Quantified language connectedness in schizophrenia-spectrum disorders

Voppel, A E; de Boer, J N; Brederoo, S G; Schnack, H G; Sommer, I E C

Published in:
Psychiatry Research

DOI:
10.1016/j.psychres.2021.114130

Publication date:
2021

Psychiatry Research 304 (2021) 114130

Contents lists available at ScienceDirect

Psychiatry Research

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

Quantified language connectedness in schizophrenia-spectrum disorders

AE Voppel [a,*], JN de Boer [a,b], SG Brederoo [a], HG Schnack [b,c], IEC Sommer [a]

[a] Department of Biomedical Sciences of Cells and Systems, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
[b] Department of Psychiatry, UMCU Brain Center, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands
[c] Utrecht University, Utrecht Institute of Linguistics OTS, Utrecht, the Netherlands

ARTICLE INFO

Keywords:
Word similarity
Psychosis
Semantic model
Speech
Natural language processing
Biomarker

ABSTRACT

Language abnormalities are a core symptom of schizophrenia-spectrum disorders and could serve as a potential diagnostic marker. Natural language processing enables quantification of language connectedness, which may be lower in schizophrenia-spectrum disorders. Here, we investigated connectedness of spontaneous speech in schizophrenia-spectrum patients and controls and determine its accuracy in classification. Using a semi-structured interview, speech of 50 patients with a schizophrenia-spectrum disorder and 50 controls was recorded. Language connectedness in a semantic word2vec model was calculated using consecutive word similarity in moving windows of increasing sizes (2–20 words). Mean, minimal and variance of similarity were calculated per window size and used in a random forest classifier to distinguish patients and healthy controls. Classification based on connectedness reached 85% cross-validated accuracy, with 84% specificity and 86% sensitivity. Features that best discriminated patients from controls were variance of similarity at window sizes between 5 and 10. We show impaired connectedness in spontaneous speech of patients with schizophrenia-spectrum disorders even in patients with low ratings of positive symptoms. Effects were most prominent at the level of sentence connectedness. The high sensitivity, specificity and tolerability of this method show that language analysis is an accurate and feasible digital assistant in diagnosing schizophrenia-spectrum disorders.

1. Introduction

Schizophrenia-spectrum disorders (henceforth: SSD) include a complex variety of psychiatric illnesses that affect approximately 2–3% of the population (Rössler et al., 2005). Language and speech disturbances are one of the key diagnostic features of SSD (American Psychiatric Association, 2013). Clinicians routinely use descriptions of spoken language in their mental health examinations, such as tangentiality, incoherence and 'word salad'. Language abnormalities have been investigated extensively in patients with SSD (Chaika, 1990; Covington et al., 2005; DeLisi, 2001; Kuperberg, 2010). These studies show that greatest difficulties arise at the level of semantics (meaning) and syntax (grammar). Language abnormalities have recently gained traction due to their possible use for classification of diagnosis; for reviews, see Corcoran and Cecchi, 2020; de Boer et al., 2020a. Given the multi-facetedness of language, research on this topic is broad and there is as of yet little overlap between methodologies or approaches.

An overarching way to look at language disturbances in SSD is the conceptualization of an impairment in 'connectedness' of language (Covington et al., 2005). Connectedness in language can be measured at multiple levels and dimensions. Language is connected syntactically, through its structure or grammar, as well as semantically given that words with related meaning occur within the same sentence. Furthermore, connectedness is present at a word-to-word level, as well as across sentences. For example, the famous sentence "Colorless green ideas sleep furiously" is syntactically correct, but semantically not connected, making the sentence nonsensical (Chomsky, 1957).

Language disturbances in SSD can thus be understood as disruptions in connectedness. For example, sentences that are tangential can be described as having reduced connectedness across sentences. While the individual sentences may be correct, the connection between them is vague at best. Incoherence can be understood as word-level disconnectedness; as word-to-word connections are also disturbed. From this viewpoint, it makes sense that language disturbances in SSD arise at both the semantic and the syntactic level (Chaika, 1990; Covington et al., 2005; DeLisi, 2001; Kuperberg, 2010); since this is where connectedness is found in language.

In clinical practice, incoherence and tangentiality is scored by clinicians using subjective rating scales such as the PANSS (Kay et al., 1987) or the TALD (Kircher et al., 2014). Recent advances in natural

* Corresponding author.
E-mail address: A.e.voppel@umcg.nl (A. Voppel).

https://doi.org/10.1016/j.psychres.2021.114130
Received 16 March 2021; Received in revised form 13 July 2021; Accepted 16 July 2021
Available online 22 July 2021

language processing have enabled quantification of connectedness in language, both at sentence and at word-level (Elvevåg et al., 2010; Elvevåg et al., 2007), with evidence that these quantifications are as sensitive as or better than clinical scales (Elvevåg et al., 2007; Tang et al., 2021). A seminal study by Bedi et al. has shown that measures derived from language can be used to predict conversion to psychosis in individuals at ultra-high risk for psychosis (Bedi et al., 2015). Since then, several studies have followed the same approach to differentiate patients from healthy controls (Bar et al., 2019; Corcoran et al., 2018; Spencer et al., 2021), to measure symptomatology (Holshausen et al., 2014; Mota et al., 2012; Pauselli et al., 2018) and predict conversion to psychosis in high-risk groups (Gupta et al., 2018; Kayi et al., 2017; Rezaii et al., 2019). Sources of language assessed include not only speech but also social media posts (Birnbaum, 2019) or specific subjects like dream reports (Mota et al., 2014). These types of analyses are sensitive enough to detect differences in first-degree relatives (Elvevåg et al., 2010), moreover, they are associated with social and neurological features (Palaniyappan et al., 2019). For overviews, see recent reviews by Corcoran et al., 2020; Corcoran and Cecchi, 2020; de Boer et al., 2020a.

As noted, reduced connectedness of language can take place both on a word-to-word level and across sentences, implicating the importance of the size of the context or "window" of language around the examined word or phrase. A previous meta-analysis by our group indicates that semantic space models perform best on sentences, rather than on individual words from for example semantic verbal fluency tasks (de Boer et al., 2018). Therefore, in the present study we used a semi-structured interview to collect spontaneous speech. Previous work showed excellent performance of language models on psychosis classification (Corcoran et al., 2018; Rezaii et al., 2019). However, some included other language measures (e.g. sentence length, topics) in addition to connectedness measures in their final model, or examined subsets of word connectedness features.

Here, we use language connectedness in SSD to classify SSD patients and healthy controls using spontaneous speech. Following previous research, we expect to find significant group differences in language connectedness, applicable for classification, with the most informative features in the sentence window range (de Boer et al., 2018). Language connectedness is calculated using a word2vec semantic space model. To fully acknowledge the different levels and dimensions of reduced connectedness, we applied word2vec across multiple window sizes, mitigating limitations in previous studies.

2. Methods

2.1. Participants

Language recordings were obtained from 50 SSD participants and 50 healthy controls at the University Medical Center Utrecht. All patients had a diagnosis of schizophrenia, psychosis NOS, schizophreniform or schizoaffective disorder. Diagnoses were made by the treating physician and confirmed using the CASH or the MINI diagnostic interview (Andreasen et al., 1992). Symptom severity was assessed using the Positive And Negative Syndrome Scale (PANSS) (Kay et al., 1987; Sheehan et al., 1998). Antipsychotic medication usage was calculated as chlorpromazine equivalents (Leucht et al., 2014); none of the patients were antipsychotic naïve. The inclusion criteria for healthy controls were the absence of a psychiatric diagnosis and history, and no family history of psychiatric disorders.

All participants were adult native Dutch speakers. To prevent participants focusing on their speech, participants were informed that the interview involved the analysis of 'general experiences'; only after completion of the interview were participants told that the research also investigates their speech and produced language. Before enrollment, all participants gave written informed consent. The study was approved by the institutional review board of the University Medical Center Utrecht.

2.2. Interview procedure

To elicit spontaneous speech, we performed a semi-structured, open-ended, neutral-topic interview. Prompts such as "Can you tell me about your experiences at the dentist's?" and "What would you do if you would win a million dollars?" were used as prompts. Occasional questions by the interviewer for further information regarding a topic were used to encourage the participant to continue talking. The interview was designed to reflect normal day-to-day speech about non-pathological topics in an ecologically valid dialogue setting. All interviews were performed by researchers trained for the interview procedures. For a full list of questions asked, see Table S1. Topics which might have excessive emotional valence for participants (e.g. health-related subjects) were avoided. The same procedures were followed for participants in both groups, with questions presented in a semi-randomized order.

The interview was recorded using two AKG-C544l head-worn cardioid microphones, one for the participant's and one for interviewer's speech. The interview was digitally recorded to a TASCAM DR40 solid state recording device, with a sampling rate of 44,100 kHz with 16-bit quantization. An automatic second recording was made with a decreased volume of six decibels to prevent clipping.

The participants' speech, once recorded, was transcribed following the CLAN—CHILDES transcription protocol for analysis (Brundage and Bernstein Ratner, 2018; MacWhinney, 2000). The interview was transcribed by researchers blind to the participant condition. After transcription, filled pauses such as 'uhm' were removed. No other preprocessing was performed; specifically, words were not stemmed and repetitions or interjections were not removed.

2.3. Word2vec model

A widely used technique to quantitatively assess connectedness in language is that of semantic space models (Landauer et al., 1998; Mikolov et al., 2013a). In these approaches, words are represented as vectors in multidimensional space. The vectors are created based on the central assumptions underlying all semantic space models; the meaning of a word is determined by its context, and similar words appear in similar contexts. The context can be defined as a (small) number of words that surround the target word, such as paragraphs or even stories. In such models, word representations are mathematical calculations associated with each word. The vectors that are used are multidimensional, in which each dimension corresponds to a 'feature' of the word. Examples of such features could be 'furry', 'pet', 'running'; attempting to grasp the meaning of 'cat'. These features are thought to have either a semantic or syntactic interpretation and are called 'word features' (Turian et al., 2010). These vectors can be used to quantify the similarity between words or sentences compared to their context. Consider the sentence "I sat on a bench" where "I sat on a" is the context of "bench". The similar phrase "I sat on a chair" has "chair" and "bench" occurring in the same context, making them semantically similar. Repeated over multiple varied sentences, target words can be quantified as more or less likely to occur with their context. The resulting trained model of connections can be used to quantify similarity between novel sentences. The resulting measures of similarity can be taken to indicate whether a word or phrase is properly connected to the context (being the nearby word or words).

Here, a distributed word semantic space model was trained on the Corpus Gesproken Nederlands (CGN; *Corpus Spoken Dutch*), a large (5654,644 words) Dutch corpus of spoken language (van Eerten, 2007). The trained word2vec model was created using the *gensim* software package with 300 dimensions, making use of the skip-gram method (Řehůřek and Sojka, 2010). Using the trained model, transcribed interviews were vectorized, yielding a 300-dimensional vector for each word.

As outlined above, language disturbances related to meaning and coherence, can be conceptualized as reductions in connectedness.

Semantic space models aim to capture connectedness in language by calculating 'similarity'. Word-to-word similarity can be computed by calculating the cosine similarity between corresponding vectors. This results in a number between −1 and 1, with −1 representing large distance or low overlap and 1 representing small distance or high overlap. Highly similar vectors between words (high cosine similarity) thus indicate high connectedness. To determine the connection of a word within its context, all cosine angles between individual words within a given window were computed; these cosine angles were averaged, giving a single average similarity value of word connectedness within the window. The window was then slid one word further on the participant's transcribed speech, repeatedly until the end of the interview, giving a per-interview set of connectedness values. This sliding window approach is well suited for spoken language, since sentences rarely are demarcated and punctuation is not available. Because reduced connectedness in language can occur over different ranges, per-participant sets of all windows of size 2 till 20 were computed. By examining the cosine distance of the embeddings of two consecutive words (comparing the cosine angle of two single word embeddings in window size=2) as well as the larger windows (up to two standard deviations above the average length of a Dutch sentence; Wiggers and Rothkrantz, 2007), both word-to-word and sentence-to-sentence level connectedness could be evaluated.

This resulted in the calculation of the following sets of variables:

1) Mean similarity. Mean similarity is defined as the mean word similarity per moving window, averaged over all calculated windows for an interview. This is repeated per window size, resulting in mean similarities for window sizes 2–20. For example, for an interview with a total of four words and a window size of two, three similarities are calculated (similarity of word 1 and 2, word 2 and 3, and word 3 and 4) of which the mean is then calculated.
2) Minimal similarity. Minimal similarity is defined as the lowest similarity in the set of calculated similarities throughout the interview. This is repeated for each different window size between 2 and 20. For example, if, for a given window size, similarities over an entire interview are 0.7, 0.5, and 0.8, minimal similarity is 0.5.
3) Variance in similarity. The variance σ^2 is calculated over all similarities an interview, using the formula $\sigma^2 = \sum_n^{(X-\bar{\mu})^2}$ where X is the similarity, μ is the mean of similarities of the interview and n is the number of similarities per interview. This measure is also calculated per window size 2–20 and serves as a measure of how the distribution of word similarities over an entire interview is shaped, specifically the width of the distribution. For example, if similarities over an interview are 0.7, 0.5 and 0.8, variance is 0.016, being the square of the standard deviation of similarities.

2.4. Statistical analysis and classification

Group demographic differences were assessed using Chi-square tests for binary variables, and ANOVAs for continuous variables. Using a random forest classifier with mean, minimum and variance of word similarity over windows of window sizes from 2 to 20 as features, participants were classified as belonging to either the SSD group or to the healthy control group. During model training 10-fold cross-validation was employed, repeated three times as an additional measure against overfitting on spurious signals in the testing fold (Vanwinckelen and Blockeel, 2012). Connectedness windows were assessed through ranking the Gini coefficient as measures of feature importance for the resulting classifier. We report ranked Gini coefficients both for all features combined and for mean, minimal and variance of similarity separately. To assess the possible confounding factors of chlorpromazine dose, years of education and their relation to classification features, Pearson's correlations were performed.

3. Results

3.1. Demographics

No significant differences were found between patients and healthy controls in age or sex (Table 1). Years of education showed a significant difference between participants with SSD and healthy controls; however, parental years of education did not significantly differ. Mean duration of illness was 2.6 (5.5) years, showing that this sample mainly consisted of early stage patients. Mean total PANSS score was 53.2 (12.6), indicating that most patients were in remission (Fig. 1).

3.2. Classification and window sizes of connectedness

Using a random tree forest binary classification algorithm based on mean, minimal and variance in connectedness, a mean accuracy of 85% was achieved. Sensitivity was 84% (71%−92% confidence interval) and specificity reached 86% (81%−95% confidence interval). Area under the curve-receiver operating characteristic (AUC-ROC) of the classifier

Table 1
Demographic characteristics.

Category		SSD patients (n = 50)	Healthy controls (n = 50)	Statistics
Age				
Years	M (SD)	29.2 (9.1)	31.5 (12.4)	$F = 1.10$, $p = 0.298$
Sex				
Male	n (%)	38 (76)	42 (84)	χ^2=1.00, $p = 0.227$
Years of education				
Participant	M (SD)	12.9 (2.8)	14.5 (2.4)	$F = 9.630$, $p = 0.003$
Parental	M (SD)	11.9 (3.5)	12.3 (3.4)	$F = 0.207$, $p = 0.650$
Transcript size				
Number of words	M (SD)	1443 (535)	1643 (812)	$F = 4.195$, $p = 0.043$
Illness duration				
Years	M (SD)	2.6 (5.48)		
Chlorpromazine dose				
Milligram equivalent	M (SD)	386.3 (268.4)		
Diagnosis				
Psychosis NOS	n (%)	22 (44)		
Schizophrenia	n (%)	19 (38)		
Schizoaffective	n (%)	6 (12)		
Schizophreniform	n (%)	3 (6)		
PANSS				
Positive	M (SD)	11.3 (4.4)		
	Range	7–25		
Negative	M (SD)	14.5 (5.0)		
	Range	7–28		
General	M (SD)	27.4 (7.1)		
	Range	16–47		
Total	M (SD)	53.2 (12.6)		
	Range	30–91		

Legend: M: mean, SD: standard deviation, PANSS: Positive And Negative Syndrome Scale.

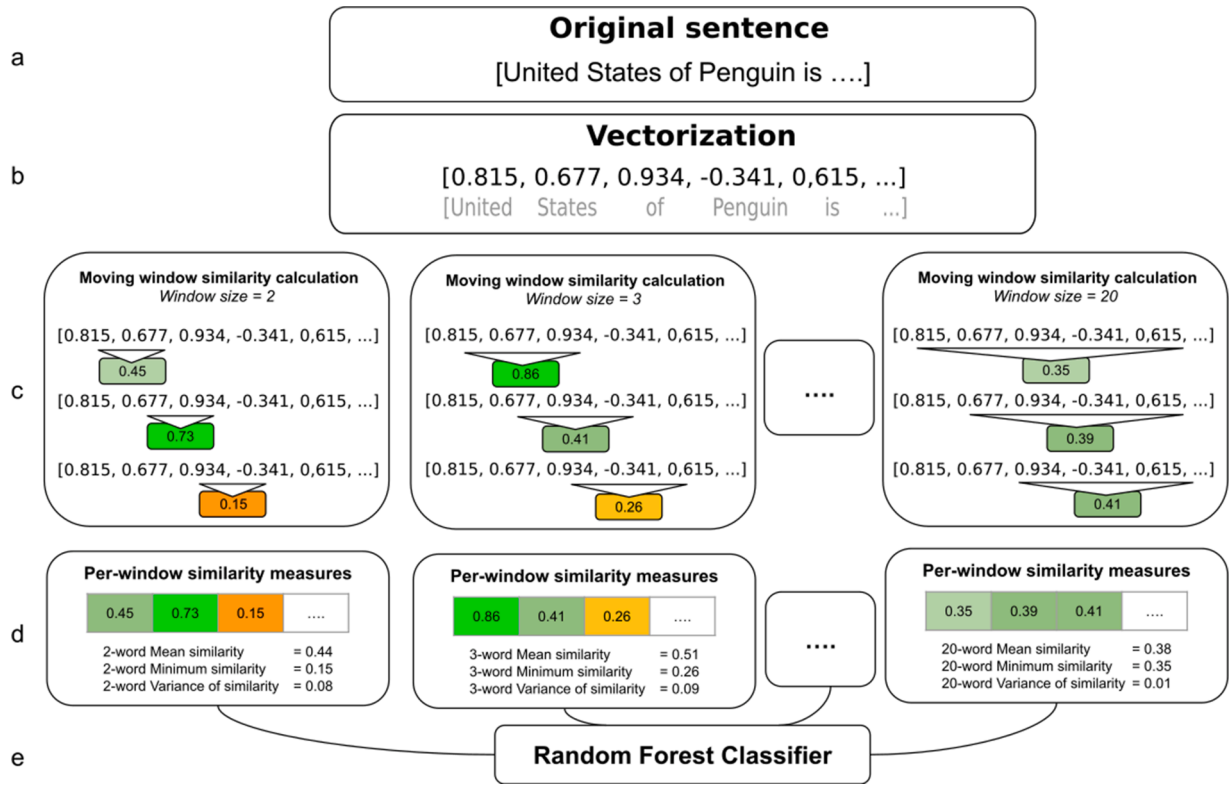

Fig. 1. Processing procedure from transcribed text to machine learning classifier features. (a): Original transcribed sentence is vectorized (b) using a word2vec model. Note that per-word numbers in b) and c) are notational and that each number reflects a 300-dimensional vector. Using a moving window approach with windows sized 2–20 (c) per-window similarity scores are calculated for each subject, for each window size until the end of the transcript. From these similarity scores, a per-subject mean similarity score, the minimum similarity score and the variance of similarity across the transcript are calculated (d). These measures for all windows sizes are then used as features for a random forest classifier (e).

was 0.88, with a 0.81–0.95 confidence interval, see Fig. 2a.

To investigate different word similarity measures, the random forest classifier's Gini importance was calculated to assess the value of each feature in our random forest classifier. Gini importance scores for each measure in the model are shown in Fig. 2b. See Fig. 3 for the combined Gini feature importance scores per similarity measure. For descriptive statistics of variance in similarity, see supplemental table S2. No significant Pearson's correlations were found between window ranges of variance and years of education (Supplemental table S3), word count (table S4) or dosage of antipsychotic medication (table S5), all $p>0.05$.

4. Discussion

The aim of the current study was to thoroughly investigate connectedness in language and its use for classification of diagnosis in language produced by participants with SSD. Using a word2vec model applied to transcriptions of recorded semi-spontaneous speech, we show connectedness in language as a robust feature suitable to classify SSD participants and healthy controls. Features of connectedness fit for classification were found over word ranges of varying window sizes in minimum and especially variance of word similarity, with features most informative for classification for variance at window sizes of 5–10 words.

Our results show that word connectedness features can be used to accurately classify participants as belonging to either patients or healthy controls with sensitivity 84% and specificity 86% using word2vec. These sensitivities and specificities are comparable to blood- or imaging-based markers which attain mean accuracies of 80.3% in classification (Schwarz et al., 2010; Zeng et al., 2018). We note that our participants were in remission, and scored low on positive symptom severity, highlighting the sensitivity of connectedness as a marker of SSD. As such, our findings support the applicability of spontaneous speech as a phenotypic, quantifiable marker for classification of SSD.

Our findings support previous research in which word connectedness was used to differentiate between healthy controls and participants with SSD (Cecchi., 2016; Iter et al., 2018b). Our results are comparable to these studies in accuracy and did not include other measures such as syntactic complexity. That we found such high accuracies using one tool might be partially explained by differences between distributed word representation types. Distributional word representations are based on co-occurrence matrices (e.g. Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997), Latent Dirichlet Allocation (LDA) (Blei et al., 2003)). In contrast, distributed tools use neural network language models to learn word representation, such as word2vec (Mikolov et al., 2013b, 2013a). Distributed word representations aim at capturing both semantic and syntactic information, and are better at preserving linear relations between words. In particular word2vec has been shown to outperform LSA in multiple studies (Glasgow et al., 2016; Villegas et al., 2016), although there is evidence that LSA is better suited for small corpora (Altszyler et al., 2016).

The range over which word connectedness was an informative measure for classification shows the importance of taking this methodological consideration into account. On closer examination, we notice that the features deemed most informative in the classifier correspond to variance in sentence-length windows (5–10 words), confirming earlier findings and other research investigating connectedness over windows (Corcoran et al., 2018; de Boer et al., 2018). These window sizes correspond to an average sentence length of spoken Dutch (Wiggers and Rothkrantz, 2007). The increased variance for participants with SSD at these values (See Table S2) indicate that subjects with SSD have more variation in word connectedness over sentence-size ranges. Minimum word connectedness, while included in our model, was a less

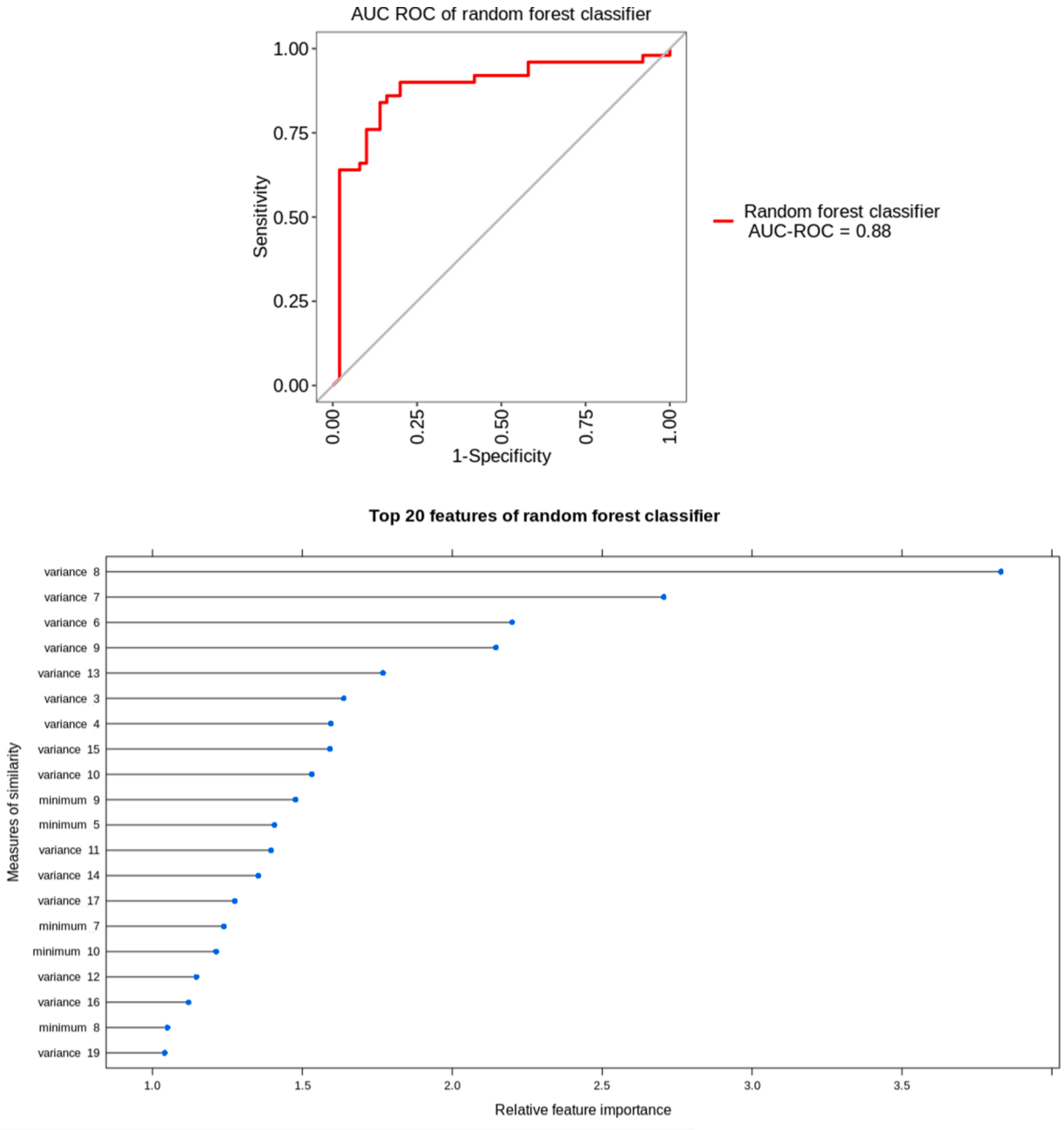

Fig. 2. a. Area under the curve – receiver operating characteristic of random forest classifier.
b: Relative feature importance for the trained random tree forest classifier.

informative feature than variance. Previous studies used different ways to invoke language production, such as spontaneous speech, picture retelling tasks or written text. We noted that for each variety of produced language the part of language suitable for classification differs per task. Other research has used non-sentence level tasks such as semantic verbal fluency to assess coherence of participants (Nicodemus et al., 2014). Therefore it would be prudent to further investigate differences between different language elicitation types, and explore a certain range of word connectedness windows for specific approaches. Attention should also be paid to our finding of variance of connectedness across the 5–10 word range, as previous research has found minimum or median features of connectedness or coherence to be informative (Bedi et al., 2015). A possible explanation for this finding could be related to our interview procedure. When a person answered a certain question and did not elaborate further on the topic, the next question would be asked, thereby introducing a change of topic. Given that patients typically give shorter answers to a question, two consecutive 'sentences' in patients will most likely be answers to at least two different questions. The controls were more likely to expand on a certain question spontaneously, thereby staying on topic for several consecutive windows.

Moreover, previous research by our group showed that the mean length of an utterance was a strong predictor of the integrity of the white matter language tracts in both patients with SSD and healthy controls (de Boer et al., 2020b). Previous studies have also shown that there is a relation between language connectedness and brain activity (Palaniyappan et al., 2019; Tagamets et al., 2014). It could thus be that a reduced integrity of the language tracts, underlies why patients give shorter answers, and might thus explain why this is such a strong predictor of group status in the current study. Other research has examined the hierarchical temporal and topological features of speech

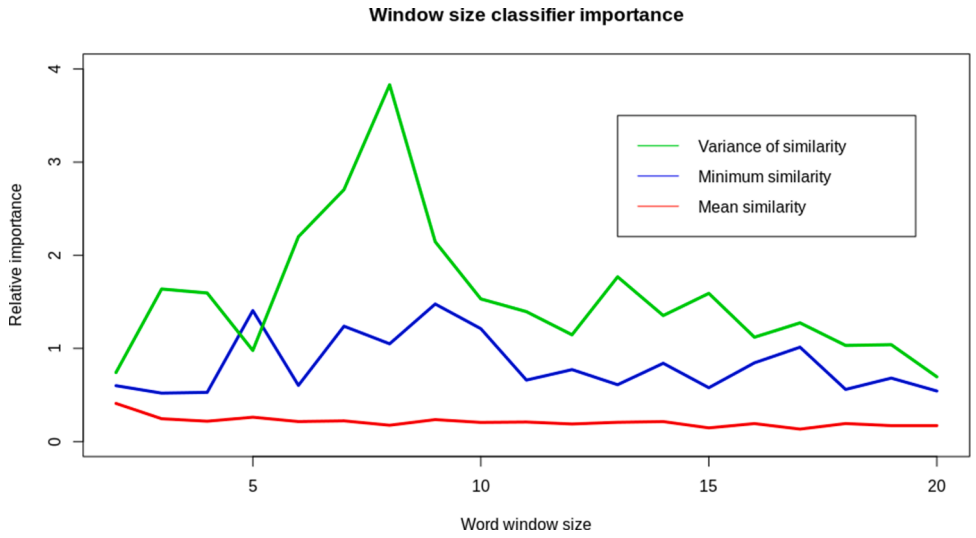

Fig. 3. Gini feature importance of similarity measures, ordered based on Gini scores. Lower values indicate lower importance in the binary classifier group. The numbers on the y-axis refer to window sizes. Note that there is substantial overlap in information over adjacent windows sizes, with a window of size 8 containing all the words of a window with size 7 and one additional word.

comprehension and production using functional imaging (Huth et al., 2016; Lerner et al., 2011). Applying these methods to subjects with SSD could inform us of the biological mechanisms of impaired coherence in speech. The classification results achieved using our approach can be seen as an initial step towards future research of word connectedness for SSD monitoring, diagnosis and prognosis. While we here explored word connectedness in depth, other sources of information should not be disregarded. To develop a precise phenotype of pathology, serving as a marker not only for the presence of pathology in general but one which is also useful to differentiate between disorders, a multitude of features and large samples are probably necessary due to overlapping features (i.e. reduced affect in both schizophrenia and depression) as well as overlap in clinical symptoms (i.e. depressive symptoms in psychosis-spectrum disorders; de Boer et al., 2020a). Such approaches should use different features of language such as phonetics, syntactic markers, or other semantic measures such as semantic density depending on the language characteristic of each disorder (Rezaii et al., 2019; for reviews, see Corcoran and Cecchi, 2020; de Boer et al., 2020a).

The progress in natural language processing and computational linguistics opens up opportunities for quantitative research of these linguistic and phonetic features; research from our group on phonetic markers in speech was similarly successful in classifying a different sample of patients from controls with a high accuracy based on a handful of acoustic features such as pausation characteristics (de Boer et al., 2020b).

4.1. Limitations and strengths

Since language is a high-level cognitive phenomenon, a variety of factors other than clinical group status could have an impact on its features. While we were able to match groups based on age and gender, years of education differed significantly between controls and participants with SSD. This group difference might partially confound our results as higher education has an effect on vocabulary (Milton and Treffers-Daller, 2013). However, the educational difference is not unexpected in SSD as the disorder usually manifests during the age at which education takes place (DeLisi, 1992). Using a correlational analysis, we did not find evidence of a relation between any of the calculated variance in connectedness and education in SSD participants (see supplemental Table S3), further making it unlikely that this group difference had an effect.

Methodological considerations should include the observation that participants with SSD produce less speech compared to healthy controls, in shorter sentences (Thomas et al., 1996) (see also Table 1). When assessing language connectedness with a moving window approach, short utterances lead to an increased percentage of different utterances within a calculated window, which is a possible explanation of our finding of increased variance in word connectedness in language produced by participants with SSD. Controlling for this confounding factor is difficult due to the absence of punctuation in spoken language, making utterance delineation a subjective process. Although we found no significant relation between the total amount of words spoken and variance of connectedness (table S4), and we find significant group differences over all window sizes, including windows where the overlap of utterances is small such as size 2 (table S2), sentence length could still play a role in these analyses.

As all our patients used antipsychotic medication, we cannot exclude an effect of medication on word and sentence level connectedness. Recent research from our group has shown that some measures of speech from subjects with SSD are influenced by antipsychotic medication (de Boer et al., 2020). While we cannot exclude antipsychotic medication as a confounder for analysis regarding semantic connectedness, we observed no significant correlation between any variance measures and the chlorpromazine equivalent dosage of (Table S5, all $p>0.05$).

While we employed repeated cross-validation to prevent model overfitting, the current study lacked an independent test sample. As a further limitation, we note that while transcribers were blinded to group status, group status could sometimes be inferred from speech produced by the participant (e.g. while talking about their last birthday, a participant might mention they were admitted).

Strengths of the current study include a relatively large sample of participants compared to previous speech research, using methodologically comprehensive word connectedness over a range of window sizes. Our findings provide support for the generalizability of previous findings in English and Spanish (Corcoran et al., 2018) to other languages (i.e. Dutch), which future research can expand on in order to position quantitative language as a reliable cross-linguistic biomarker

Concluding, word connectedness in language is an accurate and specific feature present over a range of window sizes, that can assist in the classification of patients with schizophrenia spectrum disorder and controls. High classification accuracy using machine learning classifiers can be achieved using language connectedness. By adding other language parameters such as speech acoustics or syntax, even higher accuracies are probably feasible. We found optimal discriminative ability

in the window size of 5–10 words. As the patients of our sample were mostly in remission of psychosis, this method appears to be sensitive even to low symptom levels. Our results add to the mounting evidence that a multitude of quantifiable linguistic measures are affected in schizophrenia spectrum disorders. Combining and fine-tuning these measures can help to accurately classify psychiatric disorders in a fast, non-invasive, reliable way.

Declaration of Competing Interest

IEC Sommer is a consultant to Gabather and received research support from Janssen Pharmaceuticals Inc. and Sunovion Pharmaceuticals Inc.

The other authors report not conflicts of interest.

Acknowledgments

The authors are grateful to all participants and wish to thank all research interns for their help with data collection and preparation.

Financial support

IEC Sommer received the TOP grant from The Netherlands Organization for Health Research and Development (ZonMW project: 91213009).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.psychres.2021.114130.

References

Altszyler, E., Sigman, M., Slezak, D.F., 2016. Comparative Study of LSA Vs Word2vec embeddings in Small corpora: a Case Study in Dreams Database. arXiv Prepr. arXiv1610.01520 1–14.

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders (DSM-5®). American Psychiatric Pub.

Andreasen, N.C., Flaum, M., Arndt, S., 1992. The Comprehensive Assessment of Symptoms and History (CASH): an instrument for assessing diagnosis and psychopathology. Arch. Gen. Psychiatry 49, 615–623.

Bar, K., Zilberstein, V., Ziv, I., Baram, H., Dershowitz, N., Itzikowitz, S., Vadim Harel, E., 2019. Semantic Characteristics of Schizophrenic Speech 84–93.

Bedi, G., Copelli, M., Javitt, D.C., Carrillo, F., Sigman, M., Slezak, D.F., Ribeiro, S., Cecchi, G.A., Corcoran, C.M., 2015. Automated analysis of free speech predicts psychosis onset in high-risk youths. npj Schizophr. 1, 15030. https://doi.org/10.1038/npjschz.2015.30 https://doi.org/.

Birnbaum, M.L., 2019. Detecting relapse in youth with psychotic disorders utilizing patient-generated and patient-contributed digital data from Facebook. npj Schizophr 1–9. https://doi.org/10.1038/s41537-019-0085-9 https://doi.org/.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Brundage, S.B., Bernstein Ratner, N., 2018. A Clinician's Complete Guide to CLAN and PRAAT 1–43.

Cecchi, G., 2016. A computational linguistics approach for prodromal psychosis. Neuropsychopharmacology 41, S97–S98. https://doi.org/10.1038/npp.2016.239 https://doi.org/.

Chaika, E.O., 1990. Understanding Psychotic speech: Beyond Freud and Chomsky. Charles C Thomas, Publisher.

Chomsky, N., 1957. Syntactic Structures. Mouton Publishers, The Hague.

Corcoran, C.M., Carrillo, F., Fernández-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. https://doi.org/10.1002/wps.20491 https://doi.org/.

Corcoran, C.M., Cecchi, G.A., 2020. Using Language Processing and Speech Analysis for the Identification of Psychosis and Other Disorders. Biol. Psychiatry Cogn. Neurosci. Neuroimaging 5, 770–779. https://doi.org/10.1016/j.bpsc.2020.06.004 https://doi.org/.

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. Schizophr. Res.

Covington, M.A., He, C., Brown, C., Naci, L., McClain, J.T., Fjordbak, B.S., Semple, J., Brown, J., M.A., 2005. Schizophrenia and the structure of language: the linguist's view. Schizophr. Res. 77, 85–98. https://doi.org/10.1016/j.schres.2005.01.016 https://doi.org/.

de Boer, J., Brederoo, S.G., Voppel, A.E., Sommer, I.E.C., 2020a. Anomalies in language as a biomarker for schizophrenia. Curr. Opin. Psychiatry 1. https://doi.org/10.1097/yco.0000000000000595, 0https://doi.org/.

de Boer, J., van Hoogdalem, M., Mandl, R., Brummelman, J., Voppel, A., Begemann, M., van Dellen, E., Wijnen, F., Sommer, I., 2020b. Language in schizophrenia: relation with diagnosis, symptomatology and white matter tracts. NPJ Schizophr. https://doi.org/10.1038/s41537-020-0099-3. Epub ahead of printhttps://doi.org/.

de Boer, J., Voppel, A.E., Begemann, M.J.H., Schnack, H.G., Wijnen, F., Sommer, I.E.C., 2018. Clinical use of semantic space models in psychiatry and neurology: a systematic review and meta-analysis. Neurosci. Biobehav. Rev. 93, 85–92. https://doi.org/10.1016/j.neubiorev.2018.06.008 https://doi.org/.

de Boer, J.N., Voppel, A.E., Brederoo, S.G., Wijnen, F.N.K., Sommer, I.E.C., 2020. Language disturbances in schizophrenia: the relation with antipsychotic medication. npj Schizophr. 6, 1–9. https://doi.org/10.1038/s41537-020-00114-3 https://doi.org/.

DeLisi, L.E., 1992. The significance of age of onset for schizophrenia. Schizophr. Bull. 18, 209–215. https://doi.org/10.1093/schbul/18.2.209 https://doi.org/.

DeLisi, L.E., 2001. Speech disorder in schizophrenia: review of the literature and exploration of its relation to the uniquely human capacity for language. Schizophr. Bull. 27, 481–496. https://doi.org/10.1093/oxfordjournals.schbul.a006889 https://doi.org/.

Elvevåg, Brita, 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. J. Neurolinguist. 23, 270–284. https://doi.org/10.1016/j.jneuroling.2009.05.002 https://doi.org/.

Glasgow, K., Roos, M., Haufler, A., Chevillet, M., Wolmetz, M., 2016. Evaluating semantic models with word-sentence relatedness. arXiv 1–8.

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 https://doi.org/.

Gupta, T., Hespos, S.J., Horton, W.S., Mittal, V.A., 2018. Automated analysis of written narratives reveals abnormalities in referential cohesion in youth at ultra high risk for psychosis. Schizophr. Res. 192, 82–88. https://doi.org/10.1016/j.schres.2017.04.025 https://doi.org/.

Holshausen, K., Harvey, P.D., Elvevåg, B., Foltz, P.W., Bowie, C.R., 2014. Latent semantic variables are associated with formal thought disorder and adaptive behavior in older inpatients with schizophrenia. Cortex 55, 88–96.

Huth, A.G., De Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L., 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532, 453–458. https://doi.org/10.1038/nature17637 https://doi.org/.

Iter, D., Yoon, J., Jurafsky, D., 2018. Automatic Detection of Incoherent Speech For Diagnosing Schizophrenia. pp. 136–146. https://doi.org/ 10.18653/v1/w18-0615.

Kay, S.R., Fiszbein, A., Opler, L.A., 1987. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr. Bull. 13, 261–276. https://doi.org/10.1093/schbul/13.2.261 https://doi.org/.

Kayi, E.S., Diab, M., Pauselli, L., Compton, M., Coppersmith, G., 2017. Predictive linguistic features of schizophrenia. In: *SEM 2017 - 6th Jt. Conf. Lex. Comput. Semant. Proc, pp. 241–250.

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 https://doi.org/.

Kuperberg, G.R., 2010. Language in schizophrenia part 1: an introduction. Lang. Linguist. Compass 4, 576–589.

Landauer, T., Dumais, S., 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol. Rev. 104, 211–240.

Landauer, T., Foltz, P., Laham, D., 1998. An introduction to latent semantic analysis. Discourse Process 25, 259–284. https://doi.org/10.1080/01638539809545028 https://doi.org/.

Lerner, Y., Honey, C.J., Silbert, L.J., Hasson, U., 2011. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. J. Neurosci. 31, 2906–2915. https://doi.org/10.1523/JNEUROSCI.3684-10.2011 https://doi.org/.

Leucht, S., Samara, M., Heres, S., Patel, M.X., Woods, S.W., Davis, J.M., 2014. Dose equivalents for second-generation antipsychotics: the minimum effective dose method. Schizophr. Bull. 40, 314–326.

MacWhinney, B., 2000. The CHILDES project: Tools for Analyzing talk: Volume I: Transcription format and Programs, II. The database.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013a. Efficient estimation of word representations in vector space. Arxiv 1–12. https://doi.org/10.1162/153244303322533223 https://doi.org/.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013b. Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems. pp. 3111–3119.

Milton, J., Treffers-Daller, J., 2013. Vocabulary size revisited: the link between vocabulary size and academic achievement. Appl. Linguist. Rev. 4, 151–172. https://doi.org/10.1515/applirev-2013-0007 https://doi.org/.

Mota, N.B., Furtado, R., Maia, P.P.C., Copelli, M., Ribeiro, S., 2014. Graph analysis of dream reports is especially informative about psychosis. Sci. Rep. 4, 1–7. https://doi.org/10.1038/srep03691 https://doi.org/.

Mota, N.B., Vasconcelos, N.A.P., 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. https://doi.org/10.1371/journal.pone.0034928 https://doi.org/.

Nicodemus, K.K., Elvevåg, B., Foltz, P.W., Rosenstein, M., Diaz-Asper, C., Weinberger, D. R., 2014. Category fluency, latent semantic analysis and schizophrenia: a candidate gene approach. Cortex 55, 182–191. https://doi.org/10.1016/j.cortex.2013.12.004 https://doi.org/.

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. Prog. Neuro-Psychopharmacol. Biol. Psychiatry 88, 112–120. https://doi.org/10.1016/j.pnpbp.2018.07.007 https://doi.org/.

Pauselli, L., Halpern, B., Cleary, S.D., Ku, B., Covington, M.A., Compton, M.T., 2018. Computational linguistic analysis applied to a semantic fluency task to measure derailment and tangentiality in schizophrenia. Psychiatry Res. 263, 74–79. https://doi.org/10.1016/j.psychres.2018.02.037 https://doi.org/.

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 https://doi.org/.

Rössler, Wulf, Salize, Hans Joachim, van Os, Jim, Riecher-Rössler, Anita, 2005. Size of burden of schizophrenia and psychotic disorders. European Neuropsychopharmacology 15(4), 399–409. https://doi.org/10.1016/j.euroneuro.2005.04.009.

Schwarz, E., Izmailov, R., Spain, M., Barnes, A., Mapes, J.P., Guest, P.C., Rahmoune, H., Pietsch, S., Markus Leweke, F., Rothermundt, M., Steiner, J., Koethe, D., Kranaster, L., Ohrmann, P., Suslow, T., Levin, Y., Bogerts, B., van Beveren, N., McAllister, G., Weber, N., Niebuhr, D., Cowan, D., Yolken, R.H., Bahn, S., 2010. Validation of a blood-based laboratory test to aid in the confirmation of a diagnosis of schizophrenia. Biomark. Insights 2010, 39–47. https://doi.org/10.4137/bmi.s4877 https://doi.org/.

Sheehan, D., Janavs, J., Baker, R., Harnett-Sheehan, K., Knapp, E., Sheehan, M., Lecrubier, Y., Weiller, E., Hergueta, T., Amorim, P., 1998. MINI-Mini International neuropsychiatric interview-english version 5.0. 0-DSM-IV. J. Clin. Psychiatry 59, 34–57.

Spencer, T.J., Thompson, B., Oliver, D., Diederen, K., Demjaha, A., Weinstein, S., Morgan, S.E., Day, F., Valmaggia, L., Rutigliano, G., De Micheli, A., Mota, N.B., Fusar-Poli, P., McGuire, P., 2021. Lower speech connectedness linked to incidence of psychosis in people at clinical high risk. Schizophr. Res. 228, 493–501. https://doi.org/10.1016/j.schres.2020.09.002 https://doi.org/.

Tagamets, M.A., Cortes, C.R., Griego, J.A., Elvevåg, B., 2014. Neural correlates of the relationship between discourse coherence and sensory monitoring in schizophrenia. Cortex 55, 77–87. https://doi.org/10.1016/j.cortex.2013.06.011 https://doi.org/.

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 https://doi.org/.

Thomas, P., Leudar, I., Napier, E., Kearney, G., Ellis, E., Ring, N., Tantam, D., 1996. Syntactic complexity and negative symptoms in first onset schizophrenia. Cogn. Neuropsychiatry 1, 191–200. https://doi.org/10.1080/135468096396497 https://doi.org/.

Turian, J., Ratinov, L., Bengio, Y., Turian, J., 2010. Word Representations: a Simple and General Method for Semi-supervised Learning. In: Proc. 48th Annu. Meet. Assoc. Comput. Linguist., pp. 384–394 https://doi.org/10.1.1.301.5840.

van Eerten, L., 2007. Corpus gesproken Nederlands. Ned. Taalkd. 12, 194–215.

Vanwinckelen, G., Blockeel, H., 2012. On estimating model accuracy with repeated cross-validation. 21st Belgian-Dutch Conf. Mach. Learn. 39–44.

Villegas, M.P., José, M., Ucelay, G., Fernández, J.P., Álvarez-Carmona, M.A., Errecalde, M.L., Cagnina, L.C., Garciarena Ucelay, M.J., 2016. Vector-based word representations for sentiment analysis: a comparative study. XXII Congr. Argentino Ciencias la Comput. (CACIC 2016) 785–793.

Wiggers, P., Rothkrantz, L.J.M., 2007. Exploratory Analysis of Word Use and Sentence Length in the Spoken Dutch Corpus, in: Matoušek, V., Mautner, P. (Eds.), Text, Speech and Dialogue. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 366–373.

Zeng, L.L., Wang, H., Hu, P., Yang, B., Pu, W., Shen, H., Chen, X., Liu, Z., Yin, H., Tan, Q., Wang, K., Hu, D., 2018. Multi-Site Diagnostic Classification of Schizophrenia Using Discriminant Deep Learning with Functional Connectivity MRI. EBioMedicine 30, 74–85. https://doi.org/10.1016/j.ebiom.2018.03.017 https://doi.org/.

R. Řehůřek, P. Sojka, "Software framework for topic modelling with large corpora", Proc. LREC Workshop New Challenges for NLP Frameworks, pp. 45–50, May 2010.