Schizophrenia Research 266 (2024) 183–189

Contents lists available at ScienceDirect

Schizophrenia Research

journal homepage: www.elsevier.com/locate/schres

Natural language processing for defining linguistic features in schizophrenia: A sample from Turkish speakers

Tuğçe Çabuk [a], Nurullah Sevim [b], Emre Mutlu [c], A. Elif Anıl Yağcıoğlu [d], Aykut Koç [e], Timothea Toulopoulou [f,g,h,*]

[a] *Department of Psychology, National Magnetic Resonance Research Center (UMRAM) & Aysel Sabuncu Brain Research Center, Bilkent University, Bilkent, 06800 Ankara, Turkey*
[b] *Department of Electrical and Electronics Engineering, National Magnetic Resonance Research Center (UMRAM), Bilkent University, Bilkent, 06800 Ankara, Turkey*
[c] *Department of Psychiatry, Hacettepe University, Faculty of Medicine, Sıhhiye, 06230 Ankara, Turkey*
[d] *Department of Psychiatry, Hacettepe University, Faculty of Medicine, Sıhhiye, 06230 Ankara, Turkey*
[e] *Department of Electrical and Electronics Engineering, National Magnetic Resonance Research Center (UMRAM), Bilkent University, Bilkent, 06800 Ankara, Turkey*
[f] *Department of Psychology, National Magnetic Resonance Research Center (UMRAM) & Aysel Sabuncu Brain Research Center, Bilkent University, Bilkent, 06800 Ankara, Turkey*
[g] *1st Department of Psychiatry, National and Kapodistrian University of Athens, Athens, Greece*
[h] *Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, USA*

ARTICLE INFO

Keywords:
Linguistics
Speech
Natural language processing
Schizophrenia
Turkish
Machine learning
Psychosis

ABSTRACT

Natural language processing (NLP) provides fast and accurate extraction of features related to the language of schizophrenia. We utilized NLP methods to test the hypothesis that schizophrenia is associated with altered linguistic features in Turkish, a non-Indo-European language, compared to controls. We also explored whether these possible altered linguistic features were language-dependent or -independent. We extracted and compared speech in schizophrenia (SZ, N = 38) and healthy well-matched control (HC, N = 38) participants using NLP. The analysis was conducted in two parts. In the first one, mean sentence length, total completed words, moving average type-token ratio to measure the lexical diversity, and first-person singular pronoun usage were calculated. In the second one, we used parts-of-speech tagging (POS) and Word2Vec in schizophrenia and control. We found that SZ had lower mean sentence length and moving average type-token ratio but higher use of first-person singular pronoun. All these significant results were correlated with the Thought and Language Disorder Scale score. The POS approach demonstrated that SZ used fewer coordinating conjunctions. Our methodology using Word2Vec detected that SZ had higher semantic similarity than HC and K-Means could differentiate between SZ and HC into two distinct groups with high accuracy, 86.84 %. Our findings showed that altered linguistic features in SZ are mostly language-independent. They are promising to describe language patterns in schizophrenia which proposes that NLP measurements may allow for rapid and objective measurements of linguistic features.

1. Introduction

Language is a crucial tool known by healthcare professionals for diagnosing schizophrenia (SZ) (de Boer et al., 2020). Furthermore, collecting language data is easy and inexpensive. Traditionally, expert opinions, clinical ratings, and manual linguistic analyses have been used for analysing language. Although informative, they have some limitations. The results of clinical ratings can be affected by incomplete response sets and clinical judgments, which often lack precision. Furthermore, in manual linguistic analyses, researchers must put significant effort into analysing even a single person's data making large-scale studies challenging to conduct (Corcoran et al., 2020). Nowadays, linguistic markers are more trackable with the advances in natural language processing (NLP) based technologies (Foltz et al., 2016; Corcoran et al., 2018; Rezaii et al., 2019). NLP provides fast and accurate extraction of features related to language (de Boer et al., 2020; Voppel et al., 2021).

Many different NLP techniques have been used to identify language

* Corresponding author at: Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, USA.
E-mail addresses: tugce.cabuk@bilkent.edu.tr (T. Çabuk), eanil@hacettepe.edu.tr (A.E.A. Yağcıoğlu), aykut.koc@bilkent.edu.tr (A. Koç), timothea.toulopoulou@mssm.edu (T. Toulopoulou).

https://doi.org/10.1016/j.schres.2024.02.026
Received 28 June 2023; Received in revised form 26 December 2023; Accepted 17 February 2024
Available online 27 February 2024

changes in SZ. In one of the first studies, Elvevåg et al. (2007) applied Latent Semantic Analysis (LSA) to quantify coherence. They found that the patients who got a high score on the Assessment of Thought, Language, and Communication (TLC) lost coherence more quickly. LSA could discriminate SZ patients from healthy control participants (HC) with 80–82 % accuracy while detecting decreased cohesion (Elvevåg et al., 2007). In a recent study, Tang et al. (2021) utilized word-level and sentence-level analyses such as Parts of Speech tagging (POS) and Bidirectional Encoder Representations from Transformers (BERT). The results showed that NLP measurements of language disturbances in SZ could discriminate better between HC and SZ compared to clinical ratings alone (Tang et al., 2021). In another study, the usage of quantitative speech variables such as utterances and words was higher in SZ than HC, and these variables could classify, based on machine learning algorithms, SZ with 90–100 % specificity and 80–90 % sensitivity (Tan et al., 2021). NLP is also critically important in predicting conversion to psychotic disorders in the studies of clinically high-risk (CHR) populations. According to these studies, less usage of complementizers and reduced sentence length which correlated with negative symptom severity (Bedi et al., 2015) and low semantic density (Rezaii et al., 2019) could predict psychosis in CHR population with high accuracy. Other successful techniques in NLP can be ordered in graph analysis (Mota et al., 2017) and automated metaphor detection (Gutiérrez et al., 2017).

However, all of these studies were done in Indo-European languages. In this study, we applied multiple NLP methods for defining linguistic features of Turkish, a non-Indo-European language, in SZ to see whether the results of the NLP analyses in Turkish are language-independent or language-dependent. Firstly, we calculated mean sentence length, total completed words, moving average type-token ratio (MATTR), an index of variability in lexicon usage, and average first-person singular pronoun usage (i.e., “I”/*ben*) as they served more consistent results for discriminating SZ from HC (Hitczenko et al., 2021). We hypothesized that the mean sentence length and MATTR in SZ would be lower, but total completed words will be higher than in HC, and first-person singular pronoun usage in SZ will be significantly more common than in HC. In addition, mean sentence length, total completed words, MATTR, and average first-person singular pronoun usage (i.e., *ben*) were compared with Thought and Language Disorder Scale (TALD) (Kircher et al., 2014). We expected that the results of these parameters would be correlated with TALD results. Secondly, POS and a methodology utilizing Word2Vec were used as an exploratory analysis in schizophrenia. In POS, the number of uses for linguistic units (i.e., noun, verb, and adjective) is counted. In Word2Vec, word embeddings or vectors are constructed in multidimensional space depending on the context of the speech. The vectors close to each other mean that their corresponding words are semantically similar (e.g., The vector of *queen* is closer to the vector of *king* than the vector of *spoon*). Our methodology using Word2Vec via comparing linguistic features of schizophrenia and control samples enables us to analyse the semantic structure of the data with high effectiveness, such as calculating semantic similarity. Also, *K-Means* clustering algorithm can discover hidden patterns and structures as an exploratory data analysis algorithm. Bambini et al. (2022) used K-Means clustering algorithm for the first time in schizophrenia and detected two sub-groups in the sample, which shows distinctive linguistic profiles. However, they used only schizophrenia samples. Here, we used the first Word2Vec with *K-Means* clustering algorithm in schizophrenia patients and in controls to assess whether K-Means could accurately differentiate between schizophrenia and healthy control speech patterns in our sample. As there is no consistent set of NLP predictors in the results of the studies conducted with POS and Word2Vec (Tang et al., 2021; Tan et al., 2021; Voppel et al., 2021; Ziv et al., 2021; Corona-Hernández et al., 2022), we broadly hypothesized that linguistic features in SZ would be significantly different compared to HC.

2. Methods

2.1. Participants

Two groups of participants were included in the study: schizophrenia (SZ), and healthy control (HC). The SZ sample consisted of 38 outpatients of the Community Mental Health Center of Etimesgut Şehit Sait Ertürk State Hospital and Hacettepe University Hospital between September 2019 and December 2022. The patients who met the DSM-5 criteria for SZ were enrolled in the study. The patients were clinically stable and had not undergone electroconvulsive therapy (ECT) in the last year. To determine the disease severity and clinical studies, the scoring of the Clinical Global Impression Rating Scale (Guy, 1976), and the age of illness onset were collected from the patient's charts. Daily doses of antipsychotics were converted into chlorpromazine equivalents using an equivalency table provided by Gardner et al. (2010). The HC group comprised 38 carefully matched for age, gender, and education to SZ participants. The inclusion criteria of the healthy control group comprised no psychiatric history, no family history of psychiatric illness, no traumatic brain injury or any other neurological disease, no somatic disease, and no psychotropic drug usage.

The exclusion criteria for all participants included any history of neurological or chronic somatic disorder, mental retardation, alcohol or substance abuse or dependence, or head injury. All participants were right-handed, native speakers of Turkish, and Turkish nationality. All participants provided written informed consent. Twenty-four SZ and 24 HC of participants from the whole sample had previously been included in another study (Çabuk et al., 2023); however, the current study utilized one common data collection tool and different analysis techniques.

2.2. Clinical assessment

The Thought and Language Disorder Scale (TALD) was developed by Kircher et al. (2014) and adapted to the Turkish language by Mutlu et al. (2019) (TALD-TR). The Turkish language version of TALD showed adequate and effective psychometric properties with high Cronbach's Alpha values and high inter-rater reliability parallel to the original version (Mutlu et al., 2019; Kircher et al., 2014).

The TALD is a 30-item semi-structured scale with operationalized definitions of symptoms and symptom ratings, directed questions, and examples for discriminating objective and subjective symptoms from each other. The factor analysis of TALD revealed four factorial structures, namely Objective Positive (i.e., circumstantiality, derailment), Subjective Negative (i.e., poverty of thought, blocking), Objective Negative (i.e., poverty of speech, concretism), and Subjective Positive factors (i.e., thought interference, pressure/rush of thought) (Kircher et al., 2014).

The interview of TALD comprises two parts. Description of daily life events (e.g., “What do you like to do on the weekends?”) or interests such as hobbies (e.g., “Which kind of movies do you like to watch?”) are discussed by open-ended questions in the first section (Section A) so that the rater can detect objective symptoms and further query to code symptom severity. The subjective symptoms are questioned in detail in the second section (Section B). The rater assesses each subjective symptom following standard procedures (Kircher et al., 2014) and explores and notes the presence of any subjective symptoms (e.g., for rush/pressure of thoughts: “Do you sometimes feel that so many ideas come into your mind quickly, one after the other or even at the same time, that you lose control of your thoughts”). If necessary, the rater can ask more questions or give additional explanations to be sure that questioned subjective phenomena are understood.

In this study, the responses to sections *A* and *B* were audio recorded in a 20-min session. All participants gave informed consent for their interview sessions to be voice recorded. The rating was completed immediately after the interview by an expert speech and language therapist (TÇ) who was blind to the participants' clinical assessment. The grading of each item, such as derailment, tangentiality, and perseveration, was analysed according to the definitions of that item in the manual (Kircher et al., 2014; Mutlu et al., 2019).

2.3. Speech sampling

Speech was elicited using a 20-min semi-structured interview reflecting sections *A* and *B* of TALD. The questions were designed to elicit spontaneous speech with general issues and thought processing (please see: *2.1. Clinical assessment*). All interviews were conducted by an expert speech and language therapist (TÇ). The elicited verbal samples were audio-recorded and later transcribed. The transcriber (TÇ) was blind to the participants' group. The ethics committee of Bilkent University approved the procedures.

2.4. Natural language processing analyses for hypothesis-driven part

In the first part of analysis, where NLP techniques were used, total completed words and first-person singular pronoun usage were averaged over the participants. The standard deviations of these statistics were also considered to identify the distinct trends between SZ and HC groups. Moreover, the average length of the participants' sentences was analysed. Lastly, the Moving average type-token ratio (MATTR) to identify variability in lexicon usage was calculated for all participants (Covington and McFall, 2010) by dividing the number of distinct words by the total number of words in a specified moving window whose length was 50. For example, when the window length was selected as 50, the type-token ratio (TTR) for words from 1 to 50 was calculated. Then, the TTR was calculated for words from 2 to 51, then from 3 to 52, and so on until the end of the transcribed speech data. In TTR, we did not use lemmatization. For the final score, calculated TTRs were averaged (Fergadiotis et al., 2015). All analyses for this section were done in the Natural Language Tool Kit (NLTK) library of Python programming language (Fig. 1).

2.5. Natural language processing analyses for exploratory part

More advanced methods were utilized in the second part of the NLP analyses. First, through parts-of-speech tagging (POS), the number of uses for parts of speech (i.e., noun, verb, and adjective) was calculated per 100 words for each participant. A few units are used in basic sentences, such as one noun and one verb (i.e., “Ben geldim/I came”) whereas various parts of speech are contained in complex sentences (i.e., “Ben geldim ama bu kötü ortam beni rahatsız etti/I came here, but this bad atmosphere disturbed me”). A pre-trained transformer model, RoBERTa (Liu et al., 2019), was used for the POS tagging task (Fig. 1).

Second, the Word2Vec embeddings were used to obtain more detailed information about the transcribed speech data (Mikolov et al., 2013) (Fig. 1). These embeddings are vector representations of words in a high-dimensional space. In this high-dimensional space, the word vectors carry semantic information acquired through training with large textual data. The core assumption underlying these semantic vector representations is that the meaning of a word can be derived from its context. Here ‘context’ reflects the other words close to the target word. By traversing all the textual data, the semantic similarity between words can be determined by inspecting how similar the contexts for these words are.

To assess the semantic similarity of each subject's speech, the mean similarity of the transcribed speech data was defined. The mean similarity was calculated by taking the average of word similarity between consecutive words in the participant's speech. The similarity metric in this case was the cosine similarity between word vectors. Cosine similarity between two word's vectors was between -1 and 1. A cosine similarity of 1 means these two words are completely similar whereas a cosine similarity of -1 means there is no relationship between two words. Here, we used the term *semantic similarity* rather than *semantic coherence* when employing cosine similarity (Alonso-Sánchez et al., 2022). Moreover, a methodology with a clustering algorithm was also utilized for this study by using Word2Vec embeddings trained on Turkish corpus (Köksal, 2018). The embedding vectors for each word in a participant's speech were retrieved. The stop words were defined according to the NLTK default stop words and were removed before this procedure. Later, to have one distinct vector for each participant, the word vectors from each participant's speech were summed up, known as the “Bag of Words” approach. Bag of Words is a feature extraction technique from text data where the word occurrences are considered. Since it is easy to implement, intuitive and flexible, this technique is widely preferred in NLP applications (Mikolov et al., 2013). Bag of Words approach was also utilized in detection and classification of schizophrenia through speech data (Castellani et al., 2012; Rezaii et al., 2019) (Fig. 1).

The output of Bag of Words would be a distinct and individual vector

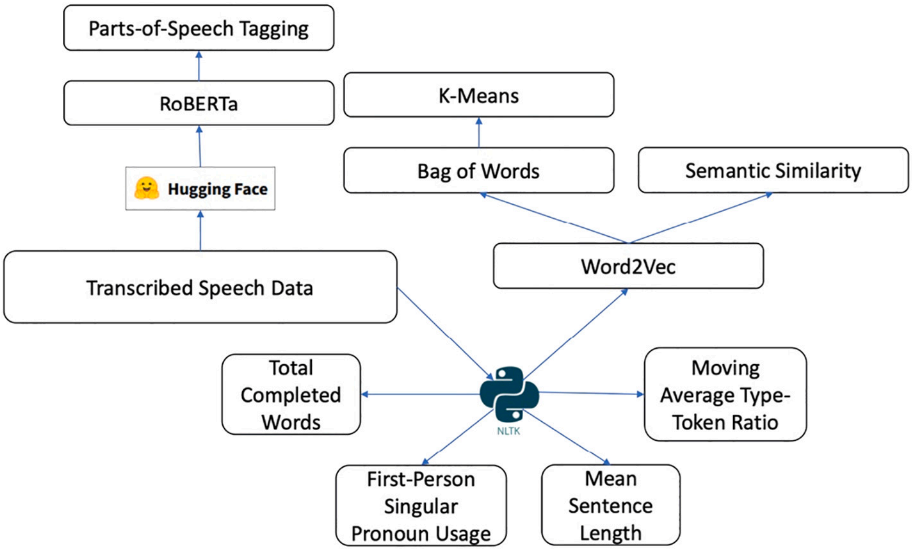

Fig. 1. The visualization of the methodology that includes Natural Language Processing techniques. Hypothesis-driven analyses part consist of total completed words, first-person singular pronoun usage, mean sentence length, and moving average type-token ratio. Parts-of-speech tagging and Word2Vec (semantic similarity and K-Means) are for exploratory analyses part.

for each participant, called 'document embeddings'. An unsupervised clustering algorithm, "K-Means" (MacQueen, 1967), was applied to these document embeddings to evaluate how much information these document embeddings carry regarding the group membership of the participant (SZ vs. HC) (Fig. 1). In other saying, our main aim was not a develop generalizable classifier but rather to use the K-Means to describe the dataset that we have. For this aim, we used n = 2 clusters in the K-Means algorithm as we have two groups (SZ and HC).

2.6. Statistical analyses

The SPSS version 22 (SPSS Inc., Chicago, Illinois, USA) was used to compare demographic and clinical characteristics. The continuous variables were checked using probability plots, histograms, and the Kolmogorov–Smirnov test to assess the normality of data distributions. In the demographic and clinical characteristics, the Chi-square test was performed to analyse categorical variables, and independent samples *t*-test was performed to analyse continuous variables. For the NLP analyses (parts 1 and 2), independent samples *t*-test and Mann Whitney *U* were used for normally distributed and non-normally distributed data, respectively. Pearson's correlations were performed to assess the possible correlations between NLP measurements and TALD (Total) and its four factors scores. The number of eigenvalues equal to or greater than 1 was determined as 3 using principal component analysis using the values of the variables in the hypothesis-driven NLP analyses part and POS, part of exploratory NLP analyses part (MacHado, 2007). Thus, the adjusted significance threshold α was determined as 0.018 and 0.014 for the hypothesis-driven NLP analyses part and POS, part of exploratory NLP analyses part, respectively.

3. Results

3.1. Demographics and clinical characteristics

Demographic and clinical characteristics are summarized in Table 1. The groups did not differ in age, education and gender. HC showed higher paid employment and marriage rates. SZ were rated worse in TALD (Total) and its factor dimension scores.

3.2. Natural language processing results for hypothesis-driven part

Table 2 shows the differences between groups for all variables in NLP analyses (i.e. hypothesis-driven). Apart from total completed words, which did not differ between schizophrenia patients and healthy controls, the remaining variables did differ significantly in schizophrenia patients compared to healthy controls. Mean sentence length and MATTR were lower, but average first-person singular pronoun usage was higher in schizophrenia patients than in healthy controls. We found significant negative correlations between TALD scores, mean sentence length, and MATTR. Average first-person singular pronoun usage (i.e. *ben*) was correlated with both TALD subjective negative and objective negative scores (Table 3).

3.3. Natural language processing results for exploratory part

Table 4 shows the differences between groups for all variables in POS. Among all variables (Adjective, Adverb, Coordinating Conjunction, Determiner, Noun, Verb, Pronoun), Coordinating conjunction did differ significantly in schizophrenia patients compared to healthy controls. Schizophrenia patients produced fewer coordinating conjunctions.

Schizophrenia patients had higher semantic similarity than healthy controls ($U = 527$, $p = 0.043$). The document embeddings were fed to the K-Means algorithm. The formed clusters matched with the true labels with 86.84 % accuracy. Fig. 2 shows the participants whose true labels matched and whose did not match with the clusters that were formed by K-Means. In order to visualize the participants' clusters on a 2-D plane, t-SNE (Hinton and Roweis, 2003) technique was used to lower the dimension of the document embeddings. Looking at the figure, HC and SZ groups were also clearly separable through inspection. However, some of the participants from SZ group were closer to the HC group, so K-Means algorithm could not determine the correct cluster for them.

Table 1
Demographic and clinical characteristics of patients with schizophrenia and healthy control group ($N = 76$).

	SZ	HC	$\chi 2$	df	*p*
	$N = 38$ (%)	$N = 38$ (%)			
Gender					
Male	21 (55.3)	20 (52.6)	0.053	1	0.81
Marital status					
Married	10 (26.3)	23 (60.5)	9.05	1	**0.003****
Paid employment					
Yes	18 (47.4)	30 (78.9)	8.14	1	**0.004****
			t	**df**	***p***
Age	38.82 ± 8.16	37.97 ± 7.96	0.45	74	0.65
Education (years)	10.66 ± 3.21	12.11 ± 3.60	−1.84	74	0.07
TALD (Total)	25.79 ± 1.97	1.97 ± 2.42	10.64	74	**0.001***
TALD (Objective Positive)	5.26 ± 5.26	0.66 ± 1.77	5.11	74	**0.001***
TALD (Subjective Positive)	2.55 ± 2.11	0.24 ± 0.54	6.54	74	**0.001***
TALD (Objective Negative)	5.92 ± 4.87	0.32 ± 0.48	6.98	74	**0.001***
TALD (Subjective Negative)	11.92 ± 7.94	0.76 ± 1.12	8.56	74	**0.001***
Duration of illness (years)	17.3 ± 7.85	N/A			
Age at illness onset	21.61 ± 6.86	N/A			
Medication (ECPZ-mg)	504.39 ± 220.29	N/A			
CGI-Illness severity	3.66 ± 1.02	N/A			

$\chi 2$ = Chi-square, df = degree of freedom, *t:* *t*-test, SD = Standard deviation, TALD = Thought and Language Disorder Scale, ECPZ = the equivalent dose of chlorpromazine, The CGI = Clinical Global Impression Scale-Severity, N/A = Not applicable, Bold values indicate significance either at level *$p < 0.001$ or **$p < 0.005$.

4. Discussion

In this study, we applied multiple NLP techniques to find linguistic features in Turkish, in a non-Indo European language, related to schizophrenia patients and found almost similar significant differences with other studies done in Indo-European languages (Manschreck et al., 1985; de Boer et al., 2020; Hitczenko et al., 2021; Tang et al., 2021; Alonso-Sánchez et al., 2022). NLP measurements reported here appear to be sensitive to detecting 'the language of schizophrenia' with high accuracy and significance level. Mean sentence length is about sentence complexity. Greater length indicates more complex sentences related to syntax (de Boer et al., 2020). SZ used less complex sentences compared to HC with lower length. As we predicted, mean sentence length was negatively correlated with TALD scores. These findings support a "simplified syntax model" (Thomas et al., 1987; Özcan and Kuruoğlu, 2018; de Boer et al., 2020) in schizophrenia. As SZ in this study had relatively high negative thought and language symptoms (TALD-Subjective Negative *M:* 11.92 *SD:* 7.94; TALD-Objective Negative *M:* 5.92 *SD:* 4.87), we found strong correlations between TALD (Total) score and mean sentence length. However, the correlation results of TALD four factors scores showed that not only negative factors but also positive factors scores had negative correlations with mean sentence length. The MATTR calculates lexical diversity (Covington and McFall, 2010). Parallel with the literature (Manschreck et al., 1981; Manschreck et al., 1985), SZ speech had low lexical diversity, pointing to increased

Table 2
The results of hypothesis-driven NLP analyses part.

Variable	SZ ($N = 38$)	HC ($N = 38$)	Test statistics	p	Effect size
Fluency					
Mean Sentence Length	4.681 ± 1.492	6.571 ± 1.684	5.178	**0.001**^{*a}	1.19
Lexical Richness					
Total Completed Words	410 (210–820)	415 (240–540)	0.197	0.843^{b}	0.022
Moving Average Type-Token Ratio	0.814 (0.766–0.845)	0.839 (0.817–0.866)	2.639	**0.008**^{*b}	0.303
Grammar					
Average First-Person Singular Pronoun Usage (i.e. *ben*)	6 (3–16)	3 (1–6)	3.076	**0.002**^{*b}	0.353

Bold values indicate significance at level $*p < 0.018$. ^{a}: p value was obtained from t-test. ^{b}: p value was obtained from Mann Whitney U test.

Table 3
Correlations of the TALD factors and variables in hypothesis-driven NLP analyses part ($N = 76$).

Variable	TALD	TALD	TALD	TALD	
	Objective Positive	Subjective Positive	Objective Positive	Subjective Positive	TALD Total
	r, p	*r, p*	*r, p*	*r, p*	*r, p*
Fluency					
Mean Sentence Length	**−0.282, 0.014***	**−0.319, 0.005****	**−0.269, 0.019***	**−0.312, 0.006****	**−0.367, 0.001****
Lexical Richness					
Total Completed Words	0.002, 0.989	0.096, 0.412	0.094, 0.421	−0.055, 0.637	0.072, 0.530
Moving Average Type-Token Ratio	**−0.234, 0.042***	**−0.284, 0.013***	−0.170, 0.141	**−0.300, 0.008****	**−0.296, 0.009****
Grammar					
Average First-Person Singular Pronoun Usage (i.e. *ben*)	0.183, 0.114	**−0.284, 0.013***	**0.253, 0.028***	0.116, 0.319	0.263, 0.020

TALD = Thought and Language Disorder Scale, $*p < 0.05$ $**p < 0.01$.

Table 4
The results of parts-of-speech tagging (exploratory NLP analyses part).

Variable	SZ ($N = 38$)	HC ($N = 38$)	Test statistics	p	Effect size
Adjective	4.10 (2.44–6.92)	6.21 (4.78–8.04)	2.348	0.019^{a}	0.269
Adverb	8.56 (4.41–11.89)	9.62 (7.70–12.00)	1.158	0.247^{a}	0.133
Coordinating Conjunction	7.17 (3.90–8.57)	9.14 (6.67–12.00)	2.717	**0.007**^{*a}	0.312
Determiner	2.83 (1.71–4.12)	3.39 (2.45–4.00)	1.434	0.152^{a}	0.165
Noun	15.54 ± 7.61	18.51 ± 6.98	−1.772	0.081^{b}	0.406
Verb	14.94 (7.30–23.33)	14.51 (10.00–19.35)	−0.431	0.666^{a}	0.049
Pronoun	7.55 ± 3.62	7.55 ± 3.31	0.002	0.998^{b}	0.000

Adjectives are words that typically modify nouns and specify their properties or attributes, e.g., big/*büyük,* one/*bir,* blue/*mavi. Adverbs* are words that typically modify verbs for such categories as time, place, direction, or manner. They may also modify adjectives and other adverbs, e.g., she read *well/güzel* okudu. *Coordinating conjunctions* are words that link words or larger constituents without syntactically subordinating one to the other and express a semantic relationship between them, e.g., and/*ve,* or/*ya da,* but/*ama. Determiners* are words that modify nouns or noun phrases and express the reference of the noun phrase in context, e.g., this/*bu,* that/*şu,* a-an/*bir. Nouns* are a part of speech typically denoting a person, place, thing, animal or idea, e.g., girl/*kız,* cat/*kedi,* tree/*ağaç. Verbs* are members of the syntactic class of words that typically signal events and actions, can constitute a minimal predicate in a clause, and govern the number and types of other constituents which may occur in the clause, e.g., run/*koş,* eat/*yedi. Pronouns* are words that substitute for nouns or noun phrases, whose meaning is recoverable from the linguistic or extralinguistic context, e.g., I/*ben,* you/*sen,* everybody/*herkes* (Universal Dependencies, 2023). Bold values indicate significance at level $*p < 0.014$. ^{a}: p value was obtained from Mann Whitney U test. ^{b}: p value was obtained from t-test.

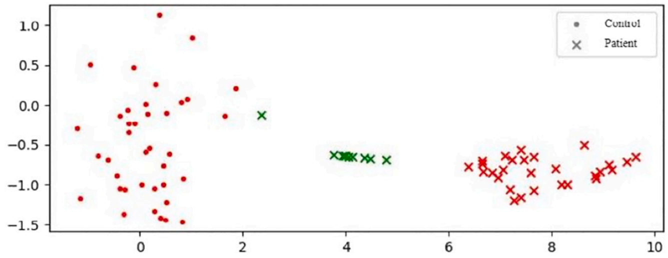

Fig. 2. The visualization of document embeddings where t-SNE algorithm was used to project the vectors to a lower dimensional space. The green marks show the participants to which K-Means did not assign a cluster that matched with their true label and the red marks show the participants whose clusters matched with their true label.

repetitiousness.

Many other studies have noted differences in first-person singular pronoun usage, carrying information about 'the self' (Maatz, 2014; Birnbaum et al., 2019; Hitczenko et al., 2021; Tang et al., 2021; Lundin et al., 2023). In the analysis of the Twitter language of SZ users, Birnbaum et al. (2019) found increased use of first-person singular and plural pronouns among posts from SZ. In an exploratory study, Tang et al. (2021) highlighted that a high percentage of participants with schizophrenia spectrum disorders preferred first-person singular pronouns in their speech. The current study also shows that SZ used significantly more first-person singular pronouns than HC and this result was correlated with negative factors scores of TALD. This finding could result from social factors relevant in SZ such as social isolation and a lower sense of social connectedness (Lundin et al., 2023) or suicidal thoughts and behaviors across diagnoses (Fineberg et al., 2016; Homan et al., 2022).

In POS, SZ used fewer coordination conjunctions (i.e., and, or, but) than HC. Less usage of coordinating conjunctions in SZ suggests that SZ used less complex sentences, and this is correlated with negative symptoms (Hitczenko et al., 2021). We are the first to report a significant decrease in coordinating conjunction among SZ. Our methodology utilizing Word2Vec showed that SZ had higher semantic similarity than

HC. This could depend on the process in lexical selections of SZ. Alonso-Sánchez et al. (2022) found in their current study that higher semantic similarity in SZ could be the result of failure in interference control as they detected a significant association between semantic similarity and Stroop Task scores. In this sense, it could be proposed that lexical selections of SZ can contribute to redundant discourse with reduced information content (Alonso-Sánchez et al., 2022). K-Means clustering algorithm could differentiate between SZ and HC into two distinct groups in the context of semantic similarity with high accuracy, 86.84 %. The algorithm in this methodology is sensitive to the content of the transcribed speech data rather than the true labels (SZ vs. HC), which is important in training supervised learning algorithms. However, as seen in Fig. 2, it missed some patients. This phenomenon implies that schizophrenia should be regarded as a spectrum where some patients show fewer deficits than others in their speech.

The limitations of our study are the following. The number of participants is modest. Having only clinical participants in a stable phase of schizophrenia (i.e., patients with more pronounced negative symptoms) may limit the generalizability of our findings to this population. We could not assess antipsychotic side-effects with a rating scale, and not control premorbid intelligence, which can affect the results in the patients. K-Means could capture hidden language patterns in the speech as it does not need pre-established labels of the groups (i.e., SZ and HC). However, because of this exact reason, the language patterns that distinguished the groups could become vague and hard to interpret in the light of psychopathology. It is not possible to say clearly on what basis the algorithm makes the decisions in the speech samples. In this current digital era, ethical points of representatives of using artificial intelligence to deal with language patterns, such as data responsibility and privacy, explainability, and trust should also be considered. Future studies should aim to collect larger samples, to compare the results of participants in different phases of the illness (e.g., first-episode schizophrenia), and to conduct studies in different languages. The last one is also important because using various language samples with different linguistic backgrounds can enrich NLP literature in SZ (Hitczenko et al., 2021).

To our knowledge, this is one of the most comprehensive studies evaluating language disturbances in SZ via using numerous NLP measurements together, and this is the first in the Turkish language with many aspects, such as utilizing POS and Word2Vec. The studies analysing 'the language of schizophrenia' by applying NLP or, in a more general sense, computational methods, and manual linguistic analyses are very limited in Turkish language. In one of the first studies, Mete et al. (1993) examined the language content in acute phase psychosis by using computer content analytic procedure and found that the speech content of Turkish patients with SZ is considerably similar to the speech content of American subjects, which was previously investigated, but certain dissimilarities appeared to reflect the impact of culture on the manifestations of SZ. Recently, Çokal et al. (2022) focused on referential noun phrases in Turkish-speaking SZ with and without formal thought disorder by performing manual linguistic analyses. Their findings confirmed that the language in SZ manifests through specific linguistic effects in the referential structure of meaning as mediated by grammar. Within this framework, our findings are promising to describe the language in schizophrenia, showing that the results of the NLP analyses in Turkish are mostly language-independent (e.g. lower mean sentence length and lexical diversity, higher usage first-person singular pronoun, higher semantic similarity in schizophrenia patients and discrimination of schizophrenic language with high accuracy via utilizing Word2Vec with K-means clustering algorithm) with some possible language-dependent exceptions (e.g. coordinating conjunctions) as we do not have any finding in other languages and propose that NLP measurements may allow for rapid and objective measurements of linguistic features, many of which are hard to detect by human raters.

Contributors

Conceived and designed the experiments: TÇ & TT. Collected data: TÇ, EM. Analysed data: NS, TÇ, AK. Wrote the first draft of the paper: TÇ. Contributed to and critically reviewed the manuscript: TT, NS, EM, AK, EY.

Funding

This work was supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) (Grant Number: 120K729) to TT.

CRediT authorship contribution statement

Tuğçe Çabuk: Conceptualization, Data curation, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Nurullah Sevim:** Formal analysis, Writing – review & editing. **Emre Mutlu:** Data curation, Investigation, Writing – review & editing. **A. Elif Anıl Yağcıoğlu:** Data curation, Methodology, Writing – review & editing. **Aykut Koç:** Formal analysis, Methodology, Writing – review & editing. **Timothea Toulopoulou:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing.

Declaration of competing interest

None of the authors has any conflicts of interest to report.

Acknowledgment

We want to thank all the participants in this study and appreciate their willingness to help. We also thank Cennet Çabuk, who helped transcribe the speech data, and María Francisca Alonso-Sánchez, who kindly shared her opinions about the result of semantic similarity. A. Koç acknowledges the 2023 BAGEP Award.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.schres.2024.02.026.

References

Alonso-Sánchez, M.F., Ford, S.D., MacKinley, M., Silva, A., Limongi, R., Palaniyappan, L., 2022. Progressive changes in descriptive discourse in First Episode Schizophrenia: a longitudinal computational semantics study. Schizophrenia 8, 1–9. https://doi.org/10.1038/s41537-022-00246-8.

Bambini, V., Frau, F., Bischetti, L., Cuoco, F., Bechi, M., Buonocore, M., Agostoni, G., Ferri, I., Sapienza, J., Martini, F., Spangaro, M., Bigai, G., Cocchi, F., Cavallaro, R., Bosia, M., 2022. Deconstructing heterogeneity in schizophrenia through language: a semi-automated linguistic analysis and data-driven clustering approach. Schizophrenia 8. https://doi.org/10.1038/s41537-022-00306-z.

Bedi, G., Carrillo, F., Cecchi, G.A., Slezak, D.F., Sigman, M., Mota, N.B., Ribeiro, S., Javitt, D.C., Copelli, M., Corcoran, M., 2015. Automated analysis of free speech predicts psychosis onset in high-risk youths. NPJ Schizophr. 1 https://doi.org/10.1038/npjschz.2015.30.

Birnbaum, M.L., Ernala, S.K., Rizvi, A.F., Arenare, E., Van Meter, R., De Choudhury, M., Kane, J.M., 2019. Detecting relapse in youth with psychotic disorders utilizing patient-generated and patient-contributed digital data from Facebook. NPJ Schizophr. 5, 1–9. https://doi.org/10.1038/s41537-019-0085-9.

Çabuk, T., Mutlu, E., Toulopoulou, T., 2023. Thought and language disorder as a possible endophenotype in schizophrenia: evidence from patients and their unaffected siblings. Schizophr. Res. 254, 78–80. https://doi.org/10.1016/j.schres.2023.02.005.

Castellani, U., Rossato, E., Murino, V., Bellani, M., Rambaldelli, G., Perlini, C., Tomelleri, L., Tansella, M., Brambilla, P., 2012. Classification of schizophrenia using feature-based morphometry. J. Neural Transm. 119, 395–404. https://doi.org/10.1007/s00702-011-0693-7.

Çokal, D., Palominos-Flores, C., Yalınçetin, B., Türe-Abacı, Bora, E., Hinzen, W., 2022. Referential noun phrases distribute differently in Turkish speakers with schizophrenia. Schizophr. Res. https://doi.org/10.1016/j.schres.2022.06.024.

Corcoran, C.M., Carillo, F., Fernandez-Slezak, D., Bedi, G., Klim, C., Javitt, D.C., Bearden, C.E., Cecchi, G.A., 2018. Prediction of psychosis across protocols and risk cohorts using automated language analysis. World Psychiatry 17, 67–75.

Corcoran, C.M., Mittal, V.A., Bearden, C.E., Gur, E., Hitczenko, K., Bilgrami, Z., Savic, A., Cecchi, G.A., Wolff, P., 2020. Language as a biomarker for psychosis: a natural language processing approach. Schizophr. Res. 226, 158–166. https://doi.org/10.1016/j.schres.2020.04.032.

Corona-Hernández, H., de Boer, J.N., Brederoo, S.G., Voppel, A.E., Sommer, I.E.C., 2022. Assessing coherence through linguistic connectives: analysis of speech in patients with schizophrenia-spectrum disorders. Schizophr. Res. https://doi.org/10.1016/j.schres.2022.06.013.

Covington, M.A., McFall, J.D., 2010. Cutting the gordian knot: the moving-average type-token ratio (MATTR). J. Quant. Linguist. 17, 94–100. https://doi.org/10.1080/09296171003643098.

de Boer, J.N., van Hoogdalem, M., Mandl, R.C.W., Brummelman, J., Voppel, A.E., Begemann, M.J.H., van Dellen, E., Wijnen, F.N.K., Sommer, I.E.C., 2020. Language in schizophrenia: relation with diagnosis, symptomatology and white matter tracts. NPJ Schizophr. 6, 1–10. https://doi.org/10.1038/s41537-020-0099-3.

Elvevåg, B., Foltz, P.W., Weinberger, D.R., Goldberg, T.E., 2007. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. Schizophr. Res. 93, 304–316. https://doi.org/10.1016/j.schres.2007.03.001.

Fergadiotis, G., Wright, H.H., Green, S.B., 2015. Psychometric evaluation of lexical diversity indices: assessing length effects. J. Speech Lang. Hear. Res. 58 (3), 840–852. https://doi.org/10.1044/2015_JSLHR-L-14-0280. Jun; PMID: 25766139; PMCID: PMC4490052.

Fineberg, S.K., Leavitt, J., Deutsch-Link, S., Dealy, S., Landry, C.D., Pirruccio, K., Shea, S., Trent, S., Cecchi, G., Corlett, P.R., 2016. Self-reference in psychosis and depression: a language marker of illness. Psychol. Med. 46, 2605–2615. https://doi.org/10.1017/S0033291716001215.

Foltz, P.W., Rosenstein, M., Elvevåg, B., 2016. Detecting clinically significant events through automated language analysis: quo Imus? NPJ Schizophr. 2, 15054. https://doi.org/10.1038/npjschz.2015.54.

Gardner, D.M., Murphy, A.L., O'Donnell, H., Centorrino, F., Baldessarini, R.J., 2010. International consensus study of antipsychotic dosing. Am. J. Psychiatry 167, 686–693. https://doi.org/10.1176/appi.ajp.2009.09060802.

Gutiérrez, E.D., Corlett, P.R., Corcoran, C.M., Cecchi, G.A., 2017. Using automated metaphor identification to aid in detection and prediction of first-episode schizophrenia. EMNLP 2017 - Conf. Empir. Methods Nat. Lang. Process. Proc. 2923–2930. doi:10.18653/v1/d17-1316.

Guy, W. (Ed.), 1976. ECDEU Assessment Manual for Psychopharmacology. US Department of Heath, Education, and Welfare Public Health Service Alcohol, Drug Abuse, and Mental Health Administration, Rockville, MD.

Hinton, G., Roweis, S., 2003. Stochastic neighbor embedding. Adv. Neural Inf. Process. Syst.

Hitczenko, K., Mittal, V.A., Goldrick, M., 2021. Understanding language abnormalities and associated clinical markers in psychosis: the promise of computational methods. Schizophr. Bull. 47, 344–362. https://doi.org/10.1093/schbul/sbaa141.

Homan, S., Gabi, M., Klee, N., Bachmann, S., Moser, A.M., Duri', M., Michel, S., Bertram, A.M., Maatz, A., Seiler, G., Stark, E., Kleim, B., 2022. Linguistic features of suicidal thoughts and behaviors: a systematic review. Clin. Psychol. Rev. 95, 102161 https://doi.org/10.1016/j.cpr.2022.102161.

Kircher, T., Krug, A., Stratmann, M., Ghazi, S., Schales, C., Frauenheim, M., Turner, L., Fährmann, P., Hornig, T., Katzev, M., Grosvald, M., Müller-Isberner, R., Nagels, A., 2014. A rating scale for the assessment of objective and subjective formal thought and language disorder (TALD). Schizophr. Res. 160, 216–221. https://doi.org/10.1016/j.schres.2014.10.024.

Köksal, A., 2018. AKOKSAL/Turkish-word2vec: pre-trained word2vec model for Turkish. GitHub. https://github.com/akoksal/Turkish-Word2Vec.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Lundin, N.B., Cowan, H.R., Singh, D.K., Moe, A.M., 2023. Lower cohesion and altered first-person pronoun usage in the spoken life narratives of individuals with schizophrenia. Schizophr. Res. https://doi.org/10.1016/j.schres.2023.04.001.

Maatz, A., 2014. Use of the first-person pronoun in schizophrenia. Br. J. Psychiatry 205, 409. https://doi.org/10.1192/bjp.205.5.409.

MacHado, A.M.C., 2007. Multiple testing correction in medical image analysis. J. Math. Imaging Vis. 29, 107–117. https://doi.org/10.1007/s10851-007-0034-5.

MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. In: Le Cam, L.M., Neyman, J. (Eds.), Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. University of California Press, California, pp. 281–297.

Manschreck, T.C., Maher, B.A., Ader, D.N., 1981. Formal thought disorder, the type token ratio, and disturbed voluntary motor movement in schizophrenia. Br. J. Psychiatry 139, 7–15.

Manschreck, T.C., Ames, D., Maher, B.A., Hoover, T.M., 1985. Repetition in schizophrenic speech. Lang. Speech 28, 255–268. https://doi.org/10.1177/002383098502800303.

Mete, L., Schnurr, P.P., Rosenberg, S.D., Oxman, T.E., Doganer, I., Sorias, S., 1993. Language content and schizophrenia in acute phase Turkish patients. Soc. Psychiatry Psychiatr. Epidemiol. 28, 275–280. https://doi.org/10.1007/BF00795907.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. In: 1st Int. Conf. Learn. Represent. ICLR 2013 - Work. Track Proc, pp. 1–12.

Mota, N.B., Copelli, M., Ribeiro, S., 2017. Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance. NPJ Schizophr. 3, 1–10. https://doi.org/10.1038/s41537-017-0019-3.

Mutlu, E., Yazıcı, M.K., Barışkın, E., Ertuğrul, A., Gürel, C., Gürkan, Ş., Göka, E., Yağcıoğlu, A.E.A., 2019. Examination of formal thought disorder and its clinical correlates with the Turkish Version of the Thought and Language Disorder Scale (TALD-TR) in schizophrenia. Compr. Psychiatry 93, 7–13. https://doi.org/10.1016/j.comppsych.2019.06.003.

Özcan, A., Kuruoğlu, G., 2018. Sentence length of Turkish patients with schizophrenia. International Journal of Psycho-Educational Sciences. 7, 68–73.

Rezaii, N., Walker, E., Wolff, P., 2019. A machine learning approach to predicting psychosis using semantic density and latent content analysis. NPJ Schizophr. 5 https://doi.org/10.1038/s41537-019-0077-9.

Tan, E.J., Meyer, D., Neill, E., Rossell, S.L., 2021. Investigating the diagnostic utility of speech patterns in schizophrenia and their symptom associations. Schizophr. Res. 238, 91–98. https://doi.org/10.1016/j.schres.2021.10.003.

Tang, S.X., Kriz, R., Cho, S., Park, S.J., Harowitz, J., Gur, R.E., Bhati, M.T., Wolf, D.H., Sedoc, J., Liberman, M.Y., 2021. Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. NPJ Schizophr. 7, 1–8. https://doi.org/10.1038/s41537-021-00154-3.

Thomas, P., King, K., Fraser, W.I., 1987. Positive and negative symptoms of schizophrenia and linguistic performance. Acta Psychiatr. Scand. 76, 144–151. https://doi.org/10.1111/j.1600-0447.1987.tb02877.x.

Voppel, A.E., de Boer, J.N., Brederoo, S.G., Schnack, H.G., Sommer, I.E.C., 2021. Quantified language connectedness in schizophrenia-spectrum disorders. Psychiatry Res. 304, 114130 https://doi.org/10.1016/j.psychres.2021.114130.

Ziv, I., Baram, H., Bar, K., Zilberstein, V., Itzikowitz, S., Harel, E.V., Dershowitz, N., 2021. Morphological characteristics of spoken language in schizophrenia patients - an exploratory study. Scand. J. Psychol. 63 (2), 91–99.

Update

Schizophrenia Research

Volume 274, Issue , December 2024, Page 178–179

DOI: https://doi.org/10.1016/j.schres.2024.09.017

Contents lists available at ScienceDirect

Schizophrenia Research

journal homepage: www.elsevier.com/locate/schres

Corrigendum

Corrigendum to “Natural language processing for defining linguistic features in schizophrenia: A sample from Turkish speakers” [Schizophr. Res. 266 (2024) 183–189]

Tuğçe Çabuk [a], Nurullah Sevim [b], Emre Mutlu [c], A. Elif Anıl Yağcıoğlu [c], Aykut Koç [b], Timothea Toulopoulou [a,d,e,*]

[a] Department of Psychology, National Magnetic Resonance Research Center (UMRAM) & Aysel Sabuncu Brain Research Center, Bilkent University, Bilkent, 06800 Ankara, Turkey
[b] Department of Electrical and Electronics Engineering, National Magnetic Resonance Research Center (UMRAM), Bilkent University, Bilkent, 06800 Ankara, Turkey
[c] Department of Psychiatry, Hacettepe University, Faculty of Medicine, Sıhhiye, 06230 Ankara, Turkey
[d] 1st Department of Psychiatry, National and Kapodistrian University of Athens, Athens, Greece
[e] Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, USA

The authors regret the following errors detected in Table 3:

- Page 187, Table 3, second variable name “Subjective Positive” should be “Subjective Negative”
- Page 187, Table 3, third variable name “Objective Positive” should be “Objective Negative”

You can find below the corrected Table 3.

DOI of original article: https://doi.org/10.1016/j.schres.2024.02.026.

* Corresponding author at: Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, USA.
E-mail address: timothea.toulopoulou@mssm.edu (T. Toulopoulou).

https://doi.org/10.1016/j.schres.2024.09.017

Available online 23 September 2024

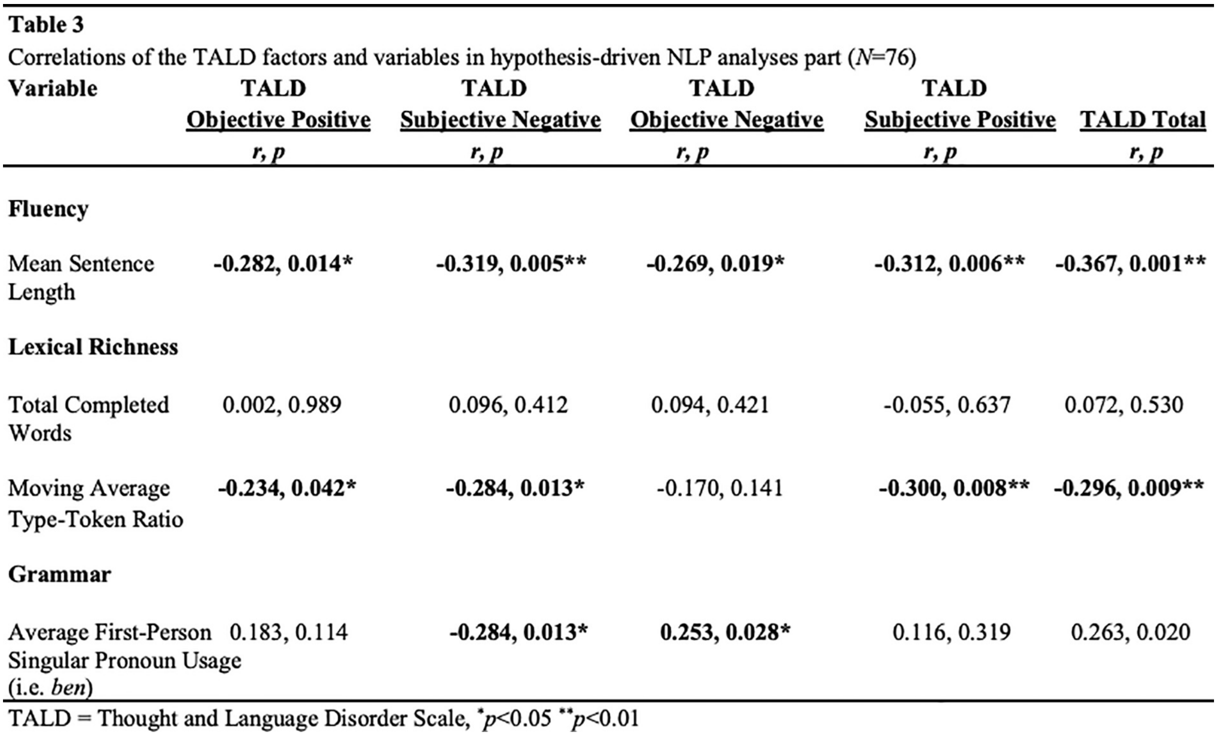

Table 3

Correlations of the TALD factors and variables in hypothesis-driven NLP analyses part (*N*=76)

Variable	**TALD**	**TALD**	**TALD**	**TALD**	
	Objective Positive	**Subjective Negative**	**Objective Negative**	**Subjective Positive**	**TALD Total**
	r, p	***r, p***	***r, p***	***r, p***	***r, p***
Fluency					
Mean Sentence Length	**-0.282, 0.014***	**-0.319, 0.005****	**-0.269, 0.019***	**-0.312, 0.006****	**-0.367, 0.001****
Lexical Richness					
Total Completed Words	0.002, 0.989	0.096, 0.412	0.094, 0.421	-0.055, 0.637	0.072, 0.530
Moving Average Type-Token Ratio	**-0.234, 0.042***	**-0.284, 0.013***	-0.170, 0.141	**-0.300, 0.008****	**-0.296, 0.009****
Grammar					
Average First-Person Singular Pronoun Usage (i.e. *ben*)	0.183, 0.114	**-0.284, 0.013***	**0.253, 0.028***	0.116, 0.319	0.263, 0.020

TALD = Thought and Language Disorder Scale, $^{*}p<0.05$ $^{**}p<0.01$

The authors would like to apologise for any inconvenience caused.