ARTICLE

Exploring the Use of Natural Language Processing for Objective Assessment of Disorganized Speech in Schizophrenia

Lydia Jeong, MI, Melissa Lee, MD, Ben Eyre, BSc, Aparna Balagopalan, MSc, Frank Rudzicz, PhD, Cedric Gabilondo, MD

Objective: Measurement-based care tools in psychiatry are useful for symptom monitoring and detecting response to treatment, but methods for quick and objective measurement are lacking especially for acute psychosis. The aim of this study was to explore potential language markers, detected by natural language processing (NLP) methods, as a means to objectively measure the severity of psychotic symptoms of schizophrenia in an acute clinical setting.

Methods: Twenty-two speech samples were collected from seven participants who were hospitalized for schizophrenia, and their symptoms were evaluated over time with SAPS/SANS and TLC scales. Linguistic features were extracted from the speech data using machine learning techniques. Spearman's correlation was performed to examine the relationship between linguistic features and symptoms. Various machine learning models were evaluated by cross-validation methods for their ability to predict symptom severity using the linguistic markers.

Results: Reduced lexical richness and syntactic complexity were characteristic of negative symptoms, while lower content density and more repetitions in speech were predictors of positive symptoms. Machine learning models predicted severity of alogia, illogicality, poverty of speech, social inattentiveness, and TLC scores with up to 82% accuracy. Additionally, speech incoherence was quantifiable through language markers derived from NLP methods.

Conclusions: These preliminary findings suggest that NLP may be useful in identifying clinically relevant language markers of schizophrenia, which can enhance objectivity in symptom monitoring during hospitalization. Further work is needed to replicate these findings in a larger data set and explore methods for feasible implementation in practice.

Psych Res Clin Pract. 2023; 5:84–92; doi: 10.1176/appi.prcp.20230003

Disorganized speech is a key component in the evaluation of psychosis in schizophrenia. Tools for objective measurement remain largely unavailable as current methods of assessment rely on subjective, qualitative clinical examination. However, emerging technologies in artificial intelligence are becoming increasingly capable of performing tasks that require high-level processing. These advances have the potential to equip psychiatrists with clinical tools to capture objective markers of mental health and improve upon methods for psychiatric assessment.

Natural language processing (NLP) is a branch of artificial intelligence concerned with using computers to interpret conversational human language. NLP can be used to efficiently and inexpensively process large amounts of language data that would otherwise be too time-consuming or impractical to perform manually. Speech-to-text transcripts can be systematically dissected and rated on various language metrics, including coherence and word use patterns. We propose these computational metrics have the potential to provide objective, quantitative markers for patients with disordered speech, such as those with schizophrenia.

HIGHLIGHTS

- Natural language processing and machine learning techniques were applied to speech transcripts from patients with schizophrenia who were admitted to hospital for psychosis.
- Certain linguistic features computed by natural language processing were found to correspond to the presence and severity of specific positive and negative symptoms.
- Natural language processing may offer an objective way to measure disorganized speech and symptom severity in schizophrenia.

Several works have aimed at characterizing language disturbances in schizophrenia with evidence that language markers can discriminate psychosis from normal speech (1–5). Compared to healthy controls, people with schizophrenia more often use word approximation (e.g., 'reflector' for 'mirror' (6)), invent novel words (i.e., neologisms (7)), generate ambiguous references and pronouns (4, 8), show more repetitions in speech (9–12), utilize less connectivity between sentences (13), and display greater use of first-person singular pronouns (5, 11, 12, 14, 15). They also tend to use simpler and shorter phrases (16) and make more syntactic errors (17). In terms of clinical relevance, language markers can predict the onset of psychosis in prodromal youth with superior accuracy compared to clinical ratings (2, 3, 18).

Other works have developed computational models that capture increased levels of incoherence and tangentiality in the speech of people with schizophrenia (1–5). More recent NLP techniques include BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (Robustly Optimized BERT Pretraining Approach) (19, 20). BERT is a machine learning technique for NLP that can be used to perform a wide variety of language-based tasks, such as text prediction, question answering, and text summarization. BERT employs bidirectional context, which means that it learns information from both left to right and right to left in a text. The bi-directionality helps the model to understand the context and generate representations. Because the bidirectionality allows for greater context, they tend to perform better than traditional NLP methods. BERT was developed by training on large amounts of texts including Wikipedia and various books, and applying two methods: 1) Masked Language Model to predict randomly masked words, and 2) Next-sentence Prediction to predict the likelihood of the second sentence following the previous sentence. Applying these techniques allows the model to capture the meaning of the text accurately. Given its relative novelty, the use of BERT in psycholinguistic studies is limited to date. Besides measuring coherence (5), BERT can also be applied to measure surprisal (21), which refers to the unexpectedness of a statement given its context. Because individuals with schizophrenia tend to use language in unexpected ways and produce greater language abnormalities compared to healthy individuals, our study takes advantage of BERT's ability to quantify the level of "surprise" in the speech of schizophrenia as a potential language marker.

In this exploratory study, we apply NLP methods to investigate language markers in schizophrenia and their potential role in bringing objectivity to the measurement of psychotic symptoms in an acute clinical setting.

METHODS

Participants

Participants included seven individuals who were admitted to an adult psychiatric inpatient unit during the first wave of enrolment from 2019 to 2021. They were defined by having a documented current clinical diagnosis of schizophrenia with psychosis as the primary reason for admission. Participation excluded non-English primary speaking language, psychosis attributable to a medical or psychiatric cause other than schizophrenia (e.g., affective psychosis, substance induced psychosis), or inability to participate in verbal interviews for any reason including disability or safety considerations. Comorbid psychiatric diagnoses were present in four participants (previous opioid use disorder, cannabis use disorder, cocaine use disorder, and query seizure disorder). No diagnoses of schizophrenia were revised upon discharge and no participants dropped out of the study.

Participant demographic characteristics are shown in Table 1. The research ethics review boards at Michael Garron Hospital (784–1901–Mis–333) and the University of Toronto (#00038134) approved the study. All participants provided written informed consent after receiving a complete description of the study.

Enrolment and data collection were paused for extended periods of time in accordance with the review boards' response to the emergence of the COVID–19 pandemic. In the interest of preliminary data analysis, the study sample was limited to the first wave of enrolment. We plan to update this study with data from future enrollment cohorts.

Speech Assay

All subjects participated in a standardized 8-min interview during their first week of admission and every week thereafter until discharge from hospital. Using instructions

TABLE 1. Demographic characteristics of participants.

	Participants (*N* = 7)	
Demographic characteristic	Mean	SD
Age (years)[a]	37.71	12.61
Education level (years)[b]	16.25	0.50
Length of hospitalization (days)[c]	29	13.50
	N	%
Female	4	57
Treated with antipsychotics	7	100
Treated with antidepressants/mood stabilizers	1	14
Treated with benzodiazepines	3	43
Treated with ECT	1	14

[a] Range = 27–58.
[b] Unknown = 3.
[c] Range = 7–48.

adapted from the "free verbalization" interview method described by Reilly et al. (22), participants were asked to spontaneously talk about two open-ended interview prompts for 4 min each: 1) any non-mental health topic of their choosing, and 2) events leading up to their hospitalization. The order of prompts was switched for every other participant. Any personally identifiable information was redacted to protect the participant's privacy. Participants took part in one to five interviews each, depending on their length of hospitalization. Subject responses were audio recorded and transcribed by a third-party medical transcription service, producing a total of 22 speech-to-text samples.

Speech Analysis

Speech pre-processing. Speech samples were pre-processed and prepared for computer-based analyses, which involved eliminating noise and metadata from transcripts (e.g., special characters, timestamps), converting all characters into lowercase, removing stop words, and lemmatizing words (in other words, replacing a word to its base form, e.g., converting *"am"*, *"are"*, *"is"* to *"be"*; the purpose of lemmatization is to group words with the same root as one item).

Linguistic Features Extracted Using COVFEFE. Linguistic features were extracted at the lexical, syntactic, semantic, and pragmatic levels using the open source tool, COVFEFE (COre Variable Feature Extraction Feature Extractor). A total of 393 features were obtained (23).

Lexical analysis examines text at the level of individual words, and can measure aspects such as vocabulary richness and lexical norms (e.g., word length, frequency of word use) (24, 25). Syntactic analysis is concerned with grammar and sentence structure in a text (e.g., noun-to-verb ratio, mean length-of-sentence) (26). Semantic analysis is concerned with the literal meaning conveyed by text based on the relatedness of words, phrases, and sentences. Pragmatic analysis focuses on the implicit meaning of text in relation to its context and is a measure of language abstraction (27). For complete descriptions of all features and algorithms, please refer to Komeilli et al. (23).

Coherence-Related Features Extracted Using BERT. The term "incoherent speech" refers broadly to low semantic similarity and disorganized speech associated with formal thought disorder (1, 4). A common approach to measuring incoherence involves using sentence embeddings, which refers to NLP techniques that map sentences onto numerical vectors in space based on their semantic meaning. Semantic dissimilarity between sentences (incoherence), is derived from the distance between vectors. The embedding technique BERT was used to extract two features: next-sentence probability and surprisal.

Next-sentence probability is defined as the probability of a given sentence using the previous sentence as context (5). The average of these probabilities was computed to measure the overall coherence of the transcript. Figure 1 illustrates the process of obtaining the coherence feature.

Surprisal refers to the level of unexpectedness of a statement given its context. To calculate surprisal, we use the technique described by Li et al (21). This technique involves fitting a normal distribution to the embeddings (representation of a language) used by a BERT model. Surprisal is quantified as the likelihood of an utterance occurring according to this distribution. Utterances with higher surprisal are those with a lower likelihood of occurring. Supplementary Figures S2-S6 shows heatmaps of mean, sum, variance, skweness, and kurtosis of surprisal for each transcript samples.

Correlation of Text Features with Clinical Ratings

At each interview cross-section, patients were concurrently evaluated with the Scale for the Assessment of Positive and Negative Symptoms (SAPS/SANS) (28, 29) and the global score of the Scale for the Assessment of Thought, Language, and Communication (TLC) (30) by attending psychiatrist, author C. G. These clinical rating scales were used to gather observational data on various aspects of speech disorganization, such as tangentiality, incoherence, and illogicality. In addition to language-related subscales, global scales were included, which consisted of global rating of hallucinations, delusions, bizarre behavior, affective flattening, and avolition/apathy.

We tested whether the extracted linguistic features were associated with the clinical ratings of SAPS, SANS, and TLC scores using Spearman's correlation ($\alpha < 0.05$). Bonferroni correction was also applied due to the large number of comparisons and Lilliefors test was performed to check whether the symptom ratings displayed a normal distribution. The features with the highest F-values were selected for further analysis using the scikit-learn SelectKBest method (31).

Machine Learning Prediction of Language Symptom Severity

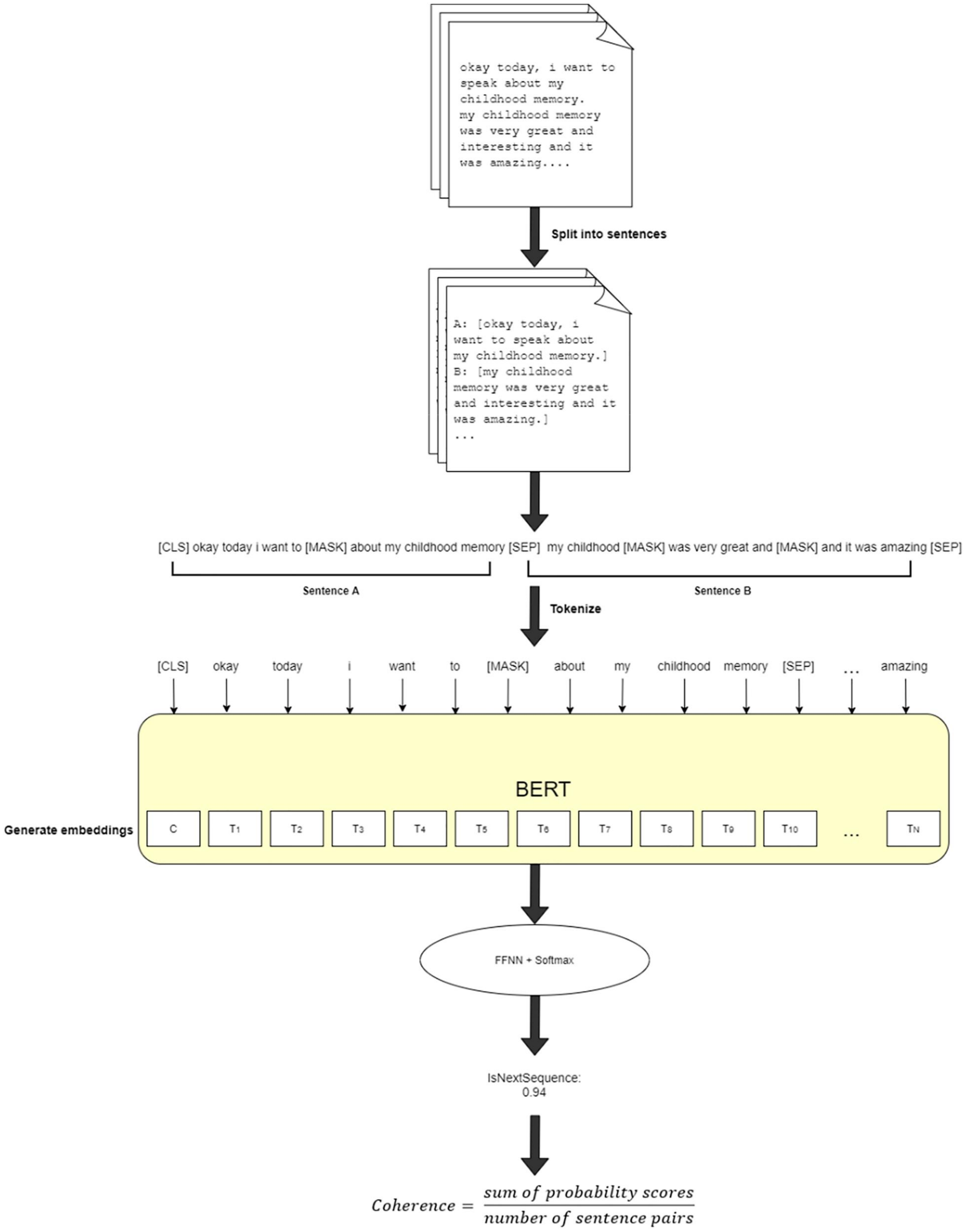
We predicted the severity of each of the included clinical rating scales using various machine learning models. We used classification algorithms with leave-one-subject-out cross validation method. For classification models, we reported F-measure and area under the curve (AUC) as evaluation metrics (32).

RESULTS

Significant Linguistic Features

Among the 393 linguistic features extracted using COVFEFE, 224 resulted in statistically significant correlations at the threshold level of $p \leq 0.05$, with 24 features surviving Bonferroni correction (df = 19 for all correlation

FIGURE 1. Pipeline for extracting sentence coherence using BERT's next-sentence prediction method. Transcripts are split into sentences, which are paired and tokenized. Each pair begins with a [CLS] token to mark the start of the sentence, and each sentence ends with a [SEP] token to mark the end of the sentence. BERT token embeddings are generated, and feed forward neural network (FFNN) and SoftMax classifier produce probability values. The coherence of a transcript is measured as the mean probability of the sequence pairs in a transcript. Scores closer to 1 indicate more likelihood of sentence B being next to A, whereas scores closer to 0 signify lower probability.



```
okay today, i want to
speak about my
childhood memory.
my childhood memory
was very great and
interesting and it
was amazing....
```

```
A: [okay today, i
want to speak about
my childhood memory.]
B: [my childhood
memory was very great
and interesting and it
was amazing.]
...
```

[CLS] okay today i want to [MASK] about my childhood memory [SEP] my childhood [MASK] was very great and [MASK] and it was amazing [SEP]

$$Coherence = \frac{sum\ of\ probability\ scores}{number\ of\ sentence\ pairs}$$

tests). Select features that had high predictive power and were relevant to previous literature are presented in Table 2. All other features that passed the Bonferroni correction are listed in Table S1 in the supplemental data. The strong correlations between certain symptoms and linguistic features are shown in Supplementary Figure S1.

In terms of BERT features, the transcripts had overall next-sentence coherence ranging from 0.89 to 1.00. Coherence scores were inversely correlated to severity of derailment ($r = -0.46$, $p = 0.033$), illogicality ($r = -0.50$, $p = 0.018$), and circumstantiality ($r = -0.44$, $p = 0.041$). With regards to surprisal, greater surprisal (more negative values) was correlated with increased pressure of speech ($r = -0.65$, $p = 0.001$), circumstantiality ($r = -0.58$, $p = 0.0043$), illogicality ($r = -0.53$, $p = 0.01$), and tangentiality ($r = -0.47$, $p = 0.026$). Lower surprisal (more positive values) was correlated with poverty of speech ($r = 0.51$, $p = 0.015$).

Machine Learning Models

Machine learning classification models were trained on various combinations of the extracted linguistic features and evaluated on their ability to predict symptom severity from SAPS/SANS and global TLC scores. The top five with the best accuracy are presented in Table 3. The best performing model was the linear Support Vector Machine, which classified the global rating of alogia (comprising poverty of speech and poverty of content as core elements) with AUC of 1.0 and F-score of 82% with 29 variables.

DISCUSSION

In this study, we explored the use of automated language analysis to identify objective markers for the positive and negative language symptoms of schizophrenia. Key findings from our analysis are discussed below.

Reduced Lexical Richness and Syntactic Complexity as Markers of Negative Language Symptoms

Participants with more severe symptoms of poverty of speech, poverty of content, and social inattentiveness generally displayed reduced lexical richness and syntactic complexity, as demonstrated by substantially lower Honoré's statistics, shorter word lengths, increased use of high-frequency words, shorter mean sentence and clause lengths, and fewer occurrences of coordination and prepositions. These findings are in accordance with previous studies that suggest an important relationship between symptoms of schizophrenia and linguistic complexity (11, 13, 16), and our work shows this relationship can be measured objectively.

TABLE 2. Correlations of SANS and SAPS symptoms with linguistic features ($p < 0.05$).

Symptoms			
SANS	Features	Correlations	*p*
Poverty of speech	Honoré statistics	−0.74	0.0001[a]
	Mean length of clause	−0.71	0.0002
	Mean length of sentence	−0.44	0.040
	Coordinate	−0.68	0.0004
	Verb phrases	0.62	0.0023
	Verbs	0.44	0.042
	Prepositions	−0.43	0.044
	BERT surprisal mean	0.51	0.015
Poverty of content of speech	Presence of proper nouns in noun phrase	−0.72	0.00017
	Clauses	0.51	0.014
Social inattentiveness	Word length	−0.63	0.0016
	Frequency of words occurring in the corpus	0.67	0.00067
SAPS	Features	Correlations	*p*
Circumstantiality	Presence of foreign words in noun phrase	−0.73	0.00012[a]
Derailment	Sentences	0.73	0.00013[a]
	Content density	−0.52	0.012
	Type-token ratio	−0.63	0.0018
	BERT next sentence probability	−0.46	0.0033
Illogicality	Pronouns	0.73	0.00011[a]
	Presence of personal pronouns in noun phrase	0.73	0.00012[a]
	First person singular pronouns	−0.61	0.0025
Incoherence	Presence of verb phrase as a standalone sentence	0.51	0.016
Tangentiality	Function words	0.49	0.020
Pressure of speech	T-units per sentence	0.76	0.00004[a]
	Words	0.64	0.0013
	BERT surprisal sum	−0.65	0.001
	Mean length of sentence	0.58	0.0042
	Type-token ratio	−0.51	0.014

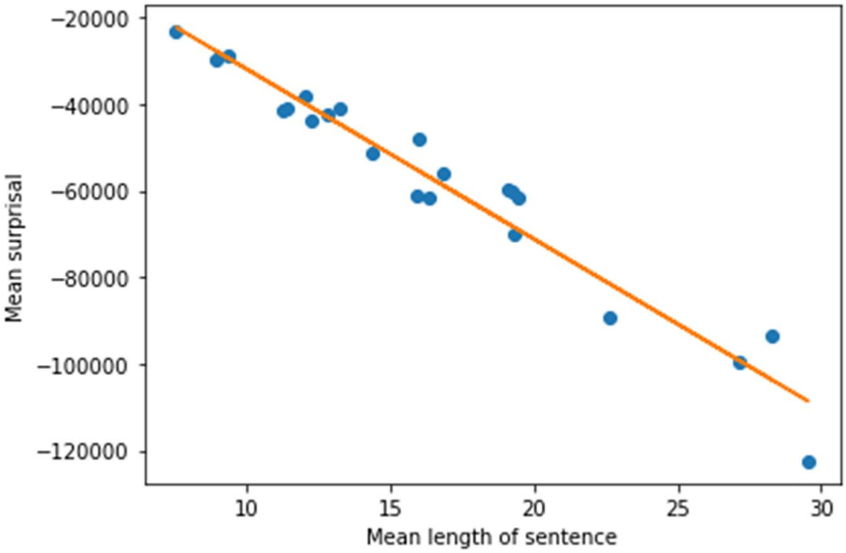
[a] Indicates significance after Bonferroni correction.

TABLE 3. Classification performance metrics on the best classifier and predictor sets.

Symptom	Best classifier	AUC[a]	F-score	Number of features
Global rating of alogia	Linear SVM	1	0.82	29
Illogicality	Decision tree	1	0.68	2
Poverty of speech	Neural net	1	0.65	60
Global rating of hallucinations	LDA	0.995	0.64	49
Global rating of avolition/apathy	Gaussian process	1	0.62	32
Social inattentiveness	Nearest neighbors	0.86	0.60	41
TLC	Naive Bayes	0.93	0.58	3
Global rating of bizarre behavior	Linear SVM	0.93	0.55	26
Global rating of affective flattening	LDA	0.995	0.54	42
Global rating of delusions	Gaussian process	1	0.53	46

[a] AUC was calculated using the train data.

FIGURE 2. Relationship between surprisal and the mean length of sentence. Plot shows strong correlation between surprisal and the mean length of sentence. Greater negative value indicates higher surprisal.



Lower Content and More Repetitions Associated with Derailment and Pressure of Speech

Participants who scored higher in derailment and pressure of speech objectively used considerably more words, clauses, and sentences, as well as longer sentences, but lower type-token ratio and content density. Type-token ratio is the number of unique words divided by the total number of words. These results support that decreased type-token ratio and content density are indicators of diminished lexical diversity and deficit of meaningful information in discourse. Additionally, lower type-token ratio despite longer utterances may be an indicator of repetitive speech. These results offer possible objective metrics for capturing aspects of disorganized speech.

Lower Next-Sentence Probability as a Marker for Incoherence, Surprisal Linked to Longer Sentences

Lower BERT next-sentence probability scores corresponded to severity of clinically observed derailment, illogicality, and circumstantiality. BERT surprisal was more prevalent among transcripts with elevated pressure of speech and attenuated poverty of speech. Our findings suggest that samples with greater speech production were more "surprising" than samples with lower amount of speech. Figure 2 demonstrates the strong correlation between surprisal and mean length of sentence. Supplementary Figure S7 shows a heatmap of surprisal for two transcript samples at a sentence level, with the longest sentence indicating greater surprisal.

Machine Learning Models

Our work on machine learning classification revealed that global rating of alogia, illogicality, poverty of speech, social inattentiveness, and global TLC score can be predicted with up to 82% accuracy (0.82 F1 score). Our model distinguishes the severity of alogia with the highest accuracy. This value is comparable to results from other NLP studies which have yielded accuracies of 69–75% (33), 78.4–87.5% (1), and 93% (4) and outperformed human raters who performed at 65.4–71.9% (1). Tang et al. (5) explored BERT's sentence-level coherence in schizophrenia spectrum disorders, demonstrating greater tangentiality among patients than controls with 87% accuracy. The particularly intriguing aspect of this work was that NLP methods revealed superior ability to detect subclinical differences (sentence coherence using BERT scores, log-odds ratio, parts-of-speech counts, and incomplete word counts) between patients and controls compared to purely human clinical ratings. Our preliminary results demonstrate that NLP can be used to predict symptom severity from speech records. This may have applications in enhancing the objectivity of psychiatric symptom assessment and mental status examination, particularly in situations where speech data may be the only information available at the time, such as, telephone assessments, voicemail, forensic evidence, etc.

Strengths and Weaknesses of the Study

This exploratory study has a few notable strengths. First, it takes into account the complex heterogeneity of schizophrenia, where the presence of symptoms varies between individuals, and symptoms may change over time. Previous studies have almost exclusively focused on dichotomous classification between patients and controls, which overlooks the diversity of symptom profiles, whereas our study investigates the severity of individual symptoms for each participant. Second, the data were collected from a real-world clinical setting and psychiatric comorbidities in addition to schizophrenia were permissible for eligibility. While this study focuses on psychosis related to schizophrenia, we eschewed overly rigid exclusion criteria to reflect the fact that patients often transcend diagnostic categories and accurate monitoring of psychotic symptom burden is no less important in the presence of comorbidities. Finally, the design of the free verbalization interview helped to minimize repetition and practice effects within subjects, and allows our method to be applied to naturalistic spontaneous speech without the need to administer a separate questionnaire or structured interview.

On the other hand, there are limitations that warrant caution in the interpretation of this exploratory study. The sample size is small compared to generic machine learning datasets; prospective research with larger groups may improve sensitivity in detecting differences in individual dimensions. The biggest unexpected challenge to recruitment was the COVID–19 pandemic, which introduced exposure risk inherent to conducting interviews for data collection. This will likely need to be considered as a potential risk in the design of future studies in this area. Furthermore, sampling bias may have also been a factor given that participants were required to be capable of consent to participate in research and willingly have their speech recorded for analysis, which may have resulted in those with more severe disorganization, or paranoia about recording devices, less likely to participate. This could account for the relatively high next-sentence coherence scores among the sample (0.89–1.00), suggestive of less severe disorganization.

Our study also relies on the assumption that the attending psychiatrist's clinical ratings reflect the patient's true mental state and symptomatology. In other words, the "gold standard" against which the language markers were compared is prone to the very same issues of subjectivity and reliability we sought to improve upon in our search for objective markers. If resources permitted, enlisting multiple raters with measured agreement would have strengthened the validity of clinical ratings as the "gold standard" comparison. Although this remains a limitation, we believe the assumption is generally justified given the psychiatrist is clinically experienced and completed the scales based on all the available clinical information up to that point in time during the subject's hospitalization. We also reduced bias from subjectivity by using scales which have been psychometrically validated for inter-rater reliability.

In terms of other limitations, our method used third-party manual transcription services to translate speech recordings into text for analysis. In practice, this would pose a cost prohibitive barrier and administrative burden for implementation in routine assessments. However, with advances in speech-to-text software, this step could be automated for real-time, point-of-care language testing and scoring. Direct analysis of the speech audio itself could also be valuable for non-verbal speech cues, such as pitch, tone, vocal inflection, and pause duration. These are often important aspects of mental status examination, and several studies have shown that non-verbal speech measures can be applied to accurately quantify negative symptoms (34–36). Therefore, combining acoustic and linguistic components might allow for better predictive power in our empirical tests.

CONCLUSION

The goal of this small exploratory study was to find linguistic markers of psychosis that can be used to objectively measure disorganized speech and symptom severity in schizophrenia. We found reduced lexical richness and syntactic complexity as characteristic of negative language symptoms (poverty of speech, poverty of content, social inattentiveness), while lower content density and more repetitions in speech as predictors of positive language symptoms (derailment, pressured speech). We also demonstrated methods for objectively measuring incoherent speech using state-of-the-art neural network models (i.e., BERT). These preliminary findings highlight the potential advantages of applying computational NLP methods as a clinical assessment tool, thus creating a framework for objective measurement-based care in schizophrenia.

AUTHOR AND ARTICLE INFORMATION

Faculty of Information, University of Toronto, Toronto, Ontario (Jeong); Department of Psychiatry, University of Toronto, Toronto, Ontario (Lee, Gabilondo); Department of Computer Science, University of Toronto, Toronto, Ontario (Eyre, Rudzicz); Vector Institute for Artificial Intelligence, Toronto, Ontario (Eyre, Rudzicz); Massachusetts Institute of Technology, Cambridge, Massachusetts (Balagopalan); Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia (Rudzicz).

Send correspondence to Dr. Lee (lissa.lee@mail.utoronto.ca).

This research study was funded by the 2019 award of the Michael Garron Hospital Educational Scholarship and Research Grant.

There are no financial relationships with commercial interests.

Received January 13, 2023; accepted April 16, 2023.

REFERENCES

1. Elvevåg B, Foltz PW, Weinberger DR, Goldberg TE. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. Schizophr Res. 2007;93(1-3):304–16. https://doi.org/10.1016/j.schres.2007.03.001
2. Bedi G, Carrillo F, Cecchi GA, Slezak DF, Sigman M, Mota NB, et al. Automated analysis of free speech predicts psychosis onset in high-risk youths. NPJ Schizophr. 2015;1(1):15030. https://doi.org/10.1038/npjschz.2015.30
3. Corcoran CM, Carrillo F, Fernández-Slezak D, Bedi G, Klim C, Javitt DC, et al. Prediction of psychosis across protocols and risk cohorts using automated language analysis. World Psychiatr. 2018;17(1):67–75. https://doi.org/10.1002/wps.20491
4. Iter D, Yoon J, Jurafsky D. Automatic detection of incoherent speech for diagnosing schizophrenia. In: *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic; Jun 2018; New Orleans, LA*. p. 136–46. https://doi.org/10.18653/v1/W18-0615
5. Tang SX, Kriz R, Cho S, Park SJ, Harowitz J, Gur RE, et al. Natural language processing methods are sensitive to subclinical linguistic differences in schizophrenia spectrum disorders. NPJ Schizophr. 2021;7(1):25. https://doi.org/10.1038/s41537-021-00154-3
6. Covington MA, He C, Brown C, Naçi L, McClain JT, Fjordbak BS, et al. Schizophrenia and the structure of language: the linguist's view. Schizophr Res. 2005;77(1):85–98. https://doi.org/10.1016/j.schres.2005.01.016
7. Solovay MR, Shenton ME, Holzman PS. Comparative studies of thought disorders. I. Mania and schizophrenia. Arch Gen Psychiatry. 1987;44(1):13–20. https://doi.org/10.1001/archpsyc.1987.01800130015003
8. Docherty NM, DeRosa M, Andreasen NC. Communication disturbances in schizophrenia and mania. Arch Gen Psychiatry. 1996;53(4):358–64. https://doi.org/10.1001/archpsyc.1996.01830040094014
9. Maher B. The language of schizophrenia: a review and interpretation. Br J Psychiatry. 1972;120(554):3–17. https://doi.org/10.1192/bjp.120.554.3
10. Manschreck TC, Maher BA, Hoover TM, Ames D. Repetition in schizophrenic speech. Lang Speech. 1985;28(3):255–68. https://doi.org/10.1177/002383098502800303
11. Hong K, Kohler CG, March ME, et al. 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; Jul 2012; Jeju Island, Korea*. p. 37–47.
12. Hong K, Nenkova A, March ME, Parker AP, Verma R, Kohler CG. Lexical use in emotional autobiographical narratives of persons with schizophrenia and healthy controls. Psychiatry Res. 2015;225(1-2):40–9. https://doi.org/10.1016/j.psychres.2014.10.002
13. DeLisi LE. Speech disorder in schizophrenia: review of the literature and exploration of its relation to the uniquely human capacity for language. Schizophr Bull. 2001;27(3):481–96. https://doi.org/10.1093/oxfordjournals.schbul.a006889
14. Maatz A. Use of the first-person pronoun in schizophrenia. Br J Psychiatry. 2014;205(5):409. https://doi.org/10.1192/bjp.205.5.409
15. Kayi ES, Diab M, Pauselli L, et al. Predictive linguistic features of schizophrenia. In: Proceedings of the 6th joint conference on lexical and computational semantics (*SEM 2017), Vancouver, Canada; Aug 2017. p. 241–50.
16. Özcan A, Kuruoglu G, Alptekin K, et al. The production of simple sentence structures in schizophrenia. Int J Arts Sci. 2017;9:159–64.
17. Fraser WI, King KM, Thomas P, Kendell RE. The diagnosis of schizophrenia by language analysis. Br J Psychiatry. 1986;148(3):275–8. https://doi.org/10.1192/bjp.148.3.275
18. Bearden CE, Wu KN, Caplan R, Cannon TD. Thought disorder and communication deviance as predictors of outcome in youth at clinical high risk for psychosis. J Am Acad Child Adolesc Psychiatry. 2011;50(7):669–80. https://doi.org/10.1016/j.jaac.2011.03.021
19. Devlin J, Chang M, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. The North American Chapter of the Association for Computational Linguistics; 2019.
20. Liu Y, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach. ArXiv 2019;abs/1907. https://doi.org/10.48550/arxiv.1907.11692
21. Li B, Zhu Z, Thomas G, et al. How is BERT surprised? Layerwise detection of linguistic anomalies. In: Proceedings of the 59th annual meeting of the Association for Computational Linguistics (ACL); Aug 2021. p. 4215–28. Online. https://doi.org/10.18653/v1/2021.acl-long.325
22. Reilly FE, Harrow M, Tucker GJ. Language and thought content in acute psychosis. Am J Psychiatry. 1973;130(4):411–7. https://doi.org/10.1176/ajp.130.4.411
23. Komeili M, Pou-Prom C, Liaqat D, Fraser KC, Yancheva M, Rudzicz F. Talk2Me: automated linguistic data collection for personal assessment. PLoS One. 2019;14(3):e0212342. https://doi.org/10.1371/journal.pone.0212342
24. Honoré A. Some simple measure of richness of vocabulary. Assoc Lit Ling Comput Bull. 1979;7:172–7.
25. Brysbaert M. New B: moving beyond Kučera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behav Res Methods. 2009;41(4):977–90. https://doi.org/10.3758/BRM.41.4.977
26. Roark B, Mitchell M, Hosom JP, Hollingshead K, Kaye J. Spoken language derived measures for detecting mild cognitive impairment. IEEE Trans Audio Speech Lang Process. 2011;19(7):2081–90. https://doi.org/10.1109/TASL.2011.2112351
27. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. J Mach Learn Res Arch. 2003;3:993–1022.

28. Andreasen NC. Scale for the assessment of negative symptoms (SANS). Iowa City: University of Iowa; 1983.
29. Andreasen NC. Scale for the assessment of positive symptoms (SAPS). Iowa City: University of Iowa; 1984.
30. Andreasen NC. Scale for the assessment of thought, language, and communication (TLC). Schizophr Bull. 1986;12(3):473–82. https://doi.org/10.1093/schbul/12.3.473
31. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12(85): 2825–30.
32. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging. 2015;15(1):29. https://doi.org/10.1186/s12880-015-0068-x
33. Xu S, Yang Z, Chakraborty D, et al. Automated verbal and nonverbal speech analysis of interviews of individuals with schizophrenia and depression. Annu Int Conf IEEE Eng Med Biol Soc. 2019:225–8.
34. Cohen AS, Alpert M, Nienow TM, Dinzeo TJ, Docherty NM. Computerized measurement of negative symptoms in schizophrenia. J Psychiatr Res. 2008;42(10):827–36. https://doi.org/10.1016/j.jpsychires.2007.08.008
35. Covington MA, Lunden SL, Cristofaro SL, Wan CR, Bailey CT, Broussard B, et al. Phonetic measures of reduced tongue movement correlate with negative symptom severity in hospitalized patients with first-episode schizophrenia-spectrum disorders. Schizophr Res. 2012;142(1-3):93–5. https://doi.org/10.1016/j.schres.2012.10.005
36. Tahir Y, Yang Z, Chakraborty D, Thalmann N, Thalmann D, Maniam Y, et al. Non-verbal speech cues as objective measures for negative symptoms in patients with schizophrenia. PLoS One. 2019;14(4):e0214314. https://doi.org/10.1371/journal.pone.0214314