Syntactic complexity of spoken language in the diagnosis of schizophrenia: a probabilistic Bayes network model

Angelica M. Silva[1*], Roberto Limongi[1,6,7*], Michael MacKinley[1,2], Sabrina D. Ford[2], Maria Francisca Alonso-Sánchez[1], Lena Palaniyappan[1,2,3,4,5]

Affiliations:

1. Robarts Research Institute, Western University, London, Ontario, Canada.
2. Lawson Health Research Institute, London, Ontario, Canada.
3. Douglas Mental Health University Institute, Department of Psychiatry, McGill University, Montreal, Quebec, Canada.
4. Department of Psychiatry, Schulich School of Medicine and Dentistry, Western University, London, Canada.
5. Department of Medical Biophysics, Western University, London Canada
6. Department of Psychology, Western University, London Canada
7. Faculty of Human and Social Sciences, Wilfrid Laurier University, Brantford Ontario

Abstract

In the clinical linguistics of schizophrenia, syntactic complexity has received much attention. In this study, we address whether syntactic complexity deteriorates within the six months following the first episode of psychosis conditional upon a consensus diagnosis of schizophrenia. We collected data from a cohort of twenty-six first-episode psychosis and 12 healthy control subjects using the Thought and Language Index interview in response to three photographs from the Thematic Apperception Test at first assessment and after six months, at the time of consensus-diagnosis. An automated labeling POS tagging along with specific syntactic processes calculated large and granular syntactic complexity indices with clause complexity as particular case of spoken language data. Probabilistic reasoning leveraging the conditional independence properties of Bayes networks revealed that consensus diagnosis of schizophrenia predicts a decrease in nominal subject per clause after experiencing a first episode of psychosis. Furthermore, it is 95.4% probable that a 50% decrease in mean nominal subjects per clause after six months is explained by the presence of first episode psychosis. Finally, a 30% decrease in this clause-complexity index after six months of experiencing a first episode of psychosis does predict with 95% probability a consensus diagnosis of schizophrenia, representing a conditional relationship between a longitudinal decrease in syntactic complexity and a diagnosis of schizophrenia. We conclude that an early drift towards linguistic disorganization/impoverishment of clause complexity—at the granular level of nominal subject per clause—is a distinctive feature of schizophrenia that decreases longitudinally identifying schizophrenia from other psychotic illnesses with shared phenomenology.

Keywords: thought disorder, progressive, longitudinal, early intervention, speech
Abstract word Count: 250
Total Word Count (main text): 3999

1. Introduction

Linguistic impoverishment is a well-known feature of the negative syndrome of schizophrenia. While extreme forms of reduced speech output (e.g., mutism) are fortunately rare, much more subtle impoverishment commonly presents as reduced verbal fluency, lack of spontaneity and diminished expression of ideas during interpersonal interactions (Alpert et al., 1997; Kircher et al., 2018; Roche et al., 2015). More focused assessments reveal other features of impoverished language from less categorical linguistic style marked by the use of fewer articles and prepositions (Silva et al., 2021), verb production deficits (Barattieri di San Pietro et al., 2022), an aberrant use of connectives (Mackinley et al., 2020), and a loss of complex morphology (Ziv et al., 2021). These abnormalities are evident very early in the course of illness, often predating overt clinical symptoms (Gooding et al., 2012). Linguistic impoverishment is also a harbinger of imminent onset of psychosis in young persons with non-specific mental health issues (Demjaha et al., 2017; Fusar-Poli et al., 2012). Among those experiencing first psychotic episode, linguistic impoverishment predicts poor functional outcomes both at baseline (Ucok et al., 2021) and over long-term follow-up (Marggraf et al., 2020; Roche et al., 2016; Yalincetin et al., 2017).

With the growth of natural language processing (NLP) (Corcoran and Cecchi, 2020; Corcoran et al., 2020; Hitczenko et al., 2020; Ratana et al., 2019), linguistic impoverishment at syntax level can be automatically quantified through syntactic complexity—the sophistication of syntactic structure seen in writing or speaking that arises from our ability to group words as phrases and embed clauses in a recursive, hierarchical fashion (Friederici et al., 2017). This complexity increases during childhood (Frizelle et al., 2018; Givón, 2009), peaks in early 20s, stays stable for most of the adult life (Nippold et al., 2013), and declines in the 7th decade (Kemper, 1987).

Studying syntactic complexity entails indentifying large and granular indices accounting for different yet interrelated thought dimensions. Large indices—number of words, length and ratio of clauses, and T-units (Beaman, 1984; Bulté and Housen, 2012)—are used as proxy of cognitive parameter distinguishing aspects of complexity that are likely to be more cognitively demanding. For example, T-units indicate the amount of independent clausal coordination in the expressed idea and provide evidence to distinguish the rule-based processes of coordination (e.g., conjoining with ‘and’) and subordination (e.g., qualifying with ‘because’). Considering variations in the use of language, granular or fine-grained indices of syntactic complexity explain both the diversity in clause subtypes (e.g., nominal subjects per clause) (Kyle and Crossley, 2018; Tavano et al., 2008) and the source of this complexity in speech or writing (Larsson and Kaatari, 2020). For instance, nominal subjects per clause or noun phrase (Np) is a granular index of syntactic complexity—a syntactic category dependent on the lexical category of Nouns—used in agentive manner. It represents either 'the doer' in a clause (a clause being a segment relating a subject and a finite verb) or the topic in the pragmatic dimension (Hurford, 2007). See Table 3 in suplementary material for a list of the granular and large indices explored in this work. Unlike written text, spoken language often does not contain defined sentences (Szmrecsányi, 2004). This makes clauses prominent features in the analysis of speech samples with built-in syntactic relationships (Biber, 1988; Biber et al., 2011). In summary, syntactic complexity refers to a construct based on large and granular measures (Norris and Ortega, 2009), with clause complexity being of particular importance in spoken language.

A large body of research underscores syntactic complexity as a key variable in schizophrenia. Morice and Ingram (1982) first demonstrated reduced syntactic complexity in

hospitalized patients with schizophrenia. This reduction in syntactic complexity would be more pronounced in chronic compared to acute stages of schizophrenia (DeLisi, 2001; Thomas et al., 1990), and in those with more pronounced negative symptoms (Barch and Berenbaum, 1997a; Bedi et al., 2015; de Boer et al., 2020b; Stanislawski et al., 2021; Thomas, 1996; Thomas et al., 1987). But some studies where granular level was not examined found no difference in syntactic complexity in schizophrenia (Sanders et al., 1995); (Barch and Berenbaum, 1997b); (Lott et al., 2002); (Perlini et al., 2012). A recent revival using NLP indicates that deficits in syntactic complexity is more pronounced in those with reduced social cognition (Minor et al., 2019) and social functioning (Voleti et al., 2019), supporting the association of reduced syntactic complexity with later stages when a diagnosis of schizophrenia is established. Measures of syntactic complexity used in the referred works are listed in the suplementary material Table 1.

NLP promises predictive utility (Palaniyappan, 2021) (i.e. the identification of illness trajectories), using linguistic markers and cohorts (FEP, mania, and CHR individuals) before the illness becomes established (Bae et al., 2021; Bazziconi et al., 2021; Bedi et al., 2015; Bilgrami et al., 2022; Corcoran et al., 2020; Corcoran et al., 2018; de Boer et al., 2020a; de Boer et al., 2020c; Elvevåg et al., 2010; Mota et al., 2017; Mota et al., 2012; Rosenstein et al., 2015; Tang et al., 2021). NLP-based assessment of reduced syntactic complexity may have clinical utility for early identification of schizophrenia. Realizing this promise requires to demonstrate *diagnostic specificity*—the ability to differentiate schizophrenia from other disorders—that have overlapping symptoms of psychosis, and *incremental utility,* i.e., the information provided by syntactic complexity is over and above what can be gathered from treatment and functioning measures that generally provide diagnostic information. To this end, we combine the theoretical basis of NLP with probabilistic reasoning.

We examine syntactic complexity of spoken language data at large (e.g., number of T-units and sentence and clause length) and granular levels (e.g., number of nominal subjects per clause) and address the relationship between longitudinal changes in syntactic complexity (six months after a first experience of an episode of psychosis) and consensus diagnosis. We studied 38 subjects over a period of 6 months, including 12 healthy control and 26 subjects with first episode psychosis (FEP) untreated at first assessment. Eighteen FEP subjects developed schizophrenia while the remaining FEP subjects developed other psychotic disorders of affective nature. We leveraged the conditional independence properties of Bayes networks (Pearl, 1988) to estimate the conditional relationship between longitudinal change of these indices and consensus diagnosis—considering concurrent effect of antipsychotic dose and changes in functional outcomes over time. Our findings support that an early drift towards disorganization/impoverishment is a distinctive feature of schizophrenia, differentiating it from other psychotic illnesses with shared phenomenology.

2. Methods

2.1 Participants

Twenty-six first-episode psychosis (FEP, 6 females) and 12 healthy control (HC, 4 females) subjects participated in the study (Table 1). This study pursues one of the pre-registered objectives of the observational study TOPSY (NCT02882204). Participants provided written informed consent conforming to the regulations of the Western University Health Sciences Research Ethics Board, London, Ontario, Canada. Patients were in the acute phase of illness (active, untreated psychosis) and recruited upon referral (irrespective of hospitalization status and before antipsychotic treatment was established) from the Prevention and Early Intervention

for Psychosis Program (PEPP) at London Health Sciences Centre, London, Ontario, Canada between April 2017 and July 2019. Based on the best estimate procedure (Leckman et al., 1982) and the Structured Clinical Interview for DSM-5 of the America Psychiatry Association (2013), after six-month, patients received a consensus diagnosis from a minimum of 3 psychiatrists (2 research psychiatrists and the primary treatment provider from the PEPP clinic), while others identified to have bipolar disorder with psychotic features, major depressive disorder with psychotic features or schizophreniform/schizoaffective disorder. HC subjects were recruited from the community through posters. They had neither personal history of mental illness nor family history of psychotic disorders. None of the participants met the criteria for substance-use disorder in the past year according to DSM-5 criteria of the American Psychiatric Association (2013) or had a history of a major head injury. Participants did not report a history of significant medical illness, the presence of intellectual/developmental disorders, or longer than 2 weeks of lifetime antipsychotic exposure.

2.2. Instruments

2.2.1. Psychiatric Symptoms

The Positive and Negative Syndrome Scale-8 Item (PANSS-8) is a condensed version of the full interview-based PANSS for psychosis with acceptable internal consistency and applied by one of the 2 research psychiatrists (Lin et al., 2018). Using the algorithm of the World Health Organization for defined daily doses (DDDs) for antipsychotic medications (World Health Organization, 2013), we derived a common unit of exposure to antipsychotics to quantify the baseline exposure. We also used the modified Social and Occupational Functioning Assessment Scale (SOFAS administered by a single rater) to assess the overall level of functioning at the time of presentation (Morosini et al., 2000). We did not administer detailed cognitive tests given the acute illness phase during which the data were gathered.

2.2.2. The Thought Language Index (TLI)

Data was collected from individuals using the TLI (Liddle et al., 2002), an interview-based instrument to assess FTD. A picture-speech task induced participants to elaborate a 1-min spontaneous speech (oral soliloquies) in response to three photographs from the Thematic Apperception Test (Murray, 1943) after hearing specific instructions: “I am going to show you some pictures, one at a time. When I put each picture in front of you, I want you to describe the picture to me, as fully as you can. Tell me what you see in the picture, describe what you see in this image, and what you think might be happening”. Responses were recorded, transcribed, and scored on eight domains integrated in two merged labels: (1) Impoverishment in Thinking which included poverty of speech, weakening of goal, preservation of ideas and (2) Disorganization in Thinking which comprised looseness, peculiar use of words, peculiar sentences, peculiar logic, and distractibility. Finally, global impoverishment in thinking and global disorganization in thinking are computed. The TLI interview and rating were completed by trained graduate-level research assistants. PANSS-8 assessments were performed in the clinical context, on the same day of the TLI interview, with blinded researcher clinical raters to participant status and linguistic scores.

2.2.3 Tool for the automatic analysis of syntactic complexity and sophistication (TAASSC)

We used TAASSC (Tool for the Automatic Analysis of Syntactic Sophistication and Complexity, https://www.linguisticanalysistools.org/taassc.html). In this tool, the syntactic

complexity analyzer (SCA) package runs three specific syntactic processes: (a) breaking each production units (e.g., sentences, clauses), (b) tokenization—a process that identify each word, and (c) automated labeling POS tagging functionalities wich employs both the Stanford neural-network dependency parser with an accuracy of tagging around 90% (Chen and Manning, 2014; Klein and Manning, 2003) and Tregex (tree query tools visualization) (Levy and Andrew, 2006) for inquiring and manipulating tree data structures. By combining two levels of syntactic complexity (i.e., large and granular), we included three large syntactic indices following the taxonomy in Lu (2010): mean length of clauses (MLC), T-units, and mean length of sentences (MLS) (Table 2, suplementary material). In addition, we used the 29 granular indices to complete clausal complexity analysis (Kyle and Crossley, 2018) (Table 3, suplementary material). To extract the indices, we used the TLI transcripts (averaged across three transcripts), and longitudinal change was computed by subtracting the index at first assessment from its value after 6 months (when consensus diagnosis was reached); negative values would indicate decrease of clause complexity over time.

3. Procedure

3.1 Bayesian Analysis

We performed a three-stage Bayesian analysis. In stage 1 (dimensionality reduction), we conducted a series of independent-samples t tests using JASP (2021) over the 32 syntactic complexity indices at first assessment—narrowing down the dimensionality space to include in stage 2 only the subset of indices that would unveil changes in syntactic complexity use given consensus diagnosis. We selected indices with Bayes Factors (BF_{10}) greater than 20. In stage 2 (independency map, I-map, identification), we identified the longitudinal change of the complexity index showing a dependency with the initial assessment and consensus diagnosis and assessed the influence of whether these dependencies were influenced by either longitudinal changes of clinical scores (SOFAS, PANSS total negative symptoms scores, and DDD) or demographic variables (sex and age). In stage three (inferences), we performed a series of inferences to explain the relationship between syntactic complexity and diagnosis. As detailed below, for the second and third stages we exploited the formal and computational properties of a directed acyclic graph, also referred to as "Bayes or beliefs network".

The graph represented a joint probability distribution over the subset of indices (from stage 1) and clinical/demographic variables represented as "nodes" (Figure 1). The graph encoded dependencies between parent (Pa) and descendant (X_i) nodes via edges indicating the directionality of the dependency. For example, in Figure 1 Pa_{xi-2} is both the parent node of the X_{i-2} and the descendant of the X_{i-1} nodes. Identifying a graph implies identifying all the independencies (an "independency map", I-map, Figure 1) held in the distribution with the form "X_i is independent of all non-descendant nodes given its parent", formally expressed as $P(X_i|Pa_{X_i}^{Network})$. This is referred to as the the formal definition of the semantics of a Bayesian network structure (Koller and Friedman, 2009). We used a prototypical constraint-based algorithm to obtain the I-map network, from which we read off the conditional independencies among nodes.

The joint distribution factorized as $P(X_1\ldots,X_N) = \prod_{i=1}^{N} P(X_i|Pa_{X_i}^{Network})$. Each factor represents a local model. For example, in Figure 1 the node x_{i-2} encodes a local probability of the form $P(x_{i-2}|Pa_{x_{i-2}})$ whereas the node $Pa_{x_{i-2}}$ encodes $P(Pa_{x_{i-2}}|x_{i-1})$. We estimated parameters via maximum likelihood estimation. From these models, we inferred (1) the probability distribution of the consensus diagnoses (Cons) given an observed longitudinal change

(Δ) of the complexity index, P(Cons|Δ complexity index), (2) the distribution of longitudinal change of the complexity index given a specific consensus diagnosis P(Δ complexity index |Cons), and (3) the probability of first assessments (Assess) if we observed a longitudinal change of the complexity index P(Assess| Δ complexity index).

4. Results

4.1. Dimensionality reduction

Five fine-grained and the three large syntactic complexity indices received a $BF_{10} > 20$ indicating a high degree of sensitivity to between-group differences, whereas the remaining indices received $BF_{10} < 4$. (Table 4, suplementary material). Table 2 reveals that whereas mean subordinating conjunctions per clause and mean nominal subjects per clause (i.e., special cases of noun phrases with an agentive role) were larger in the FEP than in the HC group, mean clausal negations per clause, mean prepositional complements per clause, and mean passive nominal subjects per clause were larger in HC than in FEP (Table 3).

4.2. I-map identification and Inferences

The network revealed that only the mean nominal subjects per clause was dependent on consensus diagnosis and independent of first assessment once the diagnosis has been observed (Figure 1). Graph directionalities (Figures 1-2) indicate that consensus diagnosis predicts the change of mean nominal subjects per clause. Reading off the local model (Figure3), consensus diagnosis of schizophrenia predicts a decrease in mean nominal subjects per clause (mean change = -0.05, Sd = 0.103). Conversely, a healthy control status predicts an increase in this index (mean change = 0.66, Sd = 0.11) whereas “other diagnosis” predicts an increase in mean nominal subjects per clause (mean change = 0.06, Sd = 0.11).

Graph directionalities also explain—backward in time—an observed decrease in mean nominal subjects per clause (Figure 1). The observed mean nominal subjects per clause across groups at first assessment was 0.57 (Sd = 0.36). By applying the chain rule of Bayes network over the local probability model, we can infer that it is 95.4% probable that the presence of first episode psychosis explains a 50% decrease in the complexity index observed after six months. Finally, a consensus diagnosis of schizophrenia predicts with 95% probabability a 30% decrease in the index after six months of experiencing a first episode of psychosis (i.e., from 0.8, Table 1, to 0.56). This finding is especially important because it ascribes a potential schizophrenia-specific diagnostic value to the mean nominal subjects per clause index, independent of the dosing of antipsychotics (DDD), clinical, and demographic variables.

5. Discussion

Syntactic complexity represents core of the morpho-syntactic machinery. It is essential for hierarchical word order (basic rules of grammar) through which we derive infinite ways to express the same information. However, during the first episode of psychosis, in some individuals who are more likely to develop the diagnostic pattern of schizophrenia, the nominal subjects per clause reduces in richness over time, despite the stability of global measures of syntactic complexity. Consider the language production from an exemplar HC subject - "*The shirtless man* (Np) who was ploughing the field (embedded relative clause) looks (governor of the dependent) sad" contrast to this with a FEP subject who was later diagnosed with schizophrenia - "*The man* (Np) looks sad he is ploughing the field". This example of how an agentive noun phrase is enriched in normal language shows that hierarchy dependency of syntactic structure is based on word choices (Chomsky, 1965; Friederici et al., 2017). This supports our focus on progressive changes in clause complexity as a marker of the course of SZ.

5.1 Computational linguistic interpretation:

The use of nominal subjects per clause is more likely observed in people experiencing a first episode of psychosis than in healthy subjects. In people with schizophrenia, this use impoverished with the passing of time and consolidation of illness.This decreasing complexity in subjects with a diagnosed of schizophrenia contrasts with an increasing complexity observed in control subjects. In natural language, nominal subjects per clause contributes to syntactic complexity via coordination.While this process is at a lower level of Chomsky's recursive-embedded clauses (Givón, 2009), in schizophrenia, a reduced nominal subject per clause also seems to reflect a loss of coordination in the word choice (e.g., "a black *and* white striped shirt"). When a reduction of information is necessary in the context of informal speech (Hughes and Allen, 2013), it is more common to see a reduction in the subject than in the object position of the clause, as perceived in patients in our sample (e.g., "*the shirtless man* looks sad*"* is more likely to be censored than "there is a man with *no shirt"*). Likewise, from the perspective of pragmatics, patients produced clauses based upon the prototypical two-unit (i.e., topic–comment) message "the sun(topic) *is* shining brightly in this photo(comment)". In this sense, the descriptive content of NPs resulted in the relocation of the descriptive comment from the NP to the verb phrase, the least complex syntactic structure (Hockett, 1963), with fewer clause connectivity-parataxis. See tye same finding in Tovar et al. (2019).

At the beginning of spontaneous speech, a rich and complex nominal subject per clause must be planned ahead because it requires longer time to encode referential information on particular occasions of language use as well as memory resources (Roll et al., 2007). Therefore, NPs are critical for the referential function of language (Halliday, 1973) and interpersonal attaching between speakers and listeners in speech acts. The outcome of referential dysfunction is said to be at the core of of clinical linguistics of schizophrenia, and our findings regarding the particular case of nominal subjects may relate to this general linguistic impairment. For more detailes on this see Fuentes-Claramonte et al. (2022); Hinzen and Rosselló (2015). Given the reports that syntactic comprehension may be unaffected in patients (Covington et al., 2005; Sanders et al., 1995), an impoverishment in generating enriched nominal subjects (*the lexical type* (e.g., "*the sad woman* on the bridge") and *indefinite and coordinated* NPs (e.g., *a man and woman…*) appears to be a specific deficit in schizophrenia (Sevilla et al., 2018).

5.2 The diagnostic specificity of syntactic complexity:

Our model showed a 95.4 % probability of being diagnosed with schizophrenia given a longitudinal decrease of clause-syntactic complexity indexed by nominal subjects per clause. This result positions this granular index as a potential marker for the early identification of schizophrenia from a group of other similarly presenting psychotic disorders and extend prior observations in a key direction. In the 1980s, when established cases of schizophrenia were compared with mania, reduced syntactic complexity separated the two illnesses (Morice and McNicol, 1986; Morice and Igram, 1983; Morice and Ingram, 1982), but this diagnostic separation was absent when patients with acute schizophrenia (Fraser et al., 1986) or those in early phase of illness (<2 years) were studied (Thomas et al., 1990). In contrast, following up a sub-sample of the acutely ill patients studied by Fraser and colleagues (Fraser et al., 1986) over 3 years, King and colleagues (King et al., 1990) reported that syntactic complexity shows a striking progressive reduction that was unique to schizophrenia, but not seen in mania. Our findings of a specific longitudinal reduction in the complexity of syntax being predictive of schizophrenia among various other diagnostic outcomes, is in line with the work of King and colleagues. Nevertheles, given the small numbers with other diagnoses, this result must be considered preliminary in terms of its diagnostic specificity.

The diagnosis-specific reduction in syntactic complexity occurred irrespective of symptomatic deterioration. This finding is reminiscent of "drift towards disorganization" described as a critical feature in the trajectory of schizophrenia (McGlashan and Fenton, 1993). Nonetheless, using antipsychotics can reduce syntactic complexity (de Boer et al., 2020c)—an effect we did not find in our sample. Additionally, diagnostic separation of schizophrenia from other disorders is more challenging during the time of first psychotic episode rather than after 2-3 years of illness (Addington et al., 2006). In fact, recently, Corcoran and colleagues did not find cross-sectional assessment of syntactic complexity to be predictive of emergence of psychosis in clinically high-risk youth followed-up over 24-30 months (Corcoran et al., 2018). Our results using a probabilistic reasoning approach in a unique cohort of first-episode indicate that a pronounced linear change in clause complexity carries sufficient diagnostic specificity even at early stages of psychosis and is unconfounded by antipsychotic dose exposure.

5.3 Limitations and future directions

This work has several strengths. We overcame the difficulty of collecting speech data from the same samples at both an untreated state in psychosis and at the time of post-intervention. We showed that DDD and SOFAS are independent of the relationship between diagnosis and change in syntactic complexity. Moreover, transcribers/raters were blind to diagnosis information. Finally, the interpretation rules of Bayesian networks are consistent with the probabilistic reasoning in clinical practice, as opposed to the traditional point-value (i.e., reject-accept) hypothesis testing with p-values and confidence intervals commonly found in between-grops (baseline-followup) studies.

We acknowledge several limitations. First, we do not know whether the reported findings would vary depending on sex, given the unbalanced distribution of this variable in our sample. Second, our focus was not a between-group comparison (i.e., HC vs. patients) but a within-group longitudinal analysis (N = 38), being group membership at the time of the first assessment a random variable. To this end, we reduced the dimensionality space to just the subset of clause complexity indices that would inform the longitudinal association. This strategy ameliorated the limitation posed by the number of data points (speech samples) otherwise needed given the number of parameters in a joint probability distribution comprising 32 variables. As defined, this longitudinal analysis reveals robustness regarding the sample size, though the number of

syntactic indices resulting from a balanced sample (i.e., with more healthy controls) might differ. Parsing the longitudinal effect of other diagnostic categories constituting other psychotic disorders will require larger samples. Finally, our speech samples were obtained from speakers of one language (English) and from one modality (spoken language), examined in a singular context (i.e, with an open ended but time-limited discursive enquiry). Corcoran and colleagues raised the possibility of complexity measures being influenced by the context of speech elicitation in psychosis (Corcoran et al., 2018), though their observation was restricted to global and not fine-grained measures that we found predictive of schizophrenia.

5.4. Conclusion

Our findings support the notion that an early drift towards linguistic disorganization/impoverishment of syntactic complexity—at the granular level of nominal subject per clause—is a distinctive feature of schizophrenia that decreases longitudinally identifying schizophrenia from other psychotic affective illnesses with shared phenomenology.

References

Addington, J., Chaves, A., Addington, D., 2006. Diagnostic stability over one year in first-episode psychosis. Schizophrenia research 86, 71-75.

Alpert, M., Kotsaftis, A., Pouget, E.R., 1997. At Issue: Speech Fluency and Schizophrenic Negative Signs. Schizophrenia Bulletin 23(2), 171-177.

American Psychiatric Association., 2013. Diagnostic and Statistical Manual of Mental Disorders. American Psychiatric Publishing, Inc, Arlington 991.

Bae, Y.J., Shim, M., Lee, W.H., 2021. Schizophrenia Detection Using Machine Learning Approach from Social Media Content. Sensors (Basel) 21(17), 5924.

Barattieri di San Pietro, C., Barbieri, E., Marelli, M., de Girolamo, G., Luzzatti, C., 2022. Processing Argument Structure and Syntactic Complexity in People with Schizophrenia Spectrum Disorders. Journal of Communication Disorders 96, 106182.

Barch, D.M., Berenbaum, H., 1997a. The effect of language production manipulations on negative thought disorder and discourse coherence disturbances in schizophrenia. Psychiatry Research 71(2), 115-127.

Barch, D.M., Berenbaum, H., 1997b. Language Generation in Schizophrenia and Mania: The Relationships Among Verbosity, Syntactic Complexity, and Pausing. Journal of Psycholinguistic Research 26(4), 401-412.

Bazziconi, P.F., Berrouiguet, S., Kim-Dufor, D.H., Walter, M., Lemey, C., 2021. Linguistic markers in improving the predictive model of the transition to schizophrenia. L'Encephale 47(5), 499-501.

Beaman, K., 1984. Coordination and subordination revisited: Syntactic complexity in spoken and written narrative discourse. Tannen (Ed.) 84, 45-80.

Bedi, G., Carrillo, F., Cecchi, G.A., Slezak, D.F., Sigman, M., Mota, N.B., Ribeiro, S., Javitt, D.C., Copelli, M., Corcoran, C.M., 2015. Automated analysis of free speech predicts psychosis onset in high-risk youths. NPJ schizophrenia 1, 15030-15030.

Biber, D., 1988. Variation across speech and writing. Cambridge University Press, Cambridge, UK.

Biber, D., Gray, B., Poonpon, K., 2011. Should We Use Characteristics of Conversation to Measure Grammatical Complexity in L2 Writing Development? TESOL Quarterly 45(1), 5-35.

Bilgrami, Z.R., Sarac, C., Srivastava, A., Herrera, S.N., Azis, M., Haas, S.S., Shaik, R.B., Parvaz, M.A., Mittal, V.A., Cecchi, G., Corcoran, C.M., 2022. Construct validity for computational linguistic metrics in individuals at clinical risk for psychosis: Associations with clinical ratings. Schizophr Res.

Bulté, B., Housen, A.D., 2012. Defining and operationalizing L2 complexity, in: Housen, A., Kuiken, F., Vedder, I. (Eds.), Dimensions of L2 performance and proficiency—Investigating complexity, accuracy and fluency in SLA. John Benjamins, Amsterdam, pp. 21–46.

Chen, D., Manning, C.D., 2014. A fast and accurate dependency parser using neural networks, in: Marton, Y. (Ed.), The 2014 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Stroudsburg, PA:, pp. 740–750.

Chomsky, N., 1965. Aspects of the theory of syntax. MIT Press, Cambridge, MA.

Corcoran, C., Cecchi, G., 2020. Using Language Processing and Speech Analysis for the Identification of Psychosis and Other Disorders. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging 5(8), 770-779.

Corcoran, C., Mittal, V., Bearden, C., E. Gur, R., Hitczenko, K., Bilgrami, Z., Savic, A., Cecchi, G., Wolff, P., 2020. Language as a biomarker for psychosis: A natural language processing approach. Schizophrenia Research.

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(1), 67-75.

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

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

de Boer, J.N., van Hoogdalem, M., Mandl, R.C.W., Brummelman, J., Voppel, A.E., Begemann, M.J.H., van Dellen, E., Wijnen, F.N.K., Sommer, I.E.C., 2020b. Language in schizophrenia: relation with diagnosis, symptomatology and white matter tracts. npj Schizophrenia 6(1), 10.

de Boer, J.N., Voppel, A.E., Brederoo, S.G., Wijnen, F.N.K., Sommer, I.E.C., 2020c. Language disturbances in schizophrenia: the relation with antipsychotic medication. NPJ Schizophr 6(1), 24.

DeLisi, L.E., 2001. Speech disorder in schizophrenia: Review of the literature and exploration of its relation to the uniquely human capacity for language. Schizophrenia Bulletin 27(3), 481-496.

Demjaha, A., Weinstein, S., Stahl, D., Day, F., Valmaggia, L., Rutigliano, G., De Micheli, A., Fusar-Poli, P., McGuire, P., 2017. Formal thought disorder in people at ultra-high risk of psychosis. BJPsych Open 3(4), 165-170.

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

Fraser, W.I., King, K.M., Thomas, P., Kendell, R.E., 1986. The Diagnosis of Schizophrenia by Language Analysis. British Journal of Psychiatry 148(3), 275-278.

Friederici, A.D., Chomsky, N., Berwick, R.C., Moro, A., Bolhuis, J.J., 2017. Language, mind and brain. Nature Human Behaviour 1(10), 713-722.

Frizelle, P., Thompson, P.A., McDonald, D., Bishop, D.V.M., 2018. Growth in syntactic complexity between four years and adulthood: evidence from a narrative task. Journal of child language 45, 1174-1197.

Fuentes-Claramonte, P., Soler-Vidal, J., Salgado-Pineda, P., Ramiro, N., Garcia-Leon, M.A., Cano, R., Arévalo, A., Munuera, J., Portillo, F., Panicali, F., Sarró, S., Pomarol-Clotet, E., McKenna, P., Hinzen, W., 2022. Processing of linguistic deixis in people with schizophrenia, with and without auditory verbal hallucinations. NeuroImage: Clinical 34, 103007.

Fusar-Poli, P., Deste, G., Smieskova, R., Barlati, S., Yung, A.R., Howes, O., Stieglitz, R.D., Vita, A., McGuire, P., Borgwardt, S., 2012. Cognitive functioning in prodromal psychosis: a meta-analysis. Arch Gen Psychiatry 69(6), 562-571.

Givón, T., 2009. The Genesis of Syntactic Complexity. Diachrony, ontogeny, and neuro-cognition, evolution. John Benjamins Publishing Company, Amsterdam:The Netherlands.

Givón, T., 2009. The genesis of syntactic complexity: diachrony, ontogeny, neuro-cognition, evolution. John Benjamins, Amsterdam.

Gooding, D.C., Ott, S.L., Roberts, S.A., Erlenmeyer-Kimling, L., 2012. Thought disorder in mid-childhood as a predictor of adulthood diagnostic outcome: findings from the New York High-Risk Project. Psychological medicine 43(5), 1003-1012.

Halliday, M.A.K., 1973. Explanation in the Functions of Language. Cambridge University Press, United Kingdom.

Hinzen, W., Rosselló, J., 2015. The linguistics of schizophrenia: thought disturbance as language pathology across positive symptoms. Frontiers in psychology 6, 971.

Hitczenko, K., Mittal, V.A., Goldrick, M., 2020. Understanding Language Abnormalities and Associated Clinical Markers in Psychosis: The Promise of Computational Methods. Schizophrenia Bulletin.

Hockett, C.F., 1963. The problem of universals in language, in: Greenberg, J.E. (Ed.), Universals of Language. MIT Press, Cambridge, MA, pp. 1–29.

Hughes, M.E., Allen, S.E.M., 2013. The effect of individual discourse-pragmatic features on referential choice in child English. Journal of Pragmatics 56, 15-30.

Hurford, J.R., 2007. The origin of noun phrases: Reference, truth and communication. Lingua 117(3), 527-542.

Kemper, S., 1987. Life-span changes in syntactic complexity. Journal of gerontology 42(3), 323-328.

King, K., Fraser, W.I., Thomas, P., Kendell, R.E., 1990. Re-examination of the language of psychotic subjects. The British journal of psychiatry : the journal of mental science 156, 211-215.

Kircher, T., Bröhl, H., Meier, F., Engelen, J., 2018. Formal thought disorders: from phenomenology to neurobiology. Lancet Psychiatry 5(6), 515-526.

Klein, D., Manning, C.D., 2003. Accurate Unlexicalized Parsing, 41st Meeting of the Association for Computational Linguistics, pp. 423-430.

Koller, D., Friedman, N., 2009. Probabilistic Graphical Models: Principles and Techniques.

Kyle, K., Crossley, S.A., 2018. Measuring Syntactic Complexity in L2 Writing Using Fine-Grained Clausal and Phrasal Indices. The Modern Language Journal 102(2), 333-349.

Larsson, T., Kaatari, H., 2020. Syntactic complexity across registers: Investigating (in)formality in second-language writing. Journal of English for Academic Purposes 45, 100850.

Leckman, J.F., Sholomskas, D., Thompson, D., Belanger, A., Weissman, M.M., 1982. Best Estimate of Lifetime Psychiatric Diagnosis: A Methodological Study. Arch Gen Psychiatry 39(8), 879-883.

Levy, R., Andrew, G., 2006. Tregex and Tsurgeon: tools for querying and manipulating tree data structures, the 5th International Conference on Language Resources and Evaluation, San Diego, Ca., pp. 2231-2234.

Liddle, P.F., Ngan, E.T.C., Caissie, S.L., Anderson, C.M., Bates, A.T., Quested, D.J., White, R., Weg, R., 2002. Thought and Language Index: An instrument for assessing thought and language in schizophrenia. The British Journal of Psychiatry 181(4), 326-330.

Lin, C.H., Lin, H.-S., Lin, S.-C., Kuo, C.C., Wang, F.-C., Huang, Y.H., 2018. Early improvement in PANSS-30, PANSS-8, and PANSS-6 scores predicts ultimate response and remission during acute treatment of schizophrenia. Acta psychiatrica Scandinavica 137.

Lott, P.R., Guggenbühl, S., Schneeberger, A., Pulver, A.E., Stassen, H.H., 2002. Linguistic analysis of the speech output of schizophrenic, bipolar, and depressive patients. Psychopathology 35(4), 220-227.

Lu, X., 2010. Automatic analysis of syntactic complexity in second language writing. International Journal of Corpus Linguistics 15(4), 474-496.

Mackinley, M., Chan, J., Ke, H., Dempster, K., Palaniyappan, L., 2020. Linguistic determinants of formal thought disorder in first episode psychosis. Early Intervention in Psychiatry n/a(n/a).

Marggraf, M.P., Lysaker, P.H., Salyers, M.P., Minor, K.S., 2020. The link between formal thought disorder and social functioning in schizophrenia: A meta-analysis. European Psychiatry 63.

McGlashan, T., Fenton, W., 1993. McGlashan TH, Fenton WS. Subtype progression and pathophysiologic deterioration in early schizophrenia. Schizophr Bull 19: 71-84. Schizophrenia bulletin 19, 71-84.

Minor, K.S., Willits, J.A., Marggraf, M.P., Jones, M.N., Lysaker, P.H., 2019. Measuring disorganized speech in schizophrenia: automated analysis explains variance in cognitive deficits beyond clinician-rated scales. Psychological Medicine 49(3), 440-448.

Morice, R., McNicol, D., 1986. Language Changes in Schizophrenia: A Limited Replication. Schizophrenia Bulletin 12(2), 239-251.

Morice, R.D., Igram, J.C., 1983. Language complexity and age of onset of schizophrenia. Psychiatry research 9(3), 233-242.

Morice, R.D., Ingram, J.C.L., 1982. Language Analysis in Schizophrenia: Diagnostic Implications. Australian & New Zealand Journal of Psychiatry 16(2), 11-21.

Morosini, P.L., Magliano, L., Brambilla, L., Ugolini, S., Pioli, R., 2000. Development, reliability and acceptability of a new version of the DSM-IV Social and Occupational Functioning Assessment Scale (SOFAS) to assess routine social funtioning. Acta Psychiatrica Scandinavica 101(4), 323-329.

Mota, N.B., Copelli, M., Ribeiro, S., 2017. Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance. NPJ schizophrenia 3, 18-18.

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(4), e34928.

Murray, H.A., 1943. Thematic Apperception Test manual. Harvard University Press, Cambridge, MA.

Nippold, M., Cramond, P., Hayward-Mayhew, C., 2013. Spoken language production in adults: Examining age-related differences in syntactic complexity. Clinical linguistics & phonetics 28.

Norris, J.M., Ortega, L., 2009. Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity. Applied Linguistics 30(4), 555-578.

Palaniyappan, L., 2021. More than a biomarker: could language be a biosocial marker of psychosis? npj Schizophrenia 7(1), 42.

Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems: Netwokrs of Plausible Inference. Morgan Kaufmann Publishers, INC, San Francisco, Ca.

Perlini, C., Marini, A., Garzitto, M., Isola, M., Cerruti, S., Marinelli, V., Rambaldelli, G., Ferro, A., Tomelleri, L., Dusi, N., Bellani, M., Tansella, M., Fabbro, F., Brambilla, P., 2012. Linguistic production and syntactic comprehension in schizophrenia and bipolar disorder. Acta Psychiatr. Scand. 1 126(363–376).

Ratana, R., Sharifzadeh, H., Krishnan, J., Pang, S., 2019. A Comprehensive Review of Computational Methods for Automatic Prediction of Schizophrenia With Insight Into Indigenous Populations. Frontiers in psychiatry 10(659).

Roche, E., Creed, L., MacMahon, D., Brennan, D., Clarke, M., 2015. The Epidemiology and Associated Phenomenology of Formal Thought Disorder: A Systematic Review. Schizophr Bull 41(4), 951-962.

Roche, E., Lyne, J., O'Donoghue, B., Segurado, R., Behan, C., Renwick, L., Fanning, F., Madigan, K., Clarke, M., 2016. The prognostic value of formal thought disorder following first episode psychosis. Schizophr Res 178(1-3), 29-34.

Roll, M., Frid, J., Horne, M., 2007. Measuring syntactic complexity in spontaneous spoken Swedish. Language and speech 50(Pt 2), 227-245.

Rosenstein, M., Foltz, P.W., DeLisi, L.E., Elvevåg, B., 2015. Language as a biomarker in those at high-risk for psychosis. Schizophrenia Research 165(2), 249-250.

Sanders, L.M., Adams, J., Tager-Flusberg, H., Shenton, M.E., Coleman, M., 1995. A comparison of clinical and linguistic indices of deviance in the verbal discourse of schizophrenics. Applied Psycholinguistics 16(3), 325-338.

Sevilla, G., Rosselló, J., Salvador, R., Sarró, S., López-Araquistain, L., Pomarol-Clotet, E., Hinzen, W., 2018. Deficits in nominal reference identify thought disordered speech in a narrative production task. PLoS One 13(8), e0201545.

Silva, A., Limongi, R., MacKinley, M., Palaniyappan, L., 2021. Small Words That Matter: Linguistic Style and Conceptual Disorganization in Untreated First-Episode Schizophrenia. Schizophrenia bulletin open 2(1), sgab010.

Stanislawski, E.R., Bilgrami, Z.R., Sarac, C., Garg, S., Heisig, S., Cecchi, G.A., Agurto, C., Corcoran, C.M., 2021. Negative symptoms and speech pauses in youths at clinical high risk for psychosis. NPJ Schizophrenia 7.

Szmrecsányi, B.M., 2004. On Operationalizing Syntactic Complexity, in: Louvain, P.U.d. (Ed.), Proceedings of the 7th International Conference on Textual Data Statistical Analysis, Louvain-la-Neuve, pp. 1031-1038.

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), 25.

Tavano, A., Sponda, S., Fabbro, F., Perlini, C., Rambaldelli, G., Ferro, A., Cerruti, S., Tansella, M., Brambilla, P., 2008. Specific linguistic and pragmatic deficits in Italian patients with schizophrenia. Schizophr. Res 102, 53–62.

Team, JASP., 2021. JASP (Version 0.16)[Computer software].

Thomas, P., 1996. Syntactic Complexity and Negative Symptoms in First Onset Schizophrenia. Cognitive Neuropsychiatry 1(3), 191-200.

Thomas, P., King, K., Fraser, W.I., 1987. Positive and negative symptoms of schizophrenia and linguistic performance. Acta psychiatrica Scandinavica 76(2), 144-151.

Thomas, P., King, K., Fraser, W.I., Kendell, R.E., 1990. Linguistic Performance in Schizophrenia: a Comparison of Acute and Chronic Patients. The British Journal of Psychiatry 156(2), 204-210.

Tovar, A., Fuentes-Claramonte, P., Soler-Vidal, J., Ramiro-Sousa, N., Rodriguez-Martinez, A., Sarri-Closa, C., Sarro, S., Larrubia, J., Andres-Bergareche, H., Miguel-Cesma, M.C., Padilla, P.P., Salvador, R., Pomarol-Clotet, E., Hinzen, W., 2019. The linguistic signature of hallucinated voice talk in schizophrenia. Schizophrenia Research 206, 111-117.

Ucok, A., Karakaş, B., Şahin, O.Ş., 2021. Formal thought disorder in patients with first-episode schizophrenia: Results of a one-year follow-up study. Psychiatry Research 301, 113972.

Voleti, R., Woolridge, S., Liss, J.M., Milanovic, M., Bowie, C.R., Berisha, V., 2019. Objective assessment of social skills using automated language analysis for identification of schizophrenia and bipolar disorder, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 1433-1437.

World Health Organization, 2013. Guidelines for ATC Classification and DDD Assignment.

Yalincetin, B., Bora, E., Binbay, T., Ulas, H., Akdede, B.B., Alptekin, K., 2017. Formal thought disorder in schizophrenia and bipolar disorder: A systematic review and meta-analysis. Schizophrenia Research 185, 2-8.

Ziv, I., Baram, H., Bar, K., Zilberstein, V., Itzikowitz, S., Harel, E.V., Dershowitz, N., 2021. Morphological characteristics of spoken language in schizophrenia patients - an exploratory study. Scandinavian journal of psychology.

Table 1. Demographic and Clinical Information

	Group at First Assessment						
	FEP						HC
SES	3.85 (1.12)						3.17 (1.70)
Assess GDIT	0.34 (0.38)						0.04 (0.07)
ΔGDIT	-0.22 (0.33)						-0.05 (0.03)
Assess GIOT	0.15 (0.19)						0.03 (0.08)
ΔGIOT	-0.05 (0.19)						-0.03 (0.08)
Assess P2	3.00 (1.55)						
ΔP2	-1.81 (1.5)						
DDD	1.36 (3.04)						
DUP	9.48 (13.75)						
Assess SOFAS	42 (13.22)						82.27 (5.69)
ΔSOFAS	19 (14.46)						-0.8 (10.26)
	Group at Consensus Diagnosis						
	SZ	Bipolar	Clinical high risk	NOS	Major depressive disorder	Schizoaffective	HC
Age	22.8 (5.5)	22 (0)	21 (0)	21 (0)	17 (0)	22 (3.46)	21.33 (3.23)
N	18	1	1	1	2	3	12
Sex (females)	3	0	0	0	0	1	6

Note: Mean (standard deviation); SES, socioeconomic status; GDIT, (TLI) global disorganization of thinking; GIOT, (TLI) global impoverishment of thinking; P2, conceptual disorganization; DUP, duration of untreated psychosis (months); DDD, defined daily dose equivalents of antipsychotic medication (total dose at the time of the first assessment); SOFAS, Social and Occupational Functioning Assessment Score; Δ, longitudinal change (6 months – first assessment) where relevant, NOS (psychosis not otherwise specified), SZ (schizophrenia), SZA (schizoaffective), Bip (bipolar), MDD (major depressive disorder), Assess (first assessment), Assess (measurement at the time of the first assessment).

Table 2. Descriptive statistics and Bayes factors against the null hypothesis of independent samples t tests relevant to the sensitive clause complexity indices.

Clause Complexity Index	Group	Mean	SD	SE	95% Credible Interval		BF_{10}
					Lower	Upper	
Mean subordinating conjunctions per clause	FEP	0.122	0.08	0.015	0.092	0.153	> 1000
	HC	0.001	0	0.001	-0.001	0.004	
Mean clausal negations per clause	FEP	0.047	0.04	0.007	0.032	0.062	363
	HC	0.129	0.08	0.022	0.082	0.177	
Mean nominal subjects per clause	FEP	0.803	0.08	0.015	0.771	0.835	> 1000
	HC	0.053	0.04	0.012	0.027	0.079	
Mean clausal prepositional complements per clause	FEP	0.001	0	<0.01	<0.01	0.003	> 1000
	HC	0.023	0.02	0.006	0.01	0.037	
Mean passive nominal subjects per clause	FEP	0.024	0.03	0.007	0.01	0.037	> 1000
	HC	0.744	0.08	0.023	0.692	0.795	
MLC	FEP	7.722	0.858	0.168	7.376	8.069	> 1000
	HC	12.98	1.955	0.564	11.738	14.221	
T-units	FEP	11.611	2.304	0.452	10.681	12.542	139
	HC	15.074	2.468	0.712	13.506	16.642	
MLS	FEP	13.496	3.296	0.646	12.165	14.828	> 1000
	HC	150.139	32.642	9.423	129.399	170.879	

Note. SD (standard deviation), SE (standard error), BF_{10} (Bayes Factor against the null hypothesis), MLS (mean length of sentences), MLC (mean length of clauses).

Table 3. Five of the fine-grained syntactic complexity indices that received a $BF_{10} > 20$

Index Name	Linguistic annotation	Examples of clausal dependent types analyzed by TAASSC
mark_per_cl	Subordinating conjunctions per clause **mark([after] all [said]**)	A subordinating conjunction that marks a subordinate clause "*So, I think we have your typical, you know, 1910 family here. Maybe**,** it's not 1910,* ***after all said".***
neg_per_cl	clausal negations per clause **neg_per_cl(do[not]neg know)**	A verb phrase that is negated *"Um, so* ***I don't*** *really know what's going on there"*
nsubj_per_cl	Nominal subjects per clause **nsubj(looks, shirtless-man)**	It can be understood as a noun phrase (NP) acting like a noun in a sentence. The argument in this NP has an agentive role, the do-er of the clause. ***"The shirtless man*** *looks sad"*
	Variant of nsubj(shining-brightly, sun)	When the verb is a copular verb (to be), the root of the NP is the complement of copular verb. It can be an adjective or noun "*Sun is* ***shining brightly*** *in this photo"*
nsubjpass_per_cl	Passive nominal subjects per clause nsubjpass**(seemed-disturbed, water)**	It refers to a noun phrase which is the syntactic subject of a passive clause. In the follow example, look at a simple present tense-passive voice clause. *"The water is not* **seemed disturbed"**
pcomp_per_cl	Clausal prepositional complements per clause **pcomp_per_cl[about she jumping]pcomp**	Clausal complement that consists of a prepositional phrase that includes a clausal prepositional object *"It looks like she is all by herself and she's looking down so she might be thinking* ***about she jumping****"*

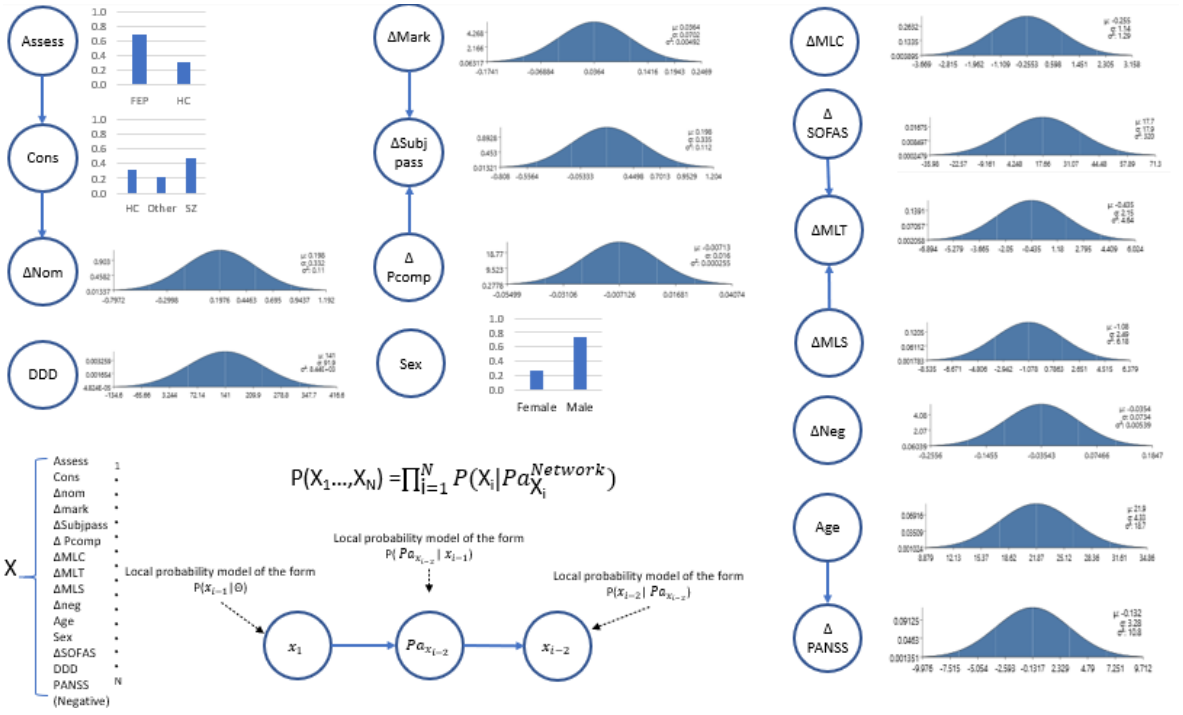

Figure 1. Bayes network with prior distributions. *Bottom left:* The calculation of conditional probability distribution using the chain rule of Bayesian networks is displayed in the bottom panel. From these models, we inferred the (posterior) probability distribution of the consensus diagnoses (Cons: $Pa_{x_{i-2}}$) given an observed longitudinal change (Δ, x_{i-2}) of various indices of interest (linguistic, clinical and demographic indices, listed as X). For example, P(Cons|Δ complexity index). The probability distributions of longitudinal change of an index of interest given a specific consensus diagnosis, and that of the first assessment grouping (Assess) if we only observed a longitudinal change of the index of interest [e.g., P(Δ complexity index |Cons) and P(Assess| Δ complexity index)] were also retrieved from this approach. *Top left*: A three-node Bayes network representing the conditional dependencies between first assessment (Assess), consensus diagnosis (Cons; i.e., SZ =schizophrenia, HC = healthy control status, or Other = bipolar disorder with psychotic features, major depressive disorder with psychotic features or schizophreniform/schizoaffective disorder), and mean number of nominal subjects per clause (ΔNom). For the sample studied here, the directionality comprised in this directed acyclic graph indicates that at the first assessment stage), the probability of FEP was 67.95%, as shown in the insert with FEP/HC bars. The consensus diagnosis was conditional upon the first assessment and explains the longitudinal change (Δ, where relevant) in the mean number of nominal subjects per clause. By the time of consensus diagnosis, the probability of schizophrenia (SZ) was 47.01%, as shown in the insert with HC/Other/Schizophrenia bars. Only ΔNom had a dependency on the consensus diagnosis (the first Directed Acyclic Graph). Change in the other four fine-grained (mean subordinating conjunctions per clause, ΔMark; passive nominal subjects per clause, ΔSubjpass; mean clausal prepositional complements per clause, ΔPcomp, and mean clausal negations per clause, ΔNeg) as well as the three large-grained (MLC, mean length of clause; MLS, mean length of sentence, and T-units) clause complexity indices were independent of the consensus diagnostic probability at follow-up (2nd and 3rd Directed Acyclic Graphs) but related to changes in social and occupational functioning (ΔSOFAS). *Bottom right:* Diagnostic status showed no conditional dependencies with changes in clinical measures of PANSS Negative syndrome though age had a relationship with the latter(4th Directed Acyclic Graph). Prior probability distributions for other variables of interest (DDD: daily dose equivalents of antipsychotics, sex) are shown beside the respective nodes.

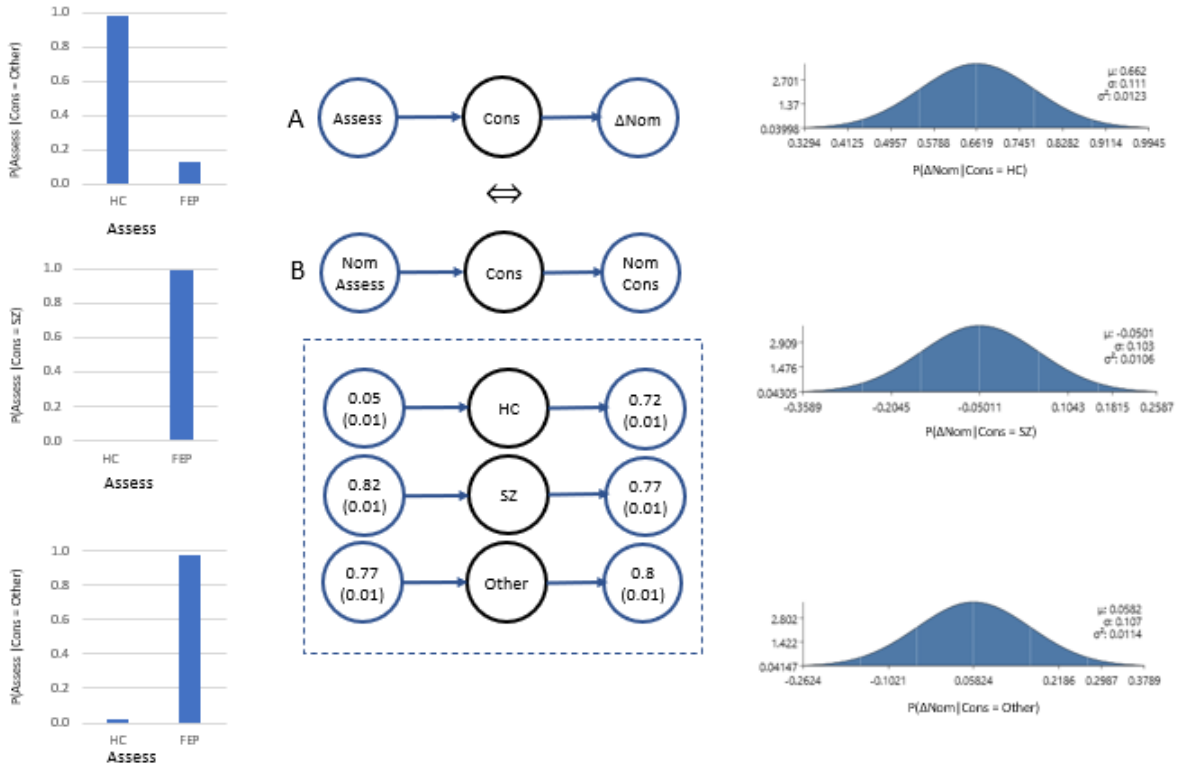

Figure 2. Inferences based on the independencies between syntactic complexity measurements given a consensus diagnosis. Two equivalent Bayes networks showing the dependencies between consensus diagnosis (Cons; SZ = schizophrenia; HC = healthy control; Other = bipolar disorder with psychotic features, major depressive disorder with psychotic features or schizophreniform/schizoaffective disorder) and longitudinal change in mean nominal subjects per clause (ΔNom). This equivalence expresses two different ways to represent that the longitudinal change of mean nominal subjects per clause does not depend on the measurement performed at first assessment, and is formalized as (ΔNom ⫫ Assess | Cons) ⇔ (Nom_Assess ⫫Nom_Cons | Cons). The "Cons" node(the circle with the black line -i.e., the middle) is the parent of both ΔNom (A network) and the mean nominal subjects per clasuse measured at the time of consensus diagnosis (Nom_Cons, B network). Furthermore the Cons node is the descendant of both the group membership at first assessment (A network) and the mean nominal subjects per clasuse measured at first assessment (Nom_Assess, B network). After knowing a consensus diagnosis, there is no need to know the measurement at first assessment to know the change at the time of consensus diagnosis (A network). This is equivalent to saying that there is no need to know the measurement at the time of first assessment to know the measurement at the time of consensus diagnosis (B network). Though equivalent, each of the networks serves specific goals. Network A allows us to estimate a smaller number of parameters whereas network B allows us to visually represent the marginal probabilities of the complexity index both at the time of first assessment and at the time of consensus diagnosis. The posterior probability distributions show that the observation at the time of consensus diagnosis, retaining a healthy control status (distribution curve at the top) predicts a notable increase in mean nominals, whereas a consensus diagnosis of schizophrenia after six months of experiencing a first episode of psychosis predicts a decrease in mean nominals (second distribution curve from top). "Other"

diagnosis predicts an increase in mean nominals of comparable magnitude to SZ (distribution curve at the bottom).

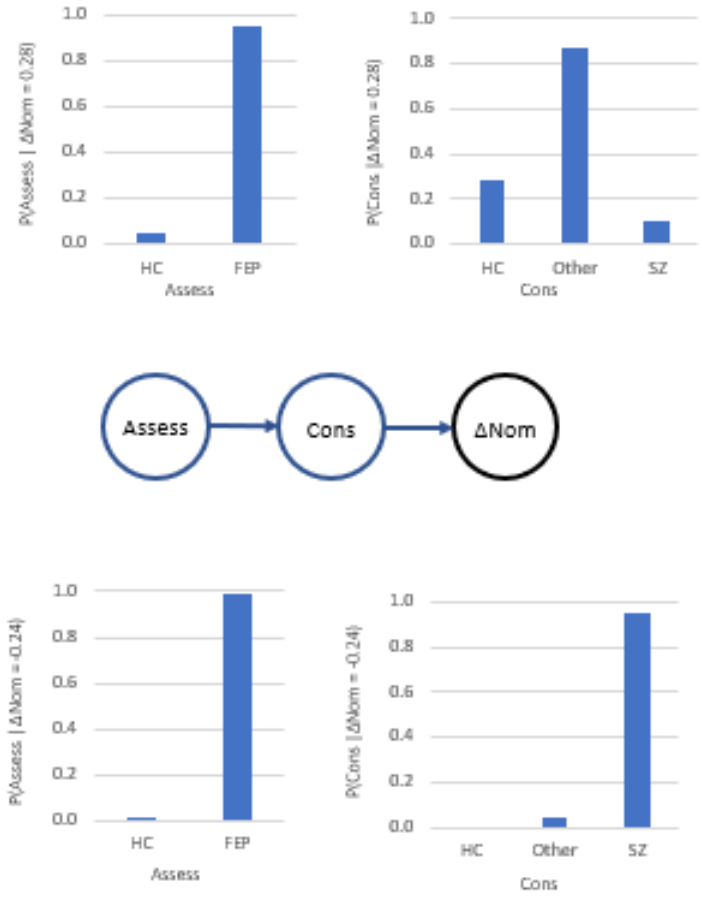

Figure 3. Inferences on first assessment (Assess) and consensus diagnosis (Cons) given a change of mean nominal subjects per clause (ΔNom). Top panel: without any prior information on clinical status, a 50% decrease in mean nominal subjects indicates with 95 % probability that the (unassessed subject) has experienced a first episode of psychosis (FEP). But here, without selecting subjects with FEP, a diagnosis of schizophrenia is less probable (10%). Bottom panel: given a subject that has experienced a first episode of psychosis, a decrease of 30% in their mean nominal subjects after six months of assessment predicts with 95 % probability a consensus diagnosis of schizophrenia. Cons; SZ, schizophrenia; HC, healthy control; Other, bipolar disorder with psychotic features, major depressive disorder with psychotic features or schizophreniform/schizoaffective disorder.