Investigation par méthodes d'apprentissage des spécificités langagières propres aux personnes avec schizophrénie

Maxime Amblard[1] Chloé Braud[2] Chuyuan Li[1]
Caroline Demily[3] Nicolas Franck[3] Michel Musiol[1,4]
(1) LORIA, UMR 7503, Université de Lorraine, CNRS, Inria, 54000 Nancy, France
{maxime.amblard, chuyuan.li}@univ-lorraine.fr
(2) IRIT, CNRS, Toulouse
chloe.braud@irit.fr
(3) Centre Hospitalier le Vinatier & UMR 5229, CNRS - Univeristé Lyon 1, Lyon, France
{caroline.demily, nicolas.franck}@ch-le-vinatier.fr
(4) ATILF, UMR 7118, Université de Lorraine, CNRS, 54000 Nancy, France
michel.musiol@univ-lorraine.fr

RÉSUMÉ

Nous présentons des expériences visant à identifier automatiquement des patients présentant des symptômes de schizophrénie dans des conversations contrôlées entre patients et psychothérapeutes. Nous fusionnons l'ensemble des tours de parole de chaque interlocuteur et entraînons des modèles de classification utilisant des informations lexicales, morphologiques et syntaxiques. Cette étude est la première du genre sur le français et obtient des résultats comparables à celles sur l'anglais. Nos premières expériences tendent à montrer que la parole des personnes avec schizophrénie se distingue de celle des témoins : le meilleur modèle obtient une exactitude de 93,66%. Des informations plus riches seront cependant nécessaires pour parvenir à un modèle robuste.

ABSTRACT

Towards automatic identification of persons with schizophrenia in controlled conversations.

We present experiments to automatically identify patients with symptoms of schizophrenia in controlled conversations between patients and psychotherapists. We merge the speech turns of each interlocutor and train basic classifiers with these data, using lexical, morphological and syntactic features. This study is the first of its kind in French and obtains results comparable to the English state-of-the-art. Our first experiments highlight that the speech of patient with schizophrenia differs from control one : the best model obtains an accuracy of 93.66%. However, richer descriptions will be needed to produce an applicable and reliable model.

MOTS-CLÉS : Dialogue, schizophrénie, apprentissage automatique.

KEYWORDS: Dialog, schizophrenia, machine learning.

1 Introduction

La schizophrénie est définie comme un trouble mental sévère (Association *et al.*, 2015). Cette maladie s'accompagne de symptômes très variables, les plus manifestes étant les idées délirantes, les halluci-

nations et le discours désorganisé. De nombreuses études convergent aujourd'hui vers l'hypothèse selon laquelle il existe une composante génétique de la pathologie schizophrénique, à tout le moins comme une condition nécessaire et non suffisante. Son étiologie demeure toutefois complexe. Pour autant, les avancées empiriques, théoriques et méthodologiques non négligeables en psychopathologie cognitive, en pragmatique linguistique et les progrès considérables, en particulier dans le domaine de l'électrophysiologie et de l'imagerie cérébrale, contribuent à une meilleure caractérisation des troubles schizophréniques (Besche-Richard *et al.*, 2018). Selon l'Organisation Mondiale de la Santé, les troubles schizophréniques touchent environ 1% de la population mondiale. En outre, les troubles cognitifs affectent 70 à 80% de la population atteinte de troubles schizophréniques (Potvin *et al.*, 2017). Pour autant les symptômes de cette pathologie, ou leurs effets, impactent largement les pratiques langagières qui apparaissent ainsi comme un bon angle pour aborder la pathologie.

L'identification automatique de patients manifestant des symptômes de la schizophrénie à partir de la production langagière, écrite ou orale, est un enjeu important dans le domaine de la santé, car cela pourrait constituer une aide décisive vers un diagnostic pour les médecins. Par ailleurs, étudier des cas de parole affectée permet d'améliorer notre compréhension du fonctionnement du langage en général, et cela devrait également permettre d'adapter des systèmes de TAL à des parties de la population qui, en plus de souffrir d'un trouble psychiatrique potentiellement désociabilisant, présentent un usage de la langue qui dévie des modèles dont nous disposons.

A terme, il serait intéressant de modéliser cet usage de la langue dans le dialogue, car c'est dans les interactions que nous attendons les déviations les plus importantes. C'est notamment les perspectives du projet SLAM [1] et son corpus en français, fondé sur des conversations contrôlées entre personnes avec schizophrénie ou témoins et psychologues qui ont été filmées et enregistrées puis retranscrites. Ces dialogues servent de base à la présente étude. Cependant, comme première approximation, nous avons choisi de modéliser le problème à partir de quasi-monologues, c'est-à-dire une fusion de l'ensemble des tours de parole de chaque interlocuteur pour chacun des dialogues considérés. À partir de cet ensemble de tour de parole, nous construisons des systèmes de classification permettant d'identifier les personnes avec schizophrénie. Cette fusion permet de construire des instances de classification contenant plus d'information qu'un simple tour de parole. Ce choix, qui exclut la parole du praticien lors de la classification, est également justifié par la spécificité de ces entretiens dans lesquels le psychologue n'est pas personnellement investi : sa mission est de maintenir l'interaction en relançant l'échange pour qu'il se poursuive le plus longtemps possible. Nous entraînons des modèles de classification sur les monologues et montrons qu'il est possible d'identifier des personnes avec schizophrénie à partir d'informations lexico-syntaxiques dans notre corpus avec une exactitude de 93,66%.

2 État de l'art

Les développements linguistiques sur le discours schizophrénique émergent avec les études de Chaika (1974); Fromkin (1975). Depuis, des indices de moins bonne maîtrise des catégories morpho-syntaxiques (*Part-Of-Speech*, *POS*) et de la syntaxe ont été étudiés (Andreasen, 1979; Fraser *et al.*, 1986; Hoffman & Sledge, 1988). Cependant, il est souvent difficile de caractériser ce qui relève de la pathologie de ce qui relève de la médication. De plus, les éléments mis en avant sont peu discriminants. Les patients ont une maîtrise moins bonne de ces niveaux linguistiques, mais ce sont leurs capacités

1. https://team.inria.fr/semagramme/fr/slam/

cognitives en général qui semblent dégradées (Docherty *et al.*, 1996). Il apparaît alors que travailler sur le discours des personnes avec schizophrénie c'est aussi travailler sur les capacités cognitives.

La détection automatique de troubles schizophréniques est un champ de recherche actif, avec des études qui se concentrent principalement sur deux types de caractéristiques : signaux biomédicaux du type électro-encéphalographie (EEG) et images de résonance magnétique (IRM) (Greenstein *et al.*, 2012; Sabeti *et al.*, 2011). Les études fondées sur les données langagières sont encore assez rares, mais un courant de recherche émerge ces dernières années dans le domaine de l'identification automatique de différents troubles comme la dépression, seule (Pestian *et al.*, 2017) ou associée à d'autres troubles comme le syndrome post-traumatique (PTSD) (Pedersen, 2015) et les pré-symptômes de la maladie d'Alzheimer (Jarrold *et al.*, 2010).

Quelques études ont été spécifiquement consacrées à la schizophrénie. Dans la première étude dédiée à ce problème, Strous *et al.* (2009) utilisent des écrits de personnes avec schizophrénie pour construire des systèmes de classification fondés sur des informations lexicales et obtiennent une exactitude de 83,3%. Ils observent des traits particuliers aux personnes avec schizophrénie comme un usage plus restreint de prépositions et une sur-représentation de la première personne. Ensuite, plusieurs études ont été menées à partir de messages écrits sur Twitter par des personnes s'auto-identifiant avec schizophrénie. Mitchell *et al.* (2015) ont collecté des données pour 174 patients (3200 *tweets*) et testé différents ensembles de traits lexicaux, comme des catégories sémantiques issues de lexique ou des clusters Brown : ils présentent des systèmes de classification (SVM) avec au mieux 82,3% d'exactitude. Cette étude est étendue dans (Birnbaum *et al.*, 2017) à partir de 1,9 millions de *tweets* collectés pour 146 patients. Ils obtiennent également des scores hauts, avec 90,0% d'exactitude, avec des traits lexicaux, notamment des catégories du LIWC (Pennebaker *et al.*, 2001). Ils observent comme précédemment une utilisation accrue des pronoms de première personne, ainsi que les termes appartenant au champ lexical de la santé.

Enfin, Kayi *et al.* (2017); Allende-Cid *et al.* (2019) ont exploré d'autres représentations notamment en se fondant sur des informations morpho-syntaxiques et syntaxiques. Allende-Cid *et al.* (2019) utilisent des textes narratifs rédigés par les sujets (13 personnes avec schizophrénie et 50 témoins) et démontrent que les catégories morpho-syntaxiques permettent déjà des performances assez hautes, 82,8% de F1. Kayi *et al.* (2017) ont eux étudié à la fois des *tweets* (174 sujets par groupe) et des textes narratifs (environ 95 sujets par groupe). Ils utilisent des informations syntaxiques, sémantiques (rôles sémantiques, LDA, clusters) et pragmatiques (sentiments), ces deux dernières se révélant particulièrement utiles pour les données issues de réseaux sociaux (81,65% en F1), tandis que, comme dans l'étude précédente, les traits morpho-syntaxiques se révèlent efficaces sur les textes narratifs (69,76% de F1).

Toutes ces études mettent en jeu des corpus différents dont les données ne sont forcément disponibles, les comparaisons entre études sont donc difficiles. Elles mettent cependant clairement en lumière le fait que les personnes atteintes de schizophrénie présentent des spécificités dans leur usage de la langue à différents niveaux, et nous testons également dans cette étude les informations lexicales, morpho-syntaxiques et syntaxiques mis en oeuvre dans des modèles de classification mais sur des données dialogiques et en français.

Différentes études ont déjà proposé des analyses du corpus du projet SLAM, sur les disfluences, les *POS* et la lexicographie (Amblard *et al.*, 2015; Amblard & Fort, 2014), ainsi que sur des aspects discursifs (Rebuschi *et al.*, 2014), notamment en s'appuyant sur une représentation de l'interaction adaptée de la S-DRT (Asher *et al.*, 2003). Ces différents travaux permettent d'avoir une modélisation fine de l'interaction, et des caractérisations sur ces niveaux linguistiques. Dans la présente étude, nous

reprenons l'analyse des usages langagiers à travers le développement d'un système de classification.

3 Origine et caractérisation des données

Dans la suite, nous appelons "entretien" le dialogue originel entre une personne avec schizophrénie (ou un témoin) et un psychologue, "document" les transcriptions de ces entretiens. Ces entretiens sont constitués des tours de parole (TDP) de chaque locuteur, et nous appelons cTDP la concaténation de tous les tours de parole d'un locuteur au sein d'un document.

Le corpus SLAM : Le corpus a été développé dans le cadre du projet SLAM[2]. Les entretiens sont réalisés en milieu hospitalier auprès de patients diagnostiqués par des médecins-psychiatres et des psychologues de l'institution d'accueil. L'entretien est associé à des tests neuropsychologiques permettant de mesurer les aptitudes des patients sur différents plans (capacité de mémoire de travail, fluence verbale, attention, vitesse motrice, fonctions exécutives, *etc*). Les interactions verbales des patients avec une psychologue sont par ailleurs enregistrées réalisent au cours d'un entretien semi dirigé. La participation des patients est libre et les éléments recueillis lors de l'expérience ne sont pas utilisés par l'équipe médicale pour le suivi du patient. Il y a donc une vraie liberté dans l'entretien. Les thématiques abordées restent simples (quotidien du patient, historique médical, anamnèse avant l'hospitalisation, *etc*). Ces entretiens sont enregistrés avec un double système d'eye-tracker permettant de travailler en parallèle sur les mouvements de regard des deux locuteurs. Les entretiens sont conduits par une psychologue qui n'est pas engagée personnellement dans le dialogue. Il ne s'agit donc pas d'une situation d'interaction symétrique du quotidien, la parole du patient se rapproche d'un monologue. Ceci explique notre choix d'extraire la production langagière du locuteur et de l'isoler comme un tout cohérent.

Description des données : Le corpus est composé de 41 documents, 18 personnes avec schizophrénie et 23 témoins pour le groupe contrôle. Une seule psychologue interroge ces 41 sujets. Chacun de ces groupes contient 15 sujets masculins, le reste (donc 3 et 8) étant féminin. Cette répartition présente donc un biais. Il est admis qu'il existe des différences significatives selon le genre (aspects cliniques et paracliniques) (Douki Dedieu *et al.*, 2012). En ce moment, la majorité des études porte surtout sur des sujets mâles et nous pensons que ces différences devront être prises en compte dans la démarche diagnostique.

De manière non surprenante, les personnes avec schizophrénie ont, en moyenne, le même nombre de TDP par document que la psychologue (199,9 vs. 200,2). Par contre, ils parlent plus (2675,5 mots par document) et leurs phrases sont plus longues (13,4 mots par phrase) par rapport à la psychologue (1814,6 mots par document, 9,1 mots par phrase). Les témoins s'expriment sensiblement plus (342 TDP et 3305 mots par document) avec des phrases plus courtes (10,5 mots par phrase). Les personnes avec schizophrénie ont par ailleurs un taux plus élevé d'utilisation de mots grammaticaux (n'appartenant pas aux catégories : nom, verbe, adverbe ou adjectif) que la psychologue ou les témoins (SCZ 56% vs. témoins 51% vs. psychologue 50%).

2. Pour Amblard *et al.* (2014), le contenu des entretiens donnait de nombreux éléments géographiques et biographiques du patient et son entourage que l'anonymisation ne suffit pas à rendre opaque, la distribution des données est difficile.

Parmi les mots les plus utilisés, que ce soit chez les personnes avec schizophrénie ou les témoins, nous observons plusieurs thématiques :

— Pour les personnes avec schizophrénie : typiquement des mots liés à la douleur comme "maladie", "hospitalisation", "hallucinations". Ce point correspond au label *Catastrophe* parmi les *top semantic features* observés par Kayi *et al.* (2017) qui présentent, dans leur étude, des traits linguistiques prédictifs des personnes avec schizophrénie à l'écrit. De ce point de vue, l'analyse empirique donne corps au contexte conversationnel au sein duquel les patients étaient amenés indirectement à évoquer les prémices de l'entrée dans la maladie. Nous constatons également une utilisation plus fréquente du mot "je" chez les personnes avec schizophrénie.

— Pour les témoins : des mots liés à l'éducation comme "master", "thèse", "licence" et à la psychologie comme "psychiatre" et "psychologue" ressortent significativement. Il se trouve que les sujets témoins sont majoritairement des étudiants de 1er ou 2me année inscrits dans une filière de sciences humaines.

Ces différences dépendent en grande partie des thématiques de conversation choisis par les sujets. Les patients sont censés parler de leur quotidien qui recoupe de fait des réalités différentes ce qui explique les différences de champ lexical. Ceci pourrait donc correspondre à un biais dans nos données.

4 Expériences

Nous présentons les résultats de classification entre locuteurs schizophrènes et non schizophrènes à partir de leurs interventions dans des dialogues transcrits.

Dans ces expériences, nous avons choisi comme première approche d'isoler les tours de parole de chaque locuteur dans les dialogues : nous extrayons et concaténons les TDP de personne avec schizophrénie (respectivement du témoin) dans le dialogue considéré et utilisons cette concaténation comme instance de classification (les cTDP du psychologue sont ici ignorées). La classification est alors traitée à partir de documents longs, contrairement à une approche à partir du dialogue (*i.e.* une instance de classification serait un tour de parole) qui se construirait sur la succession d'interventions courtes contenant trop peu d'information pour une classification précise. Cette approche nous permet d'englober tout la production d'un interlocuteur donné, mais elle a le désavantage de perdre le contexte fourni par l'autre interlocuteur, donc les éléments d'interaction. Ces aspects sont reportés à de futurs développements.

Par ailleurs, nous nous sommes concentrés sur des traits linguistiques directement extraits des transcriptions. Il est cependant certain que des informations non linguistiques seraient cruciales pour cette tâche, comme le non verbal, le genre, l'âge, ainsi que les résultats des patients obtenus aux tests neurocognitifs. Là aussi, de futurs développements intégrant le comportement oculomoteur pourrait être pris en compte.

4.1 Représentation des données

Les cTDP des personnes avec schizophrénie sont libellées comme des instances positives, et celles des témoins comme des instances négatives. Les traits sont construits à partir d'informations lexicales, morpho-syntaxiques et syntaxiques.

Type de traits	Classifieur	#Orig.	Seuil	#Sélec.	Ratio %
bow	NB	6504	9	6488	99,75
bow	SVM	6504	méd.	3254	50,03
n-gram	SVM	118473	8	98	0,08
treelet	SVM	16865	3	675	4,00
bow + treelet	NB	23369	8	11684	49,99
bow + treelet	SVM	23369	moy.	3434	14,69
bow + n-gram	SVM	124977	4	491	0,39
n-gram + treelet	SVM	135338	4	552	0,41
bow + n-gram + treelet	SVM	141842	5	257	0,18

TABLE 1: Nombre de traits à l'origine ("#orig.") et sélectionnés ("#selec") par les classifieurs

Traits lexicaux : Un document est représenté par les mots (tokens) qui le constituent sans tenir compte de l'ordre (sac de mots, bow). Cette représentation est la plus simple et sert de système de référence. Ce modèle permet par ailleurs d'identifier de potentielles préférences lexicales. Nous testons également une représentation en n-grammes sur les tokens (n-gram), afin de prendre en compte partiellement l'ordre des mots. Notons que les n-grammes peuvent contenir des mots à cheval sur différentes prises de parole d'un même locuteur, et donc encoder une partie du contexte dialogique. Nous testons des bi-grammes et des tri-grammes. Les données sont normalisées en utilisant le TF-IDF.

Traits morpho-syntaxiques et syntaxiques : Nous utilisons UDPipe [3] (Straka & Straková, 2017).Comme nos données sont dialogiques, les modèles classiques donnent d'assez mauvais résultats. Nous utilisons donc un modèle ré-entraîné sur un corpus oral du français (Spoken-French 2.5 [4]). Le pré-traitement supprime la ponctuation et segmente minimalement (par exemple, ajout d'un espace pour les apostrophes). Nous obtenons un étiquetage morpho-syntaxique et l'analyse syntaxique en dépendances correspondante. Afin d'encoder les traits syntaxiques, nous utilisons la méthode proposée dans (Johannsen *et al.*, 2015) qui consiste à extraire tous les sous-arbres d'au plus 3 *tokens* (treelet). Ces auteurs s'intéressaient, eux, aux variations syntaxiques liées au genre et à l'âge. Un treelet d'1 token correspond simplement au *POS* associé. Un treelet contenant 2 tokens est une relation typée entre une tête et un dépendant, par exemple : 'VERB→nsubj→NOUN'. Un treelet sur 3 tokens peut avoir deux formes, selon que l'on a une tête dominant deux dépendants ('NOUN←nsubj←VERB→dobj→NOUN') ou une chaîne de dépendances ('PRON←poss←NOUN←nsubj←VERB').

Sélection de traits : Les trois ensembles de traits construits correspondent à des vocabulaires larges (voir la table 1). Notre problème d'apprentissage est confronté à des données rares (41 instances) mais de dimension élevée, ce qui conduit généralement à des problèmes de sur-apprentissage et de manque de généralisation des modèles. Nous incluons une sélection des traits au cours de l'entraînement avec une méthode implémentée [5] dans scikit-learn (Pedregosa *et al.*, 2011). En calculant les coefficients (ou poids) attribués par un modèle à chaque trait et en ne conservant que ceux dont le poids est supérieur

3. http://ufal.mff.cuni.cz/udpipe
4. https://tinyurl.com/UniversalDependencies-French-S
5. feature_selection.SelectFromModel https://scikit-learn.org/

à un seuil, cette méthode permet de sélectionner les traits importants. Nous testons sans sélection (seuil `None`), puis avec un seuil correspondant à la moyenne et la médiane sur les poids obtenus, ainsi que 10 valeurs régulièrement distribuées entre $1e-5$ (la valeur par défaut dans l'implémentation utilisée) et le poids du 50^e trait le plus important. Cette valeur maximale choisie *a priori* permet de conserver au minimum 50 traits dans le modèle.

La sélection de traits nous permet de réduire drastiquement la taille du vocabulaire, comme indiqué dans la table 1. Notons que NB conduit généralement à conserver plus de traits, la distribution des coefficients étant plus continue.

4.2 Classification

Nous avons trop peu de données (41 documents) pour séparer les données entre entraînement et test, nous utilisons une validation croisée enchâssée permettant d'obtenir une estimation réaliste de l'erreur du modèle (Varma & Simon, 2006; Scheffer, 1999). Contrairement à une simple validation croisée, le modèle est choisi sur un ensemble de données différent de celui utilisé pour l'évaluation. Cette méthode repose sur deux boucles : à l'extérieur, un sous-ensemble parmi N est conservé pour l'évaluation, tandis qu'à l'intérieur, une validation croisée en M sous-ensembles permet d'optimiser le modèle (choix des hyper-paramètres et sélection de traits), le processus est répété N fois, ici $N = M = 5$.

Modèles : Nous testons différents modèles de classification implémentés dans ScikitLearn en optimisant les hyper-paramètres suivants :

— Naive Bayes : paramètre de lissage $\alpha \in V = \{0.001, 0.005, 0.01, 0.1, 0.5, 1, 5, 10, 100\}$;
— Régression logistique : norme L_2, et optimisation du coefficient de régularisation $C \in V$;
— SVM avec noyau linéaire : norme L_2, et optimisation de $C \in V \cup \{1000\}$.

Pendant le processus de validation, nous optimisons les hyper-paramètres dans les boucles intérieures. Ces hyper-paramètres restent relativement stables sur l'ensemble des expériences pour chaque modèle et chaque type de trait. Les valeurs des hyper-paramètres choisis sont, pour NB, $\alpha = 0{,}1$, avec `ngram`, $0{,}001$ avec `bow`, `treelet`, `bow+treelet`, $0{,}01$ dans les autres cas ; pour MaxEnt, $C = 100$ pour tous les traits ; pour SVM, on trouve $C = 1000$ pour `treelet`, $C = 100$ pour `bow`, `n-gram` et `n-gram+treelet`, sinon $C = 5$.

5 Résultats

5.1 Systèmes de référence

Nous reportons dans la table 2 les résultats de nos systèmes de référence. Le classifieur correspondant à la chance, qui attribue systématiquement la classe majoritaire (ici témoin), a une exactitude de $56{,}1\%$. Par ailleurs, nous avons examiné trois autres systèmes de référence : deux testant un indice simple de complexité des mots et le troisième la différence d'utilisation des déictiques "je" et "tu".

Avec l'hypothèse que les personnes avec schizophrénie utilisent des mots moins complexes, nous testons des systèmes fondés sur, simplement, la taille moyenne des mots ou, pour introduire une forme de normalisation, la taille moyenne des mots au dessus de la taille moyenne de tous les mots.

	Acc.	Prec.	Rec.
Majorité	56,10		
long. mot	49,51	17,21	11,11
$>$ long. moy. mot	52,43	37,43	22,78
ratio je/tu	**72,19**	**69,87**	**35,56**

TABLE 2: Résultats des systèmes de référence : classe majoritaire, longueur des mots, longueur des mots supérieure à la moyenne, ratio d'utilisation des déictiques "je" et "tu"

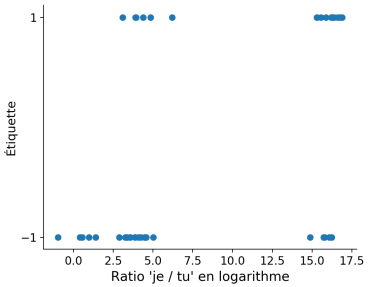

FIGURE 1: Ratio *je* / *tu* par document

Les deux classifieurs donnent des résultats inférieurs à la chance (exactitude de, resp., 49,51% et 52,43%), la longueur des mots n'est donc pas une information pertinente pour cette tâche.

Comme de précédentes études ont noté l'importance des pronoms de première personne pour la tâche, nous testons également un système fondé sur le ratio des déictiques "je" et "tu" (normalisé en logarithme) dans chacune des deux classes. Le système obtient de meilleurs résultats, largement au dessus de la chance (72,19% d'exactitude). La figure 1 montre que, dans nos données, les témoins (étiquette -1) ont un taux quasiment similaire d'utilisation de ces déictiques (majorité des documents autour de 0) tandis que les personnes avec schizophrénie (resp. 1) favorisent la première personne.

5.2 Meilleurs systèmes

La table 3 présente les résultats obtenus par les différents classifieurs pour les différents ensembles de traits testés. Le meilleur système obtient 93,66% d'exactitude (F1 92,21%), il est fondé sur l'algorithme bayésien naïf (NB) et les traits de type sac de mots (`bow`). Avec SVM, nous obtenons 90,98% d'exactitude (89,79 en F1). Ces résultats sont supérieurs à ceux présentés dans (Allende-Cid *et al.*, 2019) (87,50% en F1) qui utilisent également une représentation sac de mots et SVM mais un corpus plus large, et également supérieurs à ceux présentés dans (Birnbaum *et al.*, 2017), ceux-ci obtenant 90% d'exactitude sur des données Twitter également plus nombreuses, avec des traits de type n-grammes ($n = 1, 2$ et 3, donc correspondant ici à `bow+n-gram`) ainsi que des catégories sémantiques issues du lexique LIWC (Pennebaker *et al.*, 2001) et un classifieur de type Random Forest. Ceci pourrait indiquer que nous avons un vocabulaire plus restreint et peut-être biaisé dans nos données. Le second meilleur système, 92,20% d'exactitude (F1 90,38%), est obtenu avec ces mêmes traits lexicaux combinés aux traits syntaxiques (`bow+treelet`) et également NB.

De manière assez classique, les meilleurs scores sont obtenus avec le classifieur SVM, sauf lorsque les traits `bow` sont pris seuls ou du moins dominent (dans la combinaison avec les `treelet`, la phase de sélection a tendance a plus largement supprimé des traits de cette catégorie que des traits sac de mots, cf. table 1), auxquels cas c'est NB qui permet les meilleures performances.

Comme les différences observées entre les scores sont assez faibles et que la taille du corpus est restreinte, nous avons vérifié la significativité statistique de certains résultats : notamment, nous cherchons à vérifier si NB est vraiment supérieur à SVM ($\pm$3,66% avec `bow`, $\pm$3,42% avec `bow+treelet`), et si ces deux ensembles de traits correspondent à des performances différentes ($\pm$1,46 avec NB et $\pm$2,2 avec SVM). Nous utilisons le *test de Student* qui est interprétable avec un échantillon de très petite

Algorithme Sélection	SVM non	SVM oui	MaxEnt oui	NB oui
`bow`	90,00	90,98	87,07	**93,66**
n-`gram`	68,78	**81,71**	79,76	65,61
`treelet`	61,46	**66,83**	58,29	58,05
`bow`+n-`gram`	80,49	**88,54**	86,59	70,49
`bow+treelet`	87,07	88,78	84,88	**92,20**
n-`gram+treelet`	68,54	**80,73**	77,56	62,20
`bow`+n-`gram+treelet`	80,98	**85,85**	84,15	77,07

TABLE 3: Exactitude moyenne ("Avg Acc.") sans ou avec sélection ("SVM", "MaxEnt" et "NB") pour chaque ensemble de traits

Groupe d'échantillons		t-statistique	p-value	d de Cohen	Taille d'effet
`bow_nb`	`bow_svm`	2,74	0,01	1,23	fort
`bow+treelet_nb`	`bow+treelet_svm`	2,10	0,05	0,94	fort
`bow_nb`	`bow+treelet_nb`	1,21	0,24	0,54	moyen
`bow_svm`	`bow+treelet_svm`	1,49	0,15	0,67	moyen

TABLE 4: Résultats des Tests de Student pour la comparaison de classifieurs

taille, en particulier si la taille d'effet (*effect size*, calculée en utilisant le coefficient d de Cohen [6]) et la corrélation entre les échantillons (t-statistiques) sont suffisamment importantes (De Winter, 2013).

Les résultats sont présentés dans la table 4. Ces tests démontrent que l'algorithme NB permet effectivement des performances significativement supérieures à celles obtenues avec SVM (p-value $\leq$ 0,05). Par contre, l'algorithme étant fixé, la perte en performance observée en ajoutant les informations syntaxiques (`treelet`) n'est pas significative, cette combinaison n'apporte rien.

Le volume de données étant limité, les erreurs de classement des documents entraînent une variation importante des résultats. Ainsi, SVM a en fait systématiquement un meilleur rappel (pour `bow` : 90,56 vs 86,11 ; pour `bow+treelet` : 86,67 vs 84,44) et, pour `bow`, il classe correctement 16,3 instances de personnes avec schizophrénie contre 15,5 pour NB (sur 18) tandis que NB classe généralement correctement toutes les instances de témoins, donc la classe majoritaire.

5.3 Différents jeux de traits

Les informations lexicales semblent clairement les plus pertinentes pour la tâche : après `bow`, ce sont les n-`gram` qui conduisent aux meilleures performances (au mieux, 81,71% d'exactitude) tandis que les performances sont bien plus basses (66,83% d'exactitude, 57,30% en F1) avec les traits syntaxiques (`treelet`) pris seuls. Ce score est supérieur à la chance (56,10%) ce qui semble indiquer qu'un signal est présent, mais largement moins discriminant que l'information lexicale et apparemment non complémentaire, comme le montre l'absence d'amélioration lorsque les traits sont

6. Traditionnellement, un d autour de 0,2 est décrit comme un effet faible, 0,5 moyen et 0,8 comme fort.

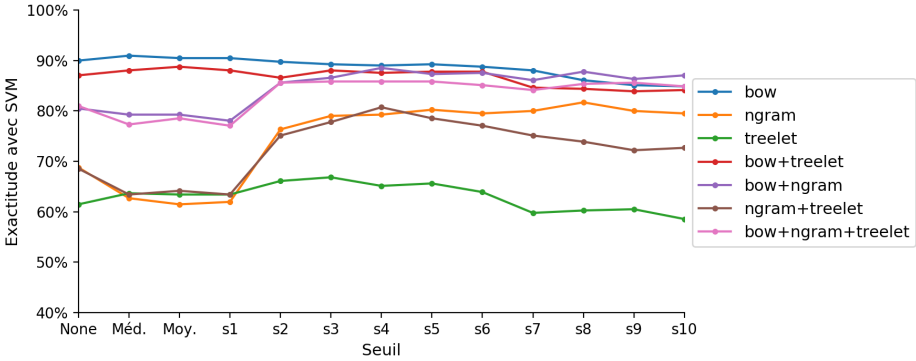

FIGURE 2: Score d'exactitude en fonction des seuils de sélection pour les traits

combinés (seconde partie de la table 3). Par contre, le système fondé sur les `treelet` obtient un score d'exactitude inférieur au système de référence utilisant le ratio de déictique, ce qui démontre l'importance de cette information pour la tâche. Kayi *et al.* (2017) rapportent des scores supérieurs (F1 de 68,48% pour les textes narratifs et 63,19% pour les tweets)) avec le même algorithme (SVM) et en ne considérant que les POS, pourtant incluses dans `treelet`. Cette baisse est attribuable soit au bruit amené par les autres traits pris en compte dans notre système (`treelet` de 2 ou 3 tokens, cf. Section 4.1) soit à la taille des données, ces auteurs disposant de plus d'instances de classification (348 tweets ou 190 textes narratifs contre 41 ici).

Dans nos expériences, combiner les traits ne conduit à aucune amélioration, voire dégrade les performances. L'information contenue dans les traits lexicaux semblent redondantes, puisque combiner `bow` et n`-gram` laissent les performances inchangées avec sélection. Sans sélection, on peut constater (cf. figure 2, seuil "None") que cette combinaison surpasse n`-gram` seul mais pas `bow` seul. La représentation par n`-gram` ne semble ici pas adapté, peut-être à cause des chevauchements entre tours de parole. Par ailleurs, comme le montre la figure 2, l'étape de sélection est cruciale, elle permet une amélioration pour tous les traits, et les meilleurs scores sont très généralement obtenus avec des seuils hauts.

6 Analyse des traits

Traits lexicaux : Afin d'évaluer la pré-dominance de certains champs lexicaux, nous utilisons le test de corrélation de Spearman pour classer les tokens et regardons ceux avec les plus hautes valeurs (p-valeur $< 0,05$ et coefficient $|\rho| > 0,3$). Nous obtenons 210 mots, une sélection est présentée dans la table 5. Les termes relatifs à la pathologie comme "maladie, traitement, médecin, diagnostic" sont positivement corrélés à la classe schizophrène, tandis que des termes relatifs à la scolarité comme "licence, thèse" et à la vie comme "vacances, bio, monde" reçoivent des poids négatifs, ce qui rejoint les observations de la Section 3. Par ailleurs, les sujets avec schizophrénie utilisent plus de références à la première personne, ce que l'on voit avec des déictiques ("j'" ("je"), "mon", "ma", "mes") ainsi que des formes d'auxiliaires ("suis" et "ai") tandis que les témoins utilisent plus de références de seconde personne ("tu", "es" et "as"). Nous évaluons également l'impact de ces traits dans les modèles : en ignorant "je" et "tu" (et les formes élidées "j'" et "t'"), nous constatons une chute faible de l'exactitude

Vocabulaire	ρ	p-value
Douleur		
maladie	0,540	$< 1e-3$
hospitalisé	0,509	$< 1e-3$
hallucinations	0,420	0,006
Éducation		
master	$-0,505$	$< 1e-3$
concours	$-0,496$	$< 1e-3$
fac	$-0,490$	0,001

Vocabulaire	ρ	p-value
Psycho		
psychologie	$-0,536$	$< 1e-3$
psychologue	$-0,453$	0,002
Déictique		
j' / je	0,635	$< 1e-5$
mon	0,613	$< 1e-5$
t' / tu	$-0,467$	0,002
nous	$-0,342$	0,028

TABLE 5: Valeurs ρ et p du test de Spearman pour les mots traits

avec NB ($-0,49\%$) mais importante avec SVM ($-6,59\%$).

Ces observations rejoignent les conclusions d'études précédentes : par exemple, Strous *et al.* (2009) argumentent qu'une utilisation plus importante des déictiques à la première personne et moins de références aux sujets à la troisième personne, accompagnée par des répétitions lexicales sont des caractéristiques de sujets renfermés sur eux-même. D'autres études ont également affirmé que l'utilisation de la première personne du singulier est associée à des états affectifs négatifs tels que la dépression (Rude *et al.*, 2004; Chung & Pennebaker, 2007). Évidemment, ce type de résultat est à apprécier relativement aux conditions contextuelles et conversationnelles dans lesquelles les données sont recueillies.

Traits syntaxiques : Pour les traits syntaxiques (voir la table 6), les verbes semblent un marqueur fort des personnes avec schizophrénie tandis que les noms apparaissent plus souvent dans le corpus des témoins. Les statistiques des *2-token* `treelet` tendent à indiquer que les personnes avec schizophrénie utiliseraient plus de groupes verbaux et moins de groupes nominaux. Ainsi le *2-token* `treelet` "VERB→aux→AUX" (par exemple : "(j')ai fait", "(c')est (pas) gagné") et "VERB→nsubj→PRON" (par exemple : "ça va", "(je) sais pas") sont les traits les plus discriminants des personnes avec schizophrénie. Une forte utilisation des auxiliaires montre aussi que les personnes avec schizophrénie parlent souvent du passé, de ce qu'ils ont fait. Côté témoins, on trouve plus de relations qui capturent des nominaux : "expl" capture des nominaux explicatifs ou pléonastiques ; les cas ("case") sont traités comme des dépendants du nom auquel ils s'attachent souvent avec des adpositions [7] (par exemple : "(fait partie) de l'expérience").

7 Conclusion

Nous avons proposé les premiers systèmes d'identification automatique de personnes atteintes de schizophrénie dans des données dialogiques et en français. Nous avons testé différentes représentations, incluant des informations lexicales, morpho-syntaxiques et syntaxiques, ainsi que différents classifieurs. Notre meilleur système utilise des informations lexicales uniquement et obtient une exactitude de 93,66 avec le classifieur NB. Cependant, l'étude des données et des modèles nous a permis

7. Dans le format conllu, l'adposition recouvre les prépositions et postpositions.

treelet	SCZ	ρ	Témoin	ρ
1-*token*	verb	0,21	noun	−0,17
2-*token*	verb→aux→aux	0,41	pron→nsubj→pron	−0,64
	verb→nsubj→pron	0,37	cconj→nsubj→pron	−0,46
	aux→advcl→verb	0,34	propn→conj→pron	−0,46
3-*token*	pron→nsubj→verb←iobj←pron	0,51	pron→obj→verb←mark←sconj	−0,66
	aux→aux→verb←obl←pron	0,49	adp→mark→verb←det←det	−0,39
	adj→advcl→verb←nsubj←pron	0,47	verb→expl→noun→case→adp	−0,36

TABLE 6: Traits typiques des classes SCZ et témoins (p-value $<0,05$ pour les *2-tokens* et *3-tokens*)

d'identifier de potentiels biais lexicaux dans notre corpus, notamment à travers un groupe témoins dont le vocabulaire est centré sur les études, et des patients habitués à décrire leur environnement médical, ce qui rend probablement nos modèles peu robustes.

Par ailleurs, nous nous limitons au contenu linguistique de l'échange sans considérer l'évolution de la phonologie, de la phonétique, ni ce qui appartient au non-verbal (position, regards, *etc.*). Cependant, ce groupe reste selon nous pertinent pour aller vers le développement d'applications ciblées, comme la détection de changement d'états chez les patients ou l'adaptation automatique d'un *chatbot* par exemple.

Dans une extension de l'étude nous souhaitons tester d'autres classifieurs comme random forest ou des perceptrons. Par ailleurs, nous souhaitons tester plus de traits certains toujours linguistiques comme les déictiques, et surtout intégrer des informations sémantiques comme des connecteurs, d'autres extra linguistiques, en particulier les résultats aux tests neuro-cognitifs. Il nous apparaît aussi primordial de nous intéresser à la production langagière dans sa dynamique en l'analysant comme un dialogue et pas seulement comme un quasi-monologue.

Les résultats obtenus montrent que si les classifications ne sont pas parfaitement opérantes, elles ouvrent vers des indices intéressants. Des analyses récentes tendent à prouver qu'il ne faut pas limiter l'analyse à la seule production des patients. De manière surprenante, il semble que les interlocuteurs des personnes avec schizophrénie adaptent leur manière de parler à leur interlocuteur. Ainsi, tout un chacun identifierait inconsciemment des défaillances chez l'autre, les spécialistes interprétant ces indices de façon diagnostique. Ainsi, un prolongement de cette étude doit s'intéresser à la production langagière du psycho-thérapeute soit en face d'un patient, soit en face d'un témoin.

Remerciements

Nous remercions les relecteurs pour leurs commentaires pertinents. Ce travail a obtenu le soutien du projet PIA "Lorraine Université d'Excellence", ANR-15-IDEX-04-LUE, ainsi que les infrastrustures du CPER LCHN (Contrat de Plan État-Région - Langues, Connaissances et Humanités Numériques). Nous remercions le Centre Hospitalier Le Vinartier pour avoir contribué de manière décisive à la mise en place de l'expérimentation.

Références

ALLENDE-CID H., ZAMORA J., ALFARON-FACCIO P. & ALONSO M. (2019). A machine learning approach for the automatic classification of schizophrenic discourse. *IEEE Access*, p. 45544–45554. DOI : 10.1109/ACCESS.2019.2908620.

AMBLARD M. & FORT K. (2014). Étude quantitative des disfluences dans le discours de schizophrènes : automatiser pour limiter les biais. In *TALN - Traitement Automatique des Langues Naturelles*, p. 292–303, Marseille, France. HAL : hal-01054391.

AMBLARD M., FORT K., DEMILY C., FRANCK N. & MUSIOL M. (2015). Analyse lexicale outillée de la parole transcrite de patients schizophrènes. *Traitement Automatique des Langues*, **55**(3), 91 – 115. HAL : hal-01188677.

AMBLARD M., FORT K., MUSIOL M. & REBUSCHI M. (2014). L'impossibilité de l'anonymat dans le cadre de l'analyse du discours. In *Journée ATALA éthique et TAL*, Paris, France. HAL : hal-01079308.

ANDREASEN N. C. (1979). Thought, language, and communication disorders : I. clinical assessment, definition of terms, and evaluation of their reliability. *Archives of general Psychiatry*, **36**(12), 1315–1321.

ASHER N., ASHER N. M. & LASCARIDES A. (2003). *Logics of conversation*. Cambridge University Press.

ASSOCIATION A. P. *et al.* (2015). *DSM-5-Manuel diagnostique et statistique des troubles mentaux*. Elsevier Masson.

BESCHE-RICHARD C., TERRIEN S., RINALDI R., VERHAEGEN F., LEFEBVRE L. & MUSIOL M. (2018). Les troubles du spectre de la schizophrénie. In C. BESCHE-RICHARD, Éd., *Psychopathologie cognitive. Enfant, adolescent, adulte*, Univers Psy, chapitre 6, p. 153–179. Dunod.

BIRNBAUM M. L., ERNALA S. K., RIZVI A. F., DE CHOUDHURY M. & KANE J. M. (2017). A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. *Journal of medical Internet research*, **19**(8), e289. DOI : 10.2196/jmir.7956.

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

CHUNG C. & PENNEBAKER J. W. (2007). The psychological functions of function words. In K. FIEDLER, Éd., *Social Communication*, volume 1, chapitre 12, p. 343–359. Psychology Press. DOI : 10.4324/9780203837702.

DE WINTER J. C. (2013). Using the student's t-test with extremely small sample sizes. *Practical Assessment, Research, and Evaluation*, **18**. DOI : 10.7275/e4r6-dj05.

DOCHERTY N. M., HAWKINS K. A., HOFFMAN R. E., QUINLAN D. M., RAKFELDT J. & SLEDGE W. H. (1996). Working memory, attention, and communication disturbances in schizophrenia. *Journal of Abnormal Psychology*, **105**(2), 212–219. DOI : 10.1037/0021-843X.105.2.212.

DOUKI DEDIEU S., OUALI U. & NACEF F. (2012). Schizophrénie et genre. In J. DALÉRY, T. D'AMATO & M. SAOUD, Éds., *Pathologies schizophréniques*, Psychatrie, p. 199–205. Lavoisier.

FRASER W. I., KING K. M., THOMAS P. & KENDELL R. E. (1986). The diagnosis of schizophrenia by language analysis. *The British Journal of Psychiatry*, **148**(3), 275–278.

FROMKIN V. A. (1975). A linguist looks at "linguist looks at 'schizophenic language' ". *Brain and Language*, **2**, 498–503. DOI : 10.1016/S0093-934X(75)80087-3.

GREENSTEIN D., WEISINGER B., MALLEY J. D., CLASEN L. & GOGTAY N. (2012). Using multivariate machine learning methods and structural MRI to classify childhood onset schizophrenia and healthy controls. *Frontiers in psychiatry*, **3**. DOI : 10.3389/fpsyt.2012.00053.

HOFFMAN R. E. & SLEDGE W. (1988). An analysis of grammatical deviance occurring in spontaneous schizophrenic speech. *Journal of neurolinguistics*, **3**(1), 89–101.

JARROLD W. L., PEINTNER B., YEH E., KRASNOW R., JAVITZ H. S. & SWAN G. E. (2010). Language analytics for assessing brain health : Cognitive impairment, depression and pre-symptomatic alzheimer's disease. In *International Conference on Brain Informatics*, p. 299–307 : Springer.

JOHANNSEN A., HOVY D. & SØGAARD A. (2015). Cross-lingual syntactic variation over age and gender. In *Proceedings of the nineteenth conference on computational natural language learning*, p. 103–112.

KAYI E. S., DIAB M., PAUSELLI L., COMPTON M. & COPPERSMITH G. (2017). Predictive linguistic features of schizophrenia. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, p. 241–250.

MITCHELL M., HOLLINGSHEAD K. & COPPERSMITH G. (2015). Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd workshop on Computational linguistics and clinical psychology : From linguistic signal to clinical reality*, p. 11–20.

PEDERSEN T. (2015). Screening twitter users for depression and ptsd with lexical decision lists. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology : from linguistic signal to clinical reality*, p. 46–53.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

PENNEBAKER J., FRANCIS M. & BOOTH R. (2001). *Linguistic inquiry and word count (LIWC)*.

PESTIAN J. P., SORTER M., CONNOLLY B., BRETONNEL COHEN K., MCCULLUMSMITH C., GEE J. T., MORENCY L.-P., SCHERER S., ROHLFS L. & GROUP S. R. (2017). A machine learning approach to identifying the thought markers of suicidal subjects : a prospective multicenter trial. *Suicide and Life-Threatening Behavior*, **47**(1), 112–121.

POTVIN S., AUBIN G. & STIP E. (2017). L'insight neurocognitif dans la schizophrénie. *L'Encéphale*, **43**(1), 15–20.

REBUSCHI M., AMBLARD M. & MUSIOL M. (2014). Using SDRT to analyze pathological conversations. Logicality, rationality and pragmatic deviances. In M. REBUSCHI, M. BATT, G. HEINZMANN, F. LIHOREAU, M. MUSIOL & A. TROGNON, Éds., *Interdisciplinary Works in Logic, Epistemology, Psychology and Linguistics : Dialogue, Rationality, and Formalism*, volume 3 de *Logic, Argumentation & Reasoning*, p. 343 – 368. Springer. DOI : 10.1007/978-3-319-03044-9_15, HAL : hal-00910725.

RUDE S., GORTNER E.-M. & PENNEBAKER J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, **18**(8), 1121–1133.

SABETI M., KATEBI S., BOOSTANI R. & PRICE G. (2011). A new approach for eeg signal classification of schizophrenic and control participants. *Expert Systems with Applications*, **38**(3), 2063–2071.

SCHEFFER T. (1999). *Error Estimation and Model Selection*. Thèse de doctorat, Technischen Universitet Berlin, School of Computer Science.

STRAKA M. & STRAKOVÁ J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, p. 88–99, Vancouver, Canada : Association for Computational Linguistics.

STROUS R. D., KOPPEL M., FINE J., NACHLIEL S., SHAKED G. & ZIVOTOFSKY A. Z. (2009). Automated characterization and identification of schizophrenia in writing. *The Journal of nervous and mental disease*, **197**(8), 585–588.

VARMA S. & SIMON R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, **7**.