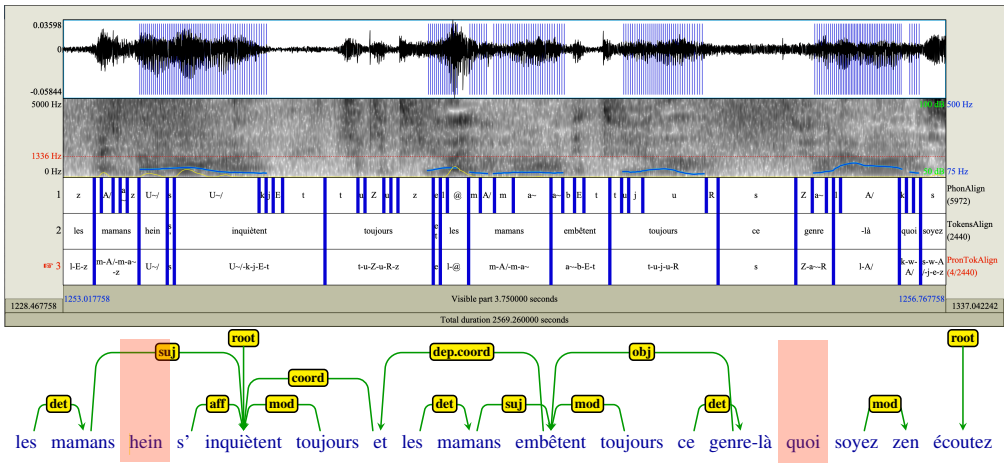

FIG. 4 – *Un énoncé (patient #44) visualisé dans PRAAT (alignement phonémique) et annoté par des dépendances syntaxiques. L'interjection primaire «hein» croise la dépendance de type sujet «mamans» → «inquiètent». L'interjection secondaire apposée «quoi» ne croise aucune relation de dépendance.*

au logiciel SPPAS (Bigi, 2015), qui produit un fichier comportant des phonèmes horodatés, des mots en orthographe standard et des mots en représentation phonémique. Cependant, SPPAS ne détecte pas les pauses. Nous utilisons donc PRAAT (Boersma et Weenink, 2001) sur une version préparée du fichier sonore pour détecter les pauses et les écarts. Puis nous les injectons dans les données de phonèmes et de mots horodatés. PRAAT fournit également des données d'énergie, de hauteur, de F1 et de F2, que nous alignons avec les phonèmes et les mots.

Dans un flux de données parallèles (voir fig. 3), nous supprimons les interjections du texte transcrit et effectuons un marquage POS sur le résultat avec Talismane (Urieli, 2013; Urieli et Tanguy, 2013), suivi d'une analyse syntaxique des dépendances effectuée par Grew (Guillaume et Perrier, 2015). Ce processus nous fournit des données CoNLL relativement propres. Nous alignons ensuite les deux flux de données (données fournies par SPPAS et données sous format CoNLL) en utilisant l'algorithme de Needleman-Wunsch tel qu'implémenté dans bioPython. Nous obtenons ainsi des données CoNLL horodatées. L'horodatage des pauses et des interjections nous sert à étudier leurs croisements avec les dépendances syntaxiques. Il est prévu que les résultats des analyses soient confrontés aux données cliniques recueillies pour chacun des patients et validés par un psychiatre clinicien de l'équipe de détection et d'intervention précoce.

3.2 Définition et justification du PIDC et du IIDC

Considérons la forêt syntaxique de dépendances[2] d'un énoncé donné. Comme on peut le voir, figure 4, dans *«les mamans s'inquiètent toujours et les mamans embêtent toujours*

2. Nous utilisons le terme de *forêt* en raison de la co-présence de plusieurs arbres syntaxiques dans le même énoncé.

nature	fréquence	DDM dép./gouv.	nature	fréquence	DDM dép./gouv.
mod	120,741	4,1937	det	85,154	1,1987
obj.p	90,400	1,7511	suj	35,402	4,2315

TAB. 2 – *Les quatre relations les plus fréquentes dans le corpus French Treebank*

ce genre-là quoi soyez zen écoutez», les mots «quoi» et «écoutez» ne sont pas connectés à l'arbre syntaxique des deux phrases coordonnées «les mamans s'inquiètent toujours» et «les mamans embêtent toujours ce genre-là» («quoi» et potentiellement «ce genre-là» pourraient aussi être considérés comme des interjections secondaires). Nous ne disposons donc pas d'un arbre syntaxique unique mais de fragments d'arbre de taille variable.

Nous supprimons toutes les interjections primaires afin d'obtenir des dépendances plus proches de l'intention du locuteur et d'éviter une mauvaise interprétation par l'analyseur syntaxique qui a été entraîné sur un corpus sans interjections. Nous introduisons la mesure IIDC (croisements d'interjections), dont le but est de quantifier le croisement des interjections, en tant qu'interstices entre les mots, avec les relations de dépendance qui relient les mots. Comme le lecteur peut le voir dans la figure 4, l'interjection primaire «hein» croise une relation de dépendance entre le nom «mamans» agissant comme sujet, et le verbe «s'inquiètent», qui est la racine du fragment d'arbre. Une autre interjection (secondaire, cette fois), «quoi», ne croise aucune relation de dépendance puisqu'elle est située entre des arbres syntaxiques distincts dans la forêt. Nous faisons de même pour les pauses : le PIDC (croisements de pauses) est une mesure du croisement des pauses (c'est-à-dire des silences internes aux répliques de chaque patient) en tant qu'interstices entres les mots avec les relations de dépendance.

Notre hypothèse est la suivante : les croisements d'interjections et les croisements de pauses peuvent servir d'indicateurs de la désorganisation linguistique du patient. Nous nous intéressons donc (1) au nombre de dépendances traversant des interjections ou des pauses et (2) aux étiquettes des relations de dépendance croisées. Nous définissons ainsi quatre mesures :

$$\mathrm{PIDC} = (\#\mathrm{c}) \times \frac{\text{durée de pause}}{\text{durée de l'énoncé}}, \qquad \mathrm{PIDC}_S = (\#\mathrm{c\ dans\ S}) \times \frac{\text{durée de pause}}{\text{durée de l'énoncé}},$$
$$\mathrm{IIDC} = (\#\mathrm{c}) \times \frac{\text{durée de l'interjection}}{\text{durée de l'énoncé}}, \qquad \mathrm{IIDC}_S = (\#\mathrm{c\ dans\ S}) \times \frac{\text{durée de l'interjection}}{\text{durée de l'énoncé}},$$

où #c est le nombre de croisements, et S est un ensemble de relations de dépendance.

Nous calculons par la suite les valeurs de PIDC_S et d'IIDC_S pour quatre ensembles spécifiques de relations de dépendances : {det,suj}, {det}, {obj.p} et {suj}. Ces dernières sont les quatre relations les plus fréquentes dans le corpus French Treebank (Abeillé et al., 2003) (voir tableau 2). Malgré sa fréquence élevée, nous n'avons pas sélectionné la dépendance «mod» (modificateur) pour la raison suivante : les mots gouvernés (dans 26% des cas, un adjectif ; dans 22% des cas, une préposition ; dans 20% des cas, un adverbe ; dans 18% des cas un nom) peuvent être assez éloignés de leur gouverneur et donc l'existence d'une pause ou d'une interjection entre gouverné et gouverneur n'est pas nécessairement significative. Au contraire, la dépendance «obj.p» (objet prépositionnel) est en fait l'équivalent d'un *cas de gouvernance* (pour les langues à cas) et donc, selon Osborne (2019, p. 142), elle serait techniquement plutôt morphologique que syntaxique. Elle est très stable en termes de partie de discours (86% de ses gouvernés sont des noms) et la distance entre gouverneur et gouverné est assez faible (1,7511 en moyenne). Sa nature morphologique et ses caractéristiques positionnelles nous amènent à

	pauses			interjections		
dépendances	comorbidités	ρ	p-valeur	comorbidités	ρ	p-valeur
{obj.p}	**ADD**	**0,8660**	0,0054	**ADD**	0,8248	0,0117
{det}	**ADD**	0,7735	0,0254	**ADD**	0,7285	0,04
{det,suj}	**ADD**	0,5770	0,1340	**ANX**	−0,8247	0,0117
{suj}	**ANX**	−0,5086	0,1980	**ANX**	**−0,8450**	0,0080
toutes	**THY**	0,7042	0,0512	**THY**	−0,6730	0,0671

TAB. 3 – *Corrélation de Spearman sur les comorbidités des trois groupes **THY**, **ANX** et **ADD***

formuler l'hypothèse que le croisement d'une interjection ou d'une pause avec une dépendance obj.p est susceptible de révéler une désorganisation mentale.

La dépendance «det» (déterminant) est également candidate à révéler une désorganisation : la liste des déterminants est très réduite et ils sont très proches de leur gouverneur (1,1987 en moyenne, plus petite distance moyenne des relations). Enfin, la relation «suj» est importante, malgré sa distance élevée entre gouverneur et gouverné (4,2315 en moyenne), puisque (sauf à l'impératif) les verbes français possèdent nécessairement des sujets.

3.3 Résultats et analyse

Nous avons effectué un test de corrélation de Spearman sur les valeurs de comorbidité des trois groupes **THY**, **ANX** et **ADD** par rapport aux croisements de pauses/interjections calculés. Le tableau 3 présente les résultats les plus pertinents avec leur p-valeur.

Pour {obj.p} et {det} nous obtenons un comportement similaire pour les pauses et les interjections, même si ces deux phénomènes paralinguistiques sont bien distincts (et mesurés de manière différente, cf. fig. 3). Nous remarquons également que les pauses ou interjections traversant la dépendance {obj.p} constituent un indicateur très fort ($\rho > 0{,}82$) du groupe **ADD**, avec une significativité élevée ($p = 0{,}012$). La dépendance {det} a également un comportement cohérent ($\rho \approx 0{,}75$, avec $0{,}025 \leq p \leq 0{,}04$) et cible, de nouveau, le groupe **ADD**. Les valeurs des autres dépendances révèlent des comportements différents : alors que les pauses croisant {det,suj} ou {suj} donnent des résultats non significatifs ($p > 0{,}13$), les interjections croisant {det,suj} et {suj} donnent des résultats très élevés, mais ciblent négativement le groupe **ANX** ($\rho < 0{,}824$ avec $p = 0{,}012$). Ces résultats peuvent être résumés comme suit :

— les membres du groupe **ADD** ont tendance à placer des pauses ou des interjections entre la préposition et le nom gouverné ou entre le déterminant et le nom qui le gouverne ;

— les membres du groupe **ANX** ont tendance à placer des interjections entre le déterminant et le nom qui le gouverne, ou entre le sujet et le verbe qui le gouverne.

Le premier résultat peut refléter la forte prévalence des comportements addictifs chez les patients à risque de psychose (Valmaggia et al., 2014). Il montre que le croisement d'une interjection ou d'une pause entre préposition et nom ou entre déterminant et nom est susceptible de révéler une désorganisation mentale qui est un des symptômes psychotiques souvent retrouvés chez les patients à risque, elle est caractéristique de la schizophrénie (Fusar-Poli et al., 2013). Le second résultat peut s'expliquer par une tendance des patients anxieux à éviter de laisser des blancs, notamment dans le cadre d'une conversation où l'individu est soumis au jugement de son interlocuteur, de manière similaire aux patients bègues (Iverach et Rapee, 2014).

4 Conclusion

Ces résultats montrent qu'il est possible d'utiliser le traitement du langage naturel combiné avec des données paralinguistiques pour explorer les comorbidités psychiatriques. Les dépendances et leurs croisements avec les pauses et les interjections semblent être particulièrement indiquées à cette fin. Nous comptons poursuivre l'exploration de marqueurs linguistiques et paralinguistiques afin d'identifier des marqueurs pertinents pour la pratique clinique.

Références

Abeillé, A., L. Clément, and F. Toussenel (2003). Building a treebank for French. In *Treebanks*, pp. 165–187. Kluwer.

Anderson, K. (2019). Towards a public health approach to psychotic disorders. *Lancet Public Health 4*(5), e212-3.

Audibert, N., V. Aubergé, and A. Rilliard (2005). The prosodic dimensions of emotion in speech: the relative weights of parameters. In *European Conference on Speech Communication and Technology*, pp. 525–528. ISCA.

Bazziconi, P., C. Lemey, L. Bleton, and M. Walter (2017). CEVUP program: An analytical epidemiological cohort study. *European Psychiatry 41*, S729.

Bedi, G., F. Carrillo, G. Cecchi, D. Fernández-Slezak, M. Sigman, N. Mota, S. Ribeiro, D. Javitt, M. Copelli, and C. Corcoran (2015). Automated analysis of free speech predicts psychosis onset in high–risk youths. *npj Schizophrenia 1*, 15030.

Bigi, B. (2015). SPPAS – Multi-lingual Approaches to the Automatic Annotation of Speech. *The Phonetician – International Society of Phonetic Sciences 111–112*, 54–69.

Boersma, P. and D. Weenink (2001). PRAAT, a system for doing phonetics by computer. *Glot International 5*(9-10), 341–347.

Corcoran, C., F. Carrillo, D. Fernández-Slezak, G. Bedi, C. Klim, D. Javitt, C. Bearden, and G. Cecchi (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry 17*(1), 67–75.

Elvevåg, B., P. W. Foltz, D. R. Weinberger, and T. E. Goldberg (2007). Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophr. Res 93*, 304–316.

Fusar-Poli, P. et al. (2013). The psychosis high-risk state: a comprehensive state-of-the-art review. *JAMA Psychiatry 70*, 107–120.

Guillaume, B. and G. Perrier (2015). Dependency parsing with graph rewriting. In *International Conference on Parsing Technologies*, pp. 30–39.

Haralambous, Y., C. Lemey, P. Lenca, R. Billot, and D.-H. Kim-Dufor (2020). Using Dependency Syntax-Based Methods for Automatic Detection of Psychiatric Comorbidities. In *Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments*, pp. 142–150.

Hays, D. (1960). Grouping and dependency theories. In *Proceedings of the National Symposium on Machine Translation*, pp. 257–266. UCLA.

Hays, D. (1964). Dependency theory: A formalism and some observations. *Language 40*, 159–525.

Hays, D. (1967). *Introduction to computational linguistics*. Macdonald & co.

Hoffman, R. E., S. Stopek, and N. C. Andreasen (1986). A comparative study of manic vs schizophrenic speech disorganization. *Arch. Gen. Psychiatry 43*, 831–838.

Howes, O. D. and R. McCutcheon (2017). Inflammation and the neural diathesis-stress hypothesis of schizophrenia: A reconceptualization. *Transl. Psychiatry 7*, 1024.

Iverach, L. and R. M. Rapee (2014). Social anxiety disorder and stuttering: current status and future directions. *J. Fluency Disord. 40*, 69–82.

Le Glaz, A., Y. Haralambous, D.-H. Kim-Dufor, P. Lenca, R. Billot, R. Taylor, J. Marsh, J. DeVylder, M. Walter, S. Berrouiguet, and C. Lemey (2021). Machine learning and natural language processing in mental health: Systematic review. *Journal of Medical Internet Research 23*(5), e15708.

Lim, J., G. Rekhi, A. Rapisarda, M. Lam, M. Kraus, R. Keefe, et al. (2015). Impact of psychiatric comorbidity in individuals at ultra high risk of psychosis - findings from the longitudinal youth at risk study (lyriks). *Schizophr. Res. 164*, 1–3.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science 9*, 159–191.

McGlashan, T. H. and J. O. Johannessen (1996.). Early detection and intervention with schizophrenia: rationale. *Schizophr. Bull. 22*, 201–222.

Moore, E., M. Clements, J. Peifer, and L. Weisser (2003). Analysis of prosodic variation in speech for clinical depression. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, Volume 3, pp. 2925–2928.

Mota, N. B., N. A. P. Vasconcelos, N. Lemos, A. C. Pieretti, O. Kinouchi, G. A. Cecchi, M. Copelli, and S. Ribeiro (2012). Speech graphs provide a quantitative measure of thought disorder in psychosis. *PLOS ONE 7*(4), e34928.

Olsen, K. A. and B. Rosenbaum (2006). Prospective investigations of the prodromal state of schizophrenia: assessment instruments. *Acta Psychiatr. Scand. 113*, 273–282.

Osborne, T. (2019). *A Dependency Grammar of English*. John Benjamins.

Pruessner, M., A. E. Cullen, M. Aas, and E. F. Walker (2017). The neural diathesis–stress model of schizophrenia revisited: An update on recent findings considering illness stage and neurobio. and methodol. complexities. *Neurosci. Biobehav. Rev. 73*, 191–218.

Roca, M., M. Gili, M. Garcia-Garcia, J. Salva, M. Vives, J. Garcia Campayo, and A. Comas (2009). Prevalence and comorbidity of common mental disorders in primary care. *J. Affect. Disord. 119*(1–3), 52–58.

Rössler, W., H. Joachim Salize, J. van Os, and A. Riecher-Rössler (2005). Size of burden of schizophrenia and psychotic disorders. *European Neuropsychopharmacology 15*, 399–409.

Scherer, K. R. and T. Bänziger (2004). Emotional expression in prosody: a review and an agenda for future research. In *The Speech Prosody Conference*.

Sheehan, D. V., Y. Lecrubier, K. H. Sheehan, P. Amorim, J. Janavs, E. Weiller, T. Hergueta, R. Baker, and G. C. Dunbar (1998). The Mini–International Neuropsychiatric Interview

(M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM–IV and ICD–10. *J. Clin. Psychiatry 59 Suppl 20*, 22–33.

Silber-Varod, V., H. Kreiner, R. Lovett, Y. Levi-Belz, and N. Amir (2016). Do social anxiety individuals hesitate more? The prosodic profile of hesitation disfluencies in Social Anxiety Disorder individuals. In *Proceedings of the Speech Prosody Conference*, pp. 1211–1215.

Tanana, M., K. A. Hallgren, Z. E. Imel, D. C. Atkins, and V. Srikumar (2016). A comparison of natural language processing methods for automated coding of motivational interviewing. *J Subst Abuse Treat. 65*, 43–50.

Tesnière, L. (1959). *Éléments de syntaxe structurale*. Paris: Klincksieck.

Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Ph. D. thesis, Université de Toulouse II le Mirail.

Urieli, A. and L. Tanguy (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*, pp. 188–201.

Valmaggia, L. R., F. L. Day, C. Jones, S. Bissoli, C. Pugh, D. Hall, S. Bhattacharyya, O. Howes, J. Stone, P. Fusar-Poli, M. Byrne, and P. McGuire (2014). Cannabis use and transition to psychosis in people at ultra–high risk. *Psychological Medicine 44*(12), 2503–2512.

van den Broek, E. L. (2004). Emotional prosody measurement (EPM): a voice–based evaluation method for psychological therapy effectiveness. *Stud. Health Technol. Inform. 103*, 118–125.

Yung, A. R., H. P. Yuen, P. D. McGorry, L. J. Phillips, D. Kelly, M. Dell'olio, S. M. Francey, E. M. Cosgrave, E. Killackey, C. Stanford, K. Godfrey, and J. Buckby (2005). Mapping the onset of psychosis: The comprehensive assessment of at–risk mental states. *Aust N Z J Psychiatry 39*(11-12), 964–971.

Summary

Co-morbidities are very frequent in mental health and represent a major therapeutic issue as well as a lever for a better understanding of physiopathological mechanisms. Some diseases, such as schizophrenia, develop progressively and at different rates from one individual to another, with symptoms gradually increasing in intensity and specificity. In these cases, clinicians seek to identify early on the warning symptoms and aggravating comorbidities, in order to propose interventions that maximize the therapeutic effects. Speech, and thus language, is a key element they use during consultations to understand the psychological state of patients, and automatic language analysis systems can provide an aid to assessment. We propose such an aid, targeting the detection of comorbidities and based on dependency grammars and paralinguistic indicators such as pauses and interjections, which are shown to be relevant.

Keywords: psychiatric comorbidities, natural language processing, dependency grammars, schizophrenia, paralinguistics