RESEARCH REPORT

Prediction of psychosis across protocols and risk cohorts using automated language analysis

Cheryl M. Corcoran[1,2], Facundo Carrillo[3,4], Diego Fernández-Slezak[3,4], Gillinder Bedi[2,5,6], Casimir Klim[2,5], Daniel C. Javitt[2,5], Carrie E. Bearden[7], Guillermo A. Cecchi[8]

[1]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA; [2]New York State Psychiatric Institute, New York, NY, USA; [3]Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina; [4]Instituto de Investigación en Ciencias de la Computación, Universidad de Buenos Aires, Buenos Aires, Argentina; [5]Department of Psychiatry, Columbia University Medical Center, New York, NY, USA; [6]Centre for Youth Mental Health, University of Melbourne, and Orygen National Centre of Excellence in Youth Mental Health, Melbourne, Australia; [7]Department of Psychiatry and Biobehavioral Sciences and Psychology, University of California Los Angeles; Semel Institute for Neuroscience and Human Behavior, Los Angeles, CA, USA; [8]Computational Biology Center - Neuroscience, IBM T.J. Watson Research Center, Ossining, NY, USA

Language and speech are the primary source of data for psychiatrists to diagnose and treat mental disorders. In psychosis, the very structure of language can be disturbed, including semantic coherence (e.g., derailment and tangentiality) and syntactic complexity (e.g., concreteness). Subtle disturbances in language are evident in schizophrenia even prior to first psychosis onset, during prodromal stages. Using computer-based natural language processing analyses, we previously showed that, among English-speaking clinical (e.g., ultra) high-risk youths, baseline reduction in semantic coherence (the flow of meaning in speech) and in syntactic complexity could predict subsequent psychosis onset with high accuracy. Herein, we aimed to cross-validate these automated linguistic analytic methods in a second larger risk cohort, also English-speaking, and to discriminate speech in psychosis from normal speech. We identified an automated machine-learning speech classifier – comprising decreased semantic coherence, greater variance in that coherence, and reduced usage of possessive pronouns – that had an 83% accuracy in predicting psychosis onset (intra-protocol), a cross-validated accuracy of 79% of psychosis onset prediction in the original risk cohort (cross-protocol), and a 72% accuracy in discriminating the speech of recent-onset psychosis patients from that of healthy individuals. The classifier was highly correlated with previously identified manual linguistic predictors. Our findings support the utility and validity of automated natural language processing methods to characterize disturbances in semantics and syntax across stages of psychotic disorder. The next steps will be to apply these methods in larger risk cohorts to further test reproducibility, also in languages other than English, and identify sources of variability. This technology has the potential to improve prediction of psychosis outcome among at-risk youths and identify linguistic targets for remediation and preventive intervention. More broadly, automated linguistic analysis can be a powerful tool for diagnosis and treatment across neuropsychiatry.

Key words: Automated language analysis, prediction of psychosis, semantic coherence, syntactic complexity, high-risk youths, machine learning

(World Psychiatry 2018;17:67–75)

Language offers a privileged view into the mind: it is the basis by which we infer others' thought processes, such that disorganized language is considered to reflect disorder in thought. Language disturbance is prevalent in schizophrenia and is related to functional disability, given that an individual needs to think and speak clearly in order to maintain friends and a job[1]. In schizophrenia, the speaker "violates the syntactical and semantic conventions which govern language usage", yielding reduction in syntactic complexity (concrete speech, poverty of content) and loss of semantic coherence, e.g. the disruption in flow of meaning in language (derailment, tangentiality)[2]. This language disturbance is an early core feature of schizophrenia, evident in subtle form prior to initial psychosis onset, in cohorts of both familial[3] and clinical[4-7] high-risk youths, as assessed using clinical ratings.

Beyond clinical ratings, there has been an effort to characterize early subtle language disturbances in clinical high-risk (CHR) individuals using linguistic analysis, with the aim of improving prediction. Bearden et al[8] applied manually coded linguistic analyses to brief speech transcripts in a CHR cohort, finding that both semantic features (illogical thinking) and reduction in syntactic complexity (poverty of speech) predicted psychosis onset with an accuracy of 71%, as compared with 35% accuracy for clinical ratings. Psychosis onset was also predicted by reduced referential cohesion, such that the use of pronouns and comparatives ("this" or "that") frequently did not clearly indicate who or what was previously described.

While this manual linguistic approach appears to be superior to clinical ratings in psychosis prediction, it depends on predefined measures that may not capture other subtle language features. Therefore, we have used automated natural language processing methods to analyze speech in CHR cohorts. These are probabilistic linguistic analyses based on the computer's acquisition of vocabulary (semantics) and learning of grammar (syntax) through machine-learning algorithms trained on very large bodies of text, enabled by exponential increases in computing power, and the flood of text that arrived with the Internet.

For semantics, a common approach is latent semantic analysis, in which a word's meaning is learned based on its co-occurrence with other words, inspired by theories of vocabulary acquisition[9,10]. In this analysis, each word is assigned a multi-dimensional semantic vector, such that the cosine between word-vectors represents the semantic similarity between words. Grouping of successive word-vectors can be used to estimate the semantic coherence of a narrative.

Latent semantic analysis has been applied to speech in schizophrenia, finding an association of decreased semantic coherence with clinical ratings of thought disorder and functional impairment, and with abnormal task-related activation in language circuits[11,12].

Table 1 Demographic features of the two samples

	UCLA site				NYC site	
	CHR+ (N=19)	CHR– (N=40)	CTR (N=21)	FEP (N=16)	CHR+ (N=5)	CHR– (N=29)
Age at baseline (years, mean±SD)	17.3 ± 3.7	16.4 ± 3.0	18.0 ± 2.8	15.8 ± 1.7[a]	22.2 ± 3.4	21.2 ± 3.6
Gender (% male)	89.5	55.0[b]	61.9[b]	68.7	80.0	65.5
Ethnicity (% Caucasian)	63.1	50.0	66.7	62.5	40.0	37.9
Parental socio-economic status (Hollingshead index, mean±SD)	4.4 ± 2.1[a]	4.4 ± 1.7[a]	5.7 ± 1.4	4.9 ± 1.8	NA	NA

Significant differences at $p<0.05$ level: [a]vs. CTR, [b]vs. CHR+

UCLA – University of California Los Angeles, NYC – New York City, CHR+ – clinical high-risk subjects who converted to psychosis during follow-up, CHR– – clinical high-risk subjects who did not convert to psychosis during follow-up, CTR – healthy controls, FEP – subjects with first-episode psychosis, NA – not available

For syntax, part-of-speech tagging is used to determine sentence length and rates of usage of different parts of speech[13,14].

In an earlier proof-of-principle study in a narrative-based protocol with a small CHR cohort, we used both latent semantic analysis and part-of-speech tagging, with machine learning, to identify a classifier of psychosis that comprised minimum semantic coherence, shortened sentence length, and a decrease in the use of determiner pronouns (e.g., "that" or "which") to introduce dependent clauses[15]. These three features were correlated with but outperformed clinical ratings in prediction of psychosis.

In the present study, we applied the same automated natural language processing approach with machine learning, including latent semantic analysis and part-of-speech tagging, to the larger CHR prompt-based protocol speech dataset that Bearden et al previously analyzed using manually coded linguistic methods[8].

We hypothesized that a classifier trained with the larger prompt-based protocol dataset[8] would be highly accurate (~80%) in predicting psychosis onset when tested intra-protocol as well as when retested in the narrative-based protocol[15] (cross-protocol). We also hypothesized that the automated and manual linguistic features derived from the training dataset would be correlated with one another.

We further tested the ability of the classifier to discriminate speech in adolescents with recent-onset psychosis from normal speech, as a putative early illness marker.

METHODS

Participants

Participants at the University of California Los Angeles (UCLA) site included 59 CHR individuals. They were defined by meeting criteria for one of three prodromal syndrome categories, as assessed by the Structured Interview for Prodromal Syndromes/Scale of Prodromal Symptoms (SIPS/SOPS)[16]: a) attenuated positive symptoms, b) brief intermittent psychotic symptoms, or c) a substantial drop in social/role functioning in conjunction with a schizotypal personality disorder diagnosis or a first-degree relative with a psychotic disorder. Of these subjects, 19 developed a psychotic disorder within two years ("converters", CHR+) and 40 did not (CHR–). Transition to psychosis was determined using the SIPS/SOPS "presence of psychosis" criteria. Transcripts from UCLA were also available for 16 recent-onset psychosis patients and 21 healthy individuals similar in demographics, recruited from local schools and the community.

Participants at the New York City (NYC) site included 34 CHR individuals, defined by meeting the above SIPS/SOPS criteria. Of these subjects, five developed psychosis within 2.5 years (CHR+) according to SIPS/SOPS criteria, and 29 did not (CHR–).

The demographic features of the two samples are presented in Table 1. The institutional review boards at New York State Psychiatric Institute/Columbia University and UCLA approved the study, and informed consent was obtained from all participants (parental consent with assent for minors).

Speech assay

UCLA (prompt-based protocol dataset)

Speech was elicited using Caplan's "Story Game", in which participants retell and then answer questions about a story they hear ("what do you like about it?"; "is it true?"), and then construct and tell a new story[17]. Speech samples were transcribed and de-identified, which means that proper nouns such as names were substituted.

Manual linguistic analyses included administration of the Kiddie Formal Thought Disorder Rating Scale (K-FTDS) and the Caplan modification of the Halliday and Hassan approach to analysis of cohesion[17]. The K-FTDS scores included frequency counts of illogical thinking, loose associations, and poverty of content. Cohesion categories included referential (pronomial, demonstrative and comparative – "this", "that"), conjunction

("and", "but", "because") and unclear/ambiguous[17]. This dataset was used to analyze intra-protocol prediction accuracy.

NYC (narrative-based protocol dataset)

Open-ended narrative interviews of about one hour were obtained by interviewers trained by an expert in qualitative research methods. Prompts queried impact of life changes experienced, and expectations for the future[18]. This dataset was used to study cross-protocol prediction accuracy.

Speech analyses

Speech pre-processing

The speech transcripts were pre-processed and prepared for computer-based analyses. We used the Natural Language Toolkit, which is an open source program available on the Internet (NLTK; http://www.nltk.org). First, punctuation (e.g., commas, periods) was discarded, words were tokenized (identified as parts of speech), and then each transcript was parsed into phrases, using rules of grammar in English. Words were then converted to the roots from which they are inflected, or lemmatized, using the NLTK WordNet lemmatizer.

The resulting pre-processed speech data yielded for each transcript a series of lemmatized words, maintaining the original order in which they were spoken, without punctuation and in lower case.

Latent semantic analysis

Latent semantic analysis[9,10] was used to convert each transcript from a series of words into a series of semantic vectors, maintaining the original order of the transcribed text. In this analysis, a high-dimensional semantic vector is assigned to each word in the lexicon based on its co-occurrence with other words in a very large corpus of text, specifically the Touchstone Applied Science Associates (TASA) corpus, a collection of educational materials.

Automated analysis provides a construction of meaning in language that resembles what the human mind does, i.e. to learn the meaning of words in terms of prior experience with those words in different contexts. The computer "learns" the meaning of words computationally, by scanning a very large corpus of text and determining the frequency of co-occurrence of each word with every other word in the lexicon. Words that co-occur more frequently are considered to have greater semantic similarity (e.g., "cat"/"dog" vs. "cat"/"pencil"), and the direction of their vectors will be more aligned. Aggregates of words (e.g., sentences) have semantic vectors that are the sum of semantic vectors for all the words they contain. Semantic coherence between words, or between aggregates (e.g., successive sentences), can be indexed by calculating the cosine between successive semantic vectors (from −1.0 for incoherence to 1.0 for coherence).

As the narrative-based protocol in NYC was open-ended, yielding mean uninterrupted responses of 130 words for CHR– and 182 words for CHR+, there had been sufficient free speech for analysis of semantic coherence at the sentence level in our prior study[15]. However, the prompt-based study at UCLA[8] led to much briefer responses (mean uninterrupted response <20 words; insufficient number of sentences for analysis), such that a k-level measure of semantic coherence was used instead, which computes word-to-word variability at "k" inter-word distances, with k ranging from 5 to 8[19]. As in our prior study[15], we calculated typical statistical measures for each of the k-level measures of coherence, such as mean, standard deviation, minimum, maximum, and 90th percentile (less sensitive to outliers than the maximum), also "normalized" or adjusted for sentence length.

Part-of-speech tagging analyses

Just as each word in every transcript was assigned a semantic vector, each word was also tagged in respect to its grammatical function, using the POS-Tag procedures in the open-access Natural Language Toolkit (www.nltk.org) in reference to a hand-tagged corpus called the Penn Treebank[13]. For example, the sentence "The dog is near the fence" would be tagged as ("The", "DT"), ("dog", "NN"), ("is", "VBZ"), ("near", "IN"), ("the", "DT"), ("fence", "NN"), where DT is the tag for determiners, NN for nouns, VBZ for verbs, and IN for prepositions.

The Penn Treebank has thirty-six part-of-speech tags, which include types of nouns, verbs, adjectives, adverbs, determiners, prepositions and pronouns. For each transcript, we calculated the frequency of use for each grammatical function.

Machine learning classification

The machine learning algorithm classifies speech by whether it is characteristic of individuals who will develop psychosis, as opposed to those who will not. It does this by learning the underlying patterns in a subset of transcripts and then in an iterative fashion, predicting the classification (psychosis or no) in new transcripts not used during the learning phase.

The machine learning analysis was circumscribed to the eleven speech variables that were significantly different between CHR+ and CHR– in the UCLA cohort (nine semantic coherence features and two syntactic elements – frequencies of comparative adjectives and possessive pronouns), plus three variables that predicted psychosis in our prior study[15], including WH-family ("which", "what", "whom") determiners, pronouns and adjectives. The list of these fourteen features used for analyses is provided in Table 2. Each transcript had a vector comprised of these fourteen variables.

We then performed singular value decomposition (which is a type of factor analysis based on linear algebra) on the fourteen features in these transcript vectors, adding the UCLA healthy control sample data to have a better understanding of the intrinsic structure of the speech data. We chose the top four factors that best discriminated transcripts from CHR+ vs. CHR–. A logistic regression model was then trained on the four

factors to classify CHR+ vs. CHR–, using an iteration of learning on a subset and prediction in left-out samples.

Cross-site validation

The same fourteen features were extracted from the NYC data, and aligned to the UCLA features using a simple global coordinate "Procrustean" transformation[20,21], similar to spatial registration in brain imaging[22], that includes scaling (in size), rotation and translation in Euclidean space. This minimized the difference in covariance of the two datasets, while maintaining the relative position among data points.

We further implemented a convex hull embedding method used in our prior study[15] to create a three-dimensional space (the top three factors) to model the accuracy of the classifier derived from the UCLA cohort in discriminating CHR+ from CHR– in the transformed NYC cohort. A convex hull of a set of points is the minimal convex polyhedron that contains them.

Correlations of text features with demographics, clinical ratings and manual features

We tested whether the fourteen identified text features were associated with age, gender, ethnicity (Caucasian/non-Caucasian) and parental socio-economic status[23]. We then assessed whether these text features were correlated with clinical ratings or with the three manually-coded linguistic measures (illogical thought, poverty of content and referential cohesion) that predicted psychosis onset in the UCLA cohort in the earlier study[8]. We calculated the canonical correlation between automated and manual text variables, which is the correlation between two sets of variables obtained from the same individuals.

Utility of the classifier in discriminating psychosis from normal speech

As an independent validation, we determined the accuracy of the CHR speech classifier in discriminating speech from the 21 healthy volunteers and 16 recent-onset psychosis patients ascertained at UCLA, who were also administered the same prompt-based protocol to elicit speech samples. The idea was that healthy controls should have a speech similar to that of CHR–, while recent-onset psychosis patients should have a speech similar to CHR+.

RESULTS

Machine learning classification

Of the four factors in the machine learning classifier, the first three highlighted semantic features, respectively weighted for maximum semantic coherence, variance in semantic coherence, and minimum semantic coherence, while the fourth factor was weighted for frequency of use of possessive pronouns (Figure 1).

Table 2 Syntactic and semantic features used for predictive modeling

Description	Example
a. Adjective, comparative	"braver", "closer", "cuter"
b. Possessive pronoun	"her", "his", "mine", "my", "our", "ours", "their", "your"
c. WH-determiner	"that", "which", "what"
d. WH-pronoun	"that", "what", "which", "who", "whom"
e. WH-adverb	"how", "however", "whenever", "why"
f. Minimum coherence at 5-level, normalized	
g. Minimum coherence at 5-level	
h. 90th percentile coherence at 5-level	
i. Maximum coherence at 6-level	
j. Mean coherence at 7-level, normalized	
k. Standard deviation coherence at 7-level, normalized	
l. 90th percentile at 7-level	
m. Standard deviation coherence at 7-level	
n. 90th percentile at 8-level	

A k-level measure of semantic coherence was used, which computes word-to-word variability at "k" inter-word distances, with k ranging from 5 to 8

The accuracy of the ensemble of these four factors in classifying psychosis outcome in the UCLA cohort was 83% using the logistic regression classifier. The post-hoc analysis yielded an area under the curve (AUC) of 0.87 in the receiver operating characteristic (ROC) curve (Figure 2).

So, a classifier comprising decreased semantic coherence, greater variance in that coherence, and reduced usage of possessive pronouns ("her", "his", "mine", "my", "our", "ours", "their", "your") was highly accurate in predicting subsequent psychosis onset.

Cross-site validation

When this UCLA machine-learning classifier was applied to the original NYC speech data, after Procrustean transformation[20,21,24], it significantly discriminated CHR with respect to psychosis onset ($p<0.05$ upon label randomization), with a true negative ratio of 0.82 (24/29) and a true positive ratio of 0.60 (3/5), that is, an overall accuracy of 0.79. With logistic regression, the UCLA classifier yielded an AUC of 0.72 for the transformed NYC cohort speech data (Figure 2).

In order to compare with our previous study[15], we created a three-dimensional projection of data using the top three factors

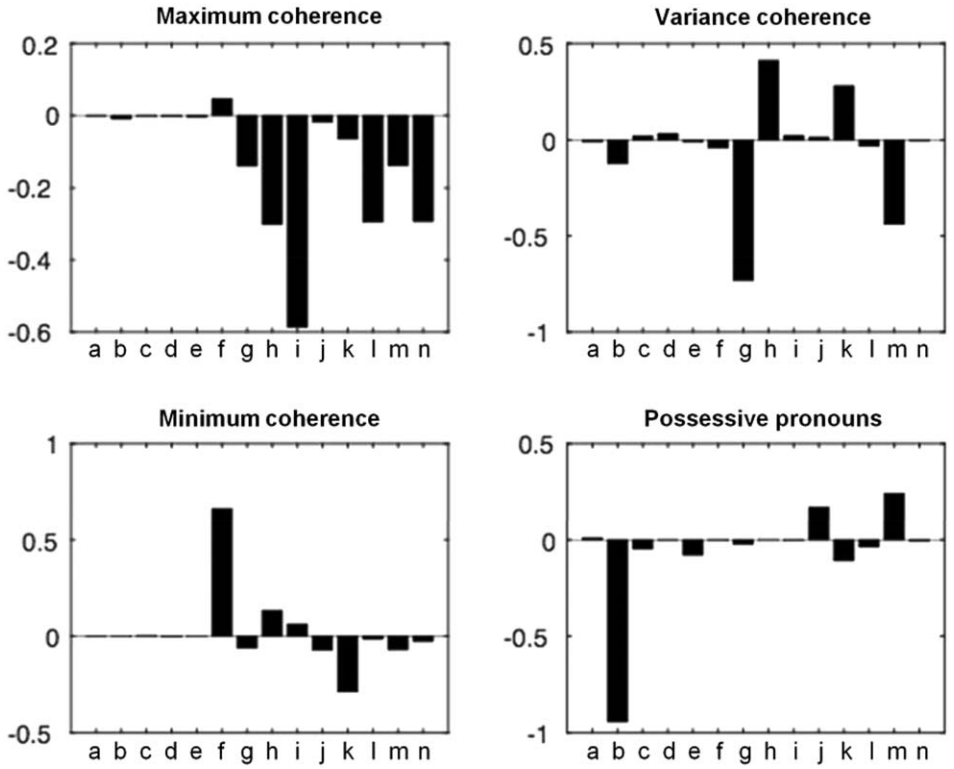

Figure 1 The four-factor University of California Los Angeles (UCLA) machine learning classifier of psychosis outcome. Factors are aggregates of weighted syntactic (a-e) and semantic coherence (f-n) features, as listed in Table 2. The first three factors are weighted toward semantic features (maximum, variance and minimum), and the fourth factor is weighted toward a syntactic feature (possessive pronouns). Y axes show factor weights.

identified from the UCLA CHR speech dataset. This yielded convex hulls that excluded 11 of 19 CHR+ in the UCLA cohort (i.e., 8/19 false negatives) (Figure 3A), indicating that the logistic regression classifier (with all four factors) was more accurate. Using the same three factors from the UCLA classifier, the convex hull of CHR– in NYC excluded three of five CHR+ (Figure 3B). Of note, there was substantial overlap in the convex hulls of CHR– individuals for both the UCLA and NYC speech datasets (Figure 3C).

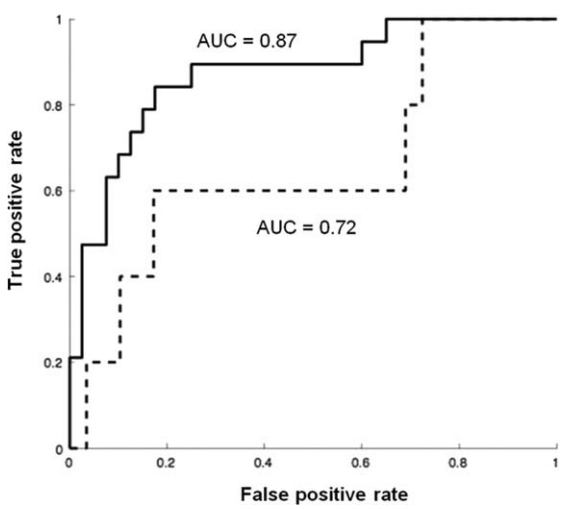

Figure 2 Receiver operating characteristics (ROC) for the University of California Los Angeles (UCLA) clinical high-risk (CHR) classifier of psychosis outcome as applied to the UCLA dataset (solid line) and to the realigned New York City (NYC) dataset (dotted line). AUC – area under the curve.

Correlations with demographics, clinical ratings and manual linguistic features

Among demographic features, age was significantly associated with three of the semantic coherence variables, specifically the 90% order variables for 5-level ($p=0.002$), 7-level ($p=0.01$) and 8-level ($p=0.004$), suggesting increasing semantic coherence with age. By contrast, there were no associations of automated text variables with gender, ethnicity, or parental socioeconomic status[23].

There was no significant association between automated analysis text features and SIPS/SOPS clinical ratings (total positive and total negative). However, the canonical correlation between the fourteen text features identified here, and the three manual linguistic features (illogical thought content, poverty of content and referential cohesion) that predicted psychosis onset in the earlier study[8], was large and highly significant, with $r=0.71$, $p<10^{-6}$.

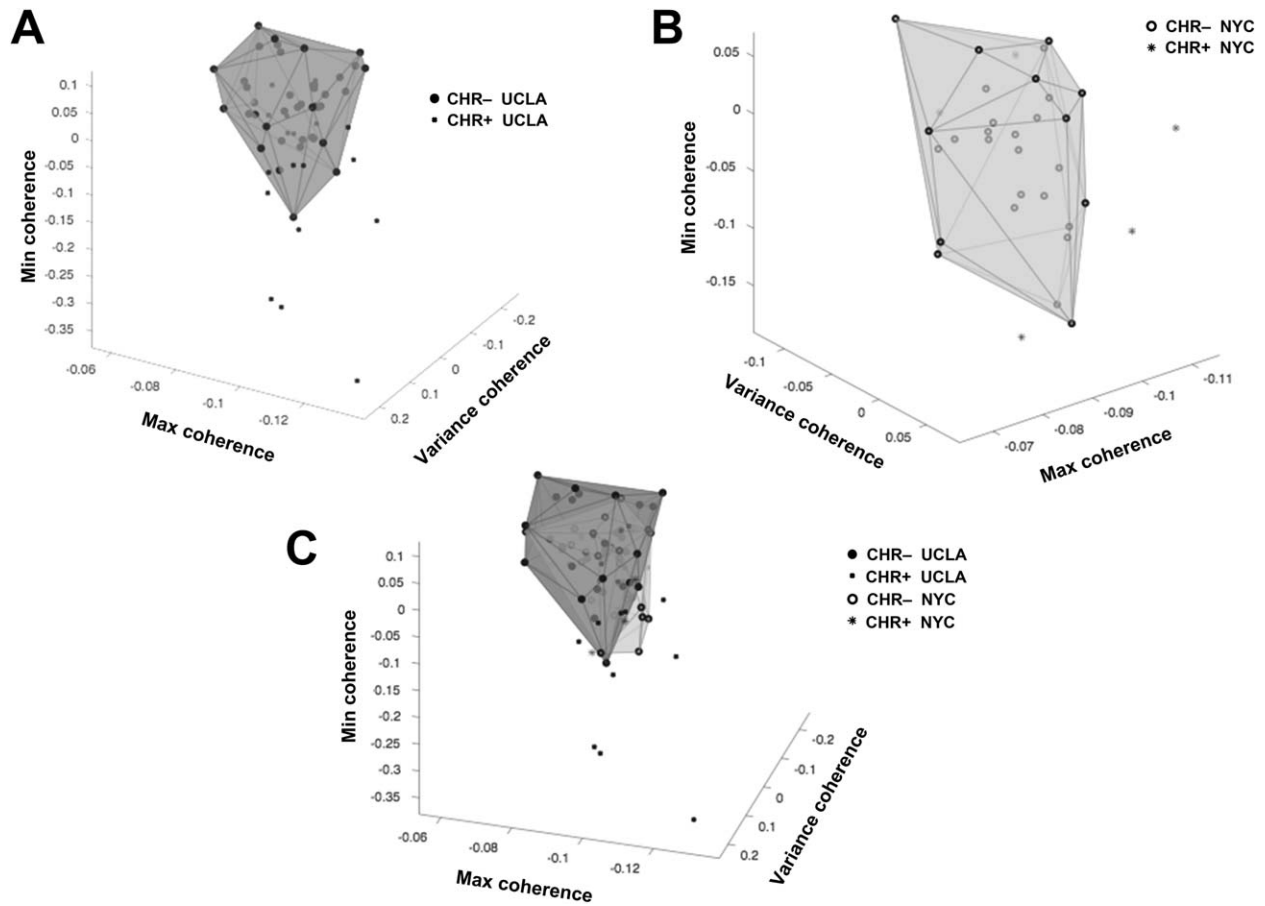

Figure 3 Projection of the top three factors for the University of California Los Angeles (UCLA) and New York City (NYC) clinical high-risk (CHR) cohorts. These factors were weighted for semantic coherence features. A. Convex hull of non-converters (CHR–) in UCLA, with 11 of 19 converters (CHR+) outside of the hull. B. Convex hull of CHR– in NYC, with 3 of 5 CHR+ outside the hull. C. Data in A and B (all CHR) shown together to demonstrate extent of overlap in language properties.

Utility of the classifier in discriminating psychosis from normal speech

A 72% accuracy was obtained with the logistic regression classifier when applied to the speech dataset of healthy controls and recent-onset psychosis patients at UCLA.

Singular value decomposition three-factor representation excluded 11 of 16 recent-onset psychosis patients from the convex hull defined by the data points of healthy volunteers, yielding a true positive rate of 0.69 (Figure 4A). There was spatial overlap between the convex hulls that contained healthy controls and CHR– individuals (Figure 4B).

DISCUSSION

Using automated natural language processing methods with machine learning to analyze speech in a CHR cohort, we generated a classifier comprising decreased semantic coherence, greater variance in that coherence, and reduced usage of possessive pronouns which was highly accurate in predicting subsequent psychosis onset.

This classifier had an intra-protocol accuracy of 83% in the training dataset, and a cross-protocol accuracy of 79% when applied to transcripts from a second independent CHR cohort (test dataset)[15], demonstrating significant transfer of predictability, despite disparate methods of speech elicitation[8,15]. Further, this same classifier discriminated the speech of recent-onset psychosis patients from that of healthy individuals with 72% accuracy, suggesting that its discriminatory power was relatively robust across illness stages, as has been found for clinical ratings of thought disorder[1,6]. Finally, the predictive automated and manual linguistic features were highly correlated in the cohort, providing evidence of concurrent validity.

It has long been observed that language in schizophrenia is characterized by a disturbance in semantic coherence, with Kraepelin describing *Sprachverwirrtheit* (e.g., confused speech)[25], and Bleuler highlighting a "loosening of associations" in language as a primary feature of schizophrenia[26]. Later, Andreasen operationalized decreased semantic coherence as positive thought dis-

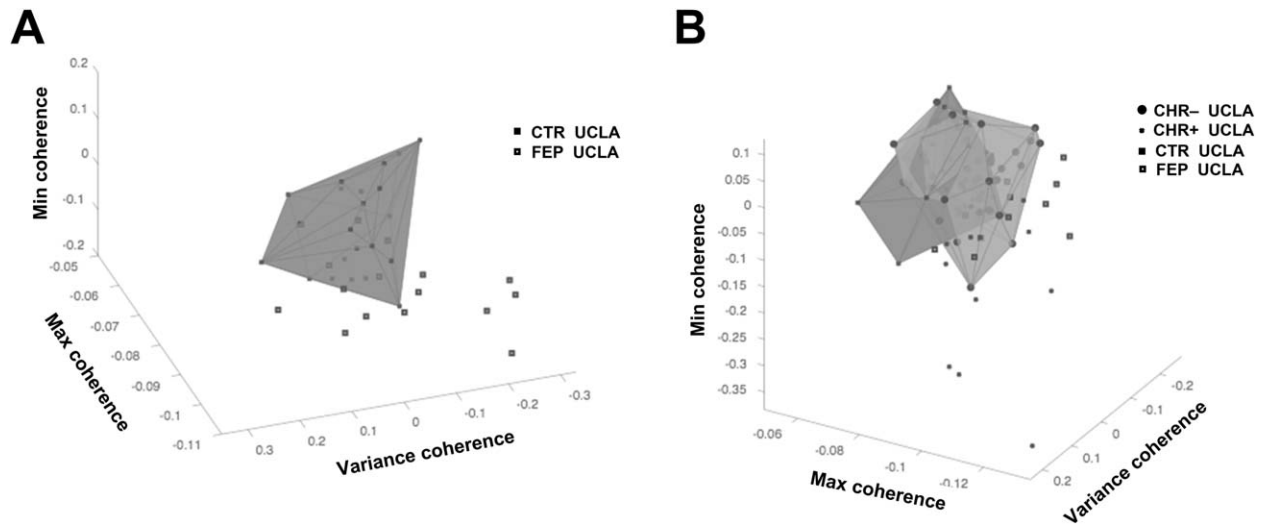

Figure 4 Projection of the top three factors for University of California Los Angeles (UCLA) first-episode psychosis (FEP) patients and healthy controls (CTR). A. Convex hull of healthy controls (CTR) with 11 of 16 FEP patients outside the hull. B. Overlap of convex hulls for FEP vs. CTR, and converters (CHR+) vs. non-converters (CHR–).

order[27]. Hoffman applied manual discourse analysis to transcribed speech from schizophrenia patients, finding a reduction in semantic coherence[28], a finding replicated later using computer-assisted discourse analysis[29].

It has only been in the last decade that natural language processing linguistic corpus-based analyses, specifically latent semantic analysis, have been applied to language production in schizophrenia, finding decreases in semantic coherence that correlate with clinical ratings, functional impairment, and task-related activation in language circuits[11,12]. Now, in the two CHR studies to date, latent semantic analysis with machine learning has shown decreased semantic coherence to predict subsequent psychosis onset.

Disturbance in syntax is also well-documented in schizophrenia. Errors of pronomial reference in schizophrenia speech were described three decades ago by Hoffman[30], a finding since replicated by other investigators using word classification/count strategies[29,31]. In the present study, using part-of-speech tagging, we identified decreased use of possessive pronouns as prognostic for psychosis onset, accounting for most of the weight of the fourth factor in the classifier. This is consistent with prior manual linguistic analysis in this same cohort, which identified decreased referential cohesion as predictive of psychosis[8], such that the use of pronouns and comparatives ("this" or "that") frequently did not clearly indicate who or what was previously described.

More commonly found in schizophrenia speech is a reduction in syntactic complexity[27,32], typically operationalized as shorter sentences, and most evident when open-ended narrative is elicited[12,30,31,33]. In our prior small natural language processing study[15], we found two measures of syntactic complexity – shorter sentences and reduced use of determiner pronouns that introduce dependent clauses – to be both predictive of psychosis and highly correlated with negative symptoms. In the present study, the failure of sentence length to predict psychosis in the training dataset may be a consequence of the brief and structured responses that were elicited (<20 mean words per response)[12], as compared with prior studies (>120 mean words/response[15], ~800 words/response[12] and >10 sentences/response[30]).

In both of our CHR studies, we have created convex hull classifications in which speech datapoints for non-converters (CHR–) were inside the hull, while those with emergent psychosis (CHR+) were outside. A similar convex hull was generated for healthy controls using the CHR classifier, with recent-onset psychosis patients largely outside the hull. Together, these findings suggest that pre-psychotic and psychotic language is deviant from a constrained hull of relatively normal language in respect to semantics and syntax.

As yet, this normal pattern of language, as characterized by automated natural language processing methods, remains poorly understood, including in a developmental context, as both semantic and syntactic complexity increase in adolescence and young adulthood[34]. Of note, the premise that processes underlying normal language production and comprehension are relatively homogeneous is supported by a body of work by Hasson, showing alignment of brain activation time courses across normal individuals (intersubject coherence) during both listening and speaking[35].

Our finding of strong correlations between automated and manual linguistic variables provides evidence of concurrent validity for the natural language processing approach. Automated natural language processing methods are far more rapid and less expensive than manual linguistic approaches, and

can be more readily adapted for research and ultimately in the clinic.

Beyond language semantic analysis and part-of-speech tagging, speech and language can also be evaluated in respect to speech graphs[36], prosody, pragmatics, metaphoricity[37], and for discourse or conversations among interlocutors. Automated natural language processing analyses have also been used to characterize other disturbances in behavior, including intoxication from drugs of abuse[38] and Parkinson's disease[39], such that this technology holds promise for medicine more broadly. Finally, automated approaches can be extended to other behavior, such as facial expressions of emotion[40]. Overall, automated speech analysis is a powerful but inexpensive technology that can be used in psychiatry for diagnosis, prognosis and estimates of treatment response.

The main limitations in the present study include sample size, and remaining gaps in our knowledge in respect to what is normal across development for automated linguistic variables, and how normal and deviant language can be mapped to underlying neural circuits. Further, different methods of speech elicitation were used in the two cohorts, such that sentence-level coherence could not be estimated for the training dataset due to brevity of responses, requiring the use of "k-level" methods to characterize semantic coherence, and an alignment transformation of data for cross-protocol validation. In ongoing studies, we are using open-ended interviews to elicit free natural speech for analysis, so that we can measure semantic coherence at the sentence level, and better capture measures of syntactic complexity.

Overall, we demonstrate the utility and validity of using automated natural language processing methods to characterize subtle disturbances in semantics and syntax across stages of psychotic disorder. This technology has the potential to improve prediction of psychosis outcome among adolescents and young adults at clinical high risk, and may have broader implications for medical research and practice at large.

ACKNOWLEDGEMENTS

This research was supported by the US National Institute of Mental Health (R01 MH 107558; R03 MH 108933 02), the New York State Office of Mental Health, and a NARSAD/BBRF Young Investigator Award and Miller Family Term Chair to C.E. Bearden. These funding sources had no role in the design and conduct of the study; collection, management, analysis and interpretation of the data; preparation, review or approval of the manuscript; and the decision to submit the manuscript for publication.

REFERENCES

1. Roche E, Creed L, MacMahon D et al. The epidemiology and associated phenomenology of formal thought disorder: a systematic review. Schizophr Bull 2015;41:951-62.
2. Andreasen NC, Grove WM. Thought, language, and communication in schizophrenia: diagnosis and prognosis. Schizophr Bull 1986;12:348-59.
3. Gooding DC, Ott SL, Roberts SA et al. Thought disorder in mid-childhood as a predictor of adulthood diagnostic outcome: findings from the New York High-Risk Project. Psychol Med 2013;43:1003-12.
4. Nelson B, Yuen HP, Wood SJ et al. Long-term follow-up of a group at ultra high risk ("prodromal") for psychosis: the PACE 400 study. JAMA Psychiatry 2013;70:793-802.
5. Addington J, Liu L, Buchy L et al. North American Prodrome Longitudinal Study (NAPLS 2): the prodromal symptoms. J Nerv Ment Dis 2015;203:328-35.
6. DeVylder JE, Muchomba FM, Gill KE et al. Symptom trajectories and psychosis onset in a clinical high-risk cohort: the relevance of subthreshold thought disorder. Schizophr Res 2014;159:278-83.
7. Cornblatt BA, Carrion RE, Auther A et al. Psychosis prevention: a modified clinical high risk perspective from the Recognition and Prevention (RAP) program. Am J Psychiatry 2015;172:986-94.
8. Bearden CE, Wu KN, Caplan R et al. Thought disorder and communication deviance as predictors of outcome in youth at clinical high risk for psychosis. J Am Acad Child Adolesc Psychiatry 2011;50:669-80.
9. Landauer TK, Dumais ST. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol Rev 1997;104:211-40.
10. Landauer TK, Foltz PW, Laham D. An introduction to latent semantic analysis. Discourse Process 1998;25:259-84.
11. Elvevag B, Foltz PW, Weinberger DR et al. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. Schizophr Res 2007;93:304-16.
12. Elvevag B, Foltz PW, Rosenstein M et al. An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. J Neurolinguistics 2010;23:270-84.
13. Santorini B. Part-of-speech tagging guidelines for the Penn Treebank Project. 3rd revision. Philadelphia: Department of Computer and Information Science, University of Pennsylvania, 1990.
14. Bird S. Natural language processing and linguistic fieldwork. Comput Linguist 2009;35:469-74.
15. Bedi G, Carrillo F, Cecchi GA et al. Automated analysis of free speech predicts psychosis onset in high-risk youths. NPJ Schizophr 2015;1:15030.
16. Miller TJ, McGlashan TH, Rosen JL et al. Prodromal assessment with the structured interview for prodromal syndromes and the scale of prodromal symptoms: predictive validity, interrater reliability, and training to reliability. Schizophr Bull 2003;29:703-15.
17. Caplan R, Guthrie D, Fish B et al. The Kiddie Formal Thought Disorder Rating Scale: clinical assessment, reliability, and validity. J Am Acad Child Adolesc Psychiatry 1989;28:408-16.
18. Ben-David S, Birnbaum ML, Eilenberg ME et al. The subjective experience of youths at clinically high risk of psychosis: a qualitative study. Psychiatr Serv 2014;65:1499-50.
19. Mandera P, Keuleers E, Brysbaert M. How useful are corpus-based methods for extrapolating psycholinguistic variables? Q J Exp Psychol 2015;68:1623-42.
20. Shönemann P. A generalized solution of the orthogonal procrustes problem. Psychometrika 1966;31:1-10.
21. Haxby JV, Guntupalli JS, Connolly AC et al. A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron 2011;72:404-16.
22. Ashburner J, Firston K. Rigid body registration. In: Penny W, Friston K, Ashburner J et al (eds). Statistical parametric mapping: the analysis of functional brain images. Cambridge: Academic Press, 2007:49-62.
23. Mollica RF, Milic M. Social class and psychiatric practice: a revision of the Hollingshead and Redlich model. Am J Psychiatry 1986;143:12-7.
24. Jorge-Botana G, Olmos R, Luzon JM. Word maturity indices with latent semantic analysis: why, when, and where is Procrustes rotation applied? Wiley Interdiscip Rev Cogn Sci (in press).
25. Kraepelin E. Psychiatrie. Ein Lehrbuch für Studierende und Ärzte. Leipzig: Barth, 1899.
26. Bleuler E. Dementia Praecox oder Gruppe der Schizophrenien. Leipzig: Deuticke, 1911.
27. Andreasen NC. Thought, language, and communication disorders. I. Clinical assessment, definition of terms, and evaluation of their reliability. Arch Gen Psychiatry 1979;36:1315-21.
28. Hoffman RE, Stopek S, Andreasen NC. A comparative study of manic vs schizophrenic speech disorganization. Arch Gen Psychiatry 1986;43:831-8.
29. Noel-Jorand MC, Reinert M, Giudicelli S et al. A new approach to discourse analysis in psychiatry, applied to a schizophrenic patient's speech. Schizophr Res 1997;25:183-98.
30. Hoffman RE, Hogben GL, Smith H et al. Message disruptions during syntactic processing in schizophrenia. J Commun Disord 1985;18:183-202.

31. Buck B, Penn DL. Lexical characteristics of emotional narratives in schizophrenia: relationships with symptoms, functioning, and social cognition. J Nerv Ment Dis 2015;203:702-8.
32. Kuperberg GR. Language in schizophrenia Part 2: What can psycholinguistics bring to the study of schizophrenia... and vice versa? Lang Linguist Compass 2010;4:590-604.
33. Andreasen NC. Thought, language, and communication disorders. II. Diagnostic significance. Arch Gen Psychiatry 1979;36:1325-30.
34. Nippold MA, Ward-Lonergan JM, Fanning JL. Persuasive writing in children, adolescents, and adults: a study of syntactic, semantic, and pragmatic development. Lang Speech Hear Serv Sch 2005;36:125-38.
35. Silbert LJ, Honey CJ, Simony E et al. Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. Proc Natl Acad Sci USA 2014;111:E4687-96.
36. Mota NB, Vasconcelos NA, Lemos N et al. Speech graphs provide a quantitative measure of thought disorder in psychosis. PLoS One 2012;7:e34928.
37. Gutierrez ED, Shuotva E, Marghetis T et al. Literal and metaphorical senses in compositional distributional semantic models. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, 2016:183-93.
38. Bedi G, Cecchi GA, Slezak DF et al. A window into the intoxicated mind? Speech as an index of psychoactive drug effects. Neuropsychopharmacology 2014;39:2340-8.
39. Garcia AM, Carrillo F, Orozco-Arroyave JR et al. How language flows when movements don't: an automated analysis of spontaneous discourse in Parkinson's disease. Brain Lang 2016;162:19-28.
40. Baker JT, Pennant L, Baltrušaitis T et al. Toward expert systems in mental health assessment: a computational approach to the face and voice in dyadic patient-doctor interactions. iproc 2016;2:e44.

DOI:10.1002/wps.20491