ARTICLE IN PRESS
SCHRES-08799; No of Pages 9
Schizophrenia Research xxx (xxxx) xxx
Contents lists available at ScienceDirect
Schizophrenia Research
journal homepage: www.elsevier.com/locate/schres

Language as a biomarker for psychosis: A natural language processing approach

Cheryl M. Corcoran [a], Vijay A. Mittal [b], Carrie E. Bearden [c,i,j], Raquel E. Gur [d], Kasia Hitczenko [h], Zarina Bilgrami [a], Aleksandar Savic [e], Guillermo A. Cecchi [f], Phillip Wolff [g,*]

[a] Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[b] Department of Psychology, Northwestern University, Evanston, IL, USA
[c] Department of Psychiatry and Biobehavioral Sciences, University of California Los Angeles, CA, USA
[d] Brain Behavior Laboratory, Neuropsychiatry Division, Department of Psychiatry, Philadelphia, PA 19104, USA
[e] Department of Diagnostics and Intensive Care, University Psychiatric Hospital Vrapce, Zagreb, Croatia
[f] Computational Biology Center—Neuroscience, IBM T.J. Watson Research Center, Yorktown Heights, NY, USA
[g] Department of Psychology, Emory University, Atlanta, GA, USA
[h] Department of Linguistics, Northwestern University, Evanston, IL, USA
[i] Department of Psychology, Semel Institute for Neuroscience and Human Behavior, Brain Research Institute, University of California Los Angeles, CA, USA
[j] Department of Psychology, University of California Los Angeles CA USA

ARTICLE INFO

Article history:
Received 28 February 2020
Received in revised form 22 April 2020
Accepted 24 April 2020
Available online xxxx

Keywords:
Psychosis
Automated language analysis
Natural language processing
Machine learning
Semantic coherence
Discourse coherence
Referential coherence
Semantic density
Latent semantic analysis
Digital phenotyping
Psychosis risk
Clinical high risk
Ultra high risk
Schizophrenia

ABSTRACT

Human ratings of conceptual disorganization, poverty of content, referential cohesion and illogical thinking have been shown to predict psychosis onset in prospective clinical high risk (CHR) cohort studies. The potential value of linguistic biomarkers has been significantly magnified, however, by recent advances in *natural language processing* (NLP) and *machine learning* (ML). Such methodologies allow for the rapid and objective measurement of language features, many of which are not easily recognized by human raters. Here we review the key findings on language production disturbance in psychosis. We also describe recent advances in the computational methods used to analyze language data, including methods for the automatic measurement of discourse coherence, syntactic complexity, poverty of content, referential coherence, and metaphorical language. Linguistic biomarkers of psychosis risk are now undergoing cross-validation, with attention to harmonization of methods. Future directions in extended CHR networks include studies of sources of variance, and combination with other promising biomarkers of psychosis risk, such as cognitive and sensory processing impairments likely to be related to language. Implications for the broader study of social communication, including reciprocal prosody, face expression and gesture, are discussed.

1. Introduction

Language and speech are the primary sources of data for clinicians to diagnose and treat mental disorders. They provide a rich source of information about the organization and content of thought, and they are easy and inexpensive to collect. Traditionally, language and speech have been analyzed through expert opinion, clinical ratings and manual linguistic analyses. While informative, these approaches have limitations. Expert opinion can be influenced by subjective appraisal. Clinical ratings can be restricted by incomplete response sets. Clinical judgments often lack precision because they are based on ordinal scales. Manual linguistic analyses can yield finer-grain distinctions than those afforded by clinical observations, but the effort required to conduct such studies is usually so high that they cannot be practically applied in large-scale studies, much less clinical settings. The close connection between language and higher-order thought processes entails that language and speech may offer one of the most informative collections of

* Corresponding author at: Emory University, 36 Eagle Row, Atlanta, GA 30322, USA.
E-mail address: pwolff@emory.edu (P. Wolff).

https://doi.org/10.1016/j.schres.2020.04.032
0920-9964/© 2020 Elsevier B.V. All rights reserved.

Please cite this article as: C.M. Corcoran, V.A. Mittal, C.E. Bearden, et al., Language as a biomarker for psychosis: A natural language processing approach, Schizophrenia Research, https://doi.org/10.1016/j.schres.2020.04.032

features for predicting mental illness (Elvevåg et al., 2016), but unless these features can be extracted quickly and reliably, the promise of this approach cannot be practically realized.

Language features are becoming more trackable. Computational methods from artificial intelligence and natural language processing (NLP) currently allow for the immediate and accurate extraction of linguistic features. Recent studies show how these features can be used to predict mental illness, even in the nascent stages of a disease (Foltz et al., 2016; Bedi et al., 2015; Corcoran et al., 2018; Mota et al., 2017; Rezaii et al., 2019). Automated analyses of language may facilitate the transition from clinical practice based on clinical judgment alone to “measurement-based care” (Insel, 2017), opening up new ways of classifying psychopathology based on objective features. Such an approach is fully compatible with the goals of The National Institute of Mental Health's Research Domain Criteria (RDoC). Language is emerging as a source of predictive features not only because the computational methods are making extraction relatively easy, but also because these methods are beginning to mine the kinds of features that are likely to be especially predictive of mental illness, including features relevant to the prediction of transition to psychosis in “clinical high risk” (CHR) individuals.

In applying NLP analytics to language and speech, replicated patterns may emerge that are characteristic for specific diagnoses or symptoms, prognostic for later outcomes, and/or markers of illness progression or treatment response, especially within psychiatric disorders. Therefore, these linguistic features or patterns may be treated as putative biomarkers that can be developed and validated, with standards of evidence established for their context of use in clinical trials (diagnostic, enrichment, stratification), in accordance with concept clearance by the National Institute of Mental Health (https://www.nimh.nih.gov/funding/grant-writing-and-application-process/concept-clearances/2014/biomarker-development-and-validation-establishing-standards-of-evidence-for-their-context-of-use-in-clinical-trials.shtml). In this concept clearance, the psychosis prodrome is considered a priority area in respect to “unmet medical need, lack of objective endpoints, reasonable development path, and traction/feasibility.”

Herein, we will 1) review key findings on language production disturbance in psychosis and schizophrenia; 2) outline procedures for collecting and analyzing language in psychosis risk, including clinical ratings, manual analyses and automated methods, with attention to harmonization and risks; 3) describe the reasonable development path for linguistic biomarker development in schizophrenia and psychosis risk and consider combinations of linguistic biomarkers with other psychosis risk biomarkers across levels of analysis (genes, molecules, circuits, physiology, cognition/behavior); and 4) describe future plans to conduct analyses at the level of the dyad, and broaden data to include prosody and face expression.

2. Language production disturbance in psychosis/schizophrenia

Disorder in thought is evident as disorganization in communication. Thought disorder has long been recognized as characteristic of psychotic disorders such as schizophrenia (Roche et al., 2015). Kraepelin described “dream speech” (Kraepelin, 2010) and Bleuler described “loosening of associations” (Bleuler, 1950) as characteristic of schizophrenia specifically. Later investigators such as Harrow (Harrow and Quinlan, 1977) and Andreasen (Andreasen and Grove, 1986) found thought disorder existed in other psychotic disorders as well. Harrow solicited speech using the Rorschach, and applied the Thought Disorder Index (TDI), which comprises clinical ratings of observed language disturbances, rated by tiers of severity and frequency of occurrence. The TDI includes 1) “minor idiosyncrasies” such as flippant responses, vagueness, peculiar verbalizations, word-finding, clangs, perseveration and incongruous combinations; 2) “distinct oddness” items such as idiosyncratic symbolism, confusion, looseness, playful confabulation and fragmentation; 3) “psychotic disruption” items such as absurd responses, confabulations and autistic logic, and 4) complete break of “reality contact”, including contamination, incoherence and neologisms (Solovay et al., 1986). The application of the TDI to responses to the Rorschach has also been used to assess thought disorder in unaffected relatives of schizophrenia patients (Levy et al., 2010) and familial risk (Metsänen et al., 2006; Gooding et al., 2012), finding disparate results.

Andreasen, on the other hand, used natural language as the basis for study. She invited patients to talk without interruption for 10 min, and they were then asked about the personal and the abstract, for 30 to 45 min. Andreasen held that thought disorder could be assessed simply from observing a “patient's speech and language behavior”, “without complicated experimental procedures” and “without any attempt to characterize the underlying cognitive processes”. Andreasen argued that in thought disorder, the speaker “violates the syntactical and semantic conventions which govern language usage” (Andreasen and Grove, 1986). Andreasen developed and validated the Scale for the Assessment of Thought, Language and Communication (TLC) (Andreasen, 1979a), which included 18 items (poverty of speech, illogicality, incoherence, clanging, neologisms, word approximations, poverty of content of speech, pressure of speech, distractible speech, tangentiality, derailment, stilted speech, echolalia, self-reference, circumstantiality, loss of goal, perseveration and blocking). The overall TLC scores discriminated diagnoses of depression from those of mania and schizophrenia, but not mania from schizophrenia. However, when “positive” and “negative” thought disorder were differentiated, TLC scores were effective in differentiated the three diagnostic categories (Andreasen, 1979b). Positive thought disorder includes decreases in semantic or discourse coherence (e.g., tangentiality, derailment, and circumstantiality), whereas negative thought disorder includes poverty of speech and content. Overall, there was equivalent positive thought disorder among patients with mania or schizophrenia, whereas negative thought disorder was most severe in schizophrenia patients (Andreasen, 1979b). Further studies, including meta-analyses (Yalincetin et al., 2016), have largely confirmed Andreasen's heuristic, showing that positive thought disorder is evident across diagnoses, with greater “negative” thought disorder in schizophrenia than in mood disorders. This is consistent with prognostic studies as well. The TLC was applied to videotaped semi-structured interviews with school-aged children of patients with schizophrenia or affective disorder, who were asked about family, friends, school and leisure activities (Gooding et al., 2013). While positive thought disorder ratings predicted psychosis, negative thought disorder was predictive specifically of schizophrenia, but not mood disorders with psychosis. Accuracy in prediction of diagnosis a decade later was as high as 92%, suggesting these are early core features of illness that predate psychosis onset.

In respect to this heuristic of positive and negative thought disorder, Barch and Berenbaum theorized that negative thought disorder is due to difficulty in generating a discourse plan, whereas positive thought disorder is due to difficulty in maintaining a discourse plan and monitoring ongoing content of speech. To test these hypotheses, they manipulated factors in eliciting speech in schizophrenia patients, including varying context before stories to influence generation of a discourse plan, and varying the question type to influence maintenance of a discourse plan (Barch and Berenbaum, 1997). They operationalized negative thought disorder as reduced verbosity (number of words) and syntactic complexity (mean number of dependent clauses per independent clause), and increased pause length. They used the TLC to count instances of positive disorder or disturbance in discourse coherence (e.g. “tangential responses”, “loss of goal”, “derailment”, “non-sequiturs” and “distractible speech”) in schizophrenia patients, adjusted for speech output. To further index discourse coherence, they included measures of referential cohesion, which refers to the use of language features that tie or link ideas between phrases or sentences (Halliday and Hasan, 2014). Referential cohesion can be pronominal (“Joe” is later referred to as “he”), demonstrative (“the girl” can be later referred to as this girl), comparative (“this” is contrasted with “that”). Overall, Barch and

Please cite this article as: C.M. Corcoran, V.A. Mittal, C.E. Bearden, et al., Language as a biomarker for psychosis: A natural language processing approach, Schizophrenia Research, https://doi.org/10.1016/j.schres.2020.04.032

Berenbaum found that low context (fewer directions) yielded speech characterized by more negative thought disorder, whereas low structure of questions (e.g. vague topic) yielded speech characterized by more positive thought disorder, all within the same individuals. Their experimental manipulation showed that indications of thought disorder are context-dependent and more evident when auxiliary conversational structure by the interviewer is less present.

2.1. Positive thought disorder

Reduction in discourse or semantic coherence, as operationalized by the TLC as positive thought disorder, has been assessed in schizophrenia and related psychotic disorders through the NLP analytic of latent semantic analysis (LSA). LSA rests on the premise that word meaning is a function of the relationship of each word to every other word in the lexicon (Landauer and Dumais, 1997; Landauer et al., 1998). The key insight in LSA is that word meanings are implicit in distributions of frequencies across contexts. LSA begins with the construction of a term-document matrix. The rows in the matrix correspond to individual words and the columns to documents, or otherwise, contexts. Cells in the matrix are filled with frequency counts (the number of times a word appears in a given context) weighted by the relative importance of each frequency, as specified in the tf-idf algorithm (Robertson, 2004). To reduce noise and increase generalization, distribution of frequencies across contexts is projected into a lower (300–400) dimensionality space using single-value decomposition. The semantic space created by LSA can specify, for example, that the words “sofa” and “couch” are highly similar in meaning. Their high similarity stems from the tendency of the words to appear in the same contexts, even if they rarely appear together in the same context because doing so would be redundant. More formally, the relative similarity between any two words can be assessed in terms of the cosine of the angle between the vectors (or word embeddings) associated with each word. New approaches to constructing word embeddings have recently appeared, such as Word2Vec (Mikolov et al., 2013a; Mikolov et al., 2013b), GloVe (Pennington et al., 2014) and more recently BERT (Devlin et al., 2018). The principles behind these new approaches are similar to those of LSA given that word embeddings are derived from distributions over linguistic contexts. Once word meanings are available, they can be combined to create representations for sentences. Vectors for sentences are calculated by summing the vectors associated with each word in the sentence. Sentence vectors, in turn, can be used to measure semantic coherence at the discourse level by simply measuring the cosine between adjacent sentences.

Elvevåg and colleagues were among the first to use LSA to compute discourse coherence in language elicited from schizophrenia patients and healthy volunteers using a variety of language tasks; patients were stratified based on TLC thought disorder ratings (Elvevåg et al., 2007). In schizophrenia patients with high TLC ratings, LSA detected more unusual word associations and less semantic similarity among animals successively named in a verbal fluency task. The language samples in this study were obtained in interviews in which participants were prompted to “Tell me the story of Cinderella/Romeo and Juliet” or “Why do some people believe in God?” or “What would someone need to do to do their laundry?” Using a “moving windows” method and computing successive cosines of phrases in respect to the initial prompt, patients with high TLC ratings lost coherence more quickly. They also had less coherence with other participants' responses. In Elvevåg et al. (2010), schizophrenia patients and their unaffected relatives were asked to talk about whatever came to mind, such as what they did yesterday or what they would be doing. LSA analyses were able to discriminate the two groups of paticipants with 86% accuracy. Decreased LSA semantic coherence also characterizes older schizophrenia patients (Holshausen et al., 2014), in whom it is related to poor adaptive functioning, independent of demographics and other symptoms.

2.2. Negative thought disorder

Negative thought disorder in schizophrenia and psychosis risk may plausibly be indexed through other NLP analytics, such as part-of-speech (POS) tagging (Santorini, 1990), by assessing semantic density as an index of poverty of content (Rezaii et al., 2019) and also through the use of speech graph analysis (Mota et al., 2012; Mota et al., 2017). In respect to POS tagging, just as every word in a text can be ascribed a semantic vector using LSA, every word in a text can also be labeled or “tagged” in respect to its grammatical function (Bird, 2009; Santorini, 1990), again learning from a large corpus of text. Once words are tagged, indices of syntactic complexity can be determined, including sentence length determined using rules of grammar, and frequency of types of “complementizer” words such as “that” and “which”, which can be used to introduce dependent clauses. Reduced sentence length and “complementizer” usage comprised part of an NLP classifier that predicted psychosis onset in one CHR cohort, and which was correlated with negative symptom severity (Bedi et al., 2015),

Poverty of content is a feature of negative thought disorder characteristic of schizophrenia (Andreasen, 1979b), and is predictive of psychosis onset (Rezaii et al., 2019; van Rooijen et al., 2017). Rezaii et al. (2019) showed how this indicator of psychosis, which they describe as *low semantic density*, could be identified through the computational technique of *vector unpacking*. The technique of vector unpacking starts with a sentence vector, which is a vector created by adding together and normalizing the vectors associated with the words in a sentence. It also begins with a large inventory of vectors for most of the words used in a given language. The technique uses an optimization algorithm known as “gradient descent” to discover the linear combination of weighted word vectors from this inventory that best approximates the observed sentence vector. When there is minimal semantic overlap among words in a sentence, all the words in the sentence vector are usually recovered. However, when the semantics of the words in a sentence overlap in meaning, the number of meaning vectors needed to create the sentence is less than the number of content words, resulting in a reduction in semantic density. Rezaii et al. (2019) showed how this technique, in combination with analysis of the speaker's content, could be used to predict psychosis onset among CHR individuals with high accuracy.

Patterns in language connectedness, that is the proximity in the discursive order of words regardless of content and syntax, offer yet another predictor of psychosis. Language connectedness, in particular its complexity, can be assessed using graph theory (Sigman and Cecchi, 2002). Graphs can be created from language by treating the words as nodes and the connections between successive words in a narrative as edges (Mota et al., 2012; Mota et al., 2017). Indices include the size of the strongly connected sub-graphs or components within the speech graph. Such sub-graphs can be used to discriminate the sparse speech of schizophrenia from that of manic psychosis. It can also be used to predict the emergence of first-episode psychosis, as well as account for the variance in negative symptoms within six months of onset (Mota et al., 2017). A normative developmental trajectory has been identified for these indices of complexity, showing early deviation for patients with psychosis (Mota et al., 2018). Further, these speech graph features have been correlated in psychosis with cortical gyrification, degree centrality in resting state functional connectivity, processing speed and clinical ratings of thought disorder (Palaniyappan et al., 2019). Speech graph methods hold promise for understanding language disturbance in CHR patients.

3. Procedures for collecting and analyzing language in psychosis risk

3.1. Clinical ratings

In schizophrenia and other psychotic disorders, language has been assessed in clinical interviews using the Positive and Negative

Please cite this article as: C.M. Corcoran, V.A. Mittal, C.E. Bearden, et al., Language as a biomarker for psychosis: A natural language processing approach, Schizophrenia Research, https://doi.org/10.1016/j.schres.2020.04.032

Syndrome Scale (PANSS) (Kay and Opler, 1987), and in many psychosis risk cohorts, its derivative, the Structured Interview for Prodromal Syndromes/Scale of Prodromal Symptoms (SIPS/SOPS) (Miller et al., 1999), as the primary way to evaluate "conceptual disorganization". Similarly, for psychosis risk, the Comprehensive Assessment of At-Risk Mental States (CAARMS) (Yung et al., 2005) assesses "disorganized speech" through both subjective review and objective rating. These items primarily assess circumstantiality and tangentiality, akin to Andreasen's "positive thought disorder" rubric, with the PANSS and SIPS/SOPS capturing Andreasen's "negative thought disorder" through negative symptom items such as "emotional expression" and "ideational richness".

Interestingly, the SIPS/SOPS "disorganized communication" item (e.g. P5) has consistently predicted psychosis onset in psychosis risk cohorts (DeVylder et al., 2014; Nelson et al., 2013; Demjaha et al., 2012; Addington et al., 2015; Cornblatt et al., 2015) including as a stable elevated trajectory over time in one medium sized cohort (N = 100, 26 converters), with a hazard of >2.2 (DeVylder et al., 2014). The predictive power was subsequently confirmed in the North American Prodrome Longitudinal Study (NAPLS) consortium (N = 764) (Addington et al., 2015), with an increased hazard of 8.0 for the same cutoff (SIPS rating > 2) at one NAPLS site, (N = 92, 25 converters), carrying the greatest weight in their predictive model (Cornblatt et al., 2015).

3.2. Manual linguistic analyses

Beyond clinical ratings, manual linguistic methods are used to assess disorder in thought. In an assessment of language abnormalities in speech transcripts, Bearden and colleagues showed that later transition to psychosis in CHR individuals was predicted by increased frequency of illogical thinking, with accuracy of 71%, compared with 35% for clinical ratings (Bearden et al., 2011). Poverty of content and decreased referential cohesion also predicted psychosis onset. In this study, the Story Game was used to elicit speech samples. The Story Game entails listening to two brief audiotaped stories, and then retelling each story, also answering sets of open-ended questions about the stories, such as what the participant liked about the story; it also entails creating a new story about one of four topics (e.g. "an unhappy child"). The Story Game was designed to be "an ecologically valid assessment of natural speech", and has been validated and used across a number of conditions in children and adolescents, including autism and schizophrenia spectrum. The Story Game is rated using the Kiddie Formal Thought Disorder Rating Scale (K-FTDS), yielding frequency counts of instances of language disturbance adjusted for amount of speech produced (Caplan et al., 1989). Other than illogical thinking and poverty of content, other disturbances included looseness of associations and incoherence, which had low base rates in this risk cohort. For K-FTDS ratings, "illogical thinking" comprises a failure in reasoning or contradiction, and "poverty of content" describes a failure to elaborate, whereas "loose associations" were abrupt unpredictable topic changes, and "incoherence" was scrambled syntax (Bearden et al., 2011).

In this same study by Bearden and colleagues, transcripts were evaluated for cohesion, which refers to language features that tie or link ideas between phrases or sentences (Halliday and Hasan, 2014). Referential cohesion can be pronominal ("Joe" is later referred to as "he"), demonstrative ("the girl" can be later referred to as this girl), comparative ("this" is contrasted with "that"). Reduction in referential cohesion can be indexed by the number of unclear or ambiguous references, adjusted for number of words; it can be elicited in schizophrenia by the use of unstructured (vs. structured) questions (Barch and Berenbaum, 1997). Decreased referential cohesion in response to the Story Game in CHR individuals predicted both later schizophrenia outcome and impairment in role function at follow-up. Likewise, poverty of content also predicted later schizophrenia outcome, as well as impairment in social function at follow-up (Bearden et al., 2011).

3.3. Natural language processing in psychosis risk cohorts

Numerous studies have shown how manual analyses of natural language can be used to identify thought disorder. In practice, however, manual analyses are difficult to implement. Such challenges have led to the use of automated NLP methods in studying language patterns related to psychosis risk: a partial inventory of such methods is shown in Table 1. These approaches are often used in combination. For example, LSA has been paired with POS tagging to evaluate discourse coherence and syntactic complexity, respectively and together, broadly following the heuristic established by Andreasen and used by investigators such as Barch and Berenbaum. Semantic density, as an index of poverty of content, has been paired with actual semantic content (the meaning of the words themselves) to predict transition to psychosis. The use of more than one technique suggests that language-based assessments of thought disorder might be most effective when the different techniques are combined.

In a small proof-of-principle study, NLP was used with machine learning to determine baseline patterns that might predict later psychosis onset among CHR individuals (Bedi et al., 2015). In this small study, LSA and POS tagging analytics were applied to open-ended narrative of 30–45 min elicited using qualitative interviewing techniques. A machine learning classifier with high predictive power for psychosis onset was identified that comprised minimum semantic coherence from one phrase to the next, phrase length and usage of "determiners" such as "which" and "that" as "complementizers", which introduce dependent clauses (Bedi et al., 2015). This classifier was correlated with positive and negative SIPS symptoms but outperformed them in classification accuracy. "Minimum semantic coherence" was validated by using it to index induced parametric scrambling of classic literary texts. Further, in a post hoc analysis of the classifier in an independent sample collected by Mota et al. (2012), the classifier distinguished the

Table 1
Natural language processing (NLP) techniques used in the assessment of psychosis.

Property of language	NLP technique	Outcome
Discourse coherence	Latent semantic analysis, word2vec, GLoVE Automatically represent each sentence as a vector and compare similarity of neighboring sentences using cosine similarity	Discriminates schizophrenia from HC (Elvevåg et al., 2007) Predicts psychosis onset (Bedi et al., 2015; Corcoran et al., 2018)
Syntactic complexity	Syntactic parsing and Part-of-speech (POS) tagging Automatically measures phrase structure, sentence length and frequency of part-of-speech classes (e.g. nouns, verbs, pronouns)	Predicts psychosis onset (Bedi et al., 2015; Corcoran et al., 2018)
Poverty of content	Vector unpacking Automatically measure semantic density: number of vectors needed to reconstruct sentence meaning	Predicts psychosis onset (Rezaii et al., 2019)
Referential coherence	Coh-Metrix Tool that applies POS-tagging and compares number of morphological roots shared across sentences	CHR vs. HC group difference (Gupta et al., 2018)
Metaphorical language	Use word2vec and neural network to automatically tag each word as literal or metaphorical	Discriminates first episode psychosis from HC (Gutiérrez et al., 2017)
Language connectedness	Speech graph analysis	Discriminates mania from schizophrenia and predicts schizophrenia diagnosis at 6 months (Mota et al., 2012, 2017)

language of schizophrenia patients from that of healthy individuals in a Brazilian cohort, after Portuguese transcripts were translated to English, suggesting the classifier might be robust across illness stages and across languages.

A similar approach, using LSA and POS tagging, with machine learning, was applied to the Story Game transcripts that Bearden and colleagues used to show that illogical thinking, poverty of content and decreased referential cohesion were predictive of psychosis onset in a CHR cohort (Corcoran et al., 2018). As speech was elicited using a more structured paradigm, and responses were briefer (<20 mean words per response at UCLA vs. >150 words per response in NYC), there was insufficient free speech for sentence-level analysis of coherence. Hence, semantic coherence was measured using k-level or "skip-gram" measures, which computes word-to-word variability at "k" inter-word distances. Five semantic and nine syntactic features were used for machine learning classification, with singular value decomposition (SVD) used for reduction of dimensions in the training set. A four-factor solution was found for the classifier, with the top three weighted toward coherence, and the fourth weighted for syntax, specifically possessive pronouns ("complementizers" did not add to the model). The receiver operating characteristic (ROC) for the classifier had a within-set area under the curve (AUC) of 0.88, consistent with high accuracy (Corcoran et al., 2018).

Further, to show cross-validation across sites this machine learning classifier derived from the Story Game CHR dataset was applied verbatim to the dataset from the small proof-of-principle CHR study, after first applying a Procrustean global transform (rigid translation and rotation in Euclidean space) to minimize distortions, given difference in length of responses. In cross-validation, the classifier had an AUC of 0.71 in predicting psychosis in the second independent CHR longitudinal dataset (Corcoran et al., 2018).

A third NLP study of psychosis prediction in a CHR cohort used a different approach that examined semantic content and also used an innovative approach to evaluate "poverty of content" through the measure of "semantic density", which reflects the number of core ideas within a sentence (Rezaii et al., 2019). A skip-gram version of Word2Vec was used. Like LSA, Word2Vec learns the meaning of words from scanning a large corpus of text (in this study the New York Times corpus), but does so using moving windows, so that neural networks are trained to predict and learn words within the context of other words within a moving window. The skip-gram version of Word2Vec predicts the surrounding words based on the central word in the window. Vector unpacking, as described, was used to calculate the number of meaning vectors needed to reconstruct the meaning of a sentence, or "semantic density." These procedures were applied to transcripts of a standardized clinical interview, along with a measure of participants' semantic content during the interview (e.g. what they tended to talk about). Overall, lower semantic density in speech, along with greater use of words related to sounds and voices, was predictive of psychosis transition with an accuracy of 90%. Work is being done to determine the cross-site validation of this machine learning classifier and its components of semantic density and content.

Additional studies have focused on group differences in language between CHR and healthy individuals (Gupta et al., 2018). One study evaluated referential cohesion, which as described earlier refers to language features that tie or link ideas between phrases or sentences (Halliday and Hasan, 2014), and which was found by Bearden and colleagues to predict later psychosis onset in CHR individuals, using manual linguistic analyses. In this study, the Coh-Metrix tool was used to assess referential cohesion, and was applied to written narrative descriptions elicited by a visual prompt. Coh-Metrix first applies part-of-speech (POS) tagging, and then identifies roots and morphological forms to identify relational connections (e.g. referential cohesion) across the text. CHR individuals showed less referential cohesion, which was associated with severity of positive and disorganization symptoms, and lower verbal learning scores (Gupta et al., 2018).

Yet another study used an NLP approach to evaluate the use of token- or word-level "metaphor" across stages of psychotic illness (Gutiérrez et al., 2017). Patients with schizophrenia have long been known to use words in an idiosyncratic or bizarre manner, with Andreasen noting examples of "watches" being referred to as "time vessels" and "gloves" as "hand shoes" (Andreasen, 1986). In the 1990's, Billow and colleagues noted increased frequency of deviant (but not coherent) metaphorical speech in schizophrenia patients (Billow et al., 1997). Similar to other studies, skip-gram Word2Vec was used with a neural network to tag each word or token as literal or metaphorical, in respect to a large metaphor corpus. This was complemented by automated sentiment analysis, which rates words from 1 (very negative) to 5 (very positive), computing sentiment (and its coherence) at the word and phrase level. Speech was elicited using qualitative interviewing methods. A classifier that used all of these features, plus gender and age, discriminated first episode psychosis from healthy controls and had an accuracy of 84% (beyond 75% accuracy for metaphor usage alone); this best classifier tagged 85% of all CHR individuals within a dataset, including all future converters and most CHR non-converters (Gutiérrez et al., 2017). This suggests that this approach of NLP assessment of metaphor and sentiment may be useful for screening, if replicated.

4. Linguistic biomarkers: reasonable development path and mechanistic studies, in tandem with other CHR biomarkers

4.1. Reasonable development path

A reasonable development path for a risk biomarker consists of initial validation and identification of sources of variability, tests of reproducibility and reliability, and mechanistic studies, and then for next steps, standardization of protocols for use as a prognostic marker and target in clinical trials, with attention to sensitivity/specificity, traction/feasibility, acceptability, cost, utility and regulatory "context of use" determined by field trials.

Among the linguistic biomarkers of psychosis risk evaluated thus far, semantic coherence reduction has been cross-validated across risk cohorts. It also may have among the most traction for consideration with other biomarkers, as it has been evaluated during the last decade in schizophrenia cohorts with genetic and circuit-level units/levels of the RDoC matrix. A preliminary study suggested associations of LSA semantic coherence with SNPs in the Disrupted in schizophrenia 1 (DISC-1) gene (Nicodemus et al., 2014). In respect to circuits, semantic coherence measured from free discourse on "religious belief" was associated during a word monitoring task with increased modality-specific activation in auditory and visual regions, and in superior/middle temporal regions in schizophrenia patients; by contrast, semantic coherence in healthy individuals was associated only with activation in executive regions during this same task (Tagamets et al., 2014). These findings suggest normal reliance on prefrontal regions for fluency and coherence, with potential compensation from more sensory regions in schizophrenia. This is consistent with the finding that abnormal activation of superior temporal gyrus during discourse processing predicts psychosis transition among CHR individuals (Sabb et al., 2010).

The optimal parameters for the solicitation of speech are not yet known and are necessary for harmonization across studies. Both speech graph and latent semantic analyses have been applied in schizophrenia to brief narratives of several sentences over fewer than 5 min, including recall of dreams and memories (Mota et al., 2012), and descriptions of free will or how to do laundry (Elvevåg et al., 2007). More subtle differences in CHR individuals may require longer transcripts (Corcoran et al., 2018). Some investigators capitalize on analyzing diagnostic interviews, which can provide the opportunity to evaluate symptom content (Rezaii et al., 2019).

A further limitation is that most NLP studies in schizophrenia and CHR have focused on transcripts of English, except for Mota's speech

Please cite this article as: C.M. Corcoran, V.A. Mittal, C.E. Bearden, et al., Language as a biomarker for psychosis: A natural language processing approach, Schizophrenia Research, https://doi.org/10.1016/j.schres.2020.04.032

graph analyses in Portuguese and de Boer et al. (2020)'s analyses of language disturbances in Dutch. However, studies are underway with speakers of other languages, such as Chinese and Spanish, which will help establish the degree to which the language patterns observed in English generalize across languages.

4.2. Mechanisms

Language production and comprehension rely on canonical circuits that involve superior temporal (and inferior frontal) regions. Reductions in discourse coherence and complexity may be related to pathology in the language circuit. In one study of schizophrenia patients, abnormal activation in superior temporal regions during a word-monitoring task was associated with decreased LSA coherence (Tagamets et al., 2014). In another study of CHR patients, increased activation during a naturalistic discourse processing paradigm was observed in a network of language-associated brain regions, with increased activation in superior temporal and inferior temporal gyri specifically predictive of later psychosis transition (Sabb et al., 2010).

A strategy for understanding the circuit dysfunction that underlies language disturbance in schizophrenia and CHR may be neuroimaging during the production and comprehension of natural language itself as has been done by Hasson and colleagues, finding in normal individuals the synchronized recruitment of an extensive bilateral network (Silbert et al., 2014). Scrambling of stories heard at the word (1 ± 0.5 s), sentence (8 ± 3 s) and paragraph (38 ± 17 s) levels shows a normative expansion in topography of intersubject synchronization over longer time windows of intact speech (Lerner et al., 2011). Hasson has postulated a hierarchy of temporal receptive windows for language that reflect a topography from basic unimodal sensory regions (shorter windows) to higher-order processing cortical areas (longer windows). He has found that a temporal receptive window of ~10 s (i.e. sentence length in English) is needed for reliable activation in middle and superior temporal regions when listening to a narrative. It may be that schizophrenia and CHR patients will have a disruption in this topography, especially in superior temporal regions, which may be correlated with NLP indices. For example, if information processing breaks down over the 8–-10 second time frame of a sentence, it is plausible that an individual may go off track (reduced coherence) or pause without elaboration (reduced complexity). Kuperberg has theorized that in schizophrenia, there is a breakdown in the hierarchical generative framework of language, in which normally, higher-level inferences constrain interpretation of sensory information and are updated based on prediction error (Brown and Kuperberg, 2015): novel computational approaches will be needed to test this.

Language disturbances similar to those seen in SZ can be readily induced by NMDA receptor antagonists such as ketamine, suggesting the language disturbance seen in SZ and CHR may reflect underlying glutamatergic dysfunction. Language production requires “chunking” or grouping of contextually related stimuli, and the formation of “nested tree structures”, processes that involve superior temporal and inferior frontal regions (Dehaene et al., 2015). Related to this, deficits in perceptual grouping of visual stimuli (contour integration, visual closure) are also associated in schizophrenia with thought disorder (Uhlhaas et al., 2006), and likewise can be disrupted by ketamine.

Language disturbance in CHR individuals may be related to information processing deficits in hub-like regions such as the superior temporal gyrus (Collin et al., 2018). However, it is plausible that it is related also to deficits in sensory processing, specifically auditory mismatch negativity event-related potentials, as well as to deficits in basic cognitive functions such as processing speed, working memory and verbal fluency. These sensory processing and cognitive deficits have been documented in both schizophrenia and CHR individuals, in whom they are associated with an increase in risk for psychosis. The proposal to study language together with them is discussed in the next section.

4.3. Combinations with other biomarkers

Linguistic biomarkers of psychosis risk can also be assessed in combination with other known replicated biomarkers of psychosis risk, including deficits in cognitive and sensory processing, and potentially genetics and imaging markers. They can also be evaluated alone and in combination to predict other outcomes in the pluripotent CHR population, including poor functional outcome, onset of other disorders, and remission and recovery. For example, in psychosis risk studies, replicated predictors of psychosis onset include slowing of processing speed, and reductions in verbal fluency and verbal memory (Seidman et al., 2010; Fusar-Poli et al., 2012); these cognitive domains also are part of the NAPLS risk calculator for psychosis onset (Cannon et al., 2016). While impairments in these cognitive domains would be expected to be related to language impairments in psychosis risk states, the associations of these with reduction in semantic coherence and other language features (e.g., semantic density, referential cohesion, use of complementizers in dependent clauses) is in need of further investigation.

Among psychosis risk biomarkers in CHR cohorts, physiological measures of auditory processing, including auditory P300s and mismatch negativity (MMN), are among the most replicated predictors of later psychosis onset (Bodatsch et al., 2011; Perez et al., 2014; Van Tricht et al., 2015; Shaikh et al., 2012; Hamilton et al., 2017), and also have predictive power for functional outcomes (Hamilton et al., 2019). Auditory MMN is the event-related potential that occurs in response to a tone deviant from a series of tones, typically in duration, frequency or intensity. Thus far, NLP linguistic biomarkers have not yet been evaluated in respect to auditory processing in psychosis and its risk states, though we would expect these to be associated. By contrast, specific language impairment, a heterogeneous disorder observed in children, has been consistently associated with reductions in auditory mismatch negativity (MMN) potentials (Kujala and Leminen, 2017). In specific language impairment, auditory MMN is normalized by language exercise and remediation (Dacewicz et al., 2018), which suggests a remediation strategy for abnormal auditory processing and language in psychosis and its risk states.

Other identified risk biomarkers for psychosis include polygenic risk scores (Perkins et al., 2020), progressive reduction in gray matter (Cannon et al., 2015), and abnormal patterns of resting state functional connectivity (Anticevic et al., 2015; Colibazzi et al., 2017; Cao et al., 2018). The association of these with linguistic risk biomarkers is not yet known, nor their combined predictive power for various outcomes, including psychosis, functional impairment and remission.

Overall, a research agenda for large-scale CHR networks is to evaluate combinations of promising risk biomarkers for varied outcomes, including the onset of psychotic and other disorders, for functional outcome, and for remission and recovery. These biomarker studies can be used to refine practical biomarker use in the context of precision medicine, enable stratification and case enhancement for clinical trials, and elucidate mechanisms to provide targets for preventive intervention.

5. Future plans to conduct analyses at the level of the dyad, and broaden data to include prosody and face expression

This review has focused on the analysis of language structure and content in psychosis. But spoken language is much more than the words that are said. Indeed, there have been several studies that indicate that acoustic features may yield important and distinct clues about etiology of psychosis as well as tools for early identification and treatment tracking. For example, computational work has shown that individuals with schizophrenia exhibit less variability in their pitch (i.e. fundamental frequency, F0) and first two formants (F1–F2), which are resonant frequencies that are determined by the shape of the vocal tract during speech (Covington et al., 2012; Bernardini et al.,

Please cite this article as: C.M. Corcoran, V.A. Mittal, C.E. Bearden, et al., Language as a biomarker for psychosis: A natural language processing approach, Schizophrenia Research, https://doi.org/10.1016/j.schres.2020.04.032

2016). In addition, research indicates individuals with psychosis pause more, speak at a slower rate, and have reduced intensity and vowel space area compared to healthy controls (e.g. Martínez-Sánchez et al., 2015; Compton et al., 2018; Arevian et al., 2020). This is particularly relevant as one recent study showed that CHR individuals have increased pauses in speech, similar to what is seen in schizophrenia (Stanislawski et al., 2019). In addition to quantitatively measuring flat affect or aprosody, acoustics could also reflect motor control deficits, which are common in the prodrome (Dean et al., 2018), as successful speech production requires complex motor coordination. For example in patients with Parkinson's Disease, acoustic measures such as the stability of syllable durations, rate of change of speech, inappropriate voicing of consonants, (e.g. pronouncing /p/ in a more /b/-like way) have shown particular promise (e.g. Karlsson et al., 2020), and could be useful for identifying individuals at high risk for psychosis as well. Although some of these measures have involved manual annotation in the past, recent advances in speech technologies allow for these measures to be automatically and objectively measured (Segal et al., 2019; Shrem et al., 2019), and to be used on a wider scale.

Importantly, impairment in social function is a key feature of psychosis and its risk states that may account for most of the morbidity of these syndromes. This impairment in social function is likely multifactorial in etiology, though may be largely accounted for by impairments in social communication. Thus, we must look at language and speech within dyads, and in the context of gesture and face emotion expression. Beginning work in this area shows that individuals at risk for psychosis have abnormal turn-taking (Sichlinger et al., 2019) as well as blunted facial affect during interview (Gupta et al., 2019). This may also be expanded to including the gestures that accompany speech, which are abnormal in psychosis risk individuals as well (Mittal et al., 2006; Millman et al., 2014; Bernard et al., 2015; Osborne et al., 2017).

6. Conclusions

Computational analysis of ecological language and communication behavior, both *in vivo* and digitally via smartphones and social media, are promising avenues to pursue to understand psychosis risk and emergence, evaluated in tandem with biomarkers across genetic, physiological, circuit-based and cognitive levels of analysis. Given the close ties with other core phenomenology, links with distinct mechanisms, and ease of ascertainment and analysis, it is clear that assessment of natural language processing will be an invaluable domain for understanding and treating individuals at CHR for psychosis.

Acknowledgements

We would like to acknowledge to grants, both to the first author Cheryl Corcoran: National Institutes of Health 5R01MH107558-06 and 5R01MH115332-03

References

Addington, J., Liu, L., Buchy, L., Cadenhead, K.S., Cannon, T.D., Cornblatt, B.A., Perkins, D.O., Seidman, L.J., Tsuang, M.T., Walker, E.F., Woods, S.W., Bearden, C.E., Mathalon, D.H., McGlashan, T.H., 2015. North American Prodrome Longitudinal Study (NAPLS 2): the prodromal symptoms. J. Nerv. Ment. Dis. 203, 328–335. https://doi.org/10.1097/NMD.0000000000000290.

Andreasen, N.C., 1979a. Thought, language, and communication disorders: I. Clinical assessment, definition of terms, and evaluation of their reliability. Arch. Gen. Psychiatry 36, 1315–1321. https://doi.org/10.1001/archpsyc.1979.01780120045006.

Andreasen, N.C., 1979b. Thought, language, and communication disorders: II. Diagnostic significance. Arch. Gen. Psychiatry 36, 1325–1330. https://doi.org/10.1001/archpsyc.1979.01780120055007.

Andreasen, N.C., 1986. Scale for the assessment of thought, language, and communication (TLC). Schizophr. Bull. 12, 473–482. https://doi.org/10.1093/schbul/12.3.473.

Andreasen, N.C., Grove, W.M., 1986. Thought, language, and communication in schizophrenia: diagnosis and prognosis. Schizophr. Bull. 12, 348–359. https://doi.org/10.1093/schbul/12.3.348.

Anticevic, A., Haut, K., Murray, J.D., Repovs, G., Yang, G.J., Diehl, C., McEwen, S.C., Bearden, C., Addington, J., Goodyear, B., Cadenhead, K.S., Mirzakhanian, H., Cornblatt, B.A., Olvet, D., Mathalon, D.H., McGlashan, T.H., Perkins, D.O., Belger, A., Seidman, L.J., Tsuang, M.T., van Erp, T.G.M., Walker, E.F., Hamann, S., Woods, S.W., Qiu, M., Cannon, T.D., 2015. JAMA Psychiatry 72. https://doi.org/10.1001/jamapsychiatry.2015.0566.

Arevian, A.C., Bone, D., Malandrakis, N., Martinez, V.R., Wells, K.B., Miklowitz, D.J., Narayanan, S., 2020. Clinical state tracking in serious mental illness through computational analysis of speech. PLoS One 15 (1), e0225695.

Barch, D.M., Berenbaum, H., 1997. The effect of language production manipulations on negative thought disorder and discourse coherence disturbances in schizophrenia. Psychiatry Res. 71, 115–127. https://doi.org/10.1016/S0165-1781(97)00045-0.

Bearden, C.E., Wu, K.N., Caplan, R., Cannon, T.D., 2011. Thought disorder and communication deviance as predictors of outcome in youth at clinical high risk for psychosis. J. Am. Acad. Child Adolesc. Psychiatry 50, 669–680. https://doi.org/10.1016/j.jaac.2011.03.021.

Bedi, G., Carrillo, F., Cecchi, G.A., Fernandez-Slezak, D., 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 Schizophr. 1, 15030. https://doi.org/10.1038/npjschz.2015.30.

Bernard, J.A., Millman, Z.B., Mittal, V.A., 2015. Beat and metaphoric gestures are differentially associated with regional cerebellar and cortical volumes. Hum. Brain Mapp. 36, 4016–4030. https://doi.org/10.1002/hbm.22894.

Bernardini, F., Lunden, A., Covington, M., Broussard, B., Halpern, B., Alolayan, Y., Attademo, L., 2016. Associations of acoustically measured tongue/jaw movements and portion of time speaking with negative symptom severity in patients with schizophrenia in Italy and the United States. Psychiatry Res. 239, 253–258.

Billow, R.M., Rossman, J., Lewis, N., Goldman, D., Raps, C., 1997. Observing expressive and deviant language in schizophrenia. Metaphor. Symb. 12, 205–216. https://doi.org/10.1207/s15327868ms1203_3.

Bird, S., 2009. Natural language processing and linguistic fieldwork. Comput. Linguist. 35, 469–474. https://doi.org/10.1162/coli.35.3.469.

Bleuler, E., 1950. Dementia praecox or the group of schizophrenias (J. Zinkin, trans.). Dementia praecox or the group of schizophrenias. J. Zinkin, trans. 35, 469–474.

Bodatsch, M., Ruhrmann, S., Wagner, M., Müller, R., Schultze-Lutter, F., Frommann, I., Brinkmeyer, J., Gaebel, W., Maier, W., Klosterkötter, J., Brockhaus-Dumke, A., 2011. Prediction of psychosis by mismatch negativity. Biol. Psychiatry 69, 959–966.

Brown, M., Kuperberg, G.R., 2015. A hierarchical generative framework of language processing: linking language perception, interpretation, and production abnormalities in schizophrenia. Front. Hum. Neurosci. 9, 643. https://doi.org/10.3389/fnhum.2015.00643.

Cannon, T.D., Yoonho, C., He, G., Sun, D., Jacobson, A., van Erp, T.G.H., McEwen, S., Addington, J., Bearden, C.E., Cadenhead, K., Cornblatt, B., Mathalon, D.H., McGlashan, T., Perkins, D., Jeffries, C., Seidman, L.J., Tsuang, M., Walker, E., Woods, S.W., Heinssen, R., 2015. Progressive reduction in cortical thickness as psychosis develops: a multisite longitudinal neuroimaging study of youth at elevated clinical risk. Biol. Psychiatry 77, 147–157. https://doi.org/10.1016/j.biopsych.2014.05.023.

Cannon, T.D., Yu, C., Addington, J., Bearden, C., Cadenhead, K.S., Cornblatt, B.A., Heinssen, R., Jeffries, C.D., Mathalon, D.H., McGlashan, T.H., Perkins, D.O., Seidman, L.J., Tsuang, M.T., Walker, E.F., Woods, S.W., Kattan, M., 2016. An individualized risk calculator for research in prodromal psychosis. Am. J. Psychiatry 173, 980–988. https://doi.org/10.1176/appi.ajp.2016.15070890.

Cao, H., Chén, O.Y., Chung, Y., Forsyth, J.K., McEwen, S.C., Gee, D.G., Bearden, C.E., Addington, C., Addington, J., Goodyear, B., Cadenhead, K.S., Mirzakhanian, H., Cornblatt, B.A., Carrión, R.E., Mathalon, D.H., McGlashan, T.H., Perkins, D.O., Belger, A., Seidman, L.J., Thermenos, H., Tsuang, M.T., van Erp, T.H.M., Walker, E.F., Hamann, S., Anticevic, A., Woods, S.W., Cannon, T.D., 2018. Cerebello-thalamo-cortical hyperconnectivity as a state-independent functional neural signature for psychosis prediction and characterization. Nat. Commun. 9, 3836. https://doi.org/10.1038/s41467-018-06350-7.

Caplan, R., Guthrie, D., Fish, B., Tanguay, P.E., David-Lando, G., 1989. The kiddie formal thought disorder rating scale: clinical assessment, reliability, and validity. J. Am. Acad. Child Adolesc. Psychiatry 28, 408–416. https://doi.org/10.1097/00004583-198905000-00018.

Colibazzi, T., Yang, Z., Horga, G., Chao-Gan, Y., Corcoran, C.M., Klahr, K., Brucato, G., Girgis, R., Abi-Dargham, A., Miham, M.P., Peterson, B.S., 2017. Aberrant temporal connectivity in persons at clinical high risk for psychosis. Biol. Psychiatry Cogn. Neurosci. Neuroimaging. 2, 696–705.

Collin, G., Seidman, L.J., Keshavan, M.S., Stone, W.S., Zhenghan, Q., Zhang, T., Tang, Y., Li, H., Anteraper, S.A., Niznikiewicz, M.A., McCarley, R.W., Shenton, M.E., Wang, J., Whitfield-Gabrieli, S., 2018. Functional connectome organization predicts conversation to psychosis in clinical high-risk youth from the SHARP program. Mol. Psychiatry https://doi.org/10.1038/s41380-018-0288-x.

Compton, M.T., Lunden, A., Cleary, S.D., Pauselli, L., Alolayan, Y., Halpern, B., Bernardini, F., 2018. The aprosody of schizophrenia: computationally derived acoustic phonetic underpinnings of monotone speech. Schizophr. Res. 197, 392–399.

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

Cornblatt, B.A., Carrión, R.E., Auther, A., McLaughlin, D., Olsen, R.H., John, M., Correll, C.U., 2015. Psychosis prevention: a modified clinical high risk perspective from the recognition and prevention (RAP) program. Am. J. Psychiatry 172, 986–994. https://doi.org/10.1176/appi.ajp.2015.13121686.

Covington, M.A., Lunden, S.A., Cristofaro, S.L., Wan, C.R., Bailey, C.T., Broussard, B., Compton, M.T., 2012. Phonetic measures of reduced tongue movement correlate with negative symptom severity in hospitalized patients with first-episode schizophrenia-spectrum disorders. Schizophr. Res. 142 (1–3), 93–95.

Dacewicz, A., Szymaszek, A., Nowak, K., Szelag, E., 2018. Training-induced changes in rapid auditory processing in children with specific language impairment:

electrophysiological indicators. Front. Hum. Neurosci. 12, 310. https://doi.org/10.3389/fnhum.2018.00310.

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

Dean, D., Walther, S., Bernard, J., Mittal, V.A., 2018. Motor clusters reveal differences in risk for psychosis, cognitive functioning, and thalamocortical connectivity: evidence for vulnerability subtypes. Clin. Psychol. Sci. 6 (5), 721–734. https://doi.org/10.1177/2167702618773759.

Dehaene, S., Meyniel, F., Wacongne, C., Wang, L., Pallier, C., 2015. The neural representation of sequences: from transition probabilities to algebraic patterns and linguistic trees. Neuron 88, 2–19. https://doi.org/10.1016/j.neuron.2015.09.019.

Demjaha, A., Valmaggia, L., Stahl, D., Byrne, M., McGuire, P., 2012. Disorganization/cognitive and negative symptom dimensions in the at-risk mental state predict subsequent transition to psychosis. Schizophr. Bull. 38, 351–359. https://doi.org/10.1093/schbul/sbq088.

Devlin, J., Chang, Ming-Wei, Lee, K., Toutanova, K., 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (arXiv:1810.04805).

DeVylder, J.E., Muchomba, F.M., Gill, K.E., Ben-David, S., Walder, D.J., Malaspina, D., Corcoran, C.M., 2014. Symptom trajectories and psychosis onset in a clinical high-risk cohort: the relevance of subthreshold thought disorder. Schizophr. Res. 159, 278–283. https://doi.org/10.1016/j.schres.2014.08.008.

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

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. J. Neurolinguistics, 270–284 https://doi.org/10.1016/j.jneuroling.2009.05.002.

Elvevåg, B., Cohen, A.S., Wolters, M.K., Whalley, H.C., Gountouna, V.E., Kuznetsova, K.A., Watson, A.R., Nicodemus, K.K., 2016. An examination of the language construct in NIMH's research domain criteria: time for reconceptualization! Am. J. Med. Genet. B Neuropsychiatr. Genet. 171, 904–919. https://doi.org/10.1002/ajmg.b.32438.

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

Fusar-Poli, P., Deste, G., Smieskova, R., Barlati, S., Yung, A.R., Howes, O., Stieglitz, R., Vita, A., McGuire, P., Borgwardt, 2012. Cognitive functioning in prodromal psychosis: a meta-analysis. Arch. Gen. Psychiatry 69, 562–571. https://doi.org/10.1001/archgenpsychiatry.2011.1592.

Gooding, D.C., Coleman, M.J., Roberts, S.A., Shenton, M.E., Levy, D.L., Erlenmeyer-Kimling, L., 2012. Thought disorder in offspring of schizophrenic parents: findings from the New York high-risk project. Schizophr. Bull. 38, 263–271. https://doi.org/10.1093/schbul/sbq061.

Gooding, D.C., Ott, S.L., Roberts, S.A., Erlenmeyer-Kimling, L., 2013. Thought disorder in mid-childhood as a predictor of adulthood diagnostic outcome: findings from the New York High-Risk Project. Psychol. Med. 43, 1003–1012. https://doi.org/10.1017/S0033291712001791.

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

Gupta, T., Haase, C.M., Strauss, G.P., Cohen, A.S., Mittal, V.A., 2019. Alterations in facial expressivity in youth at clinical high-risk for psychosis. J. Abnorm. Psychol. 128, 341–351. https://doi.org/10.1037/abn0000413.

Gutiérrez, E.D., Corlett, P.R., Corcoran, C.M., Cecchi, G.A., 2017. Using automated metaphor identification to aid in detection and prediction of first-episode schizophrenia. EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings. https://doi.org/10.18653/v1/d17-1316.

Halliday, M.A.K., Hasan, K., 2014. Cohesion in English. Cohesion in English. Longman, London https://doi.org/10.4324/9781315836010.

Hamilton, H.K., Perez, V.B., Ford, J.M., Roach, B., Jaeger, J., Mathalon, D.H., 2017. Mismatch negativity but not P300 is associated with functional disability in schizophrenia. Schizophr. Bull. 44. https://doi.org/10.1093/schbul/sbx104.

Hamilton, H.K., Roach, B.J., Bachman, P.M., Belger, A., Carrion, R.E., Duncan, E., Johannesen, J.K., Light, G.A., Niznikiewicz, M.A., Addington, J., Bearden, C.E., Cadenhead, K.S., Cornblatt, B.A., McGlashan, T.H., Perkins, D.O., Seidman, L.J., Tsuang, M.T., Walker, E.F., Woods, S.W., Cannon, T.D., Mathalon, D.H., 2019. Association between P300 responses to auditory oddball stimuli and clinical outcomes in the psychosis risk syndrome. JAMA Psychiatry https://doi.org/10.1001/jamapsychiatry.2019.2135.

Harrow, M., Quinlan, D., 1977. Is disordered thinking unique to schizophrenia? Arch. Gen. Psychiatry 34, 15–21. https://doi.org/10.1001/archpsyc.1977.01770130017001.

Holshausen, K., Harvey, P.D., Elvevåg, B., Foltz, P.W., Bowie, C.R., 2014. Latent semantic variables are associated with formal thought disorder and adaptive behavior in older inpatients with schizophrenia. Cortex 55, 88–96. https://doi.org/10.1016/j.cortex.2013.02.006.

Insel, T.R., 2017. Digital Phenotyping Technology for a New Science of Behavior. JAMA 318, 1215–1216. https://doi.org/10.1001/jama.2017.11295.

Karlsson, F., Schalling, E., Laakso, K., Johansson, K., Hartelius, L., 2020. Assessment of speech impairment in patients with Parkinson's disease from acoustic quantifications of oral diadochokinetic sequences. The Journal of the Acoustical Society of America 147 (2), 839–851.

Kay, S.R., Opler, L.A., 1987. The positive-negative dimension in schizophrenia: its validity and significance. Psychiatr. Dev. 5, 79–103.

Kraepelin, E., 2010. Cien años de psiquiatría. Vertex 21, 317–320.

Kujala, T., Leminen, M., 2017. Low-level neural auditory discrimination dysfunctions in specific language impairment—a review on mismatch negativity findings. Developmental Cognitive Neuroscience 28, 65–75. https://doi.org/10.1016/j.dcn.2017.10.005.

Landauer, T.K., Dumais, S.T., 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol. Rev. 104, 211–240. https://doi.org/10.1037/0033-295X.104.2.211.

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

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

Levy, D.L., Coleman, M.J., Sung, H., Ji, F., Maahysse, S., Mendell, N.R., Titone, D., 2010. The genetic basis of thought disorder and language and communication disturbances in schizophrenia. J. Neurolinguistics 23, 176. https://doi.org/10.1016/j.jneuroling.2009.08.003.

Martínez-Sánchez, F., Muela-Martínez, J.A., Cortés-Soto, P., Meilán, J.J.G., Ferrándiz, J.A.V., Caparrós, A.E., Valverde, I.M.P., 2015. Can the acoustic analysis of expressive prosody discriminate schizophrenia? The Spanish journal of psychology 18.

Metsänen, M., Wahlberg, K.E., Hakko, H., Saarento, O., Tienari, P., 2006. Thought disorder index: a longitudinal study of severity levels and schizophrenia factors. J. Psychiatr. Res. 40, 258–266. https://doi.org/10.1016/j.jpsychires.2005.03.004.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013a. Distributed representations of words and phrases and their compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems. Volume 2, pp. 3111–3119.

Mikolov, T., Yih, W.T., Zweig, G., 2013b. Linguistic regularities in continuous spaceword representations. NAACL HLT 2013–2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Main Conference, pp. 746–751.

Miller, T.J., McGlashan, T.H., Woods, S.W., Stein, K., Driesen, N., Corcoran, C.M., Hoffman, R., Davidson, L., 1999. Symptom assessment in schizophrenic prodromal states. Psychiatr. Q. 70, 273–287. https://doi.org/10.1023/a:1022034115078.

Millman, Z., Goss, J., Schiffman, J., Mejias, J., Gupta, T., Mittal, V.A., 2014. Mismatch and lexical retrieval gestures are associated with visual information processing, verbal production, and symptomatology in youth at high risk for psychosis. Schizophr. Res. 158 (1–3), 64–68. https://doi.org/10.1016/j.schres.2014.06.007.

Mittal, V.A., Tessner, K.D., McMillan, A.L., Delawalla, Z., Trotman, H.D., Walker, E.F., 2006. Gesture behavior in unmedicated schizotypal adolescents. J. Abnorm. Psychol. 115 (2), 351.

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 https://doi.org/10.1371/journal.pone.0034928.

Mota, N.B., Copelli, M., Ribeiro, S., 2017. Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance. NPJ Schizophr. 3, 18. https://doi.org/10.1038/s41537-017-0019-3.

Mota, N.B., Sigman, M., Cecchi, G., Copelli, M., Ribeiro, S., 2018. The maturation of speech structure in psychosis is resistant to formal education. NPJ Schizophr. https://doi.org/10.1038/s41537-018-0067-3.

Nelson, B., Yuen, H.P., Wood, S.J., Kin, A., Spiliotacopoulos, D., Bruxner, A., Broussard, C., Simmons, M., Foley, D.L., Brewer, W.J., Francey, S.M., Amminger, G.P., Thompson, A., McGorry, P.D., Yung, A.R., 2013. Long-term follow-up of a group at ultra high risk ('prodromal') for psychosis the PACE 400 study. JAMA Psychiatry 70, 1008. https://doi.org/10.1001/jamapsychiatry.2013.1270.

Nicodemus, K.K., Elvevåg, B., Foltz, P.W., Rosenstein, M., Diaz-Asper, C., Weinberger, D.R., 2014. Category fluency, latent semantic analysis and schizophrenia: a candidate gene approach. Cortex https://doi.org/10.1016/j.cortex.2013.12.004 (Epub 2013 Dec 20).

Osborne, J., Bernard, J., Millman, Z., Gupta, T., Ristanovic, I., Gupta, T., Vargas, T., Mittal, V., 2017. Beat gestures frequency and postural control in youth at ultrahigh risk for psychosis. Schizophr. Res. 185, 197–199.

Palaniyappan, L., et al., 2019. Speech structure links the neural and socio-behavioural correlates of psychotic disorders. Prog. Neuro-Psychopharmacology Biol. Psychiatry. https://doi.org/10.1016/j.pnpbp.2018.07.007.

Pennington, J., Socher, R., Manning, C.D., 2014. GloVe: Global Vectors for word representation. EMNLP 2014–2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, pp. 1532–1543. https://doi.org/10.3115/v1/d14-1162.

Perez, V.B., Woods, S.W., Roach, B.J., Ford, J.M., McGlashan, T.H., Srihari, V.H., Mathalon, D.H., 2014. Automatic auditory processing deficits in schizophrenia and clinical high-risk patients: forecasting psychosis risk with mismatch negativity. Biol. Psychiatry 75, 459–469. https://doi.org/10.1016/j.biopsych.2013.07.038.

Perkins, D.O., Olde, L.L., Barbee, J., Ford, J., Jeffries, C.D., Addington, J., Bearden, C.E., Badenhead, K.S., Cannon, T.D., Cornblatt, B.A., Mathalon, D.H., McGashan, T.H., Seidman, L.J., Tsuang, M., Walker, E.F., Woods, S.W., 2020. Polygenic risk score contribution to psychosis prediction in a target population of persons at clinical high risk. Am. J. Psychiatry 177, 155–163. https://doi.org/10.1176/appi.ajp.2019.18060721.

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

Robertson, S., 2004. Understanding inverse document frequency: on theoretical arguments for IDF. J. Doc. 60, 503–520. https://doi.org/10.1108/00220410410560582.

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, 951–962. https://doi.org/10.1093/schbul/sbu129.

Sabb, F.W., van Erp, T.G., Hardt, M.E., Dapretto, M., Caplan, R., Cannon, T.D., Bearden, C.E., 2010. Language network dysfunction as a predictor of outcome in youth at clinical high risk for psychosis. Schizophr. Res. 116, 173–183. https://doi.org/10.1016/j.schres.2009.09.042.

Santorini, B., 1990. Part-of-speech Tagging Guidelines for the Penn Treebank Project (3rd Revision). 3rd Revis. Univ. Pennsylvania https://doi.org/10.1017/CBO9781107415324.004 2nd Print.

Segal, Y., Fuchs, T.S., Keshet, J., 2019. SpeechYOLO: Detection and Localization of Speech Objects (arXiv preprint arXiv:1904.07704).

Seidman, L., Giuliano, A., Meyer, E.C., Addington, J., Cadenhead, K.S., Cannon, T.D., McGlashan, T.H., Perkins, D., Tsuang, M.T., Walker, E.F., Woods, S.W., Bearden, C.E., Christensen, B.K., Hawkins, K., Heaton, R., Keefe, R.S.E., Heinssen, R., Cornblatt, B.A., 2010. Neuropsychology of the prodrome to psychosis in the NAPLS consortium. Arch. Gen. Psychiatry 67, 578–588. https://doi.org/10.1001/archgenpsychiatry.2010.66.Neuropsychology.

Shaikh, M., Valmaggia, L., Broome, M.R., Dutt, A., Lappin, J., Day, F., Woolley, J., Tabraham, P., Walshe, M., Johns, L., Fisar-Poli, P., Howes, O., Murray, R.M., McGuire, P., Bramon, E., 2012. Reduced mismatch negativity predates the onset of psychosis. Schizophr. Res. 134, 42–48. https://doi.org/10.1016/j.schres.2011.09.022.

Shrem, Y., Goldrick, M., Keshet, J., 2019. Dr. VOT: Measuring Positive and Negative Voice Onset Time in the Wild (arXiv preprint arXiv:1910.13255).

Sichlinger, L., Cibelli, E., Goldrick, M., Mittal, V.A., 2019. Clinical correlates of aberrant conversational turn-taking in youth at clinical high-risk for psychosis. Schizophr. Res. 204, 419–420. https://doi.org/10.1016/j.schres.2018.08.009.

Sigman, M., Cecchi, G.A., 2002. Global organization of Wordnet lexicon. Proc. Natl. Acad. Sci. 99, 1742–1747. https://doi.org/10.1073/pnas.022341799.

Silbert, L.J., Honey, C.J., Simony, E., Poeppel, D., Hasson, U., 2014. Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. Proc. Natl. Acad. Sci. U. S. A. 111, E4687–E4696. https://doi.org/10.1073/pnas.1323812111.

Solovay, M.R., Shenton, M.E., Gasperetti, C., Coleman, M., Kestnbaum, E., Carpenter, J.T., Holtzman, P.S., 1986. Scoring manual for the thought disorder index. Schizophr. Bull. 12, 483–496. https://doi.org/10.1093/schbul/12.3.483.

Stanislawski, E., Bilgrami, Z., Sarac, C., Cecchi, G., Corcoran, C., 2019. S19. Analyzing negative symptoms and language in youths at risk for psychosis using automated language analysis. Schizophr. Bull. 45, S312–S313. https://doi.org/10.1093/schbul/sbz020.564.

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

Uhlhaas, P.J., Phillips, W.A., Mitchell, G., Silverstein, S.M., 2006. Perceptual grouping in disorganized schizophrenia. Psychiatry Res. 145, 105–117. https://doi.org/10.1016/j.psychres.2005.10.016 (2006).

van Rooijen, G., Isvoranu, A.M., Meijer, C.J., van Borkulo, C.D., Ruhé, H.G., de Haan, L., 2017. A symptom network structure of the psychosis spectrum. Schizophr. Res. 189, 75–83.

van Tricht, M.J., Hieman, D.H., Koelman, J.T., Mensink, A.J., Bour, L.J., van der Meer, J.N., van Amelsvoort, T.A., Linszen, D.H., de Haan, L., 2015. Sensory gating in subjects at ultra high risk for developing a psychosis before and after a first psychotic episode. World J. Biol. Psychiatry. 16, 12–21. https://doi.org/10.3109/15622975.2012.680911.

Yalincetin, B., Bora, E., Binbay, T., Ulas, H., Akdede, B.B., Alptekin, K., 2016. Formal thought disorder in schizophrenia and bipolar disorder: a systematic review and meta-analysis. Schizophr. Res. 185, 2–8. https://doi.org/10.1016/j.schres.2016.12.015.

Yung, A.R., Yuen, H.P., McGorry, P.D., Phillips, L.J., Kelly, D., Dell'Olio, M., Francey, S.M., Cosgrave, E.M., Killackey, E., Stanford, C., Godfrey, K., Buckby, J., 2005. Mapping the onset of psychosis: the comprehensive assessment of at-risk mental states. Aust. N. Z. J. Psychiatry. 39, 964–971. https://doi.org/10.1111/j.1440-1614.2005.01714.x.

Please cite this article as: C.M. Corcoran, V.A. Mittal, C.E. Bearden, et al., Language as a biomarker for psychosis: A natural language processing approach, Schizophrenia Research, https://doi.org/10.1016/j.schres.2020.04.032