

A Quantitative and Qualitative Analysis of Schizophrenia Language

Amal Alqahtani^{1,2}, Efsun Kayi^{1,3}, Sardar Hamidian¹, Michael Compton⁴, and Mona Diab^{1,5}

¹The George Washington University, DC, USA

²King Saud University, Riyadh, KSA

³Johns Hopkins University Applied Physics Laboratory, Laurel, USA

⁴Medical Center, Columbia University, New York, USA

⁵Meta AI, USA

{amalqahtani, sardar}@gwu.edu, efsun.kayi@jhuapl.edu

mtc2176@cumc.columbia.edu, mdiab@meta.com

Abstract

Schizophrenia is one of the most disabling mental health conditions to live with. Approximately one percent of the population has schizophrenia which makes it fairly common, and it affects many people and their families. Patients with schizophrenia suffer different symptoms: formal thought disorder (FTD), delusions, and emotional flatness. In this paper, we quantitatively and qualitatively analyze the language of patients with schizophrenia measuring various linguistic features in two modalities: speech and written text. We examine the following features: coherence and cohesion of thoughts, emotions, specificity, level of committed belief (LCB), and personality traits. Our results show that patients with schizophrenia score high in fear and neuroticism compared to healthy controls. In addition, they are more committed to their beliefs, and their writing lacks details. They score lower in most of the linguistic features of cohesion with significant p-values.

1 Introduction

Schizophrenia is a mental illness that can disrupt thought processes and perception (Kerns and Berenbaum, 2002). It can impair people's ability to manage their emotions, and can cause motor and behavioral disorders (Elvevag and Goldberg, 2000).

Understanding and identifying the underlying signs of schizophrenia is critical in early detection and intervention before the malady becomes severely disabling if left untreated (Seeber and Cadanhead, 2005). Moreover, it is vital to support mental health practitioners as well as policymakers to eliminate barriers to treating mental illnesses such as schizophrenia.

Gradual decline in functioning and cognition are some common characteristics of schizophrenia patients. Symptoms may include delusions, which are fixed false beliefs, as well as hallucinations but

also importantly, they tend to have strong convictions regardless of the veridicality of the beliefs themselves. Another symptom that some individuals with schizophrenia exhibit is formal thought disorder (FTD), where a patient becomes unable to form coherent or logical thoughts (Kuperberg, 2010). Moreover, they suffer in some cases from lack of motivation and/or emotional response.

One way to capture mental disorders and related symptomatology is by analyzing patients' linguistic cues. Hence, we map the aforementioned symptoms to linguistic features that we can measure. To date, most of the employed measures used by clinicians measure superficial linguistic cues and they tend to be more qualitative. We hypothesize that advances in pragmatic NLP tools allow us to measure many of these symptoms via analyzing language cues used by patients. We surmise that given such tools, we help create objective quantitative measures for clinicians beyond what they are using today for diagnostics. Moreover having such tools could help them discover and codify further studies allowing for even more signals in detecting such mental health disorders.¹ Accordingly, we present the first comprehensive study of deep pragmatically oriented linguistic modeling tools for diagnostic purposes. We leverage an emotion detection model to assess the lack of emotional response. We also employ a personality detection model to measure lack of motivation, which is one of the negative symptoms they may exhibit. We use a level of committed belief detection model to identify the level of committed belief corresponding to strength of conviction. Formal Thought Disorder (FTD) is measured by using language model-based sentence scoring as well as other coherence features such as LSA, connectives, lexical diversity, syntactic complexity, word information, and level

¹Despite our focus in this work on schizophrenia, we believe that many of the tools we use here could be applicable to other mental disorders.

of linguistic specificity. Finally, we employ the *Coh-matrix* computational tool for analyzing texts for a variety of cohesion measures. (Graesser et al., 2004).

Accordingly, we investigate the following metrics: cohesion, level of committed belief, emotion, and personality and their corresponding correlation with symptomatic patients' language use. We examine both speech and text modalities, comparing patients vs. a matched set of controls.

Our results show that when patients express their emotions in writing or speech, they tend to show fear more often than other emotions. The findings also detect a neuroticism personality as they may suffer from feelings such as anger, and anxiety more frequently and severely. Furthermore, the results indicate that their writings lack specificity (details), and they are more committed to their beliefs in contrast with healthy controls. In addition, our results show that writings of healthy controls are more coherent demonstrated via the high scores of language model probabilities of their writing. To the best of our knowledge, our findings present the first set of measurable pragmatic linguistic cues that significantly correlate with contrastive mental health patients' language use that goes beyond the typical superficial metrics used in the literature to date. Our study provides a set of objective linguistic measures that can serve as metrics that further assist clinicians and policy makers in the mental health domain. The contributions of this paper are as follows:

1. It provides a comprehensive set of cognitive and linguistics *quantitative* metrics for schizophrenia patients language use;
2. We provide a translation of clinical observations of patient language use onto specific measurable linguistic cues that are mapped into advance NLP technology;
3. For the first time, our work leverages advances in the pragmatic NLP to measure patients' cognitive state (namely their levels of committed beliefs), personality traits, emotions, specificity and coherence;
4. We use LM with perplexity scores to measure both coherence and cohesion.

2 Related Work

Language provides significant insight into the content of thought. It also reflects the presence of impairments resulting from mental disorders such as schizophrenia. The predominant reflection of mental impairment for schizophrenia is the lack of coherent text or speech. Accordingly, cohesion scores were first proposed as an indicator of predicting schizophrenia (Elvevåg et al., 2007) where they used Latent Semantic Analysis (LSA) as a feature extractor. This was further amplified by (Bedi et al., 2015) where they measured the semantic coherence in disorganized speech captured by LSA, specifically where large amounts of language overlap was interpreted as coherent language. The study found that these features, together with syntactic markers of complexity, could predict later development of psychosis with 100% accuracy using a convex hull algorithm. Later, Corcoran et al. (2018) used a logistic regression model to predict the onset of psychosis using coherence as measured by LSA combined with the usage of possessive pronouns. This approach showed an accuracy of 83% in predicting the onset of psychosis with a cross-validation accuracy of 79%.

Metrics for Schizophrenia detection were investigated by (AlQahtani et al., 2019) where they used linguistic features such as referential cohesion, text ease, situation model, and readability in patients' and controls' writing or speech to classify presence or absence of the disorder. The researchers trained Support Vector Machine (SVM) and Random Forests (RF) models. The study results showed that the situation model and readability performed the best among all cohesion features for the SVM model yielding a 72% F-score in the binary classification task of detecting whether a person (through their writing or speech) is a schizophrenia patient.

Different from previous studies of schizophrenia, we propose measuring cohesion using language model perplexity. Moreover, we provide a comprehensive exploration of the language of patients relative to that of controls along the following linguistics cues: coherence, emotion, personality, level of specificity, and level of committed belief.

3 Data

Our study comprises two datasets speech, *Lab-Speech*, and written text, *LabWriting*. The data is obtained from schizophrenia patients and healthy

controls. Both datasets are described in detail in (Kayi et al., 2018).² *LabWriting* has 188 participants who are native English-speakers between the ages of 18 – 50 years, corresponding to 93 patients and 95 healthy controls. All participants are asked to write two paragraph-long essays: the first one is about their average Sunday and the second essay is about what makes them the angriest. The total number of writing samples collected from both patients and controls is 373 pieces of text.

The second dataset, *LabSpeech*, includes three questions that prompt participants to describe some emotional and social events. Patients and controls are asked to describe (1) a picture, (2) their ideal day, and (3) their scariest experience. The total number of speech script samples collected from both patients and controls is 431. Speech data is transcribed to text and a punctuation tool (Tilk and Alumäe, 2016) is used to add the missing punctuation.

3.1 Superficial Descriptive statistics

Table 1 illustrates various descriptive statistics comparing and contrasting the *LabWriting* and *LabSpeech* datasets. The results indicate that healthy controls in both datasets are more verbose (produce more words and sentences) when answering questions in both modalities, i.e. writing or speech. The mean values of the number of words and the number of sentences generated by Controls in *LabWriting* are 141 and 7, respectively. However, the mean values are lower for Patients, with 110 and 6 for the same analysis. The Patients in *LabSpeech* also score lower averages in all the descriptive features. These results are in line with a previous study (De Boer et al., 2020) that individuals with schizophrenia speak less and use less complex sentences. * in Table 1 indicates the higher results and statistically significant.

4 Pragmatic Cues

4.1 Emotion

Emotion refers to a person’s internal or external reaction to an event. This reaction can be expressed verbally, outwardly/visibly (e.g., frowning), or physiologically (e.g., crying) (Kring and Caponigro, 2010). Schizophrenia patients are often characterized as having disorganized thinking; however, according to (Kring and Elis, 2013), they still

²The authors of (Kayi et al., 2018) kindly shared the data after we obtained IRB permission.

Descriptive	LabWriting		LabSpeech	
	P	C	P	C
Avg. # words	110	141*	220	277*
Avg. #sent.	6	7*	11	14*
sent./paragraph	5.6	6.6*	11	14*

Table 1: Descriptive statistics for *LabWriting* and *LabSpeech* datasets. We present overall average number of words, overall average number of sentences and a finer grained average number of sentences per paragraph. *P* denotes patient, and *C* denotes control.

report their emotional experiences using the same general definitions of emotions (happy, sad, etc.) as persons who do not have schizophrenia. We use the EmoNet³ (Abdul-Mageed and Ungar, 2017) to obtain the eight core emotions (PL8), which are trust, anger, anticipation, disgust, joy, fear, sadness, and surprise.

4.2 Specificity

Specificity in computational linguistic measures how much detail exists in a text (Louis and Nenkova, 2011). This is an important pragmatic concept and a characteristic of any text (Li and Nenkova, 2015). We quantify this feature because schizophrenia may impacts one’s language specificity. Hence, our hypothesis is that patients tend to write less specific paragraphs which lack references to any specific person, object, or event. We use (Ko et al., 2019) to measure a sentence specificity by indicating how many details exist in each sentence. This tool generates a rate for each sentence between 0 (general sentence) and 1 (detailed sentence). We also use Coh-Metrix (Graesser et al., 2004) to measure word hyponyms (i.e., word specificity) in a text. A higher value reflects an overall use of more specific words, which increases the ease and speed of text processing.

4.3 Level of Committed belief (LCB)

In natural language, the level of committed belief is a linguistic modality that indicates the author’s belief in a given proposition (Diab et al., 2009). We measure this feature as it can detect an individual’s cognitive state. We want to explore this feature to test our hypothesis that patients with schizophrenia may hold strong beliefs towards their own propositions. We rely on a belief tagger (Rambow et al., 2016) to label each sentence with the

³<https://github.com/UBC-NLP/EmoNet>

committed belief tags as (*CB*) where someone (*SW*) strongly believes in a proposition, Non-committed belief (*NCB*) where *SW* reflects a weak belief in the proposition, and Non-Attributable Belief (*NA*) where *SW* is not (or could not be) expressing a belief in the proposition (e.g., desires, questions, etc.). There is also the *ROB* tag where *SW*'s intention is to report on someone else's stated belief, regardless of whether or not they themselves believe it. The feature values are set to a binary 0 or 1 for each *CB*, *NCB*, *NA*, and *ROB* corresponding to unseen or observed. The following text is an example from *LabWriting*.

Every Sunday I usually $\langle cb-I \rangle$ get $\langle /cb-I \rangle$ up and $\langle cb-I \rangle$ watch $\langle /cb-I \rangle$ gospel shows on TV. I $\langle cb-I \rangle$ do $\langle /cb-I \rangle$ my house chores and then $\langle cb-I \rangle$ watch $\langle /cb-I \rangle$ other things on TV. Then later on I $\langle cb-I \rangle$ go $\langle /cb-I \rangle$ down the street to the food restaurants to $\langle na-I \rangle$ eat $\langle /na-I \rangle$ something to eat.⁴

We calculate the LCB as:

$$LCB \langle tag \rangle = \frac{total \langle tag \rangle \text{ in a text}}{all \text{ LCB tags in the same text}}$$

where $\langle tag \rangle$ is one of the 4 LCB features: *CB*, *NCB*, *NA*, or *ROB*.

4.4 Personality

In psychology, personality is the distinctive sets of behaviors, cognitions, and emotional patterns that derive from biological and environmental influence (Major et al., 2000). We study the personality of patient and healthy controls in our datasets based on the famous Big-Five (Digman, 1990) personality measure, which are the following five traits: Extraversion (*EXT*), Neuroticism (*NEU*), Agreeableness (*AGR*), Conscientiousness (*CON*), and Openness (*OPN*). Neuroticism is characterized by a proclivity for negative emotions (Bono and Vey, 2007). Individuals with high scores for neuroticism experience feelings such as anxiety, worry, fear, anger, frustration, depressed mood, and loneliness (Widiger, 2009). Extraversion indicates how outgoing and social a person is (Smelser et al., 2001). A low score in extraversion means an individual prefers to stay alone. We explore personality to test our hypothesis that patients with schizophrenia are high in neuroticism (emotionally unstable),

⁴Typos are in the original text.

especially if delusional, and low in extraversion (Horan et al., 2008). We use (Kazameini et al., 2020) to predict personality traits for each text in our datasets. The model makes binary predictions of the author's personality.

5 Cohesion Linguistic Features

5.1 Information Structure (Givenness)

Latent Semantic Analysis (LSA) measures the semantic similarity/overlap between sentences or between paragraphs (Dennis et al., 2003). We use LSA to evaluate givenness, which is an information structure defined as a phenomenon where a speaker presumes that the listener is already familiar with the context of a discussion topic (Féry and Ishihara, 2016). The sentence is considered to be coherent when the average givenness score is high (Graesser et al., 2004).

5.2 Lexical Diversity

Lexical diversity of a text is a measure of unique words (types), and consequently a measurement of different words that appear in the text compared to the total number of words (tokens) in that text (Durán et al., 2004) (Johansson, 2008). Type-token ratio (TTR), i.e., the ratio of types to tokens, is the most basic metric of lexical diversity (Durán et al., 2004). When the number of types equals that of tokens in a text, all words are different, with TTR being equal to 1, and the lexical diversity of the text reaches its maximum possible value. Such a text, i.e., one with very high lexical diversity, is likely to be either low in cohesion because cohesion requires repetition of words or very short in length. After all, a naturally occurring longer text implies a greater frequency of the same word (Graesser et al., 2004).

5.3 Connectives

The use of connecting words creates cohesive links between ideas and clauses and provides clues about text organization (Graesser et al., 2004). We evaluate two types of connectives which are logic and temporal. The logic connectives are used to connect two or more ideas (such as *and*, *or*). In contrast, temporal connectives are words or phrases that are used to indicate when something is taking place (such as *first*, *until*).

5.4 Syntactic Complexity

Syntax refers to the arrangements of words and morphemes in forming larger units, such as phrases

and clauses, ultimately resulting in well-formed sentences in a language (Crowhurst, 1983). A tree-like structure, a syntactic tree, can visualize the arrangement of words in a sentence. A tree can be simple: containing basic structure like actor-action-object; or complex, larger in size, with significant number of branches, and a complicated relationship among its different parts (Graesser et al., 2004).

5.5 Word Information

All words in a sentence can be categorized as one of two types: a) Content words, such as nouns, verbs, adjectives, and adverbs, which primarily carry the semantic substance of the sentence and contribute to its meaning; and, b) Function words, such as prepositions, determiners, and pronouns, which primarily express the grammatical relationships among content words without significant semantic content (Wilks, 1998).⁵ Word Information refers to the notion that each word can be assigned a syntactic part-of-speech category and, with this assignment, be further rendered as a content or a function word, thus carrying either substantive or "inconsequential" meaning (Graesser et al., 2004).

5.6 Language Model (LM)

A Language model (LM) is the probability distribution over text (Bengio et al., 2003). To analyze coherence in free text, we propose an approach based on LMs. We use a python library *LM-scorer* (Simone, 2020) to calculate probabilities of each word in a text and score sentences. The library uses the GPT2 model (Radford et al., 2019) internally to provide a probability score for each next word. The sentence score (probability) is computed as the mean of tokens' probabilities. For a given sentence, the LM predicts a higher score for a sentence that is more grammatically correct. Performance of LMs is commensurate with word information, content words tend to have lower probabilities compared to function words.

We calculate multiple LM scores: the perplexity scores at sentence and paragraph level. Moreover, we analyze the LM probabilities (scores) across two segmentation/levels: paragraph level and sentence level. We compare the performance of both levels using the means of statistical hypothesis testing.

⁵We contend that this view is controversial since function words are critical to the meaning of utterances, however we would like to emphasize the qualitative difference between content words and function words.

5.6.1 Analysis at Paragraph level

1. **Mean Sentence Probability:** For a given sentence, the LM predicts a higher score/probability for a sentence that is more grammatically and logically sound. We calculate the mean sentences probability in a text for each observation in each group (control/patient).
2. **Median Sentence Probability:** This statistic is calculated by taking the median of the probabilities of sentences. The justification for using this score is that the median, compared to the mean, is more robust to outliers.

5.6.2 Analysis at Sentence level

1. **Sentence probabilities:** This statistic is extracted by aggregating LM individual sentence scores. Sentences scores for all patients and all controls are compared. The number of sentence probability scores analyzed is equivalent to the number of all sentences in the sample.
2. **Mean of the deltas in sentence probabilities:** By using the sentences scores, the changes between the consecutive probability scores of the sentences in the paragraphs are extracted (deltas), and their average is calculated. The total number of this statistic is equivalent to the number of instances in the dataset. Our aim here is to check if the patient group has more fluctuations in their sentence probabilities.
3. **Minimum deltas in sentence probabilities:** The minimum of changes in the sentence probabilities of consecutive sentences in each paragraph is calculated and compared. The total number of this statistic equals the number of instances in the dataset.
4. **Maximum of deltas in sentence probabilities:** Similar to the last statistic, the maximum of changes in the sentence probabilities of consecutive sentences in each paragraph are calculated and compared. The total number of this statistic equals the number of instances in the dataset.

6 Discussion of the Results

Table 2 and Table 3 illustrate the results of emotion analysis and specificity, respectively. Table 4 reports LCB averages and Table 5 summarizes

Emotion	LabWriting		LabSpeech	
	P	C	P	C
Anger	0.159	0.162	0.104*	0.095
Anticip.	0.043	0.039	0.035	0.039
Disgust	0.086	0.093	0.117	0.112
Fear	0.117*	0.103	0.181*	0.167
Joy	0.372	0.381	0.297	0.311
Sadness	0.127	0.127	0.139	0.138
Surprise	0.077	0.079	0.117	0.127
Trust	0.019*	0.015	0.010	0.010

Table 2: Emotion results. The bold values above indicate the high means, and * indicates only the statistically significant values.

personality percentages. Table 6 in appendix A summarize the values of the cohesion linguistic features: Information Structure (Givenness), Connectives, Lexical Diversity, Syntactic Complexity, Syntactic Pattern Density, and Word Information. Table 7 and Table 8 in appendix A show the values of the language model and perplexity scores. For each comparison criteria we compare the p -value to a significance level $\alpha = 0.05$ to make conclusions about our hypotheses. * is used to indicate the results with a statistically significant p -value.

1. **Descriptive features** The p -values of the total number of sentences in both datasets are significant. There is a noticeable difference between the distribution of this statistic between the two groups and it shows that Controls, on average, generate more sentences.
2. **Emotion** We hypothesise that Patients score high in fear. Our results show that Patients in both *LabWriting* and *LabSpeech* score high in fear (p -value = **0.002**) and (p -value=**0.004**), respectively. This result is consistent with a previous study (Suslow et al., 2003) which states that Patients tend to feel fear more often. Patients in *LabWriting* score high in trust, and this may be due to interviewing them in a trustful environment.
3. **Specificity** We hypothesise that Patients write less specific paragraphs. In the score of word hyponyms (Noun) as a measure of specificity, our results show that the Controls score significantly **higher** in *LabWriting* (p -value = **0.03**). Furthermore, Controls score higher in *LabSpeech*, though not significantly. Specificity

Specificity	LabWriting		LabSpeech	
	P	C	P	C
Sent. level	0.47	0.48*	0.39	0.39
Hyponym	5.85	6.06*	6.36	6.45

Table 3: Specificity results. The above table shows the average specificity at sentence level as well as word hyponyms (Noun).

LCB	LabWriting		LabSpeech	
	P	C	P	C
CB	0.52	0.51	0.60	0.58
NCB	0.013	0.020*	0.04	0.05
NA	0.45	0.46	0.34	0.35
ROB	0.008	0.010	0.010	0.012

Table 4: LCB results.

- at sentence level is also significantly **higher** in *LabWriting* for Controls (p -value = **0.009**). However, there is no difference between Controls and Patients in *LabSpeech*. It should be noted that the speech data are faithfully transcribed where pauses and filler words such as *um*, *er*, *uh* can lower the quality of the speech relative the specificity model which is trained on native textual input hence making it challenging to capture specificity.
4. **LCB** The hypothesis of this study states that Patients show more commitment to their beliefs. Table 4 shows the results of LCB. It can be noticed that Patients in both datasets score **higher** in committed belief (CB) and Controls score **higher** in Non-committed belief (NCB). It confirms our hypothesis, and these findings coincide with a previous study (Kayi et al., 2018) that patients with schizophrenia may show more commitment of their belief to propositions expressed in either modality, writing or speech.
 5. **Personality** The hypothesis of this study states that Patients score high levels of neuroticism and low levels of extraversion. Table 5 reports the results of personality analysis. The results show that Patients in both datasets score **lower** in extroversion (EXT) (p -value = **0.03**) in *LabWriting* and score **higher** in neuroticism (NEU) (p -value = **0.04**) in *LabWriting*. These results are in line with previ-

Personality	LabWriting		LabSpeech	
	P	C	P	C
EXT	34%	50%*	27%	30%
NEU	52%*	40%	35%	27%
AGR	60%	63%	81%	85%
CON	46%	54%	6.6%	7.3%
OPN	45%	40%	84%	75%

Table 5: Frequency Distribution of Personality Traits.

ous studies (Camisa et al., 2005), (Horan et al., 2008), (Smeland et al., 2017) which show that schizophrenia is associated with high levels of neuroticism and low levels of extraversion. We report all other personality traits in table 5; However, our analysis mainly focuses on neuroticism and extraversion.

- Information Structure (Givenness)** The average givenness per sentence of the schizophrenia patients is statistically significantly **lower** than that of the Controls in both *LabWriting* (p -value = **0.001**) and *LabSpeech* (p -value = **0.01**). Patients demonstrate challenges in recognizing things that others would find obvious and consequently question or repeat those. In addition, they present something that they have already mentioned earlier as completely new, compromising givenness.
- Lexical Diversity** In the metric Type-token ratio (TTR) for all words, Patients scored **higher** than Controls, with the difference being statistically significant in both *LabWriting* (p -value = **0.004**) and *LabSpeech* (p -value = **0.0001**). The higher proportion of types by Patients stems from the fact that they produce more incomplete, indistinct, inaudible, or incomprehensible words or sounds and shorter sentences and utterances, struggling to reorganize their thoughts (Hinzen et al., 2019) (Merrill et al., 2017). These non-words, particularly shorter sentences, contribute to the higher TTRs for Patients.

Schizophrenic patients, however, are known to repeat words and phrases (Manschreck et al., 1985), and hence a basic TTR in itself is not a reliable indicator for distinguishing between Controls and Patients. TTR is only possible to apply when text or speech are of equal length. We thus compute two more metrics of lexical

diversity, namely measure of textual lexical diversity (MTLD) and measure D vocabulary diversity (VocD), which allow comparison of lexical diversity of texts of unequal lengths. By these measures, we find text and speech of Controls to be lexically much more diverse, with p -values in the order of 10^{-4} .

- Connectives** In the uses of logical, temporal, and extended temporal connectives in text and speech, Controls consistently score **higher**. The difference in scores is statistically significant in all three cases of speech which are logic, temporal, and extended temporal connectives with p -values respectively **0.03**, **0.04**, and **0.03**. In *LabWriting*, the difference is, however, found to be statistically significant (p -value = **0.03**) only in the case of logical connectives. Our findings validate one of the decisive signs of schizophrenia, deficits of logical reasoning among patients (Willits et al., 2018) (Mackinley et al., 2021).
- Syntactic Complexity** In addition to phonetic anomalies in terms of more pauses, loss of prosody, and mumbled sounds, syntactic and semantic conventions that govern the formation of sentences and ultimately the language are routinely violated by schizophrenia patients (Stein, 1993). One of the manifestations of these violations is the decrease in the syntactic complexity of their writing and speech, resulting in disorganized language with poor content. According to all our three measures of syntactic complexity – SYNMEDpos, SYNMEDwrđ, and SYNMEDlem – Controls demonstrate much **higher** syntactically complex text, with statistically significant differences from Patients in all cases, except in *LabSpeech*, in which the difference is nevertheless nearly significant. These results concur with previous studies (Kayi et al., 2018) (Hinzen et al., 2019) which showed that a patient with schizophrenia alters the patterns of linguistic organization, which leads to increased syntactic errors.
- Word Information** In the usage of pronouns, our results show that Patients use the first-person pronouns, e.g., I, my, me, comparatively more, while Controls prefer first person plural, second-person, and third-person more. The differences are statistically significant in

text but not in speech. This result is in line with the previous study (Kayi et al., 2018) (Tang et al., 2021). One metric in which Controls score significantly **higher** in both *LabWriting* and *LabSpeech* is the average minimum word frequency in sentences. With Controls producing significantly longer writings or speeches, a greater frequency of words is necessary to maintain coherence and a logical flow in the text.

11. **Language Model Analysis at Paragraph-Level** We measure the mean of the probabilities of the sentences and the corresponding medians to account for outlier effects. Since both Patients and Controls produced an appreciable number of tokens per sentence, we find these probabilities lower for both groups. We are primarily interested in the comparison of the probabilities and find that the mean and median probabilities are significantly **lower** for Patients than for Controls in *LabWriting*, with mean p -values of **0.01** and median p -values of **0.02**. The findings are in line with previous studies (Kuperberg, 2010), (Hinzen and Rosselló, 2015), (De Boer et al., 2020) that schizophrenia patients often produce idiosyncratic expressions and hence less probable naturally occurring sentences.

While the probabilities in *LabSpeech* are **lower** for Patients, the differences in corresponding probabilities are not statistically significant at mean p -values of **0.524** and median p -values of **0.237**. This can be explained by the fact that Controls can exploit the time during writing better to their advantage to produce more organized and coherent text. Speech, on the other hand, is swift and spontaneous.

12. **Language Model Analysis at Sentence Level**

In line with the mean and median of the probabilities of the sentences at the paragraph level, we compute the average of the probabilities of all sentences. This metric, average sentence probabilities, is also significantly **lower** for Patients (**0.109**) than for Controls (**0.117**) with (p -value=**0.0007**). The difference in *LabSpeech* dataset, like that in the paragraph level, is again not statistically significant at (p -value=**0.175**).

The mean of changes in the sentence probabilities, computed to evaluate how strongly the sentence probabilities change from one sentence to another in a paragraph and consequently how much the sentences deviate from a coherent and logical flow, is **higher** for Controls (p -value=**0.05**) in *LabWriting*. Two other metrics related to this, the minimum and the maximum of changes in sentence probabilities, provide mixed, hence inconclusive, results. These probabilities, therefore may not be consistent indicators for the fluctuations we expected.

13. **Perplexity** Table 8 in appendix A shows the results of perplexity. We compute it at two levels: the sentence level and the paragraph level, to determine how predictable the language of Patients is compared to that of Controls. In *LabWriting*, the model is more perplexed for Patients in both levels, and the difference between the two groups is highly significant (p -value =**0.01**) at the paragraph level while (p -value =**0.00005**) at the sentence level. However, the results are not significant for *LabSpeech* for any of the two levels.

7 Conclusion

Patients with schizophrenia experience different symptoms, some of which involve problems with concentration and memory, which in return may lead to disorganization in speech or behavior. Therefore, diagnosing this disorder early and correctly is extremely important as it may help alleviate the adverse effects on patients.

Among the linguistic features of cohesion investigated in this study, we found that Patients' scores are lower, with significant p -values in information structure (givenness), lexical diversity except for Type-token ratio (TTR), connectives, and syntactic complexity in both datasets. Among the pragmatic cues, we found that Patients' score high in fear, and their personality is associated with elevated neuroticism. They also show more commitment to their beliefs, and their average specificity at sentence and word levels is lower than Controls.

In the future, we plan to expand our analysis to other related mental health disorders. We also plan to explore the pragmatically motivated linguistics features of schizophrenia in other languages.

Limitations

One of the main limitations of this study is the size of the sample, and this is due to the data privacy and the cost associated with collecting scripts written by patients with schizophrenia.

References

- Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 718–728.
- Amal AlQahtani, Efsun Kayi, and Mona Diab. 2019. Understanding cohesion in writings and speech of schizophrenia patients. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 364–369. IEEE.
- Gillinder Bedi, Facundo Carrillo, Guillermo A Cecchi, Diego Fernández Slezak, Mariano Sigman, Natália B Mota, Sidarta Ribeiro, Daniel C Javitt, Mauro Copelli, and Cheryl M Corcoran. 2015. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia*, 1(1):1–7.
- Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.
- Joyce E Bono and Meredith A Vey. 2007. Personality and emotional performance: Extraversion, neuroticism, and self-monitoring. *Journal of occupational health psychology*, 12(2):177.
- Kathryn M Camisa, Marcia A Bockbrader, Paul Lysaker, Lauren L Rae, Colleen A Brenner, and Brian F O’Donnell. 2005. Personality traits in schizophrenia and related personality disorders. *Psychiatry research*, 133(1):23–33.
- Cheryl M Corcoran, Facundo Carrillo, Diego Fernández-Slezak, Gillinder Bedi, Casimir Klim, Daniel C Javitt, Carrie E Bearden, and Guillermo A Cecchi. 2018. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, 17(1):67–75.
- Marion Crowhurst. 1983. Syntactic complexity and writing quality: A review. *Canadian Journal of Education/Revue canadienne de l’éducation*, pages 1–16.
- JN De Boer, M van Hoogdalem, RCW Mandl, J Brummelman, AE Voppel, MJH Begemann, E van Dellen, FNK Wijnen, and IEC Sommer. 2020. Language in schizophrenia: relation with diagnosis, symptomatology and white matter tracts. *npj Schizophrenia*, 6(1):1–10.
- Simon Dennis, Tom Landauer, Walter Kintsch, and Jose Quesada. 2003. Introduction to latent semantic analysis. In *25th Annual Meeting of the Cognitive Science Society. Boston, Mass.*, page 25.
- Mona Diab, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. 2009. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 68–73.
- John M Digman. 1990. Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 41(1):417–440.
- Pilar Durán, David Malvern, Brian Richards, and Ngoni Chipere. 2004. Developmental trends in lexical diversity. *Applied Linguistics*, 25(2):220–242.
- 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.
- Brita Elvevag and Terry E Goldberg. 2000. Cognitive impairment in schizophrenia is the core of the disorder. *Critical Reviews™ in Neurobiology*, 14(1).
- Caroline Féry and Shinichiro Ishihara. 2016. *The Oxford handbook of information structure*. Oxford University Press.
- Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-matrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202.
- Wolfram Hinzen, Derya Çokal, Vitor C Zimmerer, Douglas Turkington, I Nicol Ferrier, Rosemary Varley, and Stuart Watson. 2019. Disturbing the rhythm of thought: speech pausing patterns in schizophrenia, with and without formal thought disorder. *Plos One*. 2019; 14 (5): e0217404. DOI: 10.1371/journal.pone.0217404.
- Wolfram Hinzen and Joana Rosselló. 2015. The linguistics of schizophrenia: thought disturbance as language pathology across positive symptoms. *Frontiers in psychology*, 6:971.
- William P Horan, Jack J Blanchard, Lee Anna Clark, and Michael F Green. 2008. Affective traits in schizophrenia and schizotypy. *Schizophrenia bulletin*, 34(5):856–874.
- Victoria Johansson. 2008. Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working papers/Lund University, Department of Linguistics and Phonetics*, 53:61–79.
- Efsun Sarioglu Kayi, Mona Diab, Luca Pauselli, Michael Compton, and Glen Coppersmith. 2018. Predictive linguistic features of schizophrenia. *arXiv preprint arXiv:1810.09377*.

- Amirmohammad Kazameini, Samin Fatehi, Yash Mehta, Sauleh Eetemadi, and Erik Cambria. 2020. Personality trait detection using bagged svm over bert word embedding ensembles. *arXiv preprint arXiv:2010.01309*.
- John G Kerns and Howard Berenbaum. 2002. Cognitive impairments associated with formal thought disorder in people with schizophrenia. *Journal of abnormal psychology*, 111(2):211.
- Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. 2019. Domain agnostic real-valued specificity prediction. In *AAAI*.
- Ann M Kring and Janelle M Caponigro. 2010. Emotion in schizophrenia: where feeling meets thinking. *Current directions in psychological science*, 19(4):255–259.
- Ann M Kring and Ori Elis. 2013. Emotion deficits in people with schizophrenia. *Annual review of clinical psychology*, 9:409–433.
- Gina R Kuperberg. 2010. Language in schizophrenia part 1: an introduction. *Language and linguistics compass*, 4(8):576–589.
- Junyi Li and Ani Nenkova. 2015. Fast and accurate prediction of sentence specificity. In *AAAI*.
- Annie Louis and Ani Nenkova. 2011. Automatic identification of general and specific sentences by leveraging discourse annotations. In *Proceedings of 5th international joint conference on natural language processing*, pages 605–613.
- Michael Mackinley, Jenny Chan, Hanna Ke, Kara Dempster, and Lena Palaniyappan. 2021. Linguistic determinants of formal thought disorder in first episode psychosis. *Early intervention in psychiatry*, 15(2):344–351.
- Brenda Major, Catherine Cozzarelli, Mardi J Horowitz, Peter J Colyer, Lynn S Fuchs, Edward S Shapiro, Karen Callan Stoiber, Ulrik Fredrik Malt, Thomas Teo, David G Winter, et al. 2000. Encyclopedia of psychology: 8 volume set. *New York and Washington: Oxford University Press and the American Psychological Association*.
- Theo C Manschreck, Brendan A Maher, Toni M Hoover, and Donna Ames. 1985. Repetition in schizophrenic speech. *Language and Speech*, 28(3):255–268.
- Anne M Merrill, Nicole R Karcher, David C Cicero, Theresa M Becker, Anna R Docherty, and John G Kerns. 2017. Evidence that communication impairment in schizophrenia is associated with generalized poor task performance. *Psychiatry research*, 249:172–179.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Owen Rambow, Tao Yu, Axinia Radeva, Alexander R Fabbri, Christopher Hidey, Tianrui Peng, Kathleen R McKeown, Smaranda Muresan, Sardar Hamidian, Mona T Diab, et al. 2016. The columbia-gwu system at the 2016 tac kbp best evaluation. In *TAC*.
- Katherine Seeber and Kristin S Cadenhead. 2005. How does studying schizotypal personality disorder inform us about the prodrome of schizophrenia? *Current Psychiatry Reports*, 7(1):41–50.
- Primarosa Simone. 2020. lm-scorer. <https://github.com/simonepri/lm-scorer>.
- Olav B Smeland, Yunpeng Wang, Min-Tzu Lo, Wen Li, Oleksandr Frei, Aree Witoelar, Martin Tesli, David A Hinds, Joyce Y Tung, Srdjan Djurovic, et al. 2017. Identification of genetic loci shared between schizophrenia and the big five personality traits. *Scientific reports*, 7(1):1–9.
- Neil J Smelser, Paul B Baltes, et al. 2001. *International encyclopedia of the social & behavioral sciences*, volume 11. Elsevier Amsterdam.
- Johanna Stein. 1993. Vocal alterations in schizophrenic speech. *Journal of Nervous and Mental Disease*.
- Thomas Suslow, Cornelia Roestel, Patricia Ohrmann, and Volker Arolt. 2003. The experience of basic emotions in schizophrenia with and without affective negative symptoms. *Comprehensive psychiatry*, 44(4):303–310.
- Sunny X Tang, Reno Kriz, Sunghye Cho, Suh Jung Park, Jenna Harowitz, Raquel E Gur, Mahendra T Bhati, Daniel H Wolf, João Sedoc, and Mark Y Liberman. 2021. Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. *npj Schizophrenia*, 7(1):1–8.
- Ottokar Tilk and Tanel Alumäe. 2016. Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *Interspeech 2016*.
- Thomas A Widiger. 2009. *Neuroticism*, volume 11. The Guilford Press.
- Yorick Wilks. 1998. D. arnold, l. balkan, r. lee humphries, s. meijer and l. sadler. machine translation: an introductory guide. ncc blackwell, oxford, 1994.(hardback isbn 1-85554-246-3 49.95/£40.00; paperback isbn1 – 85554 – 217 – x19.95/£ 18.99) viii+ 240 pages. *Natural Language Engineering*, 4(4):363–382.
- Jon A Willits, Timothy Rubin, Michael N Jones, Kyle S Minor, and Paul H Lysaker. 2018. Evidence of disturbances of deep levels of semantic cohesion within personal narratives in schizophrenia. *Schizophrenia Research*, 197:365–369.

A Appendix

Cohesion Linguistic Features	LabWriting		LabSpeech	
	P	C	P	C
1. LSA				
Avg. givenness of each sentence	0.20	0.23*	0.31	0.32*
2. Lexical Diversity				
Type token ratio (TTR) for all words	0.63*	0.60	0.46*	0.43
MTLD lexical diversity measure for all words	59.9	68.82*	40.10	44.95*
VOC lexical diversity measure for all words	40.6	67.7*	42.73	52.20*
3. Connectives				
Score of logic connectives	47.1	53.1*	33.66	38.31*
Score of temporal connectives	24.5	26.7	12.24	14.74*
Score of extended temporal connectives	24.3	27.2	14.44	18.04*
4. Syntactic Complexity				
SYNMEDpos*	0.56	0.61*	0.66	0.68
SYNMEDwrđ*	0.73	0.81*	0.84	0.87*
SYNMEDlem*	0.71	0.79*	0.82	0.84*
5. Word Information				
Score of pronouns, first person, single form	96.6*	86.11	56.68	55.03
Score of pronouns, first person, plural form	6.3	10.2*	5.24	7.26
Score of pronouns, second person	3.37	6.22*	7.99	7.33
Score of pronouns, third person, plural form	7.90	12.25*	7.20	8.65
Avg. minimum word frequency in sentences	0.83	1.01*	1.30	1.45*

SYNMEDpos*: mean minimum editorial distance score between adjacent sentences computed from POS.

SYNMEDwrđ*: minimum editorial distance score between adjacent sentences computed from words.

SYNMEDlem*: This is the minimum editorial distance score between adjacent sentences from lemmas.

Table 6: Coh-Metrix Linguistic Features Results

Cohesion Linguistic Features	LabWriting		LabSpeech	
	P	C	P	C
1. Analysis at Paragraph level				
- Mean of probabilities of sentences	0.110	0.119*	0.106	0.107
- Median of probabilities of sentences	0.107	0.117*	0.104	0.107
2. Analysis at Sentence level				
-Sentence Probabilities	0.109	0.117*	0.103	0.106
-Mean of changes in sentence probabilities	-0.106	-0.045*	-0.060	-0.062
-Minimum of changes in sentence probabilities	-1.283	-1.036*	-1.835*	-2.107
-Maximum of changes in sentence probabilities	1.036	0.986	1.572	1.902*

Table 7: The language model scores (probabilities) across different segmentation (levels)

Levels	LabWriting		LabSpeech	
	P	C	P	C
Sentence	1.12	1.10*	1.11	1.12
Paragraph	203.9	150.4*	245.5	230.1

Table 8: Perplexity across different segmentation (levels)