Syntactic Network Analysis in Schizophrenia-Spectrum Disorders

Ciampelli, Silvia; de Boer, Janna N.; Voppel, Alban E.; Corona Hernandez, Hugo; Brederoo, Sanne G.; van Dellen, Edwin; Mota, Natalia B.; Sommer, Iris E.C.

Published in:
Schizophrenia Bulletin

DOI:
10.1093/schbul/sbac194

Publication date:
2023

Schizophrenia Bulletin vol. 49 suppl. 2 pp. S172–S182, 2023
https://doi.org/10.1093/schbul/sbac194

SUPPLEMENT ARTICLE

Syntactic Network Analysis in Schizophrenia-Spectrum Disorders

Silvia Ciampelli[*,1,2,4], Janna N. de Boer[2,3,4], Alban E. Voppel[4], Hugo Corona Hernandez[4], Sanne G. Brederoo[4,5], Edwin van Dellen[2,3], Natalia B. Mota[6], and Iris E. C. Sommer[4,5]

[1]Faculty of Science, University of Amsterdam, Amsterdam, The Netherlands; [2]Department of Psychiatry, UMC Brain Center, University Medical Center Utrecht, Utrecht, The Netherlands; [3]Department of Intensive Care Medicine, UMC Brain Center, University Medical Center Utrecht, Utrecht, The Netherlands; [4]Department of Biomedical Sciences of Cells and Systems, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands; [5]Department of Psychiatry, University Medical Center Groningen, Groningen, The Netherlands; [6]Institute of Psychiatry, Federal University of Rio de Janeiro (IPUB-UFRJ), Rio de Janeiro, Brazil

[*]To whom correspondence should be addressed: Faculty of Science, University of Amsterdam, Science Park 904, 1098 XH, Amsterdam, The Netherlands; tel: 393339538146, e-mail: s.ciampelli@umcutrecht.nl

***Background*: Language anomalies are a hallmark feature of schizophrenia-spectrum disorders (SSD). Here, we used network analysis to examine possible differences in syntactic relations between patients with SSD and healthy controls. Moreover, we assessed their relationship with sociodemographic factors, psychotic symptoms, and cognitive functioning, and we evaluated whether the quantification of syntactic network measures has diagnostic value. *Study Design*: Using a semi-structured interview, we collected speech samples from 63 patients with SSD and 63 controls. Per sentence, a syntactic representation (ie, parse tree) was obtained and used as input for network analysis. The resulting syntactic networks were analyzed for 11 local and global network measures, which were compared between groups using multivariate analysis of covariance, considering the effects of age, sex, and education. *Results*: Patients with SSD and controls significantly differed on most syntactic network measures. Sex had a significant effect on syntactic measures, and there was a significant interaction between sex and group, as the anomalies in syntactic relations were most pronounced in women with SSD. Syntactic measures were correlated with negative symptoms (Positive and Negative Syndrome Scale) and cognition (Brief Assessment of Cognition in Schizophrenia). A random forest classifier based on the best set of network features distinguished patients from controls with 74% cross-validated accuracy. *Conclusions*: Examining syntactic relations from a network perspective revealed robust differences between patients with SSD and healthy controls, especially in women. Our results support the validity of linguistic network analysis in SSD and have the potential to be used in combination with other automated language measures as a marker for SSD.**

Keywords: biomarker/language/psychosis/syntax/sex differences

Introduction

Language disturbances in patients with schizophrenia-spectrum disorders (SSD) have been observed since the disease's earliest conceptualizations,[1] and they remain one of the 5 diagnostic criteria for SSD.[2] Language anomalies in SSD have been particularly associated with problems in the content (ie, semantics) and structure (ie, syntax) of language, resulting in speech that is sometimes difficult to understand. Recent developments in natural language processing (NLP) have focused mostly on analyzing the semantic aspects of language in SSD, eg, examining the degree to which words are coherent in their context.[3–5]

Yet not only content but also structure is needed to form an intelligible sentence. Consider for instance the following sentence "Colourless green ideas sleep furiously," while being syntactically correct, it is senseless.[6] Likewise, while the individual words in "Every morning at breakfast the eggs would eat…"[7(p4)] are strongly semantically related (morning-breakfast-egg-eat), syntactically the sentence is implausible, ie, an "egg" being an inanimate object prevents from performing the act of eating.[7] In addition, the meaning of a sentence can change depending on specific syntactic relationships between words that are independent of their content.[8] For instance, the interrogative construction "*Has she bought the book?*" is different both in structure and meaning from the declarative sentence "*She has bought the book*," although the words are the same.

Changing the structure thus changes the meaning of the sentence. Although semantics and syntax are closely intertwined, these examples also show that syntax is a semi-autonomous module that can be studied separately from semantics.[6] Previous research that looked at syntax indeed found that the language of patients with SSD displays syntactic abnormalities. More specifically, a simplification of the syntactic production has consistently been found in patients with SSD,[9,10] and they produce more[11,12] and more severe[13] syntactic errors as compared to controls.

However, quantifying syntax has proven difficult since there is no standard way to measure complexity or syntactic relations between words computationally. Moreover, earlier studies that automatically analyzed syntax in SSD also included additional language measures in their models (eg, semantic measures), preventing the disentanglement between semantic and syntactic properties.[3,14] For example, a previous study applied network modeling to language in SSD and demonstrated that network analytics based on word co-occurrences could capture features of alogia and disconnected speech.[15] These models, also called speech graphs, were successfully used to predict conversion to schizophrenia in patients with the first episode of psychosis.[16] The networks used in this study modeled co-occurrence patterns between successively uttered words, without specifying the nature of the linguistic relationships between them (eg, syntactic or semantic).[17]

Here, we aimed to employ network analysis to specifically look at syntactic relationships between words in patients with SSD. In linguistics, syntactic relations within a sentence are visualized in a specific type of network, a so-called parse tree. We propose that parse trees can be studied using a formal framework for network analysis[18] since they have nodes (ie, words and syntactic categories) and edges (ie, syntactic relations between them).

Importantly, sex is a key determinant of variation in language and speech, and differences between men and women exist in syntax across life stages and languages.[19–21] These differences were shown to be relevant in SSD, as male patients had significantly lower syntactic complexity than female patients, with fewer embedded clauses.[22] In this study, we therefore specifically considered the effects that sex had on syntactic networks while controlling for age and education.

We hypothesized the networks of patients with SSD to be smaller, less connected, and less hierarchically structured compared to those of healthy controls, corresponding with simplified syntax in SSD. Furthermore, we explored the relationship of syntactic network measures with psychotic symptoms and cognition, and we evaluated the diagnostic value of syntactic network measures in distinguishing between patients with SSD and control participants.

Methods

Subjects

Sixty-three patients with SSD and sixty-three healthy controls were recruited at the University Medical Center Utrecht (UMCU) between 2015 and 2020. The current sample is a subset of the participants described previously by our group.[23,24] Participants were matched for age and sex. Considering that the onset of psychosis often occurs during educational years, we expected patients to have less education than controls, and therefore we matched both patients and controls for parental years of education. Inclusion criteria for all participants were: age > 18 years, Dutch as the native language and absence of hearing impairment or speech disorder. Patients were included if they met the criteria for a DSM-5 diagnosis of: 295.90 (schizophrenia, schizophreniform disorder, schizoaffective disorder) or 298.9 (psychotic disorder not otherwise specified). The diagnosis was established in all patients by their treating psychiatrist and was confirmed by a trained researcher using the Comprehensive Assessment of Symptoms and History interview[25] or the Mini-International Interview.[26] All comorbidities were registered. Healthy control participants were included if they had no mental health complaints and no family history of psychotic symptoms: past episodes of depression or anxiety disorders in full remission were not an exclusion criterion. Written consent forms were signed by all participants prior to participation, and participants received a small monetary compensation (10 euros). The study was approved by the ethical review board of the UMCU.

Assessment of Symptom Severity and Cognition

Psychotic symptom severity was measured with the Positive and Negative Syndrome Scale (PANSS).[27] Cognition was assessed in 61 patients and 24 healthy controls using the Brief Assessment of Cognition in Schizophrenia (BACS).[28] Individual BACS results were converted into standardized z-scores based on existing normative scores, which were controlled for age and sex.[29]

Speech Acquisition and Pre-processing

Spontaneous speech was elicited through a semi-structured interview of approximately 15 min, which was devised especially for this purpose.[30] The interview consists of open-ended questions about informal topics, such as everyday activities and sports experiences. Topics that could trigger an emotional response in participants were avoided to control for possible variations in language caused by the topics covered. Participants' speech was recorded and manually transcribed according to CHILDES-CHAT guidelines.[31] Interview transcripts were converted to plaintext for parsing and subsequent analysis. Punctuation was removed; repetitions and interjections were preserved.

Syntactic Parsing

For each subject, up to 120 sentences were randomly selected for analysis. Across groups (ie, controls vs patients) and sexes (ie, men vs women), the longest interview produced by healthy women contained 120 sentences, this being the lowest compared to other groups. To prevent oversampling, we chose 120 as a higher upper bound for the sentences of other groups as well. The average, minimum, and maximum number of sentences produced by group and sex are included in the supplements (Supplementary table S1). Each sentence was individually parsed into Alpino[32] via Docker (https://github.com/rug-compling/alpino-docker) and the output was received in Python (version 3.9).[33] Parse trees are a specific type of network that belong to the class of "directed rooted tree" and they can be analyzed according to the formal framework for network analysis.[18] In syntax, parse trees describe the internal syntactic structure of a sentence by identifying relations between higher-level structural units. In the parse tree, each place of division (ie, a word or syntactic category) was considered a "node". Lines between the nodes were considered "edges". Edges connect nodes in a directional way where lower layers depend on upper layers. The XML version of the parse trees was processed in Python with the "xml.etree.ElementTree" module and subsequently converted into a so-called "edge list," which refers to each combination of existing connections between any 2 nodes in a tree. For example, a noun phrase (NP) with 3 connections such as "The red apple" will result into an edge list of the type [1NP-2The, 1NP-3red, 1NP-4apple], where "NP" is a parent node and "The," "red," "apple" are children nodes (figure 1).

Each node was numbered according to its linguistic position in the Alpino tree, eg, the "top" node recoded as node 0. Linguistically, it is considered meaningful to number from top to bottom and left to right given that Dutch has a sinistrodextral writing system. The "top" node, which corresponds to the root of the sentence, and the "let" node, which represents punctuation, were discarded since these nodes appeared in the tree representation of all sentences, and therefore did not have additional informative value.

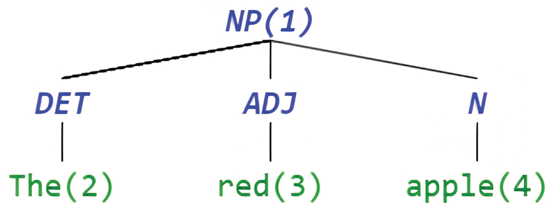


Fig. 1. Parse tree of the noun phrase "The red apple".

NP, noun phrase; DT, determiner; ADJ, adjective; N, noun.

Network Analysis

Each sentence (ie, each edge list) was imported into Cytoscape (version 3.8.2).[34] The network analyzer tool in Cytoscape was used to calculate global properties that quantify features of the entire network and local properties that capture information at the level of the node and its surrounding. Analyses were not performed on sentences containing less than 6 nodes due to the limited amount of data extractable. For each sentence, we calculated 8 network measures: nodes, leaf count and leaf fraction, degree, stress centrality, efficiency, betweenness centrality, and diameter (table 1).[35] For the maximum connected utterances (ie, with the highest edge count), we identified the largest connected component (LCC), which refers to the maximal connected subgraph containing nodes with the highest degree. Three measures were calculated for the LCC: nodes, diameter, and degree. Since the decreased verbosity of the patients resulted in networks of a smaller size (ie, lower number of nodes and leaves), we normalized the values of global network measures (ie, diameter and efficiency) by the total number of nodes in each graph.

Statistical Analyses and Classification

Demographic variables were compared between groups using independent samples *t*-tests for continuous, and chi-square analyses for categorical data. Network measures were compared between groups through multivariate analysis of covariance (MANCOVA). In the model, age and parental education were entered as covariates, while group and sex were treated as factors. The association between network measures and clinical ratings (PANSS, BACS) was assessed with Bivariate Pearson Correlation. For the variables that revealed a significant correlation with the BACS composite score, post-hoc correlational analyses were performed to see which BACS sub-domains were associated with the network variables. Correlation analyses were corrected for multiple comparisons using false discovery rate.

A Random Forest Classifier (RFC)[36] was built in Python[37] to distinguish controls from patients with SSD based on syntactic network measures. Gini coefficient was used to separately calculate the value of each syntactic network measure in relation to all other measures in the classifier. The best measures (ie, with the highest Gini coefficient) were subsequently used to build the final classifier. In the classifier, the number of tree estimators was 100, and the random state parameter was set to 5. Twentyfold cross-validation was performed, which randomly split the data set into 20 independent sections. Per iteration, a different section of the data set was held back for testing, while the remaining 19 sections were used for training. The test results obtained from each iteration were then averaged to calculate the final

Table 1. Formal Definition and Linguistic Interpretation of Network Measures

Concept	Definition	Linguistic Interpretation
Betweenness centrality	Fraction of shortest paths passing through a node	A node with high betweenness centrality is relevant in maintaining syntactic relations between other nodes
Degree	Number of edges per node	Number of syntactic relations per word and syntactic structure
Diameter	The maximum length of the shortest path between any 2 nodes in the network	The maximum distance of the relation between words and syntactic structures in the network
Efficiency	The average inverse shortest path length	The integration between words and syntactic structures
Leaf count	Number of leaves (nodes with only one connection) in the network	Number of words in the network
Leaf fraction	Fraction of leaves out of the total number of nodes in the network	Fraction of words out of the total number of nodes in the network
Nodes	Number of nodes in the network	Number of words and syntactic structures in the network
Stress centrality	The number of shortest paths passing through a node	A node with high stress centrality is a structurally central node involved in most syntactic relations

Note: Formal definitions of network measures were obtained via Cytoscape, see User Manual version 3.9.1 documentation.[33]

performance, which was based on accuracy, specificity, and sensitivity scores.

Results

Demographics

Clinical and demographic information is presented in table 2. Demographic characteristics and patients' comorbidities categorized by sex are listed in the supplements (Supplementary tables S2–4). The groups did not differ regarding sex, age, and parental education. For exploratory purposes, the effect of comorbidities on syntactic network measures was assessed, which revealed no significant effects (see supplementary results).

Correlational analyses investigating the effect of antipsychotic medication in our sample were performed. We did not find evidence of a relation between chlorpromazine equivalent dosage and PANSS (all $P > .050$) (Supplementary table S5). A negative correlation was observed between chlorpromazine equivalent dosage and three network measures, namely nodes ($r = -0.289$, $P = .027$), leaves ($r = -0.292$, $P = .015$), and efficiency ($r = -0.297$, $P = .023$), whereas a positive relation was found for diameter ($r = 0.316$, $P = .015$) (Supplementary table S6).

Network Analysis

The MANCOVA revealed significant differences between patients and controls on the global and local network measures (F(11,110) = 6.249, Pillai's trace = 0.385, $P <$ .001). Sex had a significant effect on network measures (F(11,110) = 2.184, Pillai's trace = 0.179, $P = .020$) and a positive interaction between group and sex was found (F(11,110) = 2.198, Pillai's trace = 0.180, $P = .019$), indicating that sex had a different effect on the network measures in the patients than in the controls. No main effect for age (F(11,110) = 1.816, Pillai's trace = 0.154, $P = .060$) and education (F(11,110) = 0.635, Pillai's trace = 0.60, $P = .796$) was observed. Post-hoc analyses revealed that on average patients with SSD produced networks with a lower number of nodes and leaves, lower degree, lower stress centrality, higher diameter, and lower efficiency (figure 2, Supplementary table S7). Out of the 3 measures calculated for the LCC, nodes and diameter reached statistical significance, whereas degree did not differ among the 2 groups.

Sex differences between groups revealed that both women and men in the control group had network measures that significantly differ from their counterparts in the patient group ((F(11,28) = 4.935, Pillai's trace = 0.660, $P < .001$), (F(11,74) = 2.695, Pillai's trace = 0.286, $P = .006$), respectively), however, these differences were more pronounced in women. Further post-hoc analyses showed that women on average had networks with a higher number of nodes ($F = 5.209$, $P = .024$), leaves ($F = 5.406$, $P = .022$), higher efficiency ($F = 4.343$, $P = .039$), and higher stress centrality ($F = 5.913$, $P = .016$), relative to men. Post-hoc analyses of the interaction effect revealed that the interaction between sex and group was significant for nodes, diameter, efficiency, stress centrality, degree, LCC nodes, and LCC diameter (Supplementary figure 2).

Association With Psychotic Symptoms and Cognition

A moderate correlation with the PANSS negative subscale was found for nodes, leaves, diameter, efficiency, stress centrality, and LCC node (Supplementary table S8). The BACS composite score was correlated with nodes, leaves, diameter, degree, stress centrality, efficiency, LCC nodes, and LCC diameter (Supplementary table S8). PANSS

Table 2. Demographic Characteristics of Patients With SSD and Healthy Controls

	SSD Patients (*n* = 63)	Healthy Controls (*n* = 63)	Test Statistics
Age (years)	34.1 ± 13.19	34.3 ± 13.70	*F* = 0.004, *P* = .952
Male sex	43 (68.25)	43 (68.25)	χ^2 = 0.000, *P* = 1.000
Parental education (years)	12.1 ± 2.85	12.3 ± 2.98	*F* = 0.031, *P* = .860
Illness duration (years)	9.0 ± 12.12		
Chlorpromazine dose (milligram equivalent)	236.6 ± 162.03		
Diagnosis			
Psychosis NOS	25 (35%)		
Schizoaffective disorder	7 (10%)		
Schizophrenia	29 (41%)		
Schizophreniform disorder	2 (3%)		
PANSS total	50.0 ± 13.22		
Positive	11.2 ± 4.27		
Negative	12.7 ± 4.66		
General	26.1 ± 7.05		
BACS composite *Z*-score	−1.4 ± 1.10	0.3 ± 1.36	*F* = 47.148, *P* < .001
List learning—Verbal memory	−1.0 ± 1.23	0.6 ± 1.17	*F* = 30.446, *P* < .001
Digit sequencing—Working memory	−1.0 ± 1.21	0.2 ± 1.12	*F* = 17.139, *P* < .001
Token motor task—Motor speed	−0.8 ± 1.30	−0.2 ± 1.22	*F* = 4.915, *P* = .029
Category Instances and Controlled Oral Word Association Test—Verbal fluency	−0.1 ± 1.06	0.1 ± 1.19	*F* = 15.991, *P* < .001
Symbol coding—Attention and information processing speed	−1.2 ± 0.90	0.1 ± 1.35	*F* = 27.298, *P* < .001
Tower of London—Executive function	−0.2 ± 1.37	0.3 ± 0.74	*F* = 2.882, *P* = .097

Note: Reported values are mean ± SD or *n* (%).
Note: *n*, sample size; SD, standard deviation; PANSS, Positive and Negative Syndrome Scale; NOS, not otherwise specified; BACS, Brief Assessment of Cognition in Schizophrenia.

positive and general subscales showed no significant correlations with syntactic network measures. Post-hoc correlation analyses revealed that most syntactic network measures were significantly associated with Symbol Coding and Tower of London (Supplementary table S9). LCC node was correlated with verbal fluency (*r* = 0.27, *P* = .039).

Classification

When trained on all network measures, the RFC distinguished patients with SSD from controls with 70% accuracy (sensitivity, 72%; specificity, 67%), and a receiver operating characteristics (ROC) of 0.74. The network measures that were found most important were degree (0.267), leaf count (0.190), and stress centrality (0.181). When trained with these measures, the RFC was able to distinguish patients from controls with an accuracy of 74% (sensitivity, 76%; specificity, 73%) and a ROC curve of 0.77. The classification accuracy boxplots using increasing numbers of features have been included in the supplements (Supplementary figure S1). To discard the hypothesis that these results could be explained exclusively by the differences in sentence length between the groups, we trained an RFC on the mean number of words per sentence and compared its discriminatory power to that of the model fitted with a combination of syntactic network measures. The RFC trained solely on word count showed a decrease in all evaluation metrics with an accuracy of 66% (sensitivity, 65%; specificity 68%) and a ROC of 0.69. Considering that differences in network measures between patients and controls were most pronounced in women, we trained an RFC on the women and men groups separately. The discriminatory power of the RFC trained on women improved with an accuracy of 79% (sensitivity 80%, specificity 80%) while the performance of the RFC fitted on men diminished with an accuracy of 65% (sensitivity 67%, specificity 66%).

Discussion

Our results show that patients with SSD and healthy controls differ on several syntactic network measures, indicating robust differences in syntactic relations between the groups. These differences were associated with negative symptoms and cognitive functioning. When using syntactic network measures for classification, an accuracy of 74% was reached in distinguishing patients with SSD from healthy controls. The classifier assigned a higher likelihood to an SSD diagnosis when the syntactic network is less connected (ie, degree), of smaller size (ie, leaves), and less centralized (ie, stress centrality) (Supplementary figures S2 and S3). Accuracy improved to 79% when focusing only

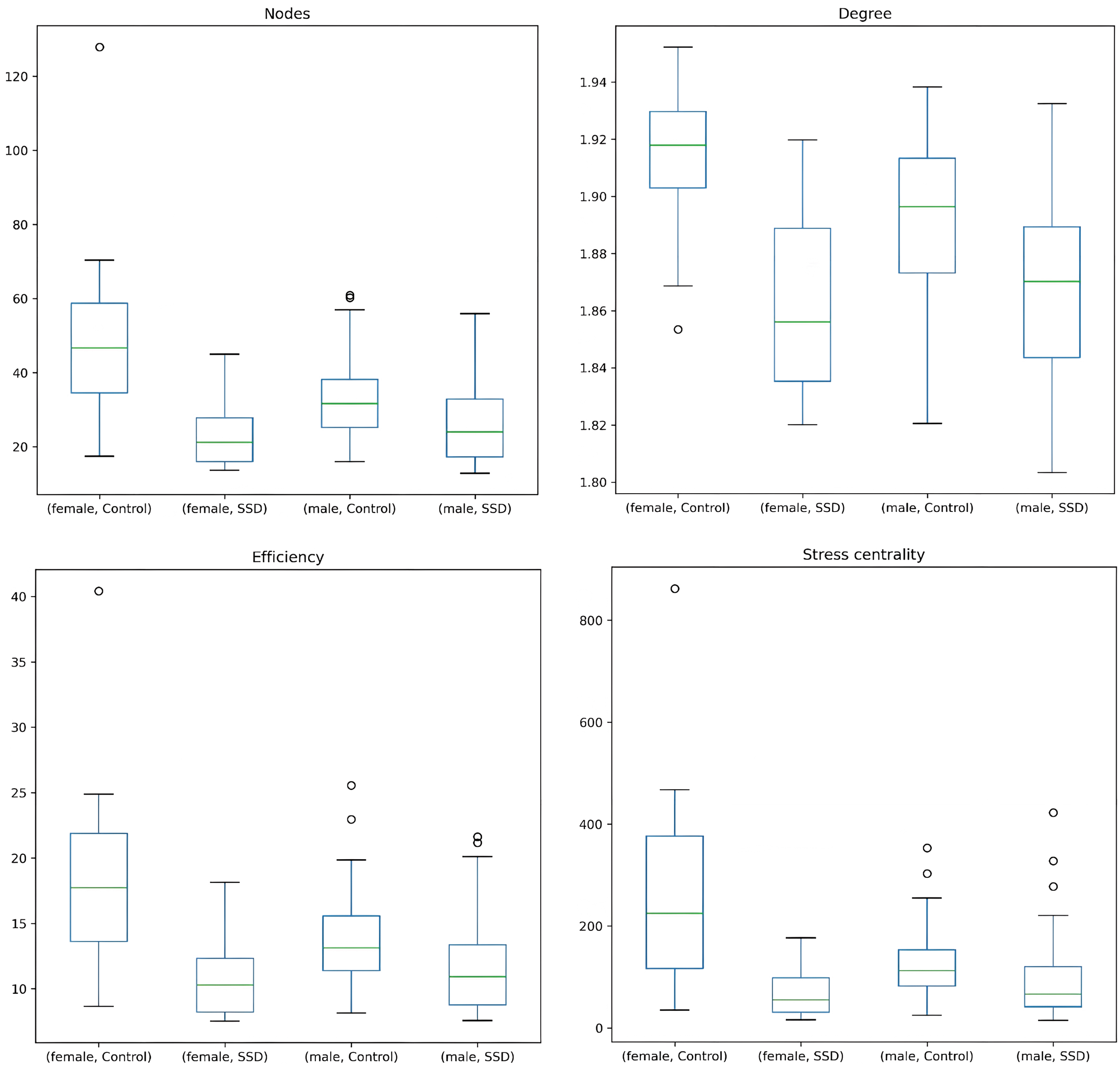


Fig. 2. Boxplot of syntactic network measures.
Nodes ($P < .001$), degree ($P < .001$), efficiency ($P < .001$), stress centrality ($P < .001$). Betweenness centrality ($P = .183$), diameter ($P < .001$), LCC nodes ($P < .001$), LCC diameter ($P = .048$). P-values refer to the significance of the difference in network measures between healthy controls and patients with SSD. For full statistical results, see Supplementary table S7.

on women, in whom we showed the anomalies in syntactic relations to be most pronounced.

Differences Between Patients With SSD and Healthy Controls

Closer examination of syntactic network measures showed that the networks of patients with SSD had a lower number of edges per node (ie, degree). Since word nodes (ie, leaves) only have 1 edge, the differences in degree between patients and controls are to be attributed to a difference in the number of edges per syntactic node (ie, syntactic structure), which might indicate lower syntactic complexity in the speech of patients. Syntactic complexity can be defined as the number of arguments and adjuncts that constitute a syntactic construct.[38] An argument is a linguistic unit whose presence is required by another expression in a sentence (example 1 in figure 3), as compared to adjuncts whose presence is optional (example 2 in figure 3).

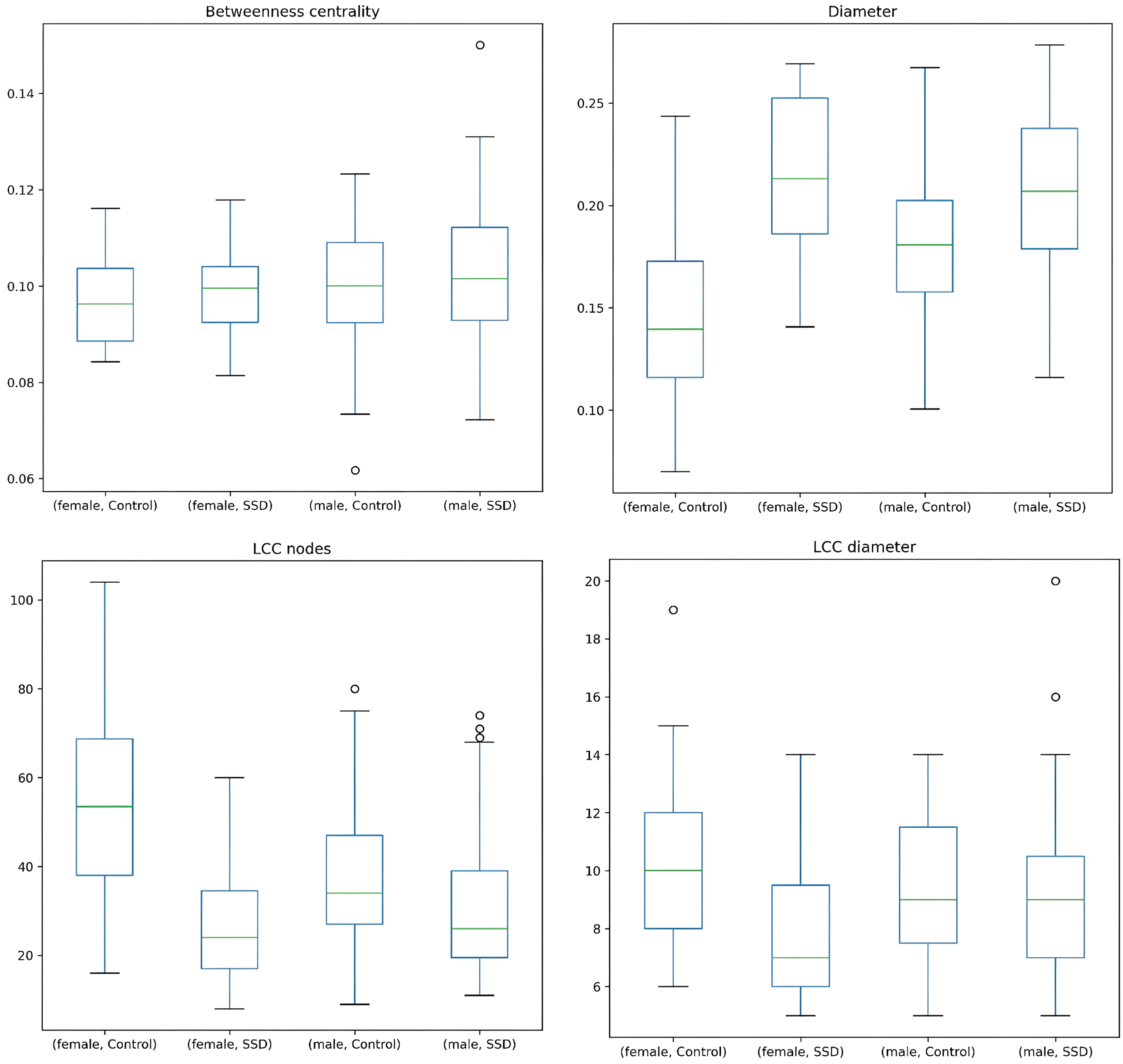
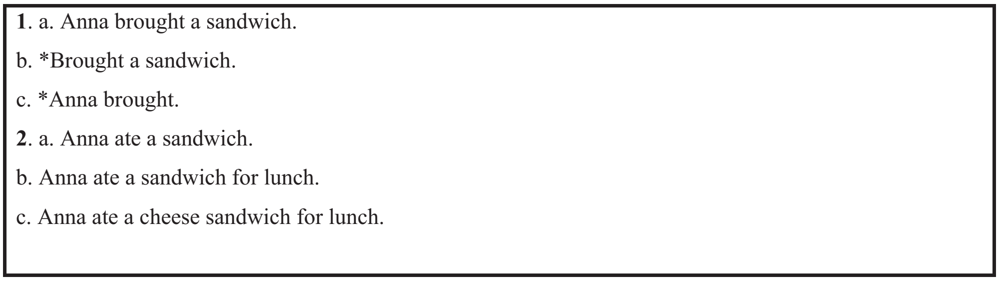


Fig. 2. Continued

The more arguments (ie, Anna) and adjuncts (ie, for lunch), constituting the syntactic structures (ie, brought/ate), the more complex the syntax is. A syntactic structure with fewer arguments and adjuncts may result in syntactic nodes with fewer edges. Another explanation for the lower degree in patients with SSD could be that their syntactic nodes are less connected due to less subordination (eg, connected clauses with "because") and clausal embedding (ie, presence of a separate clause within the main sentence),[22,39] which are also operationalizations of syntactic complexity.[40]

We further found that the network size of the patients was smaller as compared to controls' (ie, leaf count). Since network size is highly related to sentence length, small networks may reflect poverty of speech (ie, alogia), a prominent negative symptom of SSD characterized by a decrease in talking with minor elaborations.[41] Compared to the classifier trained solely on word count, the model trained on a combination of network measures reported higher accuracy in distinguishing patients with SSD from controls, suggesting that the information it captures is not redundant with the measurement of syntactic complexity based on mean length of utterance.

Our results further showed that there are fewer syntactic hubs (ie, stress centrality) in the speech of patients with SSD relative to controls. In a network, syntactic hubs are the most important nodes and are involved in most syntactic relations. The more syntactic hubs in a network, the more hierarchical the network becomes. When translated to the syntactic network, a network with many hubs thus



1. a. Anna brought a sandwich.

b. *Brought a sandwich.

c. *Anna brought.

2. a. Anna ate a sandwich.

b. Anna ate a sandwich for lunch.

c. Anna ate a cheese sandwich for lunch.

Fig. 3. Examples of syntactic co-occurrences, including arguments (1) and adjuncts (2). In (1a) noun "Anna" and noun phrase "a sandwich" are necessary arguments. In (2b) and (2c) prepositional phrase "for lunch" and noun phrase modifier "cheese" are optional adjuncts. *The asterisk indicates that the sentence is not grammatical and that, by not following syntactic constraints, its meaning is compromised.

has more hierarchical organization, possibly indicating more subordination and/or embedding. Again, this may reflect less complex syntax in the language of patients.[40]

Association With Sex

Interestingly, women accounted for the most pronounced differences in syntactic network measures between patients and controls, attaining the highest scores in the control group and lowest scores in the patient group for most measures. Healthy women generally achieve higher performance in verbal tasks than men,[42,43] which can explain the higher scores in syntactic network measures reported by healthy female participants. The low scores in the women with SSD might be related to differences in the clinical course between men and women. Women with SSD are often diagnosed at a later age and more often have comorbid affective disorders,[44] factors that may influence the advancement of the disease process compared to men, possibly also influencing syntactic complexity. While a thorough examination of these effects is beyond the scope of this article, we could observe that the presence of a concomitant anxiety disorder did not have an effect on syntactic measures.

It has previously been suggested that women are often overmedicated because antipsychotic drug dosages have been tailored to the male body, and they require lower dosages than men because of differences in pharmacokinetics and pharmacodynamics.[45] A stronger decline in syntactic abilities in women with SSD may therefore be related to relative overdosing of antipsychotic medication, especially since we found a relationship between dose and some network measures. However, further research is required to verify this hypothesis since in this study we had no information regarding drug plasma levels to confirm the suspected overdose in women.

Association With Psychotic Symptoms and Medication

Our study adds to earlier work on speech graph analysis in SSD by showing that syntactic relations, in addition to word co-occurrences, are anti-correlated with negative symptoms.[16,46] This is consistent with research that has indicated that patients with predominantly negative symptoms have a simplified syntactic production with fewer clausal embeddings.[47] Previous research suggested that this was linked to the chronicity of the illness as the ability to produce syntactically complex sentences for patients with SSD worsened as the disease progressed.[12] The present findings do not confirm these results, since age did not have a significant effect on the network measures in our samples. Furthermore, we have shown that antipsychotic drugs have an effect on language production in individuals with SSD, in line with previous research.[23] In particular, we found that higher dose was associated with syntactic networks of smaller size, lower efficiency, and longer diameter.

Association With Cognition

Our findings confirm earlier studies showing a positive relation between syntax and cognitive abilities.[48,49] Specifically, we observed that lower overall cognitive functioning is associated with smaller syntactic networks of reduced compactness (ie, diameter), lower efficiency (ie, efficiency), lower degree (ie, degree), and centrality (ie, stress centrality), in both patients and controls. Closer examination of these findings revealed that syntactic network measures were mostly associated with the domains of executive functioning, attention, and information processing speed. These results corroborate emerging literature showing a relationship between syntax and executive functioning,[50] attention,[51] and information processing speed.[52] In SSD, syntactic anomalies might occur at any phase of the grammatical encoding process due to processing speed impairment,[53] reduced inhibitory attentional control,[54] and/or planning impairments.[55] Additionally, our study supports evidence from previous observations that found no link between syntactic abilities and verbal memory.[56] The fact that most network measures were associated with general cognition and only one with verbal fluency points toward a general higher-order association with cognition, rather than a domain (language) specific association between syntax and cognition.[57]

Final Remarks

Overall, this research supports previous findings that patients with SSD have impaired syntax production[9–11] and extends these findings by showing that these anomalies are more pronounced in women than in men and that they are related to cognitive functioning and negative symptoms. Moreover, it shows the potential of network analysis as an NLP tool in using syntactic measures for differentiating patients with SSD from control participants. Since previous research has shown that a combination of different sets of linguistic features improves the discriminatory power of speech classification algorithms,[24] adding syntactic network measures to other linguistic features (eg, acoustic, semantic) could be a path worth exploring.

This study has some limitations. First, we have examined transcribed speech samples obtained using 1 elicitation method (ie, semi-structured interviews), 1 modality (ie, spoken language), and 1 language (ie, Dutch). However, the validation of syntactic network measures requires generalizability across samples, contexts, and languages.[58] Building multilingual speech banks that enable comparative studies is the first step in this direction.[59] Second, the fact that we did not find an effect of age on syntactic complexity as suggested before[60] may relate to the use of sentences as production units instead of clauses or T units as Silva et al.[60] did, or to the older age of patients in our sample. Third, since sentence length and syntactic complexity are highly interrelated (ie, shorter sentences tend to be less complex), we normalized global network measures (ie, diameter and efficiency) by the number of nodes per sentence. This way of controlling for verbosity implies a linear relation with word count, which is not always the case in language.[61] Future studies should take this into account and attempt to minimize the effects of verbosity on syntactic network measures. Lastly, we did not account for the menstrual cycle or hormone status of the female subjects, which might have impacted their verbal abilities.[62]

In conclusion, our study demonstrates that, by examining syntactic relations with network analysis, significant differences in the language of patients with SSD and controls can be found. Syntactic network measures provide a clinically meaningful way to quantify syntax and are correlated with the severity of negative symptoms and cognitive functioning. Furthermore, analyses of syntactic networks are sensitive to differences in syntax between men and women, confirming the importance of examining language characteristics in relation to sociodemographic aspects. Further research is needed to determine whether syntactic network measures can be combined with other linguistic features and used as a biomarker for SSD.

Supplementary Material

Supplementary material is available at https://academic.oup.com/schizophreniabulletin/.

Acknowledgments

We are grateful to all participants and research interns for their help with data collection and preparation.

Conflict of Interest

The authors have declared that there are no conflicts of interest in relation to the subject of this study.

References

1. Bleuler E. *Dementia praecox oder Gruppe der Schizophrenien*. Leipzig: Franz Deuticke; 1911.
2. Guha M. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. 5th ed. Arlington, VA: American Psychiatric Association; 2013.
3. Corcoran CM, Carrillo F, Fernández-Slezak D, *et al*. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry.* 2018;17(1):67–75.
4. Elvevåg B, Foltz PW, Weinberger DR, Goldberg TE. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophr Res.* 2007;93(1):304–316.
5. Voppel AE, de Boer JN, Brederoo SG, Schnack HG, Sommer IEC. Quantified language connectedness in schizophrenia-spectrum disorders. *Psychiatry Res.* 2021;304:114130–114130.
6. Chomsky N. *Syntactic Structures*. The Hague: Mouton; 1957.
7. Kuperberg GR. Language in schizophrenia part 2: what can psycholinguistics bring to the study of schizophrenia…and vice versa? *Lang Linguist Compass.* 2010;4(8):590–604.
8. Matthews PH. *Syntax*. Cambridge: Cambridge University Press; 1981.
9. Morice RD, Ingram JC. Language analysis in schizophrenia: diagnostic implications. *Aust N Z J Psychiatry.* 1982;16(2):11–21.
10. Ozcan A, Kuruoglu G, Alptekin K, *et al*. The production of simple sentence structures in schizophrenia. *Int J Arts Sci.* 2016;9(4):159–164.
11. Thomas P, King K, Fraser WI. Positive and negative symptoms of schizophrenia and linguistic performance. *Acta Psychiatr Scand.* 1987;76(2):144–151.
12. Thomas P, King K, Fraser WI, Kendell R. Linguistic performance in schizophrenia: a comparison of acute and chronic patients. *Br J Psychiatry.* 1990;156(2):204–210.
13. Hoffman RE, Sledge W. An analysis of grammatical deviance occurring in spontaneous schizophrenic speech. *J Neurolinguistics.* 1988;3(1):89–101.
14. Bedi G, Carrillo F, Cecchi GA, *et al*. Automated analysis of free speech predicts psychosis onset in high-risk youths. *Npj Schizophr.* 2015;1(1):15030.
15. Mota NB, Vasconcelos NAP, Lemoset N, *et al*. Speech graphs provide a quantitative measure of thought disorder in psychosis. *PLoS One.* 2012;7(4):e34928–e34928.
16. Mota NB, Copelli M, Ribeiro S. Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance. *Npj Schizophr.* 2017;3(1):18–10.
17. Sole RV, Corominas-Murtra B, Valverde S, Steels L. Language networks: their structure, function, and evolution. *Complexity.* 2010;15(6):20–26.
18. Van Mieghem P. *Performance Analysis of Complex Networks and Systems*. Cambridge: Cambridge University Press; 2014.

19. Tse SK, Kwong SM, Chan C, Li H. Sex differences in syntactic development: evidence from cantonese-speaking preschoolers in hong kong. *Int J Behav Dev.* 2002;26(6):509–517.
20. Cornett HE. Gender differences in syntactic development among English speaking adolescents. *Inq J.* 2014;6(3):1.
21. Mondorf B. *Gender Differences in English Syntax*. Berlin: De Gruyter; 2011.
22. DeLisi LE. Speech disorder in schizophrenia: review of the literature and exploration of its relation to the uniquely human capacity for language. *Schizophr Bull.* 2001;27(3):481–496.
23. de Boer JN, Voppel AE, Brederoo SG, Wijnen FNK, Sommer IEC. Language disturbances in schizophrenia: the relation with antipsychotic medication. *Npj Schizophr.* 2020;6(1):24–24.
24. Voppel AE, de Boer JN, Brederoo SG, Schnack HG, Sommer IEC. Semantic and acoustic markers in schizophrenia-spectrum disorders; a combinatory machine learning approach [published online ahead of print October 28, 2022]. *Schizophr Bull.* doi:10.1093/schbul/sbac142.
25. Andreasen NC, Flaum M, Arndt S. The Comprehensive Assessment of Symptoms and History (CASH). an instrument for assessing diagnosis and psychopathology. *Arch Gen Psychiatry.* 1992;49(8):615–623.
26. Sheehan DV, Lecrubier Y, Sheehan KH, *et al*. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry.* 1998;59:22–33.
27. Kay SR, Fiszbein A, Opfer LA. The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. *Schizophr Bull.* 1987;13(2):261.
28. Keefe RSE, Goldberg TE, Harvey PD, Gold JM, Poe MP, Coughenour L. The Brief Assessment of Cognition in Schizophrenia: reliability, sensitivity, and comparison with a standard neurocognitive battery. *Schizophr Res.* 2004;68(2–3):283–297.
29. Keefe RSE, Harvey PD, Goldberg TE, *et al*. Norms and standardization of the Brief Assessment of Cognition in Schizophrenia (BACS). *Schizophr Res.* 2008;102(1–3):108–115.
30. de Boer JN, van Hoogdalem M, Mandl RCW, *et al*. Language in schizophrenia: relation with diagnosis, symptomatology and white matter tracts. *Npj Schizophr.* 2020;6(1):10–10.
31. MacWhinney B. *The CHILDES Project: Tools for Analyzing Talk*. 2nd ed. Mahwah, NJ: Lawrence Erlbaum; 1995.
32. van Noord G. At last parsing is now operational. In: Actes de la 13ème conférence sur le Traitement Automatique des Langues Naturelles. Conférences invitees, Leuven, Belgium; 2006:20–42.
33. Van Rossum G, Drake FL. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace; 2009.
34. Shannon P, Markiel A, Ozier O, *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003;13(11):2498–2504.
35. *Analyzer–Cytoscape User Manual 3.9.1 Documentation*. http://manual.cytoscape.org/en/stable/. Accessed April 14, 2022.
36. Breiman L. Random forests. *Mach Learn.* 2001;45:5–32.
37. Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: machine learning in Python. *J Mach Learn Res.* 2011;12:2825–2830.
38. Hawkins JA. *A Performance Theory of Order and Constituency*. Cambridge: Cambridge University Press; 1994.
39. Çokal D, Sevilla G, Jones WS, *et al*. The language profile of formal thought disorder. *Npj Schizophr.* 2018;4(1):18–18.
40. Givón T. Markedness in grammar: distributional, communicative and cognitive correlates of syntactic structure. *Stud Lang.* 1991;15(2):335–370.
41. Fervaha G, Takeuchi H, Foussias G, Agid O, Remington G. Using poverty of speech as a case study to explore the overlap between negative symptoms and cognitive dysfunction. *Schizophr Res.* 2016;176(2–3):411–416.
42. Halpern DF. *Sex Differences in Cognitive Abilities*, 2nd ed. Lawrence Erlbaum Associates, Inc; 1992.
43. Hyde JS, Linn MC. Sex differences in verbal ability: a meta-analysis. *Psychol Bull.* 1988;104(1):53–69.
44. Brand BA, de Boer JN, Dazzan P, Sommer IEC. Towards better care for women with schizophrenia-spectrum disorders. *Lancet Psychiatry.* 2022;9(4):330–336.
45. Brand BA, Haveman Y, de Beer F, de Boer JN, Dazzan P, Sommer IEC. Antipsychotic drug use in schizophrenia spectrum disorders: towards specific treatment for women. *Psychol Med.* 2021;1:15.
46. Mota NB, Furtado R, Maia PPC, Copelli M, Ribeiro S. Graph analysis of dream reports is especially informative about psychosis. *Sci Rep.* 2014;4(1):3691–3691.
47. Thomas P, Kearney G, Napier E, Ellis E, Leudar I, Johnston M. Speech and language in first onset psychosis differences between people with schizophrenia, mania, and controls. *Br J Psychiatry.* 1996;168(3):337–343.
48. Sand Aronsson F, Kuhlmann M, Jelic V, Östberg P. Is cognitive impairment associated with reduced syntactic complexity in writing? Evidence from automated text analysis. *Aphasiology.* 2021;35(7):900–913.
49. Sung JE, Choi S, Eom B, Yoo JK, Jeong JH. Syntactic complexity as a linguistic marker to differentiate mild cognitive impairment from normal aging. *J Speech Lang Hear Res.* 2020;63(5):1416–1429.
50. White LJ, Alexander A, Greenfield DB. The relationship between executive functioning and language: examining vocabulary, syntax, and language learning in preschoolers attending head start. *J Exp Child Psychol.* 2017;164:16–31.
51. Myachykov A, Garrod S, Scheepers C. Attention and syntax in sentence production: a critical review. *Discours.* 2009;4:1–17.
52. Fernald A, Perfors A, Marchman VA. Picking up speed in understanding: speech processing efficiency and vocabulary growth across the 2nd year. *Dev Psychol.* 2006;42(1):98–116.
53. Knowles EEM, David AS, Reichenberg A. Processing speed deficits in schizophrenia: reexamining the evidence. *Am J Psychiatry.* 2010;167(7):828–835.
54. Galaverna F, Morra CA, Bueno AM. Attention in patients with chronic schizophrenia: deficit in inhibitory control and positive symptoms. *Eur J Psychiatry.* 2012;26(3):185–195.
55. Holt DV, Wolf J, Funke J, Weisbrod M, Kaiser S. Planning impairments in schizophrenia: specificity, task independence and functional relevance. *Schizophr Res.* 2013;149(1-3):174–179.
56. Meir N, Novogrodsky R. Syntactic abilities and verbal memory in monolingual and bilingual children with high functioning autism (hfa). *First Lang.* 2020;40(4):341–366.
57. Bates E, Elman J, Johnson MH, Karmiloff-Smith A, Parisi D, Plunkett K. *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press; 1996.
58. Parola A, Lin JM, Simonsen A, *et al*. Speech disturbances in schizophrenia: assessing cross-linguistic generalizability

of NLP automated measures of coherence [published online ahead of print August 1, 2022]. *Schizophr Res.* doi:10.1016/j.schres.2022.07.002.

59. Palaniyappan L, Alonso-Sanchez MF, MacWhinney B. Is collaborative open science possible with speech data in psychiatric disorders? *Schizophr Bull.* 2022;48(5):963–966.
60. Silva AM, Limongi R, MacKinley M, Ford SD, Alonso-Sánchez MF, Palaniyappan L. Syntactic complexity of spoken language in the diagnosis of schizophrenia: a probabilistic bayes network model [published online ahead of print June 22, 2022]. *Schizophr Res.* doi:10.1016/j.schres.2022.06.011.
61. Hunt KW. *Grammatical Structures Written at Three Grade Levels. NCTE Research Report No. 3*. Champaign, IL: National Council of Teachers of English; 1965.
62. Scheuringer A, Pletzer B. Sex differences and menstrual cycle dependent changes in cognitive strategies during spatial navigation and verbal fluency. *Front Psychol.* 2017;8:381–381.