Morphological characteristics of spoken language in schizophrenia patients – An exploratory study

Ido Ziv – Ph.D[1]

Heli Baram – M.A[2]

Kfir Bar – Ph.D[3]

Vered Zilberstein – M.A[4]

Shmuel Itzikobitz – Ph.D [5]

Eran V. Harel – M.D[6]

Nachum Dershowitz[7]

Running head: Morphological characteristics in schizophrenia patients

[1] Psychology Department, The College of Management – Academic Studies, Rishon LeZion, Israel

[2] Psychology Department, – Ruppin Academic Center, Ruppin, Isreal

[3] School of Computer Science, The College of Management – Academic Studies, Rishon LeZion, Israel

[4] School of Computer Science, Tel Aviv University, Tel Aviv, Israel

[5] School of Computer Science, The College of Management – Academic Studies, Rishon LeZion, Israel

[6] Be'er Ya'akov Medical Center for Mental Health, Be'er Ya'akov, Israel

[7] School of Computer Science, Tel Aviv University, Tel Aviv, Israel

Acknowledgements

The authors wish to claim for equal contribution.

The authors also thank the Research Authority of the College of Management – Academic Studies, Rishon Lezion, Israel, for supporting this research (grant # 707015).

Keywords: Mental illness, Schizophrenia, Psychosis, Language Processing, Morphology, Lexeme.

Author Note

Please address correspondence to Ido Ziv, Psychology Department, The College of Management – Academic Studies, Israel, idozi@colman.ac.il.

Corresponding author:
Dr. Ido Ziv
Psychology Department,
The College of Management Academic Studies
Rabin Boulevard #7,
Rishon LeZion, Zip code 7549071
Israel
Email: idozi@colman.ac.il
Phones: 972 528935749

Morphological characteristics of spoken language in schizophrenia patients – An exploratory study

Abstract

Psychosis is diagnosed based on disruptions in the structure and use of language, including reduced syntactic complexity, derailment, and tangentiality. With the development of computational analysis, natural language processing (NLP) techniques are used in many areas of life to make evaluations and inferences regarding people's thoughts, feelings and behavior. The present study explores morphological characteristic of schizophrenia inpatients using NLP. Transcripts of recorded stories by 49 male subjects (24 inpatients diagnosed with schizophrenia and 25 controls) about 14 Thematic Apperception Test (TAT) pictures were morphologically analyzed. Relative to the control group, the schizophrenic inpatients employed (1) a similar ratio of nouns, but fewer verbs, adjectives and adverbs; (2) a higher ratio of lemmas to token (LTR) and type to token (TTR); (3) a smaller gap between LTR and TTR; and (4) greater use of the first person. The results were cross-verified using three well-known fitting classifier algorithms (Random Forest, XGBoost and a support vector machine). Tests of prediction accuracy, precision and recall found correct attribution of patients to the schizophrenia group at a rate of between 80% and 90%. Overall, the results suggest that the language of schizophrenic inpatients is significantly different from that of healthy controls, being morphologically less complex, more associative and more focused on the self. The findings support NLP analysis as a complementary addition to the traditional clinical psychosis evaluation for schizophrenia.

Key Words: Natural Language Processing, Mental illness, Schizophrenia, Psychosis, Morphology

1. Introduction

Based on the assumption that language conveys cognitive processes like thoughts and feelings, natural language processing (NLP) techniques are used to make evaluations and inferences regarding people's thoughts, feelings and behavior (Cambria & White, 2014). For the past few decades NLP and related computational tools have been used extensively in marketing, with companies analyzing customers' social media posts and emails to generate targeted advertisements (Calvo et al., 2017). However, NLP has also been applied in research advancing the use of artificial intelligence in medicine (Peek et al., 2015). In particular, over the past decade NLP techniques such as text mining have slowly gained a presence in mental health research (Dreisbach et al., 2019). Promising research directions in the mental health realm include developing monitoring applications (Morris et al., 2018), training in mindfulness (Toivonen, Zernicke & Carlson, 2017), one-on-one text interventions (Hoermann et al., 2017), predicting depression (Imahori, 2018; Qureshi et al., 2019), predicting remission and recovery after a first psychotic episode (Koutsouleris et al., 2016; Leighton et al., 2019), and identifying suicidal ideation and suicide attempts using rule-based and hybrid machine learning approaches (Carson et al, 2019; Fernandes et al., 2018).

Though the use of NLP for mental health research and practice is growing, its application to the schizophrenia spectrum has only begun to realize its potential. The common approach used by researchers implementing NLP to study schizophrenia is to develop classifiers predicting the presence or future development of the disease, primarily by identifying signs of associative thinking (Bar et al., 2019; Bedi et al., 2015; Iter, Yoon & Jurafsky, 2018; Rezaii, Walker & Wolff, 2019). However, to the best of our awareness, there is no systematic characterization of morphological differences in the language used by patients diagnosed with schizophrenia relative to the general population. The current study as well as other studies (e.g. De Boer et al., 2020b) aims to help fill this gap.

We used a morphological part-of-speech tagger developed by Ben-Gurion University (Adler, 2007) to process short narratives delivered orally by 49 Israeli males, half of them schizophrenia inpatients and the others, members of the general population with no indications of mental illness. The narratives were elicited via the Thematic Apperception Test (TAT; Murray, 1943) and examined for signs of morphological differences between the two groups in three areas: the prevalence of different parts of speech; use of inflection; and an inward vs. outward focus (as seen in the use of the first vs. third person).

This paper offers two main contributions. First, we add to the literature on linguistic markers for schizophrenia by systematically characterizing morphological differences in the spoken language of schizophrenia patients and healthy controls. Second, we contribute to the growing use of computational tools and, specifically, natural language processing, in the study and (potentially) diagnosis of thought disorders and other forms of mental illness.

1.1 Psychosis and Linguistic Expression

Disturbances in the normal structure and use of language were originally identified as manifestations of thought disorders, such as schizophrenia, by Bleuler (1911). Language disturbances prevalent in schizophrenia involve violations of "the syntactical and semantic conventions which govern language usage" (Andreasen & Grove, 1986, p. 474), and include reduced syntactic complexity, poverty of speech, greater use of concrete terms, loss of semantic coherence, derailment (jumping from one thought to another drifting away from the topic discussed), and tangentiality (a tendency to speak around the topic). The Diagnostic and Statistical Manual of Mental Disorders–5th edition (DSM–5) (American Psychological Association, 2013) lists disorganized speech as well as psychotic states – a broad phenomenon that includes thought, perception, and language disruptions (Chan, 2001) – among the criteria for making a diagnosis of schizophrenia. Currently, only well-trained examiners, mainly psychiatrists, perform differential diagnosis for schizophrenia using a qualitative psychiatric interview or a more structured interview like the MINI Mental State Examination. Psychosis is diagnosed when specific speech disturbances are detected (Grinker, 2010).

Language disruptions in schizophrenia are in general well-understood. Many studies report deficits in grammar tasks that require syntactic processing, in both production (e.g., arranging words to form a sentence) and comprehension (e.g., the Token Test) (Condray et al., 2002). Impairments in lexical performance have also commonly been reported for both expressive and receptive tasks (Bokat & Goldberg, 2003). Such disruptions may be predicted of schizophrenia before psychotic symptoms are visible. Bearden et al. (2011) studied thought and communication disturbances as predictors of outcomes in adolescents identified as putatively prodromal for psychosis. Despite the absence of fully psychotic symptoms, they found signs of communication disturbances in putatively prodromal individuals to be qualitatively similar to those seen in schizophrenia, and predictive of both conversion to psychosis and psychosocial outcomes.

However, while abnormal lexical and syntactic functions in schizophrenia are well-documented (Covington et al., 2005), less is known about morphological disruptions. Since the seminal investigation of language in schizophrenia by Kleist (1914), who reported deficits in affected patients in multiple areas of language, we are aware of only a few brief mentions of morphological abnormalities in the schizophrenia literature (Corcoran et al., 2018; Covington et al., 2005; DeLisi et al., 1997). Morphology is concerned with the internal structure of words and their relationships. It examines how inflected forms of lexemes (units of meaning) are related (e.g., the inflected forms *runs* and *ran* are related to the base word *run*), and how new words are formed based on patterns of form–meaning correspondence between existing words (e.g., the word *run* gives rise to the new word *runner*). Given the strong association between schizophrenia and language impairments, it seems worth further investigating morphological manifestations of this disorder.

1.2 Natural Language Processing and Thought Disorders

Over the last decade, researchers have increasingly used computer science tools to study mental health in general as well as thought disruptions aiming to develop reliable approaches to mark and distinguish between mental health patients and control groups. For example, Al-Mosaiwi & Johnstone (2018) applied textual analysis to investigate absolutist thinking as a cognitive distortion of anxiety and depression. Based on textual analysis of 63 Internet forums using the Linguistic Inquiry and Word Count (LIWC) software package, they found that anxiety, depression, and suicidal ideation forums contained more absolutist words than control forums. Bedi et al. (2015) based on Elvevaag, Foltz & Rosenstein (2010) latent semantic analysis used automated speech analysis and machine learning to predict later psychosis onset among 34 youths at clinical high risk for psychosis. They found that semantic coherence, maximum phrase length and use of determiners predicted later psychosis development with high accuracy. This approach was further developed by Corcoran et al. (2018), who reported F83% accuracy in predicting the onset of psychosis in adolescents and young adults. Mota et al. (2012) employed speech graph attributes (SGA) analysis approach to represent semantic and grammatical relationships between words. Their findings showed that using a binary classifier based on those graphs enabled them to distinguish between schizophrenics and manic patients with a sensitivity and specificity of almost 94%. Mota et al. (2014) replicated SGA approach for studying of speech among mental health patients while distinguishing between dream reports of bipolar and control subjects. Further results implementing SGA also allowed to clarify schizophrenia and bipolar thought process. Looking into their speech connectedness, Palaniyappan et al. (2019) found schizophrenia patients present with less connected speech output.

Birnbaum et al. (2017) used machine learning along with clinical appraisals to identify markers of schizophrenia in social media. They extracted Twitter timeline data from 671 users with self-disclosed diagnoses of schizophrenia and built a classifier aiming to distinguish users with schizophrenia from healthy controls. Their classifier distinguished between the two groups with a mean accuracy of 88% using linguistic data alone. Similar results were found by Iter, Yoon and Jurafsky (2018).

For more details about the usage of NLP in psychiatry and the use of language processing and speech analysis for studying, identifying and diagnosing schizophrenia, we refer the reader to a number of recent reviews - Corcoran & Cecchi, (2020), Corcoran, Mittal, Bearden et al., (2020) and De Boer et al., (2020a).

1.3 The Present Study

As presented above, several research groups have developed classifiers which predict the presence or future development of schizophrenia by focusing on linguistic markers (e.g., incoherence or associative thought). However, to the best of our awareness, there is no systematic characterization of morphologic differences in language use by patients diagnosed with schizophrenia relative to healthy controls.

In the current study, we explore morphological differences in transcribed spoken language of schizophrenia male inpatients and males of similar ages drawn from the local population. Based on theories of language acquisition and what is known about language disruptions in schizophrenia, we expect to find three areas of discrepancy, as follows:

Hypothesis 1: We expect schizophrenia patients to use less complex language, with fewer verbs, adjectives, and adverbs, and more lemma to token ratio - LTR[1] (a high rate of LTR ratio means less word inflection, indicating greater poverty of language). We expect no difference in the use of nouns, which comprise the largest class of words in most languages

[1] In linguistics, the word lemma refers to the canonical or base form of an inflected word, the one that stands at the head of a dictionary definition (e.g., *ran* and *running* are inflected forms of the lemma *run*). In the present study, we consider any inflected form to be a representative of its lemma. See the Data Analysis section for a detailed description of our method.

(including Hebrew, the language of the present study) and are considered the most basic part of speech.

Hypothesis 2: We expect that the language of schizophrenia patients will include higher type to token ration – TTR, indicating higher lexical diversity-(Fergadiotis & Wright, 2011).

Hypothesis 3: The difference between LTR ratio and TTR ratio will be bigger among the control group, reflecting higher rate of morphological word inflection considered as progressive language usage.

Hypothesis 4: Illness of any kind tends to lead to a more inward focus. Hence, we expect schizophrenia patients to use the first person more than controls, and the third person less.

2. **Methods**

2.1 Participants and Procedure

Forty-nine men aged 19–63 (*M* = 32.29, *SD* = 9.66) participated in the study in exchange for approximately $8. The experimental group comprised 24 inpatients in the Be'er Ya'akov–Ness Ziona Mental Health Center in Israel who were admitted following a diagnosis of schizophrenia. Diagnoses were made by a hospital psychiatrist according to the DSM5 criteria (American Psychiatric Association, 2013) and a full psychiatric interview. The control group comprised 25 participants recruited primarily through the Facebook social media platform.

Exclusion criteria for all participants were as follows: (1) having a history of dependence on drugs or alcohol over the past year; (2) having a past or present neurologic illness; and (3) using less than 500 words total in their transcribed narratives. Control participants also scored below the threshold for subclinical diagnosis of depression and post-traumatic stress disorder (PTSD), and most scored below the threshold for anxiety (see below). Most of the schizophrenia patients scored above the threshold for borderline or mild psychosis symptoms on a standard measure (note that as this group were inpatients being treated through medication, higher scores were not expected). Three participants whose mother tongue was not Hebrew were excluded. The demographic characteristics of the sample by group are presented in Table 1.

The patients were interviewed in a close room at the health center department by one of the research team, and the non-patient participants were interviewed at a room similar to the inpatients room. Interviews lasted approximately 60 minutes. All interviews were conducted in Hebrew. The interviews were recorded and later transcribed for analysis. Participants were assured of anonymity, and told that they were free to end the experiment at any time. Ethics approval was received from the Helsinki Ethical Review Board of the Be'er Ya'akov–Ness Ziona Mental Health Center.

After providing written consent, participants were asked to tell brief stories about 14 pictures drawn from the TAT, based on four open questions (see below). Following this, participants filled in questionnaires containing the demographic questions and the scales relevant to their group attribution (control/experiment). During the analysis stage, a Hebrew-language morphological tagger (Adler, 2007) was used to extract linguistic features from the transcriptions (see under Data Analysis below).

2.2 Measures

2.2.1 Thematic Apperception Test (TAT). Participants were presented with 14 black-and-white images drawn from the TAT. The images used were TAT main clinical diagnostic picture numbers 1, 2, 3BM, 4, 5, 6BM, 7GF, 8BM, 9BM, 12M, 13MF, 13B, 14, and 3GF. The images were chosen to include a mixture of men and women, children and adults. Each picture stand by it self, presented alone and has no relation to the other pictures. Participants were asked to tell a brief story about each image based on four open questions: What led up to the event shown? What is happening at the moment? What are the characters thinking and feeling? What is the outcome of the story? The interviewer remained silent during the respondent's narration and offered no prompts or questions.

2.2.2. Control group measures

Depression. Symptoms of depression were assessed using Beck's Depression Inventory–II (BDI–II; Beck, Steer, & Brown, 1996). The BDI–II is a 21-item inventory rated on a 4-point Likert-type scale (0= "not at all" to 3= "extremely"), with summary scores ranging between 0 and 63. Beck et al. (1996) suggest a preliminary cutoff score of 14 as indicating mild depression, and scores above 19 as indicating moderate depression. The BDI–II has been found to demonstrate high reliability (Gallagher, Nies, & Thompson, 1982). We used the BDI–II Hebrew translation of Hasenson-Atzmon et al. (2016).

PTSD. Symptoms of PTSD were assessed using the PTSD checklist of the DSM–5 (PCL–5; Weathers et al., 2013). The questionnaire contains 20 items that can be divided into four subscales corresponding to the clusters B–E in the DSM–5: intrusion (five items), avoidance (two items), negative alterations in cognitions and mood (seven items), and alterations in arousal and reactivity (six items). The items are rated on a 5-point Likert-type scale (0= "not at all" to 4= "extremely"). Total scores range from 0 to 80, with a preliminary cutoff score of 38 suggested as indicating PTSD (Weathers et al., 2013). The PCL–5 has been found to demonstrate high reliability (Blevins et al., 2015). We used the PCL–5 Hebrew translation by Bensimon et al. (2013).

Anxiety. Symptoms of anxiety were assessed through the State Trait Anxiety Inventory (STAI; Spielberger et al., 1970). The STAI consists of two 20-item self-report measures. The STAI measure of state anxiety (S-anxiety) assesses how respondents feel "right now, at this moment" (e.g., "I feel at ease"; "I feel upset"), and the STAI measure of trait anxiety (T-anxiety) targets how respondents "generally feel" (e.g., "I am a steady person"; "I lack self-confidence"). For each item, respondents are asked to rate themselves on a 4-point Likert scale, ranging from

1= "not at all" to 4="very much so" for S-anxiety, and from 1="almost never" to 4= "almost always" for T-anxiety. Total scores range from 20 to 80, with a preliminary cutoff score of 40 recommended as indicating clinically significant symptoms for the T-Anxiety scale (Knight, Waal-Manning, & Spears, 1983). The STAI has been found to demonstrate high reliability (Barnes, Harp, & Jung, 2002). We used the STAI Hebrew translation of Saka and Gati (2007).

2.2.3. Experimental group measure. Psychosis symptoms were assessed by the Positive And Negative Syndrome Scale–6 (PANSS–6; Østergaard,et al., 2016). The 30-item Positive and Negative Syndrome Scale (PANSS–30) is the most widely used rating scale in schizophrenia, but is relatively long for clinical use. Items in the PANSS–6 are rated on a 7-point scale (0= “not at all” to 6= “extremely”). Total scores range from 0 to 36, with a score of 14 representing the threshold for mild schizophrenia, and scores of 10 to 14 defined as borderline disease or as remission. The PANSS–30 has been found to demonstrate high reliability (Lin et al., 2018). Østergaard found high correlation between PANSS–6 and PANSS–30 total scores (Spearman correlation coefficient=0.86). We used the PANSS–6 Hebrew translation by Katz et al. (2012). Range of positive and negative symptoms are presented in Table 1.

Controls. All participants filled in a demographic questionnaire eliciting data about their age, marital status, children, education, occupation, residence, and income.

2.3 Data Analysis: Natural Language Processing

A morphological and part-of-speech (POS) tagger developed by Ben-Gurion University (hereafter, BGU Tagger; Adler, 2007) was used to extract linguistic features from the transcriptions. The BGU Tagger overcomes the problem of ambiguity in individual words (e.g., the word *book* can be either a noun or a verb) by processing text in a context-sensitive fashion while providing in-depth structural and semantic information about every word. Hebrew is a highly inflected language, based on (usually) three-letter consonantal roots which are modified according to a set of standard patterns to yield words with different meanings (e.g., the root L-M-D yields *lamad*, learn, *limed*, teach, *lamdan*, learned person, and *melamed*, teacher). Words are then inflected for person, gender, number, and tense, and can be combined with affixes that serve as articles, conjunctions, prepositions, and possessive forms. For every word, the BGU Tagger provides information about its part of speech (e.g., noun, verb), its lemma, and (for verbs) its person (first or third) and tense (past, present, and future). The features identified by the BGU Tagger are summarized in Table 2.

The transcribed data were analyzed with the BGU Tagger as follows. For every story, we counted the number of each tested morphological feature (nouns, verbs, adjectives, etc.), and divided that sum by the total number of words in the story. Additionally, we measured the rate of two additional quantitative features: lemmas to token ratio – LTR and type to token ratio - TTR. Both LTR and TTR are used as a measure of unique vocabulary usage. The rate of LTR was calculated by counting the number of lemma or one of its inflections appeared only once (e.g., either *run* or *ran* but not both), divided by the total number of words in that story. In our case, a high rate of LTR means less word inflection, indicating greater morphological poverty of language since as the rate of LTR is higher there are less lemma that are inflected in the text. The rate of TTR was calculated by counting the instances in each story in which a given word (including inflected form) appeared only once, divided by the total number of words. A high rate of TTR reflects less repetition of identical vocabulary (words). High values of TTR indicates higher lexical diversity. Though the denominator for both LTR and TTR is identical, which is the total number words in the running text the nominators are different; LTR has the total number of unique lemmas, and TTR has the total number of unique words. The difference between LTR ratio and TTR ratio reflect the rate of morphological inflection in the text[2]. High difference between LTR ratio and TTR ratio may indicates a rich and creative use of language (using different words inflections to describe the same idea), or associative thinking where a flow of confliction thoughts results in words inflections continually interrupt the flow of thought. Only the latter is indicative of schizophrenia.

3. Results

3.1 Descriptive Analyses

As a first step, descriptive analyses were computed for the control and experimental groups. None of the participants in the control group met the criteria for moderate depression (BDI score above 19; M = 5.3, SD = 4.85) or PTSD (PCL–5 score of 38 or above; M = 7.16, SD = 8). Thirty-two percent of the control group scored above the threshold for high trait anxiety (40 or above on the STAI T-Anxiety scale; M = 38.64, SD = 6.72). This result might reflect the fact that the control group participants lived mainly in the south and north of Israel, areas that over recent years have suffered relatively intense military activity.

All participants in the experimental group were hospital inpatients being treated for schizophrenia, including with pharmaceutical drugs. The PANSS–6 was used to assess the presence of positive and negative schizophrenia symptoms in this group. Among them, 16.7% had PANSS–6 scores under 10 (the threshold for a borderline schizophrenia symptoms' diagnosis), and 50% had scores of 10 to 14. Thus, 66.7% of the inpatients were considered to be in remission from schizophrenia symptoms. The remaining 33.3% had scores somewhat above 14 (mild symptoms presence).

[2] For e.g. the sentence 'this table, my table, is green' will receive a rate of LTR = 0.83 and TTR of 1 = 0.83. However in Hebrew the words 'my table' is the inflection of table ("shulhani") – 'this table, shulhani, is green'. There for the rate in Hebrew of LTR = 0.8 and TTR =1. The difference between LTR and TTR reflect the inflection rate.

Descriptive analyses were also conducted for the 686 narratives (14 stories for each of the 49 participants). These were approximately 89 words long on average, with a range of 82 to 94, and with no difference in total sample length between the two groups t(4???7)=1.5???4, n.s[3]. With respect to the relative weight of the morphological elements, two of the elements examined (noun andthird person) appeared more often in the stories than the other elements. The prevalence of nouns and the third person is not surprising, given that the stories related to characters in pictures. Also the rate of LTR and TTR were relatively high. This may relate merely to the fact that the narratives were fairly short.

*3.1.*Main analyses

For the main analyses, we used a mixed-design 10 X 2 ANCOVA (morphological element [noun, verb, etc.] X group [experimental vs. control]). We treated the morphological language elem)ents as within-participant variables and the group as a between-participants variable. Age, education, income, marital status and place of residence were controlled and treated as covariates. None of the covariates were related.

Before conducting the analyses, we ensured there were no outliers in the data (2 SD above or below the mean; see Ratcliff, 1993). Skewness was < 1 for all variables, and all criteria relating to kurtosis were satisfactory.

The main effect for morphological elements was significant, $F(9,423) = 45.94$, $p < .0001$, $\eta p^2 = 0.92$, and as stated above expected given the pictures and language morphological structure. Though, our main interest, however, was the interaction between the various morphological elements and the groups. As expected, we found a significant dual interaction between morphological elements and the groups, $F(9,423) = 16.718$, $p \leq .0001$, $\eta p^2 = 0.59$.

[3] For a detailed study on the influence of the sample length on TTR rati and other related measurements we the reader to Fergadiotis & Wright (2011) and Fergadiotis, Wright & Green (2015).

Thus, the morphological structure of the language used by the schizophrenia patients differs from that of the control participants.

Figure 1 presents a follow-up *t*-test analysis (with Bonferroni adjustment of the alpha to 0.05/10 = 0.005). As expected, there is no difference between the groups in the frequency of nouns, the most basic and concrete element of language, t(47) = 1.4, p = 0.14. However, significant differences were found in relation to all the other morphological elements. First, the schizophrenic participants use fewer verbs t(47) = 4.37, p < .0001), adjectives t(47) = 4.06, p <.0001, and adverbs t(47) = 2.46, p <.05, than the healthy controls. This finding suggests that the former employ a relatively more basic language structure than the latter. In addition, the LTR ratio was higher among the schizophrenic participants, t(47) = 4.70, p < .0001, indicating less word inflection and, in consequence, greater poverty of language (H1). The schizophrenic patients also exhibit a higher rate of TTR, t(47) = 3.54, p < .000 (H2), demonstrating seemingly higher lexical diversity by repeating less times words previously used. However taken together with the previous results this my also indicate scattered thinking relating to non relevant or dis connected themes. This option will be elaborated in the discussion Further the difference among the LTR ratio and the TTR ratio was significantly higher for the control group, t(47) = 6.78, p < .0001, pointing to lower rate of morphological words inflection among the schizophrenic patients (H3).

Significant differences were also found in use of the first vs. third person. As expected, the schizophrenic participants favored the first person, t(47) = 3.09, p < .005 (H4). Conversely, those in the control group favored the third person, t(47) = 2.91, p < 0.01. However, this latter finding is not significant, requiring Bonferroni adjustment of the alpha.

Finally, the control group showed greater use of the past tense, t(47) = 3.67, p < .0001. No differences were found regarding use of the future tense, t(47) = 1.54, n.s.

Overall, the findings show that the language of schizophrenic patients is less complex than that of healthy individuals, with greater reliance on nouns and diminished use of word forms that add richness and color to a narrative (verbs, adjectives and adverbs). The language of schizophrenic patients also contains indicators of associative thinking. Finally, it suggests a focus on the self, with greater use of the first person.

We next used machine learning to cross-verify the results. For this purpose, a classifier was trained to distinguish between controls and patients using the extracted morphological features. Each participant was represented by all the morphological features mentioned above, calculated across the entire set of responses per participant. The value of each feature was count divided by the total number of words in the responses. To evaluate the classifier, a 10-fold cross-validation approach over the entire set of 49 participants was followed. Three different well-known fitting algorithms were explored: Random Forest (Breiman, 2001) and XGBoost (Chen & Guestrin, 2016), both based on decision trees, and a support vector machine (SVM) (Cortes & Vapnik, 1995) with a 2nd-degree polynomial kernel. Table 3 summarizes the results for each classifier algorithm with respect to prediction accuracy (per subject), precision (the fraction of true patients among all individuals identified as patients) and recall (the fraction of patients correctly identified as such). As shown in Table 3, all three fitting algorithms correctly attributed patients to the schizophrenia group at a rate of between 80% and 90%. Both alpha and beta errors are still high. However, considering the initial number of patients, these results seem promising.

4. Discussion

Currently, differential diagnosis for schizophrenia is based on psychiatrists' clinical impression relating to the symptoms described in the DSM-5. New approaches using natural-language-processing (NLP) evaluations are aiming to develop classifiers for schizophrenia diagnosis based on an algorithm looking for thought disorders expressed in language (Bar et al., 2019; Iter, Yoon & Jurafsky, 2018; Bedi et al., 2015; Rezaii et al., 2019). In the current work we suggest a complementary approach looking at the morphological language use of schizophrenia patients. Delving into the morphological component and relating to new approaches of language emergence and development, three main areas of difference between ordinary language use and that of Schizophrenia patients emerged. (A) Basic versus richness of language morphology; (B) Associative thinking; and (C) Self focus due to illness implication. Studying these issues we combined a systematic word feature analysis - using a context-sensitive, part-of-speech, BGU Tagger (Adler, 2007) - with a more traditional statistical analysis. Our results generally indicate that the language use of schizophrenics is a combination of the unique characteristics of these issues. The language use was more basic, using fewer verbs, adjectives and adverbs, gave an indication of associative thinking and was more concentrated around the self, first person use. Age, education, place of residence, marital status and income were controlled and thus are unlikely to explain the observed results. Finally, the initial classifier, morphologically based, programmed using machine learning methods, yielded a very high ratio of attribution to the healthy versus schizophrenic groups.

Since the studies learning morphology in schizophrenic patients are scarce relative to the literature on thought disruptions or language syntax, we cannot elaborate much on the morphology literature. However, our findings differentiate between basic morphology of language - dominated by more nouns, to reach morphology of language expressed by advantage of verbs, adjectives, adverbs, and a high gap between LTR and TTR forming together futile expression ability. It is important to stress that the current finding supports Covington et al.'s (2005) literature review and does not suggest any abnormal morphology in schizophrenic patients. Nevertheless, as described above morphology differences were sufficient to distinguish between the two groups. This means that the current results illustrate that though patients suffer from schizophrenia can convey on average coherent meaning they do so by using less complex morphology of language. This idea have already been addressed by several other works, such as Elvevaag et al. (2010), Bedi et al. (2015), Corcoran et al. (2018), as well as as Juola (2008) who studied methods to evaluate and assess morphological and syntactic linguistic complexity across several mainly European languages. Juola (2008) suggested that removing inflectional morphology allows one to compare linguistic complexity. Following this idea, measuring LTR ratio (lemmas counted only once for all word inflections) allowed using Juola (2008) termination to compress and compare the linguistic complexity between the two groups. Further measuring the gap between TTR and LTR, which was smaller for the schizophrenic language, indicate ratio of word inflections. As expected, schizophrenic language had a higher rate of LTR and a higher gap of LTR to TTR, meaning less word inflection, indicating morphology poverty of language.

The current results also support recent studies on Dutch speakers using diffusion tensor imaging. While the relation between language variables and the integrity of the white matter tracts was apparent for healthy controls but not for schizophrenia spectrum disorder, the letter, however, did use less complex sentences compered to matched healthy controls even in the absence of large scale white matter aberrations (De Boer et al., 2020b). Complementary SGA approach accompanied by FMRI analysis Palaniyappan et al. (2019) found that speech structure among English speakers becomes disorganized as a reduction in the integrity of the core functional hubs is observed. Taking together while schizophrenic spectrum does not show abnormal morphology the above findings and ours indicates an increasing ability to identify subtle language disturbances characteristics' of schizophrenic spectrum among several languages including English, Dutch, Hebrew and Portuguese (Mota et al, 2014).

Our third area of morphological difference between ordinary language use and that of schizophrenia patients refers to the use of first person pronouns relative to third person pronouns indicating self focus by the schizophrenia patients. Other results suggest that this trend in not unique to schizophrenia patients but is found among people who suffers from depression as well (Fineberg et al., 2016). Pairing previous and current results suggest that the use of first person is likely due to being in an illness state and not to any thought disruptions in the structure and use of language. The current results also contribute to approaches using computer based paradigm to study mental health thought disruptions, such as deviant thoughts (Al-Mosaiwi & Johnstone, 2018; Iter, Yoon & Jurafsky, 2018). We applied TTR index, to study thought disruptions in schizophrenia. Usually high TTR ratio - fewer words repeating themselves - suggests lexical diversity. However, high TTR ratio with high gap between LTR and TTR (the results of greater word inflections) may suggest creative thinking, where new words are used to convey related ideas. The opposite incidence - high TTR ratio with low gap between LTR and TTR - may point to loose, scattered, associative thinking, wherein the speaker has trouble maintaining a train of thought. The prevalence of TTR in our schizophrenic sample may suggest looser, more associative thinking. This line of thought can be supported by Fergadiotis, Wright & Green (2015) that while compering between four different techniques that suggest ways to decrease the influence of the sample length on TTR, found that even D scores - that is used to evaluate type to token ratio seemingly independently of the difference between the compared groups' sample length - is not immune to sample length and the possibility that the production of new themes might influence and increase the lexical diversity that is measured. The higher TTR ratio of our schizophrenic sample may suggest that as a result of scattered and associative thinking this group related to more themes while telling their stories of the pictures.

Producing the indexes of LTR (unique lemmas), TTR (unique words) and the gap between them is relatively simple compared to more sophisticated algorithmic methods that represent mathematically the relative linguistic distance between the words used to present an idea. However, the present study offers a new morphologic way of evaluating associative thinking that is supported by the mathematical algorithm described by Bar et al. (2019). The pattern of results appears to support previous assumptions that schizophrenia patients differ linguistically from non-patients (Condray et al., 2002; Covington et al., 2005) Also, that language analysis can contribute to diagnosis (Boer et al., 2020b).

5. Final Remarks

Several factors limit the generalizability of the current findings. First, our findings pertain to patients receiving antipsychotic medication. Previous research shows an effect of antipsychotic treatment on the structure and function of the basal ganglia (Goldman et al., 2008). As well as on the time of speaking (e.g. slower articulating rate), and clauses per utterance (e.g. increased and prolonged paused and shorter utterances with fewer clauses), related by De Boer et al. (2020c) to illness effect rather than to psychotic symptoms. To assess the effect of the medication treatment De Boer et al. (2020b) have measured the correlation between the medication dosage and PANSS positive and negative symptoms. However, the effects of medication on language processing, though may have some influence, cannot be precisely determined even by measuring the above correlations. Second - due to the moderately modest patient cohort size - we did not present any trends regarding the association between PANSS positive and negative symptoms and any of the morphological elements. These have been studied in the past using different research approaches (De Boer et al., 2020b; Mota el al., 2017). The possible morphological unique structure among PANSS positive and negative symptoms should be further studied with a larger cohort. Third the current study focused on schizophrenic patients. Therefore, the current results do not address how morphology might differ in general between mental health patients and non-patients or more specifically for schizophrenic patients who are not hospitalized. Repeated studies should examine the morphological effects of other mental health conditions. Fourth, we studied male schizophrenic patients. Morphological changes in female schizophrenic patients cannot be deduced from the current results, and should be examined in future research. Fifth, the sample size of the current study is not sufficient to statistically examine all morphological features of language, or to make the most of machine learning methods. Sixth, though the classifier succeeded in applying cases to the patient and non-patient groups at a rate better than would be achieved by chance alone, these initial results should be confirmed with a larger cohort. In the meantime, they should be received with caution.

Nonetheless, the present findings produce a relatively clear picture of morphological attributes in the Hebrew language use of male schizophrenic patients. On the basis of the present findings, several promising research directions suggest themselves. In particular, new computational methods hold promise for studying and diagnosing schizophrenia and other thought disorders through NLP, identifying how these disorders are expressed in language use. These methods may also prove fruitful for related mental classifications like depression and post-traumatic stress disorder. At a practical level, morphological testing may eventually become useful as a kind of “mental blood test” for mental health conditions.

References

* The data that support the findings of this study are openly available in https://docs.google.com/spreadsheets/d/1wQH23-25Vs5G2KC_6dvcisuUlFtBzqy9PwCBzmETrIo/edit?usp=sharing

Adler, M. (2007). Hebrew morphological disambiguation: An unsupervised stochastic word-based approach. Ben Gurion University.

Al-Mosaiwi, M., & Johnstone, T. (2018). In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. Clinical Psychological Science, 6(4), 529-542.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Andreasen, N. C., & Grove, W. M. (1986). Thought, language, and communication in schizophrenia: diagnosis and prognosis. *Schizophrenia Bulletin*, *12*(3), 473-482.

Bar*, K., Zilberstein*, V., Ziv*, I., Baram*, H., Dershowitz, N., Itzikowitz, S., Harel, E. V. Semantic Characteristics of Schizophrenic Speech. (6th Annual Computational Linguistics and Clinical Psychology Workshop, NAACL, 2019, Minneapolis, MN *the authors Claim for equal contribution).

Barnes, L., Harp, D., & Jung, W. (2002). Reliability Generalization of Scores on the Spielberger State-Trait Anxiety Inventory. *Educational and Psychological Measurement*, *62*(4), 603–618. https://doi.org/10.1177/0013164402062004005

Bearden, C. E., Wu, K. N., Caplan, R., & Cannon, T. D. (2011). Thought disorder and communication deviance as predictors of outcome in youth at clinical high risk for psychosis. *Journal of the American Academy of Child & Adolescent Psychiatry*, *50*(7), 669-680.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck Depression Inventory-II (BDI-II)* (2nd ed.). San Antonio, TX: Psychological Corporation.

Bedi, G., Carrillo, F., Cecchi, G. A., Slezak, D. F., Sigman, M., Mota, N. B., ... & Corcoran, C. M. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia,* 1, 15030.

Bensimon, M., Levine, S. Z., Zerach, G., Stein, E., Svetlicky, V., & Solomon, Z. (2013). Elaboration on posttraumatic stress disorder diagnostic criteria: A factor analytic study of PTSD exposure to war or terror. *Israel Journal of Psychiatry and Related Sciences*, *50*(2), 84–90. Retrieved from

Birnbaum, M. L., Ernala, S. K., Rizvi, A. F., De Choudhury, M., & Kane, J. M. (2017). A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. *Journal of Medical Internet Research, 19*(8), e289.

Bleuler, E. (1911). Dementia praecox: oder Gruppe der Schizophrenien. Viennese: Frnaz Deuticke.

Blevins, C. A., Weathers, F. W., Davis, M. T., Witte, T. K., & Domino, J. L. (2015). The posttraumatic stress disorder checklist for DSM-5 (PCL-5): Development and initial psychometric evaluation. *Journal of Traumatic Stress*, *28*(6), 489-498.

Bokat, C. E., & Goldberg, T. E. (2003). Letter and category fluency in schizophrenic patients: a meta-analysis. *Schizophrenia research, 64*(1), 73-78.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language*

Engineering, 23(5), 649-685.

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, 9(2), 48-57.

Carson, N. J., Mullin, B., Sanchez, M. J., Lu, F., Yang, K., Menezes, M., & Le Cook, B. (2019). Identification of suicidal behavior among psychiatrically hospitalized adolescents using natural language processing and machine learning of electronic health records. *PloS ONE,* 14(2), e0211116.

Chan, S. (2001). Kaplan & Sadock's comprehensive textbook of psychiatry. *Hong Kong Journal of Psychiatry*, 11(3), 23-25.

Chen, T., & Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). New York, NY, USA: ACM, 10(2939672.2939785).

Condray, R., Steinhauer, S. R., van Kammen, D. P., & Kasparek, A. (2002). The language system in schizophrenia: Effects of capacity and linguistic structure*. Schizophrenia Bulletin, 28*(3), 475-490.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning,* 20(3), 273-297.

Corcoran, C. M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D. C., Javitt, D. C., Bearden, C., & Cecchi, G. A. (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, *17*(1), 67-75.

Corcoran, C. M., & Cecchi, G. (2020). Using language processing and speech analysis for the identification of psychosis and other disorders. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*.

Corcoran, C. M., Mittal, V. A., Bearden, C. E., Gur, R. E., Hitczenko, K., Bilgrami, Z., Savic, A., Cecchi, G. A., & Wolff, P. (2020). Language as a biomarker for psychosis: A natural language processing approach. *Schizophrenia Research*, 226, 158-166.

Covington, M. A., He, C., Brown, C., Naçi, L., McClain, J. T., Fjordbak, B. S., ... & Brown, J. (2005). Schizophrenia and the structure of language: the linguist's view. *Schizophrenia research, 77*(1), 85-98.

De Boer, J. N., Brederoo, S. G., Voppel, A. E., & Sommer, I. E. (2020a). Anomalies in language as a biomarker for schizophrenia. *Current Opinion in Psychiatry*, *33*(3), 212-218.

De Boer, J. N., Van Hoogdalem, M., Mandl, R. C. W., Brummelman, J., Voppel, A. E., Begemann, M. J. H., ... & Sommer, I. E. C. (2020b). Language in schizophrenia: relation with diagnosis, symptomatology and white matter tracts. *NPI Schizophrenia*, 6(1), 1-10.

De Boer, J. N., Voppel, A. E., Brederoo, S. G., Wijnen, F. N. K., & Sommer, I. E. C. (2020c). Language disturbances in schizophrenia: the relation with antipsychotic medication. *NPJ schizophrenia*, 6(1), 1-9.

DeLisi, L. E., Sakuma, M., Kushner, M., Finer, D. L., Hoff, A. L., & Crow, T. J. (1997). Anomalous cerebral asymmetry and language processing in schizophrenia. *Schizophrenia Bulletin, 23*(2), 255-271.

Dreisbach, C., Koleck, T. A., Bourne, P. E., & Bakken, S. (2019). A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *International journal of medical informatics*.

Elvevaag, B., Foltz, P. W., Rosenstein, M., & DeLisi, L. E. (2010). An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. *Journal of neurolinguistics*, *23*(3), 270-284.

Fergadiotis, G., & Wright, H. H. (2011). Lexical diversity for adults with and without aphasia across discourse elicitation tasks. *Aphasiology, 25*(11), 1414-1430. https://doi.org/10.1080/02687038.2011.603898

Fergadiotis, G., Wright, H. H., & Green, S. B. (2015). Psychometric evaluation of lexical diversity indices: Assessing length effects. *Journal of Speech, Language, and Hearing Research, 58*(3), 840-852. https://doi.org/10.1044/2015_JSLHR-L-14-0280

Fernandes, A. C., Dutta, R., Velupillai, S., Sanyal, J., Stewart, R., & Chandran, D. (2018). Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. *Scientific reports, 8*(1), 7426.

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. *Psychological medicine*, 46(12), 2605-2615.

Gallagher, D., Nies, G., & Thompson, L. W. (1982). Reliability of the Beck Depression Inventory with older adults. *Journal of Consulting and Clinical Psychology*. US: American Psychological Association.

Goldman, M., Marlow-O'Connor, M., Torres, I., & Carter, C. S. (2008). Diminished plasma oxytocin in schizophrenic patients with neuroendocrine dysfunction and emotional deficits. *Schizophrenia research*, *98*(1-3), 247-255.

Grinker, R. R. (2010). In retrospect: The five lives of the psychiatry manual. *Nature, 468*(7321), 168.

Hasenson-Atzmon, K., Marom, S., Sofer, T., Lev-Ari, L., Youngmann, R., Hermesh, H., & Kushnir, J. (2016). Cultural impact on SAD: Social Anxiety Disorder among Ethiopian and former Soviet Union immigrants to Israel, in comparison to native-born Israelis. *Israel*

Journal of Psychiatry and Related Sciences, *53*(3), 48-55.

Hoermann, S., McCabe, K. L., Milne, D. N., & Calvo, R. A. (2017). Application of synchronous text-based dialogue systems in mental health interventions: systematic review. *Journal of Medical Internet Research, 19*(8), e267.

Imahori, E. (2018). Linguistic Expressions of Depressogenic Schemata. Working Papers in Applied Linguistics & TESOL, 18(2), 20-32.

Iter, D., Yoon, J., & Jurafsky, D. (2018). Automatic detection of incoherent speech for diagnosing schizophrenia. In Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic (pp. 136-146).

Juola, P. (2008). Assessing linguistic complexity. Language complexity: Typology, contact, change, 89-108.

Katz, G., Grunhaus, L., Deeb, S., Shufman, E., Bar-Hamburger, R., & Durst, R. (2012). A comparative study of Arab and Jewish patients admitted for psychiatric hospitalization in Jerusalem: The demographic, psychopathologic aspects, and the drug abuse comorbidity. *Comprehensive Psychiatry*, *53*(6), 850–853. https://doi.org/10.1016/j.comppsych.2011.11.005

Kleist, K. (1914). Aphasie und geisteskrankheit. *Münchener Medizinische Wochenschrift, 61*, 8-12.

Knight, R. G., Waal-Manning, H. J., & Spears, G. F. (1983). Some norms and reliability data for the State-Trait Anxiety Inventory and the Zung Self-Rating Depression scale. *British Journal of Clinical Psychology*, *22*(4), 245-249.

Koutsouleris, N., Kahn, R. S., Chekroud, A. M., Leucht, S., Falkai, P., Wobrock, T., ... & Hasan, A. (2016). Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. *The Lancet Psychiatry*, 3(10), 935-946.

Leighton, S. P., Upthegrove, R., Krishnadas, R., Benros, M. E., Broome, M. R., Gkoutos, G. V., ... & Fowler, D. (2019). Development and validation of multivariable prediction models of remission, recovery, and quality of life outcomes in people with first episode psychosis: a machine learning approach. *The Lancet Digital Health, 1*(6), e261-e270.

Lin, C. H., Lin, H. S., Lin, S. C., Kuo, C. C., Wang, F. C., & Huang, Y. H. (2018). Early improvement in PANSS-30, PANSS-8, and PANSS-6 scores predicts ultimate response and remission during acute treatment of schizophrenia. *Acta Psychiatrica Scandinavica*, *137*(2), 98–108. https://doi.org/10.1111/acps.12849

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 Schizophrenia*, 3(1), 1-10.

Mota, N. B., Furtado, R., Maia, P. P., Copelli, M., & Ribeiro, S. (2014). Graph analysis of dream reports is especially informative about psychosis. *Scientific reports*, *4*, 3691.

Mota, N. B., Vasconcelos, N. A., Lemos, N., Pieretti, A. C., Kinouchi, O., Cecchi, G. A., Copelli, M., & Ribeiro, S. (2012). Speech graphs provide a quantitative measure of thought disorder in psychosis. *PloS one*, *7*(4), e34928.

Morris, R. R., Kouddous, K., Kshirsagar, R., & Schueller, S. M. (2018). Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. *Journal of Medical Internet Research, 20*(6), e10148.

Murray, H. A. (1943). Thematic apperception test. Cambridge, MA, US: Harvard University Press.

Østergaard, S. D., Lemming, O. M., Mors, O., Correll, C. U., & Bech, P. (2016). PANSS-6: a brief rating scale for the measurement of severity in schizophrenia. *Acta Psychiatrica Scandinavica, 133*(6), 436-444.

Palaniyappan, L., Mota, N. B., Oowise, S., Balain, V., Copelli, M., Ribeiro, S., & Liddle, P. F. (2019). Speech structure links the neural and socio-behavioural correlates of psychotic disorders. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 88, 112-120.

Peek, N., Combi, C., Marin, R., & Bellazzi, R. (2015). Thirty years of artificial intelligence in medicine (AIME) conferences: A review of research themes. *Artificial intelligence in medicine, 65*(1), 61-73.

Qureshi, S. A., Hasanuzzaman, M., Saha, S., & Dias, G. (2019). The Verbal and Non Verbal Signals of Depression--Combining Acoustics, Text and Visuals for Estimating Depression Level. arXiv preprint arXiv:1904.07656.

Rezaii, N., Walker, E., & Wolff, P. (2019). A machine learning approach to predicting psychosis using semantic density and latent content analysis. *NPJ schizophrenia, 5*(1), 1-12.

Saka, N., & Gati, I. (2007). Emotional and personality-related aspects of persistent career decision-making difficulties. *Journal of Vocational Behavior*, *71*(3), 340–358. https://doi.org/10.1016/j.jvb.2007.08.003

Spielberger, C. D., Gorsuch, R. L., Lushene, R. E., Vagg, P. R., & Jacobs, G. A. (1970). State-trait anxiety inventory. Palo Alto.

Toivonen, K. I., Zernicke, K., & Carlson, L. E. (2017). Web-based mindfulness interventions for people with physical health conditions: systematic review. *Journal of medical Internet* *research, 19*(8), e303.

Weathers, F. W., Litz, B. T., Keane, T. M., Palmieri, P. A., Marx, B. P., & Schnurr, P. P. (2013). The PTSD checklist for DSM-5 (PCL-5). *Scale available from the National Center for PTSD at www.ptsd.va.gov.*

Figure Captions

Figure 1 – Morphological language structre as a function of group relatedness.

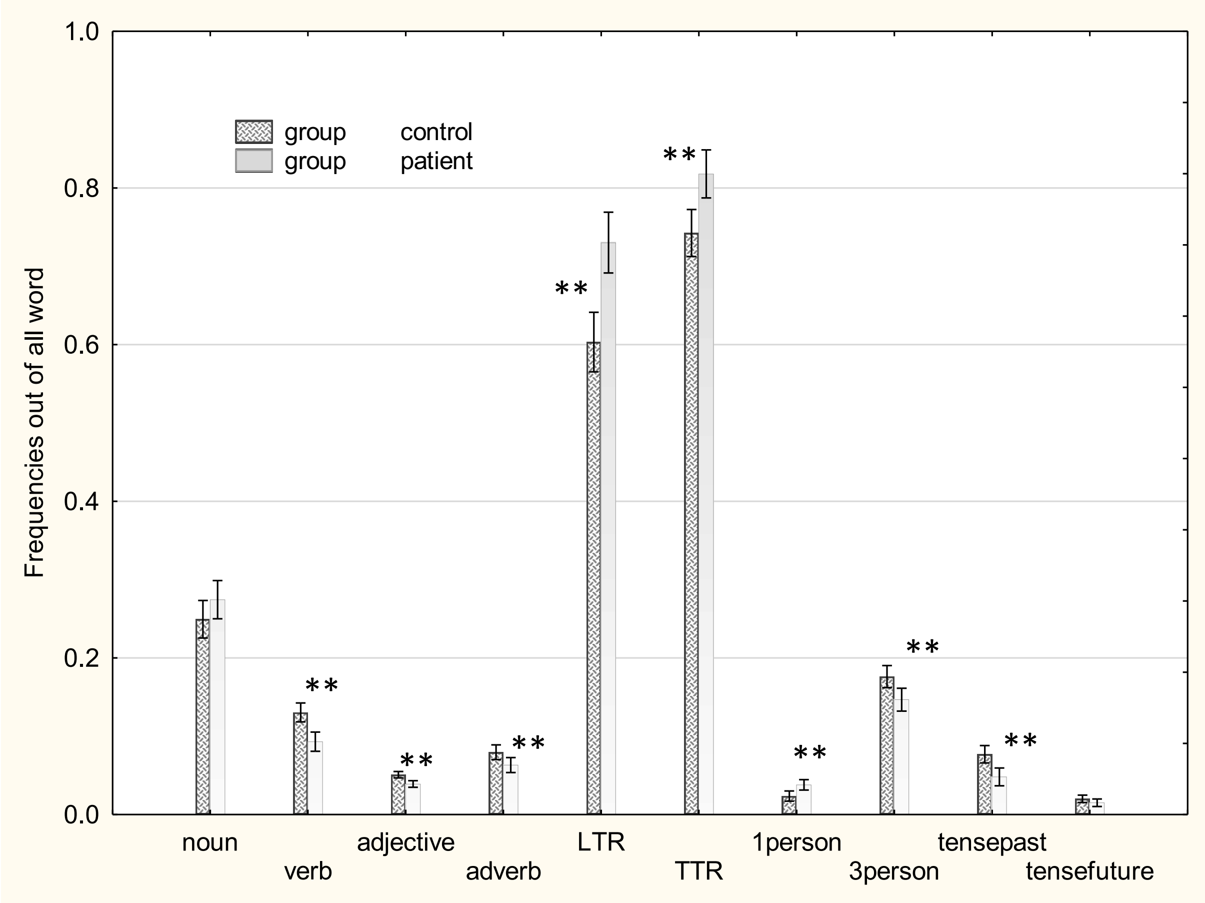

Figure 1 – Morphological language structre as a function of group relatedness.

*p< .05; **p<.005; ***p<.001.
Note. Each part of speech was analyzed separately by conducting two-tailed independent sample t-tests. Bonferroni adjustment of the alpha was conducted (α=0.05/10= 0.005).

Table 1 - *Participants' Demographic Characteristics by group*

Table 1

Demographic characteristics of participants by group

	Control	Experimental	Statistics
N	25	24	
Age (mean in years, range and sd in parentheses)	30.3 (19-49) (8.26)	38.3 (23-63) (10.43)	t=3**
Education (frequencies)			χ2 (2,49) =10.46**
Pre-high school	0	5	
High school	17	18	
Academic	8	1	
Place of residence (frequencies)			χ2 (2,49) =21.13**
Southern Israel	13	10	
Central Israel	2	11	
Northern Israel	8	1	
Jerusalem	2	2	
Marital status (frequencies)			χ2 (2,49) =4.6, *p* = .00, n.s.
Single	20	21	
Married	5	1	
Divorced	0	2	
Income (frequencies)			χ2 (2,49) =0.5, *p* = .00, n.s.
Average or lower	21	20	
Higher than average	4	4	
PANSS positive subscale		6.29 ± 3.04	
PANSS negative subscale		6 ± 2.62	
PANSS total subscale		12.29 ± 4.05	

*p< .05; **p<.005; ***p<.001

Table 2 - *Features provided by BGU tagger for every word*

Table 2

Features provided by BGU tagger for every word

Feature	Description
Part of speech	A category assigned to each word in accordance with its syntactic role. Here we are interested in the following categories: noun, verb, adjective and adverb.
Lemma	The dictionary entry of a given word. In Hebrew, this is normally the third person singular past tense for verbs, and the singular form with no affixes for nouns.
Person	First and third.
Verb tense	Past, present and future.

Table 3 - *Accuracy percentage for each classifier algorithm*

Table 3

Classifier algorithm results

	Precision	Recall	Accuracy
XGBClassifier	0.8416	0.8716	0.8333
RandomForestClassifier	0.8858	0.8783	0.8666
SVM, Poly (2)	0.8358	0.8341	0.8166