Investigating non lexical markers of the language of schizophrenia in spontaneous conversations

Chuyuan Li[1], Maxime Amblard[1], Chloé Braud[2]
Caroline Demily[3], Nicolas Franck[3], Michel Musiol[1,4]

[1] Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
[2] IRIT, Université de Toulouse, CNRS, ANITI, Toulouse, France
[3] Centre Hospitalier le Vinatier & UMR 5229, CNRS - Univeristé Lyon 1, Lyon, France
[4] ATILF, UMR 7118, Université de Lorraine, CNRS, 54000 Nancy, France
[1,4] {firstname.name}@loria.fr, [2] chloe.braud@irit.fr

Abstract

We investigate linguistic markers associated with schizophrenia in clinical conversations by detecting predictive features among French-speaking patients. Dealing with human-human dialogues makes for a realistic situation, but it calls for strategies to represent the context and face data sparsity. We compare different approaches for data representation – from individual speech turns to entire conversations –, and data modeling, using lexical, morphological, syntactic, and discourse features, dimensions presumed to be tightly connected to the language of schizophrenia. Previous English models were mostly lexical and reached high performance, here replicated (93.7% acc.). However, our analysis reveals that these models are heavily biased, which probably concerns most datasets on this task. Our new delexicalized models are more general and robust, with the best accuracy score at 77.9%.

1 Introduction

Schizophrenia is defined as a severe mental illness (APA, 2015) that comes with varied symptoms, ranging from delirium to hallucinations. Among these symptoms, there are language disorders, especially the so-called *positive thought disorder* (i.e., disorganized language output such as *derailment* and *tangentiality*) [1] and *negative thought disorder*[2] (Kuperberg, 2010). Schizophrenia affects about 1% of the world's adult population, with cognitive troubles for 70-80% of the patients (Potvin et al., 2017). Since the symptoms often affect language skills, several studies proposed using NLP techniques on patients' productions (Section 2) to identify what is affected in language, thus understand better the disease and its symptoms and how language works in general.

In this paper, we explore linguistic markers of schizophrenia through feature exploration within a classification system. We do so on spontaneous dialogues in French where all the previous work was in English and most used social media data or monologues. Replicating state-of-the-art results allows us to confirm some previous findings of specific features of the language of schizophrenia.

Our study focuses on two aspects: carefully exploring data representations and investigating preliminary modeling of dialogues, both with scarce data. Using spontaneous conversations makes for a realistic scenario – the patient is merely talking with her clinician. However, representing dialogues is not easy: we restrict ourselves to patients' speech turns, and test varied context windows to tackle data sparsity. Additionally, we compare several representations and confirm that lexicon is a good indicator, making for high-performing models with at best 93.7% (acc.). Nevertheless, our analysis demonstrates that it probably corresponds to a bias in our data caused by the constraints imposed during the collection process. Most of the datasets are likely biased the same way. This analysis led us to delexicalized models while focusing on dimensions presumed to be affected in schizophrenia: morpho-syntactic, syntactic, dialogue, and discourse information are therefore considered. Our best delexicalized model gets 77.9% (acc.) and shows the importance of morpho-syntactic information and high-level features in dialogue.

When dealing with medical data, ethical questions arise. The diagnosis of schizophrenia is complex and relies on many indices. Automatic systems could provide psychiatrists with further clues, possibly alleviating the need for the patients to go through several cognitive tests, but this is a long-reach goal. It is clear that the systems developed can not substitute for a human expert, as a diagnosis

[1] *Derailment*: spontaneous speech that tends to slip off track. *Tangentiality*: reply to a question in an oblique or irrelevant manner.

[2] Negative thought disorder are those of poverty of speech and language (known as *alogia*) and poverty of content.

Proceedings of the 2nd Workshop on Computational Approaches to Discourse, pages 20–28
November 10–11, 2021.

is a medical act. Moreover, linguistic clues, while crucial, have to be interpreted within the patient's social environment.

Contributions This study: (i) is the first in French, replicating English studies with comparable results with less data and resources; (ii) continues seminal work on schizophrenia detection in dialogues but with a focus on modeling and bias - two crucial issues for a task inherently data-scarce; (iii) reveals language features of schizophrenia, confirming psychologists' descriptions on the use of complex structures or the capacity to maintain conversation.[3]

2 Related work

Psychiatrists rely on language and speech behavior as one of the main clues in psychiatric diagnosis (Ratana et al., 2019). They found that these patients' speech tends to be less predictable (Salzinger et al., 1964, 1970; Salzinger, 1979), with a poorer vocabulary (Salzinger and Hammer, 1963; Manschreck et al., 1991). It has also been found that their productions tend to be more grammatically deviant (Hoffman and Sledge, 1988) and less syntactically complex than that of controls (Fraser et al., 1986; Morice and Ingram, 1982). At the discourse level, they associate words within a larger context than controls (Maher et al., 2005) with often more diffuse associations (Chaika, 1974; Elvevåg et al., 2007). They also present referential impairments - categorized as vagueness, missing information, or confusing reference (Rochester, 2013; Docherty et al., 1996) -, and specific discontinuities at the discourse level (Musiol and Trognon, 2000; Rebuschi et al., 2014).

On the other hand, many researchers have used NLP methods to help to identify mental disorders, such as depression (Howes et al., 2014; Guntuku et al., 2019; Sekulić and Strube, 2019), post-traumatic stress disorder (Pedersen, 2015; Kleim et al., 2018), suicide risk (Benton et al., 2017), Alzheimer's disease (Orimaye et al., 2014; Fraser et al., 2016), and autism (Goodkind et al., 2018; Sakishita et al., 2019).

For schizophrenia, previous work has mainly focused on lexical information (Mitchell et al., 2015; Hong et al., 2012; Birnbaum et al., 2017; Xu et al., 2019). Unlike ours, these studies rely on Linguistic Inquiry Word Count (LIWC) categories (Pennebaker et al., 2001) - psycho-metrically validated lexicon mapping words to psychological concepts), Latent Dirichlet Allocation (LDA) (Blei et al., 2003) - inferring topics in each document, and Brown clustering (Brown et al., 1992) - grouping contextually similar words into the same cluster. However, most of these resources are only available in English.

More recent approaches considered syntactic, semantic, and sentiment information (Kayi et al., 2017; Allende-Cid et al., 2019). Both studies show good performance with morpho-syntactic features, especially with Part-Of-Speech (POS) tags.[4] They were based on narrative texts (essays and tweets). We here demonstrate that some findings can generalize to spontaneous conversations.

Amblard et al. (2020) proposed the first study on detecting schizophrenia patients from conversations, mostly limited to lexical features. Also, close to our work, Howes et al. (2012a,b, 2013) investigated linguistic features in transcripts of conversations between patients and clinicians. The authors tried to predict patient satisfaction and adherence to treatment on the concatenation of speech turns of the patient. Inspired by the work of Howes et al. (2012b), we also use higher-level features (see Section 3) on real conversations but directly investigating a model of detecting patients with schizophrenia symptoms. Furthermore, we extend previous work by varying the length of dialogues and testing more complex features, including sequences of POS tags, finer tree representations, and dialogical information.

3 Approach

Varying dialogue size: Our data are composed of 41 dialogues with $2,811$ words, and 268 speech turns on average (when limited to patients/controls). The clinician's speech turns are ignored in all dialogues to reduce their impact on classification, but further studies should also include the interaction. First, we concatenate all the speech turns of a patient/control (**Full** setting), thus making for a large document that contains the whole context. Since the documents are long, it could be hard for the system to find regularities, especially with only a few classification instances (i.e., 41). The opposite

[3] Our code is on: `https://github.com/chuyuanli/non-lexical-markers-scz-conv`.

[4] POS tagging is a process of marking up a word in a text to a particular part of speech. Allende-Cid et al. (2019) tested two types of POS tags: a general one called meta-POS (12 labels) and a precise one POS (160 labels). Both allow performance higher than chance.

	#Doc.	#Speech T./doc.			#Word/doc.		
Setting	total	min	max	avg	min	max	avg
Indiv.	10,319	1	1	1	1	274	11
W-128	893	1	34	11	128	317	145
W-256	443	1	72	20	256	424	271
W-512	209	2	129	42	512	609	530
Full	41	76	555	268	703	6,778	2,811

Table 1: Number of documents, speech turns and words per document when varying the window.

option is to classify each speech turn individually (**Indiv.**): this leads to more instances (10,319), but we lose the context of the neighboring speech turns. Moreover, the speech turns are of varied length with an average of 11 words; some of them contain too few words to be informative. The last option is in between: we use a window of at least n words (**W-**n), always going until the end of the current speech turn, to assess the possibility of identifying distinctive features already in smaller blocks of conversation. We test $n \in \{128, 256, 512\}$, providing with middle representations (see Table 1). The number of instances is (resp.) 893, 443, and 209, with an average number of speech turns 11, 20, and 42. This configuration allows keeping some context without overwhelming the model.

Comparing representations: Existing work on schizophrenia language demonstrated the importance of lexical features. For French, as for many languages, we do not have access to a resource such as LIWC. We thus propose to simply include bag-of-words (bow) and n-grams ($n \in \{2, 3\}$) to our models as a proxy for topic identification.

Howes et al. (2012b) showed the importance of features specific to spontaneous dialogues that do involve lexicon but in a more generic way: OCR corresponds to *Open Class Repair* initiators (*pardon?*, *huh?*); *Backchannel* (BC) responses are phatic expressions (*yeah*, *hum mm*). To reflect text organization, we also include discourse features by extracting the forms (without disambiguation) corresponding to connectives (*but*, *because*, *since*) as identified in LexConn (Roze et al., 2012).

Finally, we test the two following non-lexical features: Part-Of-Speech n-grams and treelets. Allende-Cid et al. (2019) demonstrated that POS tags are effective features. We also test for larger patterns with sequences, POS n-gram with $n \in \{1, 2, 3\}$. Kayi et al. (2017) only used the dependencies as syntactic features. We extend to *treelet* features (Johannsen et al., 2015) based on the dependency parse trees: 2-*treelet* corresponds to 2 tokens with a syntactic relation between a head and a dependent, e.g., 'VERB→nsubj→NOUN', and 3-*treelet* corresponds to 3 tokens with 2 syntactic relations: could be 1 head dominates 2 dependents or a chain of dependencies, e.g., 'PRON←poss←NOUN←nsubj←VERB'. See Figure 1 for an illustration.

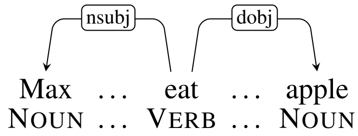

Figure 1: An example of syntactic relation represented as `treelet`.

4 Experimental setting

Data: Forty-one conversations between patients (18) or controls (23) and a psychologist come from SLAM project (Rebuschi et al., 2014; Amblard et al., 2015). The transcripts are standardized and follow a transcription guide. The groups are balanced with gender, age, intelligence quotient (IQ) score, years of studies, and three cognitive tests' results (WAIS-III, TMT, CVLT)[5]. They are free exchanges carried out in a medical setting where the psychologist is not personally involved - her main action is to maintain the exchange. Preliminary experiments showed that we could distinguish the two groups with relatively high accuracy with the clinician's data. We thus removed clinician's speech turns to reduce this impact and only focused on patients' factors. Further studies are needed to decide how to take into account the entire interaction.

Classification: We compare several classification algorithms: Support Vector Machines (SVM) (Cortes and Vapnik, 1995), Logistic Regression (LR), Random Forest (RF), Perceptron (Perc), and Naive Bayes (NB), without and with feature selection based on importance weight, all implemented in Scikit-Learn (Pedregosa et al., 2011). Hyperparameters are:

- Naive Bayes: smoothing $\alpha \in V = \{0.001, 0.005, 0.01, 0.1, 0.5, 1, 5, 10, 100\}$;

[5] WAIS-III: Wechsler Adult Intelligence Scale (WAIS) is an IQ test designed to measure intelligence and cognitive ability in adults and older adolescents. Trail Making Test (TMT) is a widely used test to assess executive abilities in patients. California Verbal Learning Test (CVLT) measures episodic verbal learning and memory.

Features	Full	Indiv.	W-512
bow	<u>**93.66**</u>	**72.43**	-
ngram	<u>85.61</u>	69.59	-
OCR	<u>60.62</u>	50.17	59.28
BC	<u>74.48</u>	54.79	67.86
Connectives	72.44	55.28	<u>73.57</u>
POS	53.66	55.80	<u>60.09</u>
2-POS	67.36	56.33	<u>71.74</u>
3-POS	71.65	56.53	<u>72.55</u>
2-treelet	69.19	56.73	<u>74.19</u>
3-treelet	66.78	55.34	<u>69.03</u>
1-2-3-POS	69.01	58.36	<u>72.67</u>
POS+2-3-treelet	66.59	57.77	<u>72.39</u>
3-POS+BC	74.93	57.46	<u>**77.86**</u>

Table 2: Best averaged accuracy for Full, Individual and W-512 (underlined: best setting for each feature).

- Logistic Regression: L_2 and regularization $C \in V$;
- SVM with linear kernel: L_2 and regularization $C \in V \cup \{1000\}$;
- Perceptron: L_2 and $\alpha \in V$;
- Random Forest: max_depth $\in \{2, None\}$;

Thresholds for feature selection are the range of 10 values equally distributed from $1e-5$ to the weight of the 50^{th} most important feature (thus allowing to keep at least 50 features), plus the mean and median of the weights.

Since our dataset is minimal, we use nested cross-validation to assess the performance of our system: tune hyper-parameters on $K-1$ folds and then evaluate on the left-out fold, repeating the whole process M times ($K = M = 5$). We report average accuracy over the M out folds. Best hyper-parameters values and algorithms are given in Appendix A.2.

5 Results

Lexical features: We compare different representations for Full and Indiv. settings - the most similar to long narrative texts or short Twitter messages. As in previous work, we found that lexical information is very effective (Table 2: *bow* and *n-gram*) with at best 93.66% in accuracy. However, analysis from previous studies suggested a potential issue: Mitchell et al. (2015) reported that health-related lexicon is more represented in the tweets dataset, and Howes et al. (2012b) that the most predictive unigrams are about conditions, treatment and, medication. We investigate our data using Spearman correlation[6] to rank lexical features and find similar results: terms linked to the condition are in top ranks for schizophrenia (*maladie [disease]*, *traitement [treatment]*, *médecin [doctor]*), while terms related to studies (*licence [bachelor]*, *thèse [PhD]*) and social life (*vacances [holidays]*, *monde [world / people]*) are correlated with controls. This finding is due to the nature of our data: patients talk about their disease with a clinician, and controls talk more about their everyday life. These features perform well because they reflect a lexical bias in data collection. However, the models will not be usable in the wild.

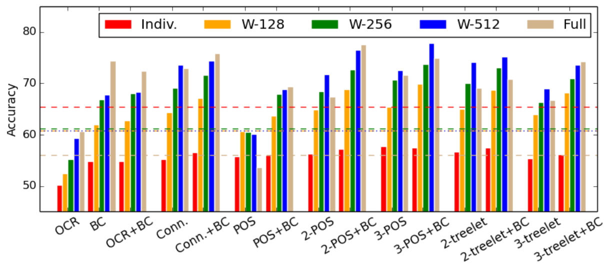

Figure 2: Accuracy for all features and window sizes. OCR: *Open Class Repair*, BC: *Backchannel response*, Conn.: *connectives*. W-n: window size.

Dialogue and discourse: Figure 2 presents results on selected subsets of non- or less- lexicalized features for the five splits of our data. Horizontal lines correspond to the majority vote baselines.

Concerning dialogue features, OCR gives poor results mostly behind the baseline, while BC is above with 74.48% (Full). Moreover, combining with BC to another feature set almost consistently allows improvements (not the case with OCR). These features are good indicators, contrary to what was reported in (Howes et al., 2012b). Note that we directly use the tokens as features rather than the proportion of BC per word, which allows more refined analysis. The most informative features for controls are phatic expressions (*ah*, *ok*, *hum-hum*, *vraiment [really]*, *c'est ça [that's right / yeah, right]*). At the same time, patients with schizophrenia are correlated with more ambiguous expressions which are also used in non-phatic contexts

[6] p-value < 0.05, coefficient $|\rho| > 0.3$

(*je comprends [I understand]*, *bien sûr [of course]*, *exactement [exactly]*), i.e., less BC responses: this supports that the patients are less prone to maintain the conversation.

Connectives also give promising results, at best 73.6%. Trend shows that controls use longer connectives (*jusqu'à ce que [until that]*, *au point de [to the point that]*) *vs.* patients (*donc [so]*, *puis [then]*). Connectives linked to the present moment are also highly correlated to schizophrenic group (*maintenant (que) [now (that)]*, *depuis que [ever since]*); this might refer to changes after treatment.

POS tags and syntax: Sequences of POS tags (2-POS and 3-POS) and of treelet (2-treelet and 3-treelet) are fully non-lexicalized features. They capture some internal structure of the interaction. We obtain our best scores with the longest sequences (3-POS, 72.55% acc., 74.34% F_1). These scores are higher than the ones reported by Kayi et al. (2017) on tweets (69.20% F_1) or essays (69.76% F_1) with simple POS tags and a lot more documents, and are very close to Allende-Cid et al. (2019) with meta-POS (75.1% in F_1): this confirms the predictive power of POS for the task.

We found that patients with schizophrenia used more verbs than controls (e.g., 2-POS such as VERB-ADP[7], 3-POS such as PRON-AUX-VERB), and, as in (Kayi et al., 2017), a higher proportion of adverbs. Precisely, we observe that the usage of adverbs of time (*parfois [sometimes], plus maintenant [not anymore], quasiment jamais [almost never]*), of place (*ici déjà [here already]*) and of frequency and manner (*beaucoup plus [much more], beaucoup mieux [much better]*) is higher than that of controls - this is possibly linked to the exchange about their (current) heath condition. On the other hand, controls employ a higher portion of linking adverbs (*enfin [finally], donc [so], quand même [anyway]*).

Syntactic features confirm these observations, the most predictive being verbal structures, followed by adverbial modifiers (*advmod*, *advcl*)[8]. This goes along with (Kayi et al., 2017), in which the top parse tag is *advmod*, and confirms clinician's descriptions on the use of less complex syntactic structures for patients with schizophrenia. Controls tend to use more complicated syntactic structures, such as those with SCONJ (subordinating conjunction) and CCONJ (coordinating conjunction), confirmed by our analysis of discourse connectives.

Context window size: Our experiments were also designed to test the impact of the context when dealing with dialogues. Figure 2 demonstrates that, in general, the larger the window, the better the scores. Individual speech turns are too small and contain no context. However, using the whole conversation most often leads to a drop in performance compared to our largest window (512 words) due to data sparsity, as we can observe for connective, n-POS and n-treelet. OCR and Backchannels do not follow this trend, meaning that they are probably less sparse.

These experiments demonstrate that using the block of conversation is relevant – the models find enough information to make accurate classification –, while allowing to increase the number of classification instances artificially.

Best algorithm: Among the 5 classifiers, NB generally performs well when dealing with word counts (in Full and Indiv.), while SVM and LR are generally better in other cases. More precisely, SVM performs better when the context window is relatively large, and the data sparsity is more pronounced (Full). At the same time, LR is better at dealing with small to medium-sized contexts (Indiv. and W-n settings). Detailed information is in the supplementary material.

6 Conclusion

We used conversations involving patients with schizophrenia in order to learn about language features associated with the disease. We compared various settings to represent dialogues and several representations to deal with data scarcity and lexical bias. Our experiments replicate performances as high as previous studies in English. Further experiments will be designed to take into account the entire interaction, probably with neural networks. We would also like to investigate the effect of adversarial loss in mitigating the bias within a neural model.

We hope that this paper will remind us of the importance of looking for bias in data and exploring higher-level, less language-dependent information to produce robust systems and draw more general conclusions on conversational data.

[7] ADP stands for adposition and it covers preposition and postposition.

[8] *advmod* is a (non-clausal) adverb or adverbial phrase; *advcl* is an adverbial clause modifier. They serve to modify a verb or other predicate.

Acknowledgements

The authors thank the anonymous reviewers for their insightful comments and suggestions. This work was supported by the PIA project "Lorraine Université d'Excellence", ANR-15-IDEX-04-LUE, as well as the CPER LCHN (Contrat de Plan État-Région - Langues, Connaissances et Humanités Numériques). We would like to thank the Centre Hospitalier Le Vinartier for having contributed decisively to the implementation of the experimentation. Experiments presented in this paper were carried out in secured nodes on the Grid'5000 testbed. We would like to thank the Grid'5000 community (see `https://www.grid5000.fr`).

References

Hector Allende-Cid, Juan Zamora, Pedro Alfaron-Faccio, and María Alonso. 2019. A machine learning approach for the automatic classification of schizophrenic discourse. *IEEE Access*, pages 45544–45554.

Maxime Amblard, Chloé Braud, Chuyuan Li, Caroline Demily, Nicolas Franck, and Michel Musiol. 2020. Investigation par méthodes d'apprentissage des spécificités langagières propres aux personnes avec schizophrénie (investigating learning methods applied to language specificity of persons with schizophrenia). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2: Traitement Automatique des Langues Naturelles*, pages 12–26.

Maxime Amblard, Karën Fort, Caroline Demily, Nicolas Franck, and Michel Musiol. 2015. Analyse lexicale outillée de la parole transcrite de patients schizophrènes. *Traitement Automatique des Langues*, 55(3):91 – 115.

American Psychiatric Association APA. 2015. *DSM-5-Manuel diagnostique et statistique des troubles mentaux*. Elsevier Masson.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162.

Michael L Birnbaum, Sindhu Kiranmai Ernala, Asra F Rizvi, Munmun De Choudhury, and John M Kane. 2017. A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. *J Med Internet Res*, 19(8):e289.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Peter F Brown, Vincent J Della Pietra, Peter V Desouza, Jennifer C Lai, and Robert L Mercer. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–480.

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

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Nancy M Docherty, Maddalena DeRosa, and Nancy C Andreasen. 1996. Communication disturbances in schizophrenia and mania. *Archives of General Psychiatry*, 53(4):358–364.

Brita Elvevåg, Peter W Foltz, Daniel R Weinberger, and Terry E Goldberg. 2007. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophrenia research*, 93(1-3):304–316.

Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. 2016. Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

William I Fraser, Kathleen M King, Philip Thomas, and Robert E Kendell. 1986. The diagnosis of schizophrenia by language analysis. *The British Journal of Psychiatry*, 148(3):275–278.

Adam Goodkind, Michelle Lee, Gary E Martin, Molly Losh, and Klinton Bicknell. 2018. Detecting language impairments in autism: A computational analysis of semi-structured conversations with vector semantics. *Proceedings of the Society for Computation in Linguistics*, 1(1):12–22.

Sharath Chandra Guntuku, Daniel Preotiuc-Pietro, Johannes C Eichstaedt, and Lyle H Ungar. 2019. What twitter profile and posted images reveal about depression and anxiety. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 236–246.

Ralph E Hoffman and William Sledge. 1988. An analysis of grammatical deviance occurring in spontaneous schizophrenic speech. *Journal of neurolinguistics*, 3(1):89–101.

Kai Hong, Christian G Kohler, Mary E March, Amber A Parker, and Ani Nenkova. 2012. Lexical differences in autobiographical narratives from schizophrenic patients and healthy controls. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 37–47.

Christine Howes, Matt Purver, Rose McCabe, Patrick GT Healey, and Mary Lavelle. 2012a. Helping the medicine go down: Repair and adherence in patient-clinician dialogues. In *Proceedings of SemDial 2012 (SeineDial): The 16th Workshop on the Semantics and Pragmatics of Dialogue*, page 155.

Christine Howes, Matthew Purver, and Rose McCabe. 2013. Using conversation topics for predicting therapy outcomes in schizophrenia. *Biomedical informatics insights*, 6:BII–S11661.

Christine Howes, Matthew Purver, and Rose McCabe. 2014. Linguistic indicators of severity and progress in online text-based therapy for depression. *ACL 2014*, page 7.

Christine Howes, Matthew Purver, Rose McCabe, Patrick Healey, and Mary Lavelle. 2012b. Predicting adherence to treatment for schizophrenia from dialogue transcripts. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 79–83.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of the nineteenth conference on computational natural language learning*, pages 103–112.

Efsun Sarioglu Kayi, Mona Diab, Luca Pauselli, Michael Compton, and Glen Coppersmith. 2017. Predictive linguistic features of schizophrenia. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 241–250.

Birgit Kleim, Andrea B Horn, Rainer Kraehenmann, Matthias R Mehl, and Anke Ehlers. 2018. Early linguistic markers of trauma-specific processing predict post-trauma adjustment. *Frontiers in Psychiatry*, 9:645.

Gina R Kuperberg. 2010. Language in schizophrenia part 1: an introduction. *Language and linguistics compass*, 4(8):576–589.

Brendan A Maher, Theo C Manschreck, Jakob Linnet, and Steven Candela. 2005. Quantitative assessment of the frequency of normal associations in the utterances of schizophrenia patients and healthy controls. *Schizophrenia Research*, 78(2-3):219–224.

Theo C Manschreck, Brendan A Maher, Jill E Rosenthal, and Jonathan Berner. 1991. Reduced primacy and related features in schizophrenia. *Schizophrenia Research*, 5(1):35–41.

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 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*, pages 11–20.

Rodney D Morice and John CL Ingram. 1982. Language analysis in schizophrenia: Diagnostic implications. *Australian and New Zealand Journal of Psychiatry*, 16(2):11–21.

Michel Musiol and Alain Trognon. 2000. *Eléments de psychopathologie cognitive: le discours schizophrène*. A. Colin.

Sylvester Olubolu Orimaye, Jojo Sze-Meng Wong, and Karen Jennifer Golden. 2014. Learning predictive linguistic features for alzheimer's disease and related dementias using verbal utterances. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 78–87.

Ted Pedersen. 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*, pages 46–53.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

James Pennebaker, Martha Francis, and Roger Booth. 2001. *Linguistic inquiry and word count (LIWC)*.

Stéphane Potvin, Ginette Aubin, and Emmanuel Stip. 2017. L'insight neurocognitif dans la schizophrénie. *L'Encéphale*, 43(1):15–20.

Randall Ratana, Hamid Sharifzadeh, Jamuna Krishnan, and Paul Pang. 2019. A comprehensive review of computational methods for automatic prediction of schizophrenia with insight into indigenous populations. *Frontiers in psychiatry*, 10:659.

Manuel Rebuschi, Maxime Amblard, and Michel Musiol. 2014. Using SDRT to analyze pathological conversations. Logicality, rationality and pragmatic deviances. In Manuel Rebuschi, Martine Batt, Gerhard Heinzmann, Franck Lihoreau, Michel Musiol, and Alain Trognon, editors, *Interdisciplinary Works in Logic, Epistemology, Psychology and Linguistics: Dialogue, Rationality, and Formalism*, volume 3 of *Logic, Argumentation & Reasoning*, pages 343 – 368. Springer.

Sherry Rochester. 2013. *Crazy talk: A study of the discourse of schizophrenic speakers*. Springer Science & Business Media.

Charlotte Roze, Laurence Danlos, and Philippe Muller. 2012. Lexconn: a french lexicon of discourse connectives. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (10).

Masahito Sakishita, Chihiro Ogawa, Kenji J Tsuchiya, Toshiki Iwabuchi, Taishiro Kishimoto, and Yoshinobu Kano. 2019. Autism spectrum disorder's severity prediction model using utterance features for automatic diagnosis support. In *International Workshop on Health Intelligence*, pages 83–95. Springer.

K Salzinger. 1979. Ecolinguistics: A radical behavior theory approach to language behavior.

Kurt Salzinger and Muriel Hammer. 1963. Some formal characteristics of schizophrenic speech as a measure of social deviance. *Annals of the New York Academy of Sciences*.

Kurt Salzinger, Stephanie Portnoy, and Richard S Feldman. 1964. Verbal behavior of schizophrenic and normal subjects. *Annals of the New York Academy of sciences*, 105(15):845–860.

Kurt Salzinger, Stephanie Portnoy, David B Pisoni, and Richard S Feldman. 1970. The immediacy hypothesis and response-produced stimuli in schizophrenic speech. *Journal of Abnormal Psychology*, 76(2):258.

Ivan Sekulić and Michael Strube. 2019. Adapting deep learning methods for mental health prediction on social media. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 322–327.

Shihao Xu, Zixu Yang, Debsubhra Chakraborty, Yasir Tahir, Tomasz Maszczyk, Yi Han Victoria Chua, Justin Dauwels, Daniel Thalmann, Nadia Magnenat Thalmann, Bhing-Leet Tan, et al. 2019. Automated lexical analysis of interviews with individuals with schizophrenia. In *Proceedings of the 9th International Workshop on Spoken Dialogue System Technology*, pages 185–197. Springer.

A Appendices

A.1 OCR and backchannel word list

In order to improve reproducibility, we give the full list of tokens used for OCR (Table 3) and backchannel (Table 5), as well as their corresponding translation in English (Table 4, Table 6). They were obtained by translating the list given by the authors of we contacted, and by adding a few additional terms specific to French.

pardon vous disiez	pardon	ah vous parler pardon
excusez-moi	excuse moi	bon je suis désolée
désolé(e)	(ah) ouais ?	ah bon ?
c'est vrai ?	c'est euh ?	hum ?
de quoi	c'est quoi ?	c'est-à-dire
euh ?	dites moi plus	mais encore

Table 3: Open class repair initiators list (French).

pardon you said	pardon	ah pardon you were saying
excuse-me	excuse me	i am sorry
sorry	(ah) yes?	ah really?
is it true?	it's euh?	huh?
of what	what is it?	which means
euh?	tell me more	but still

Table 4: Open class repair initiators list (English translation).

oui	ouais	ouais voilà
oui c'est ça	oui bah oui	oui... forcément
bah ouais	hum (hum)	muh mmh
mmh/mmhh	d'accord	ok
voilà	c'est ça	c'est vrai
c'est sûr	ça c'est clair	eh bien sûr
carrément	bien sûr	super
ok... bon	d'accord ça marche	certes
mais hein	je comprends	vraiment
bien	bon	très bien
quand même	tout à fait	certainement
exactement	tant mieux	oh
ah	ben	alors ben
ah d'accord	ah ça euh	eh bah c'est bien

Table 5: backchannel response list (French).

yes	yeah	yeah that's it
yes that's it	yes euh yes	yes... for sure
euh yeah	hum (hum)	muh mmh
mmh/mmhh	okay	ok
that's it	that's it	that's true
(yes) (for) sure	that's clear/clearly/definitely	eh of course
completely	of course	super
ok... then	all right	indeed/yes
but hein	i understand	really
good	well	very good
still	exactly	certainly/sure
exactly	all the better/so much the better	oh
ah	well	so... well
ah okay	ah (this) euh	eh well that's good

Table 6: backchannel response list (English translation).

A.2 Best scores and corresp. settings

	Full				Indiv.				W-512			
Features	Acc.	Algo	Hyper-params.	Thres.	Acc.	Algo	Hyper-params	Thres.	Acc.	Algo	Hyper-params	Thres.
bow	93.66	NB	$\alpha = 0.001$	9	72.43	NB	$\alpha = 0.1$	$1e-5$	-	-	-	-
ngram	85.61	SVM	$C = 5$	4	69.59	SVM	$C = 5$	2	-	-	-	-
OCR	60.62	LR	$C = 0.001$	$1e-5$	50.17	PERC	$\alpha = 0.001$	mean	59.28	NB	$\alpha = 0.001$	2
BC	74.48	SVM	$C = 0.1$	9	54.79	RF	$max\_depth = 2$	$1e-5$	67.86	RF	$max\_depth = None$	1
Connectives	72.44	RF	$max\_depth = 2$	2	55.28	LR	$C = 100$	5	73.57	RF	$max\_depth = 2$	median
POS	53.66	NB	$\alpha = 0.001$	5	55.80	LR	$C = 1$	$1e-5$	60.09	SVM	$C = 100$	$1e-5$
2-POS	67.36	SVM	$C = 100$	3	56.33	LR	$C = 1$	$1e-5$	71.74	LR	$C = 100$	mean
3-POS	71.65	SVM	$C = 100$	2	56.53	SVM	$C = 0.5$	$1e-5$	72.55	SVM	$C = 100$	1
2-treelet	69.19	RF	$max\_depth = 2$	$1e-5$	56.73	LR	$C = 5$	$1e-5$	74.19	LR	$C = 100$	6
3-treelet	66.78	SVM	$C = 100$	8	55.34	LR	-	$1e-5$	69.03	LR	$C = 100$	6
1-2-3-POS	69.01	SVM	$C = 1000$	1	58.36	LR	$C = 100$	$1e-5$	72.67	SVM	$C = 100$	7
POS+2-3-treelet	66.59	SVM	$C = 1000$	4	57.77	SVM	$C = 0.5$	$1e-5$	72.39	LR	$C = 100$	4
3-POS + BC	74.93	SVM	$C = 100$	8	57.46	LR	$C = 5$	$1e-5$	77.86	SVM	$C = 100$	$1e-5$

Table 7: Best scores (averaged accuracy Acc.), best algorithms (Algo), corresponding hyper-parameters (Hyper-params.) and threshold (Thres.) for full documents (Full), individual speech turns (Indiv.) and Window size of 512 tokens (W-512).