Schizophrenia Bulletin vol. 47 no. 2 pp. 344–362, 2021
doi:10.1093/schbul/sbaa141
Advance Access publication 18 November 2020

MAJOR REVIEW

Understanding Language Abnormalities and Associated Clinical Markers in Psychosis: The Promise of Computational Methods

Kasia Hitczenko*[,1], Vijay A. Mittal[2,3,4,5,6], and Matthew Goldrick[1,6]

[1]Department of Linguistics, Northwestern University, Evanston, IL; [2]Department of Psychology, Northwestern University, Evanston, IL; [3]Department of Psychiatry, Northwestern University, Chicago, IL; [4]Institute for Policy Research, Northwestern University, Evanston, IL; [5]Medical Social Sciences, Northwestern University, Chicago, IL; [6]Institute for Innovations in Developmental Sciences, Northwestern University, Evanston and Chicago, IL

*To whom correspondence should be addressed; Northwestern University, 2016 Sheridan Road, Evanston, IL 60208; tel: 847-491-5831, fax: 847-491-3770, e-mail: kasia.hitczenko@northwestern.edu

The language and speech of individuals with psychosis reflect their impairments in cognition and motor processes. These language disturbances can be used to identify individuals with and at high risk for psychosis, as well as help track and predict symptom progression, allowing for early intervention and improved outcomes. However, current methods of language assessment—manual annotations and/or clinical rating scales—are time intensive, expensive, subject to bias, and difficult to administer on a wide scale, limiting this area from reaching its full potential. Computational methods that can automatically perform linguistic analysis have started to be applied to this problem and could drastically improve our ability to use linguistic information clinically. In this article, we first review how these automated, computational methods work and how they have been applied to the field of psychosis. We show that across domains, these methods have captured differences between individuals with psychosis and healthy controls and can classify individuals with high accuracies, demonstrating the promise of these methods. We then consider the obstacles that need to be overcome before these methods can play a significant role in the clinical process and provide suggestions for how the field should address them. In particular, while much of the work thus far has focused on demonstrating the successes of these methods, we argue that a better understanding of when and why these models fail will be crucial toward ensuring these methods reach their potential in the field of psychosis.

Key words: computational linguistics/language production/automated linguistic analysis/thought disorder/speech

Individuals with psychosis have a number of impairments in cognition[1,2] and motor processes.[3–6] Language production—communicating with others through speech, written text, or sign—is a domain that is severely disrupted by these impairments.[7,8] Individuals with psychosis exhibit disorganized speech that can be off topic, drift from the original thought, or be incoherent or difficult to follow.[9,10] Speech by individuals with psychosis can be vague and repetitive, as well as reduced in quantity and syntactic and lexical complexity.[11–15] In addition, individuals with psychosis differ in their vocal characteristics from healthy individuals. For example, they often speak with a flat affect—sometimes producing emotionally intense thoughts in a disconnected way.[16,17] Many of these language disturbances are characteristic symptoms of psychosis, contribute to worse outcomes, and are evident in the early stages of psychosis, even before formal onset.[18–20]

These language disturbances are helpful in identifying individuals at high risk for and with psychosis, allowing for early intervention, as well as for tracking and predicting symptom progression.[12,18,21,22] However, several practical issues limit this area from meeting its full potential. Specifically, language is currently assessed via manual annotation by expert raters and/or clinical rating scales. These data are highly time intensive to gather—making it impractical to use these methods on a wide scale—and rely on rating scales that may be underpowered, making it difficult to pick up on anything but the most extreme versions of these impairments.

Computational methods could drastically improve the ability to use linguistic information clinically, providing a scalable method for using language in a more objective, reliable, and replicable way. In this article, we will first review how these automated, computational methods work and how they have been applied to the field of psychosis, by reviewing preliminary yet promising findings in this area. As will become clear, across many studies, using many different methods, at many different levels of language, computational methods have been shown to capture differences between individuals with psychosis and healthy controls and have been able to categorize speech samples as belonging to either group at rates of 70%–100%.[21,23–30]

However, despite their initial promise, there are substantial hurdles to overcome before computational methods can play a significant role in the clinical process. In the second part of the article, we argue that while much of the work thus far has focused on demonstrating the successes of these methods, critical evaluation of when and why these models fail will be crucial toward ensuring these methods reach their potential in the field of psychosis.

Why Care About Language? Observed Impairments in Language Production in Psychosis

In this section, we review empirical work demonstrating what abnormalities individuals with psychosis exhibit, as well as their clinical and neuropsychological correlates. We focus on 3 main types of disturbances: (1) disorganized speech (positive thought disorder), (2) poverty of speech (negative thought disorder), and (3) flat affect. Lexical abnormalities (eg, increased use of nonwords or word approximations) are also present in psychotic disorders, but as they are less frequent, less understood neuropsychologically, and less studied from a computational perspective (but see Gutiérrez et al[31]), we do not focus on them here. Within each of these 4 categories, table 1 provides definitions and examples of specific subtypes.

Positive Thought Disorder (Disorganized Speech)

Disorganized speech has been found to correlate with other positive symptoms of psychosis, primarily delusions.[54–56] While the underlying causes are not yet fully understood and may vary between individuals,[57] disorganized speech is argued to be related to deficits in semantic memory and abnormal semantic associations between words,[58–61] working memory, attention, and other executive function deficits[62] (but see Bagner et al[63]), and/or failure to incorporate linguistic context (possibly due to executive function deficits).[64–66] Neurally, the severity of disorganized speech is associated with reduced gray matter in the superior temporal and inferior frontal cortices[67] (but see Palaniyappan et al[68]) and abnormal activation in superior temporal cortex[69–71] during both free speech production and semantic priming tasks. Finally, while disorganized speech is associated with poorer outcomes and role functioning, it is often considered to be less persistent and less prognostically useful than negative thought disorder.[18,19,54,72]

Negative Thought Disorder (Poverty of Speech/Content and Reduced Syntactic Complexity)

Negative thought disorder correlates with other negative symptoms, is predictive of the age of onset of psychosis, and is prognostic of future outcomes (ie, transition to psychosis, being psychotic at follow-up)[73–77] and social role functioning.[78] It is associated with impairments in lexicosemantic retrieval[79] as well as working memory deficits. Neurally, patients who produce less complex sentences showed weaker activation in the right temporal and left prefrontal cortex,[14] and negative thought disorder is associated with gray matter reductions in the orbitofrontal and insular cortex.[67,68]

Speech and Conversation: Flat Affect and Pausing

Flat affect predicts the course and outcome of the illness 20 years after initial hospitalization[80,81] and is associated with worse quality of life[81] and poorer social functioning.[82] Studies examining individuals with flat affect as measured through facial expressivity and emotion processing show that the severity of flat affect is associated with reduced activity in the amygdala, parahippocampal gyrus, as well as multiple regions of the left prefrontal cortex.[83] Flat affect has been shown to correlate with other negative symptoms (eg, negative thought disorder).

Heterogeneity

Not all individuals with psychotic disorders exhibit these abnormalities; furthermore, some healthy individuals do: in an extreme case, one study found that 32% of healthy individuals exhibited “tangentiality” in 50 min of speech vs 50%–60% of the patient group.[84] Additionally, these abnormalities need not co-occur within individuals: studies investigating the co-occurrence of negative and positive thought disorder have found weak correlations at best ($r = .23$) and sometimes observe an inverse relation ($r = -.32$). As a result, these abnormalities should be approached dimensionally rather than as categorically present vs absent.[75]

In spite of this heterogeneity, we note that each type of language abnormality has predictive clinical value. However, language has been underused as a signal in clinical evaluations. This is likely due to the subtlety of some of these abnormalities, as well as the reliance on time-intensive manual evaluations or holistic clinical ratings. Computational methods may provide a way to capitalize on the predictive value of language abnormalities.

Table 1. Descriptions and Examples of Observed Language Production Abnormalities in Psychosis, With Corresponding Computational Methods Used to Measure Them

Language abnormality	Description/Example	Computational Methods
Disorganized Speech-I		
Derailment	Progressively moving off topic. “I always liked geography. My last teacher in that subject was Professor August A. He was a man with black eyes. I also like black eyes. There are also blue and grey eyes and other sorts, too…”[30]	Vector-based models represent sentences as lists of numbers that are compared in similarity [23,24,26,28,29,32,33] Coh-Metrix[33,34] Perplexity – how unexpected a word is given context[63] Graph-based models: represent text as word nodes connected in order, and use network connectivity measures[36,37,38,39]
Tangentiality	Providing oblique or irrelevant answers. Q: “What city are you from?” A: “Well that’s a hard question to answer… I was born in Iowa, but I know that I’m white instead of black so apparently, I came from the North somewhere and I don’t know where, you know, I really don’t know where my ancestors came from. So I don’t know whether I’m Irish or French or Scandinavian or I don’t I don’t believe I’m Polish…”[9]	
Incoherence	Essentially incomprehensible speech. “They’re destroying too many cattle and oil just to make soap. If we need soap when you can jump into a pool of water, and then when you go to buy your gasoline, my folks always thought they should get pop, but the best thing to get is motor oil, and money.”[9]	
Distractible speech	Abrupt topic shifts. “Then I left San Francisco and moved to… where did you get that tie?”[9]	
Disorganized Speech-II		
Abnormal use of referential markers	“They let *him* go, so why not me? (with no prior mention of a ‘him’); “He stabbed the dude and I kicked him.”[8,18,40,41] (ambiguous referent of ‘him’)	Automatic coreference extraction[29]
Illogicality	Non-logical conclusions. “Parents are the people that raise you. Anything that raises you can be a parent. Parents can be anything, material, vegetable, or mineral that has taught you something.”[42]	N/A
Poverty of Speech and Speech Content		
Poverty of speech content	Non-substantive, vague speech. “… I happen to be quite pleased with who I am or how I am and many of the problems that I have and have been working on I have are difficult for me to handle or to work on because I am not aware of them as problems which upset me personally.”[9]	Vector unpacking[38]
Poverty of speech	Q: “Do you think there’s a lot of corruption in the government?” A: “Yeah, seem to be.”[9]	Sentence length[26]
Reduced syntactic complexity	Higher percentage of simple sentences (i.e. ‘I jumped at the sound of his voice’ rather than ‘I jumped at the sound of his voice, which was inordinately loud in the silence’); When complex sentences are used, reduced number of clauses (2 clauses: ‘Mary expects to meet the guy’; 3 clauses: ‘Mary expects to meet the guy who called today’)[7,8,11–14]	Automatic part-of-speech tagging + counts of subordinate clauses [26,28,43]
Lexical abnormalities		
Differences in part of speech use	Decreased density of adjectives, possessive pronouns (‘her’, ‘mine’), determiners (‘that’, ‘what’)[15,26,28]	LIWC[35,44–48]
Neologisms (new words); Increased approximation use	“I got so angry I picked up a dish and threw it at the *geshinker*.”; Refer to a watch as a ‘time vessel’, gloves as ‘hand shoes’, a ballpoint point as a ‘paper skate’[9]	Automatic metaphor detection[31]
Flat affect and pausing		
Flat affect	Speech that is monotonous, emotionless, slowed, and lacking in normal variation in pitch, loudness, tone, & emphasis[16,17] Note: other individuals exhibit pressured speech – loud, rapid (>150 words per minute) speech that is difficult to interrupt[9]	Automatic measurement of pitch[7,16,17,49], formants, speech rate[25,27,49,50], loudness [25,27,77,78,50,53]
Abnormal turn-taking	More frequent and/or longer pauses[25,27], especially when turn-taking	Automatic measurement of pauses[20,25,27]

Note: LIWC, Linguistic Inquiry and Word Count.

Measuring Abnormalities in Language Production Using Automated, Computational Methods

Desiderata

The goal of the computational approaches we review is to provide quantitative measures of the severity of these language abnormalities given a speech/language sample. We evaluate this body of work for (1) construct validity, or evidence that the automated measures are indeed measuring the language abnormalities they are designed to (eg, by comparing them to human ratings, by showing that systematic changes in language lead to systematic changes in measures, or by qualitatively demonstrating the sorts of sentences that score high/low), (2) theoretical validity, (3) replicability, (4) generalizability and equity, and (5) predictive value, or evidence that these measures relate to symptoms, functional outcomes, neurocognitive measures, behavior, etc., so they can lead to targeted intervention or treatment. However, we note that equally important in these early stages of development is a critical evaluation of where models fall short of these standards and promising directions for improvement.

Obtaining Speech Samples

The first step in any analysis is obtaining the relevant data. Currently, most studies gather data in clinical or research settings (eg, in a therapy session, at an in-patient hospital, or in a research lab). Speech can come from clinical assessments or be elicited by a variety of prompts—eg, “Could you tell me about your favorite hobby and how one does it?” and “Tell me the story of Cinderella,” or prompts relating to personal experiences. The benefit of this approach is that the investigator can simultaneously collect demographic and symptom information from participants, which can lead to nuanced and particularly informative analyses. More recently, computational linguists have turned to social media, finding users who declare a psychosis diagnosis and collecting other posts of theirs on unrelated topics, which they contrast against analogous posts by individuals who do not report a diagnosis.[35,44] This method allows access to large language samples of text produced by many users, including those who may not otherwise seek help—but does not allow for analysis of speech nor for systematic clinical measurement.

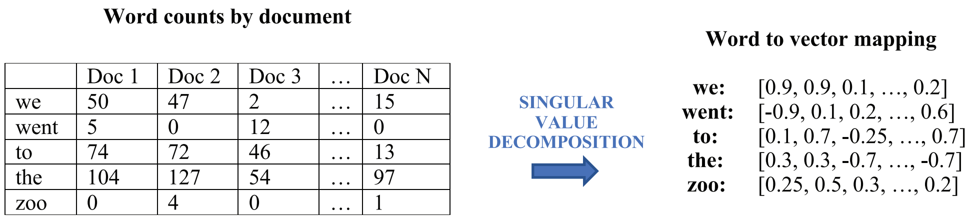
Computational Methods

With these data in hand, researchers have applied a number of computational techniques to measure a variety of linguistic abnormalities. We focus on the most studied methods but list other promising methods in table 1.

Measuring Disorganized Speech

Vector Representations Latent semantic analysis (LSA) and word embedding models (eg, word2vec and GloVe) have been used to obtain measures of disorganized speech (ie, measures of derailment, tangentiality, and coherence). These methods provide a measure of how similar phrases are to one another and have been applied to psychosis research with the idea that more coherent and less derailed speech will, on average, have phrases that are more similar to one another than less coherent texts. These methods first represent each word in the text as a vector—a list of numbers. Roughly speaking, these vectors represent the contexts in which a word is used; words with similar meaning will appear in the same context. For example, similar words (“king”/“queen”) are likely to co-occur in similar contexts and, consequently, will have similar vectors, whereas dissimilar words (“broccoli”/“shoe”) are less likely to co-occur and, consequently, will have less similar vectors (figure 1). Word vectors are combined to obtain phrase-level vectors (eg, by averaging word vectors); similarity of the phrase vectors is used to get measures of disorganization (eg, by measuring how similar adjacent sentences are or how dissimilar subsequent sentences get from the participant’s first sentence; see figure 2).

As shown in table 2, studies using these methods show considerable promise, finding that (1) disorganization scores are significantly higher for individuals with psychotic disorders than controls,[23,29,30] (2) disorganization scores correlate with manual holistic ratings of disorganization,[23,28] and (3) disorganization scores, in combination with other factors, can predict conversion to psychosis or discriminate patients vs controls with accuracies around 70%–100%, sometimes even outperforming



Word counts by document

	Doc 1	Doc 2	Doc 3	…	Doc N
we	50	47	2	…	15
went	5	0	12	…	0
to	74	72	46	…	13
the	104	127	54	…	97
zoo	0	4	0	…	1

Word to vector mapping

we:	[0.9, 0.9, 0.1, …, 0.2]
went:	[-0.9, 0.1, 0.2, …, 0.6]
to:	[0.1, 0.7, -0.25, …, 0.7]
the:	[0.3, 0.3, -0.7, …, -0.7]
zoo:	[0.25, 0.5, 0.3, …, 0.2]

Fig. 1. Latent semantic analysis. Each cell in the table represents the number of times a particular word (eg, ‘we’) occurs in a particular document (eg, Document 1). These counts undergo singular value decomposition to arrive at one vector corresponding to each word in the lexicon (of length 100–500). Words that co-occur in similar documents will have similar vectors, whereas words that occur in different documents will not.

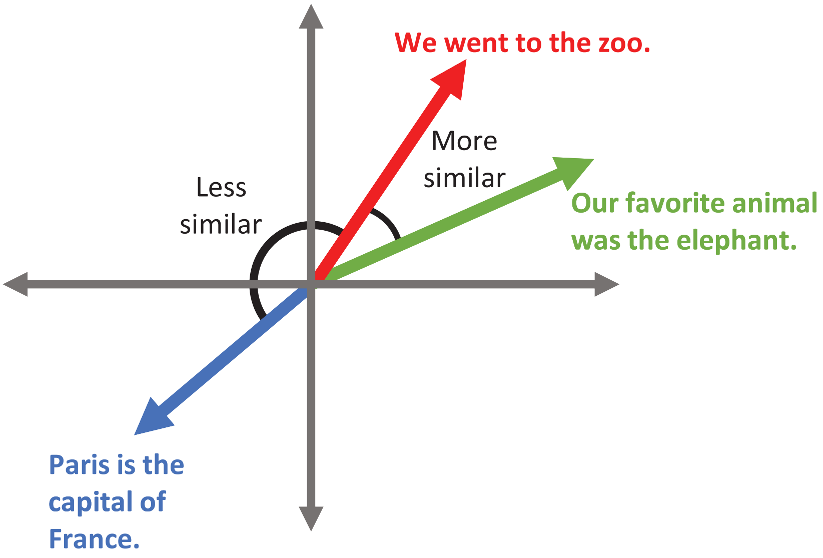
Step 1 **Obtain vectors for each word (e.g. using LSA in Figure 1)**

we:	[0.9, 0.9, 0.1, -0.1, …, 0.2]	Paris:	[-0.4, -0.4, 0.2, -0.1, …, 0.2]
went:	[-0.9, 0.1, 0.2, -0.3, …, 0.6]	is:	[-1.0, 0.1, 0.4, 0.15, …, 0.1]
to:	[0.1, 0.7, -0.25, 0.1, …, 0.7]	capital:	[-0.5, -0.4, 0.7, -0.3, …, 0.3]
the:	[0.3, 0.3, -0.7, 0.4, …, -0.7]	of:	[0.1, -0.8, 0.1, -0.9, …, -0.6]
zoo:	[0.25, 0.5, 0.3, 0.6, …, 0.2]	France:	[-0.3, -0.6, 0.2, -0.1, …, 0.8]

Step 2 **Combine word vectors to obtain phrase-level vectors (e.g. by averaging word vectors)**

We went to the zoo.	**[0.25, 0.5, -0.07, 0.14, …, 0.2]**
Our favorite animal was the elephant.	**[0.5, 0.3, -0.1, 0.5, …, -0.25]**
Paris is the capital of France.	**[-0.3, -0.3, 0.15, -0.14, …, 0.02]**

Step 3 **Compare phrase vector similarity using cosine between vectors.**
Note: The vectors below are represented using only their first two dimensions.


We went to the zoo.
More similar
Less similar
Our favorite animal was the elephant.
Paris is the capital of France.

"We went to the zoo. Our favorite animal was the elephant" would be judged as more coherent than **"We went to the zoo. Paris is the capital of France."**

Fig. 2. Explanation of how vector-based computational approaches can be used to measure components of thought disorder (coherence, derailment, tangentiality) from text.

classifications based on clinical symptoms scales (ie, SIPS).[28,29,26]

Despite this initial promise, this research area faces key challenges. As of yet, no consistent measure of disorganization has yielded reliable findings across multiple studies (eg, the implementations of Bedi et al[26] and Elvevåg et al[23] did not work on the sample of Iter et al[29]). It is hard to interpret these inconsistent results as studies have used different categorization methods (table 2; "Classification") and different ways of obtaining word vectors, have studied different subsets of measures (table 2; "Study"), and have applied them to small, heterogeneous, sometimes poorly controlled samples (eg, age in Iter et al[29]; table 3).

Existing measures of disorganization are also difficult to interpret. While these computational measures correlate with human judgments, it is still unclear what aspect(s) of the complex construct of disorganization these dissimilarity measures reflect. These interpretive difficulties are critical because measures of disorganization sometimes *do not* correlate with positive symptoms (but see Bedi et al[26])

Table 2. Key Findings From Reviewed Studies

Study	Group means	Classification accuracy	Demographics	Symptoms and Functioning	Human judgments	Neuroimaging
			Disorganized Speech			
Elvevag et al.[23]: SZ High ThD SZ Low ThD Control Disorganized speech	High ThD > Low ThD High ThD > Controls	N/R	N/R	N/R	R=0.44 with blind human ratings of tangentiality (scale 1–7)	N/R
Bedi et al.[26] CHR+ CHR- Disorganized speech Determiner density Max. phrase length	N/R	Convex hull: 100%	N/R	Canonical correlation: none with full SIPS; R=0.57 (p=0.046) when restricted to SIPS positive + negative	N/R	N/R
Corcoran et al.[28] CHR+ CHR- FEP Controls Disorganized speech (mean, variance) Possessive pronouns	N/R	Logistic regression: 83% (UCLA), 79% (NYC) (true positive 60%; true negative 82%), 72% (psychosis vs. HC; true positive 69%)	Significant positive association with age No association with gender, ethnicity, or parental SES	No association with SIPS/SOPS clinical ratings (total positive, total negative)	Canonical correlation: R=0.71 ($p<10^{-6}$) between 14 automated and 3 manual ratings (illogicality, poverty of content, referential cohesion)	N/R
Iter et al.[29] SZ/SZA Controls Disorganized speech Ambiguous pronouns	SZ > HC	Random forest: 93% Logistic regression: 86%	N/R	Disorganized speech: N/R Ambiguous pronoun use: R = 0.732 with SANS (p < .01)	Disorganized speech: N/R Ambiguous pronoun use: R = 0.749 with Global Thought Disorder (p<.01)	N/R
Just et al.[30] SZ/SZA (PTD) SZ/SZA (no PTD) Controls Disorganized speech	PTD > no PTD > HC	N/R	N/R	Negative correlation with SAPS PTD (incoherence, tangentiality, derailment) + delusions No association with SANS	N/R	N/R
Gupta et al.[33] UHR Controls Coh-Metrix	HC > CHR for local cohesion (repetition of stem words across sentences)	N/R	N/R Controlled for general intelligence	Lower stem overlap correlated with increased positive symptoms (R = -0.31; p = 0.04); negative symptoms (R = -0.33; p = 0.03); disorganized symptoms (R = -0.31, p = 0.04).	Cognitive correlations: Verbal learning (R = 0.44; p = 0.004); No associations with working memory. Argument overlap correlates with sustained attention (r = 0.32; p = 0.04)	N/R

Table 2. Continued

Study	Group means	Classification accuracy	Demographics	Symptoms and Functioning	Human judgments	Neuroimaging
Poverty of speech and speech content						

Rezaii et al.[38]
CHR+
CHR-
Semantic density (Poverty of content)
	N/R	Logistic regression:
		Training 86.7% (sens. = 0.428; spec. = 1);
		Holdout 80% (sens. = 0.6; spec. = 1);
			No correlation with IQ, age, sex, sentence length (race N/R)	R = -0.446 (p < 0.01) with SIPS negative symptoms	R = 0.42 (p < 0.001) with human judgments of semantic density	

Mota et al.[85] Schizophrenia Mania Healthy controls Graph connectivity measures	SZ > M (in almost all measures) SZ > HC in some measures (e.g. number of loops), but mostly n.s.	Naïve Bayes: SZ vs. M: 94% (sens/spec =94%) SZ vs. HC: 87.5% (sens = 87.5%; spec = 87.5%	N/R	PANSS: n.s. BPRS: n.s.	N/R	N/R
Mota et al.[37] Schizophrenia Bipolar disorder I Healthy controls Graph connectivity measures	SZ > HC, SZ > BP; fewer but still significant differences when controlling for verbosity	Naïve Bayes: SZ vs. HC: AUC = 0.94 SZ vs. BP: AUC = 0.77 (comparable to categorizing using scales)	N/R	Graph measures correlated with PANSS (total, negative, flat affect, poor contact, difficulties with abstract thought, less fluent speech), BPRS (emotional retraction)	N/R	N/R
Mota et al.[68] Schizophrenia Bipolar disorder Healthy controls Graph connectivity measures	Disorganization Index: SZ > BP, HC	Naïve Bayes: SZ vs. (HC + BP): 91.7%, Sens: .92, spec = .76 SZ vs. HC: N/R	No correlation with age, education, or antipsychotics dosage (trends w/ education; e.g. r = 0.6)	Disorganization metric correlated with PANSS negative symptoms (r=.94)	N/R	N/R
Palaniyappan et al.[39] Schizophrenia Bipolar disorder Graph connectivity measures	SZ < BP: SZ produced less connected reports	N/R	N/R	Speech connectedness was associated with psychometric evaluation of thought disorder (r = 0.6), global functioning (r = 0.59), processing speed (r=0.54)	N/R	Speech connectedness associated with structural/ functional brain markers: degree centrality, cortical gyrification

Table 2. Continued

Study	Group means	Classification accuracy	Demographics	Symptoms and Functioning	Human judgments	Neuroimaging
			Flat affect + pausing			
Rapcan et al.[25] Schizophrenia Healthy controls Pausing, pitch	Pausing: SZ > HC Pitch variability: n.s.	Discriminant analysis: 79.4% (sens.: 75.21%, spec.: 83.62%)	Length of illness vs. proportion of silence (r=.43)	Utterance duration vs. BPRS (r=.46) Total utterance duration vs. SANS (r=.46)	N/R	N/R
Martínez-Sánchez et al.[27] Schizophrenia Healthy controls Pausing, speech rate, pitch	Pausing: SZ > HC Speech rate: SZ < HC Pitch variability: SZ < HC Variation in pitch timing: SZ < HC F0: n.s.	Discriminant analysis: 93.8%; holdout 87.5%	Years of chronicity correlated with prosodic variables e.g. phonation trajectories (r=-0.422, p = 0.013)	Intensity of voice correlated with sympoms (r=-0.346, p=0.02) No other associations with positive, negative, or total symptoms	No correlation with BPRS flat affect score	N/R
Covington et al.[52] FEP Pitch, F1-F2	N/R	N/R	N/R	Pitch: n.s. F2 variability correlated with total negative symptoms (r=-0.45), including flat affect (r=-0.398), lack of spontaneity and flow of conversation (r = -0.523), and motor retardation (r = -0.488)	N/R	N/R
Bernardini et al.[51] Schizophrenia Pitch, F1, F2, Speaking time (%)	N/R	N/R	Pitch, F2 variability, speaking time: all greater for women F0 inversely correlated with age	SD of F1 and F2 inversely correlated with positive symptoms SD F2: inversely correlated with SANS total (in Italian sample, not US), avolition, apathy score Pitch variability inversely correlated with SANS total, alogia, affective flattening	N/R	N/R
Compton et al.[86] Psychosis (w/ and w/o aprosody) Healthy controls	Pitch variability, F2 variability, Loudness: SZ(aprosody) < HC in some tasks	N/R	Medication dosage correlated with reduced variability in pitch, F1, F2	N/R	N/R	N/R

Table 2. Continued

Study	Group means	Classification accuracy	Demographics	Symptoms and Functioning	Human judgments	Neuroimaging
			Lexical Abnormalities + Metaphor Use			
Gutiérrez et al.[31] Schizophrenia Healthy controls CHR+ CHR- Metaphor Sentiment analysis	SZ (6.3%) > HC (5.2%) in metaphor use (p<.001) Sentiment analysis: n.s.	Support vector or convex hull: Metaphor: 75% (F-score = 0.789) Sentiment: 68.8% (F-score = 0.732)	When used to predict group: Gender: 65.6% (F-score = .70) Age: 59.4% (F-score = .63) Corrs: N/R	N/R	N/R	N/R
			Exploratory Analyses			
Mitchell et al.[35] Self-report psychosis Controls LIWC Topic Modeling	SZ ≠ HC in topics used (e.g. negative/positive emotion, auxiliary verbs, leisure)	Support vector: 82.3% Maximum entropy: 81.9%	N/R	N/R	N/R	N/R
Zomick et al.[45] Self-report psychosis Controls LIWC	SZ ≠ HC in topics used on Reddit (e.g. negative emotion, function words, word count)	Logistic regression: 82%	N/R	N/R	N/R	N/R

Note: AUC, area under the curve; BP, bipolar disorder; BPRS, Brief Psychiatric Rating Scale; FEP, first episode psychosis; HC, healthy controls; LIWC, Linguistic Inquiry and Word Count; M, mania; n.s., not significant; N/R, not reported; NYC, subset of data collected in NYC; PANSS, The Positive and Negative Syndrome Scale; PTD, Positive thought disorder; SANS, Scale for the Assessment of Negative Symptoms; sens, sensitivity; SES, socioeconomic status; spec, specificity; SZ, schizophrenia; SZA, schizoaffective disorder; ThD, thought disorder; UCLA, subset of data collected at UCLA.

Table 3. Demographic and Other Participant Information From Reviewed Studies

Study	Groups	N (Sex)	Age	Race/Ethnicity	Medication
		Disorganized Speech			
Elvevag et al[23]	Schizophrenia, schizoaffective	26 (27% F)	33.77 (7.63)	N/R	Yes (92%)
	Healthy controls	25 (60% F)	35.44 (12.94)	N/R	No
Bedi et al.[26]	CHR+ (converted)	5 (20% F)	22.2 (3.4)	40% Caucasian	Yes (20%)
	CHR- (did not convert)	29 (34% F)	21.2 (3.6)	38% Caucasian	Yes (21%)
Corcoran et al.[28]	CHR+	19 (11% F)	17.3 (3.7)	63.1% Caucasian	N/R
	CHR-	40 (45% F)	16.4 (3.0)	50.0% Caucasian	N/R
	Controls	21 (38% F)	18 (2.8)	66.7% Caucasian	N/R
	First episode psychosis	16 (31% F)	15.8 (1.7)	62.5% Caucasian	N/R
Iter et al.[29]	Schizophrenia, schizoaffective	9 (20% F)	29.5 (N/R)	N/R	N/R
	Healthy controls	5 (0% F)	40.3 (N/R)	N/R	N/R
Just et al.[30]	Schizophrenia, schizoaffective (w/ positive thought disorder	10 (20% F)	48.1 (12.2)	N/R	Yes (90%)
	Schizophrenia, schizoaffective (w/o positive thought disorder)	10 (50% F)	45.7 (11.7)	N/R	Yes (100%)
	Healthy controls	10 (50% F)	44.5 (13.8)	N/R	No
Gupta et al.[33]	UHR	43 (42% F)	19.33 (1.44)	67% Caucasian, 16% Central/South American, 7% First Nations, 5% Asian, 5% Other	Yes
	Healthy controls	41 (56% F)	18.76 (2.63)	73% Caucasian, 12% Asian, 10% Central/ South American, 2% Black, 3% Other	No
		Poverty of speech and speech content			
Mota et al.[36]	Schizophrenia	8 (13% F)	35.8 (8.7)	N/R	Yes
	Mania	8 (0% F)	42.3 (9.6)	N/R	Yes
	Healthy controls	3 (38% F)	32.3 (10.9)	N/R	No

Table 3. Continued

Study	Groups	N (Sex)	Age	Race/Ethnicity	Medication
Mota et al.[85]	Schizophrenia	20 (20% F)	34.1 (9.6)	N/R	Yes
	Bipolar Disorder I	20 (30% F)	38.2 (12.4)	N/R	Yes
	Healthy controls	20 (55% F)	35.1 (10.9)	N/R	Yes (e.g. antidepressants)
Mota et al.[37]	Schizophrenia	11 (18% F)	14.6 (2.6)	N/R	Yes
	Bipolar Disorder	10 (73% F)	15.3 (3.8)	N/R	Yes
	Healthy controls	21 (55% F)	15.4 (3.6)	N/R	No
Palaniyappan et al.[39]	Schizophrenia	34 (15% F)	32.9 (8.9)	N/R	Yes
	Bipolar Disorder	22 (36% F)	34.6 (10.4)	N/R	Yes
Rezaii et al.[38]	CHR+ (training set)	7 (43% F)	20.6 (5.6)	14% Caucasian, 57% Black, 28.6% Other	Yes
	CHR- (training set)	23 (52% F)	21.5 (4.5)	35% White, 47.8% Black, 17.4% Other	Yes
	CHR+ (test set)	5 (0% F)	22.9 (3.4)	20% Caucasian, 20% Black, 60% Other	Yes
	CHR- (test set)	5 (0% F)	22.8 (2.3)	40% Caucasian, 20% Black, 40% Black	Yes
		Flat affect and pausing			
Martínez-Sánchez et al.[27]	Schizophrenia	45 (13F)	39.49 (10.89)	N/R	Yes (100%)
	Healthy control	35 (13F)	35.34 (10.48)	N/R	No
Covington et al.[52]	First-episode psychosis	25 (24% F)	23.8 (4.4)	88% Black	Yes (92%)
Bernardini et al.[51]	Schizophrenia (Italian)	20 (35% F)	35.4 (11.2)	95% Caucasian, 5% Black	N/R
	Schizophrenia (American)	20 (35% F)	33.6 (10.1)	90% Black, 5% Hispanic, 5% White	N/R
Rapcan et al.[25]	Schizophrenia	39 (31% F)	42.3 (13.5)	N/R	Yes
	Healthy controls	18 (55% F)	40.5 (12.9)	N/R	No

Table 3. Continued

Study	Groups	N (Sex)	Age	Race/Ethnicity	Medication
Compton et al.[86]	Schizophrenia, psychotic disorder NOS, schizophreniform (aprosody)	25 (32% F)	29.9 (9.6)	80% Black, 4% Caucasian, 16% Other	Yes
	Schizophrenia, psychotic disorder NOS, schizophreniform (no aprosody)	29 (21% F)	32.3 (9.2)	83% Black, 10% Caucasian, 7% Other	Yes
	Healthy controls	102 (45% F)	33.7 (9.3)	71% Black, 17% Caucasian, 12% Other	No
Lexical abnormalities					
Gutiérrez et al.[31] (see Bedi et al.[26] for CHR)	Schizophrenia	17 (24% F)	39	N/R	N/R
	Healthy controls	15 (53% F)	35	N/R	N/R
Exploratory analyses					
Mitchell et al.[109]	Schizophrenia (self-report)	174	N/R	N/R	N/R
	Healthy controls	174	N/R	N/R	N/R
Zomick et al.[85]	Schizophrenia (self-report)	159	N/R	N/R	N/R
	Healthy controls	159	N/R	N/R	N/R

Note: N/R, not reported; NOS, not otherwise specified; UHR, ultra-high risk.

—but *do* correlate with other confounds such as age[28] (older participants exhibited less disorganized speech) and sentence length[29] (shorter sentences are rated as more disorganized). Relatedly, measures of disorganization are typically just one of many other variables in categorization models, so it is difficult to quantify the unique contribution of disorganized speech. While these methods show considerable promise, more validation work is clearly needed.

Measuring Poverty of Speech and Content

Word Graphs Mota et al[36,37,85] have used word graphs to measure differences in speech between individuals with schizophrenia, mania, and healthy controls. The structure of speech is represented by linking word nodes based on their order and then using established measures of graph connectivity and complexity (eg, number of nodes/edges/loops and length of longest path) to obtain thought disorder scores. These measures show group differences between schizophrenia, bipolar disorder, and control participants[36,85]; correlate with negative symptoms[37,85]; can predict the presence of psychosis 6 months later[38]; and relate to differences in neural measures.[39] The performance of these measures is impressive; however, it is not yet clear what abnormalities these measures reflect (ie, positive thought disorder vs negative thought disorder), and how theoretically valid they are.

Vector Unpacking Rezaii et al[38] used a method called vector unpacking to automatically measure poverty of speech content (vague, repetitive, or non-substantive speech). They examined whether sentence vectors could be well approximated by other vectors composed of fewer words (eg, the meaning of *The president flew to China on a plane* is well approximated by *The president flew to China*; the corresponding sentence vectors are likely to be very similar). This measure could categorize which adolescents at clinical high risk (CHR) for psychosis would convert to psychosis with accuracy exceeding 80%,[38] correlated with negative symptoms and nonexpert human ratings, and was shown to outperform related measures such as idea density (roughly the density of content words) and information value (roughly the average sentence vector length). This measure shows particular promise as it was individually tested on a held-out dataset and was well-validated against clinical scales and human judgments; future work should test its generalizability.

Syntactic Parsing Speech by individuals with psychosis often exhibits reduced syntactic complexity.[11–14,87] This has primarily been studied by automatically tagging each word in a text with its part-of-speech information (eg, noun and verb) and counting the number of subordinated clauses individuals use.[43,88,89] For example, in addition to the semantic coherence measures described above, Bedi et al[26] found that reduced density of determiner pronouns (eg, "that," "what," and "whatever"), reflecting fewer subordinated clauses, was associated with worse symptom severity. Similarly, Corcoran et al[28] showed that reduced possessive pronoun (eg, "her," "his," and "mine") counts improved performance of their model of CHR conversion. However, these measures have not been considered independently of disorganization measures, so the relative role that each plays is not yet clear.

Measuring Flat Affect and Abnormal Pausing

The methods described above focus on what is said and, thus, can work off of written transcripts of speech. To measure flat affect, researchers have used automated methods to analyze *how* individuals speak, studying the acoustic characteristics of their vocal productions, as well as pausing behavior. We briefly review some promising results (eg, classifying psychosis vs control samples at 70%–94% accuracy[25,27]) here. However, we note that a recent meta-analysis[49] has documented substantial heterogeneity in the findings across both computationally oriented and manual annotation studies, making it clear that there is much work to be done in this area.

Researchers have automatically measured mean pitch (ie, fundamental frequency, F0), as well as pitch variability, of speech by individuals with psychosis vs healthy controls. Some have found that individuals with psychosis have reduced pitch variability relative to controls[27,86] and that within the psychosis group, reduced pitch variability is associated with worse negative symptoms.[49] However, other studies have not found this relationship.[25,52] Other studies have automatically measured the mean and variance of formant values (a measure of spectral properties of speech, largely determined by the shape of the vocal tract). Some studies found that individuals with psychosis exhibit decreased variability in the first 2 formant values, and that decreased variability in formants is associated with worse negative symptoms,[51,52,86,90] but others have failed to replicate these findings.[53] Additional work in this area has shown that individuals with schizophrenia speak at a slower rate,[27,50,90] show less variability in syllable timing,[27] and show decreased variability in loudness/intensity.[25,86] In addition to acoustic differences, individuals with psychosis have also shown abnormal conversational turn-taking relative to controls, pausing more often and for longer.[25,27] Between-turn pauses have also been associated with worse positive symptoms in youth at high-risk for psychosis but showed no significant differences between high risk and control participants.[20]

Meta-analyses of this body of work have documented substantial heterogeneity in the results. Across 5 studies, Cohen et al[91] found no meaningful differences between patients and controls after controlling for sociodemographic and contextual factors. Parola et al[49] reviewed 55 studies (1254 schizophrenia and 699 controls) and found modest, variable effects of pause duration, pitch variability, spoken time, speech rate, and number of pauses (with some evidence of publication bias).

Recent literature has begun investigating the puzzling discrepancy between the size of group differences as measured by acoustic measures vs clinical ratings of blunt affect (the construct that these acoustics are thought to measure). Researchers have suggested that these measures operate at "different resolutions," with clinical ratings providing holistic measures of an entire interaction, while acoustic measures zoom in on sub-portions. This may allow for a more nuanced understanding of flat affect, though more work needs to be done to validate this suggestion.[92]

Additional factors could contribute to the heterogeneity in findings. Acoustic analyses currently require that speech be recorded under very good conditions, such that different recording conditions can make different studies incomparable. In addition, much of the work on vocal characteristics has attempted to measure flat affect; however, other factors that have not been accounted for could lead to voice differences. For example, some individuals with psychosis exhibit motor difficulties, which would likely affect their articulations, and in the sample of Andreasen and Grove,[84] between 16% and 32% of individuals with schizophrenia exhibited pressured speech, which would have the opposite impact on vocal productions than flat affect. The heterogeneity in results could simply reflect the heterogeneity in mechanisms involved, so a more systematic, hypothesis-driven study is required to tease these factors apart and better understand what these measures reflect.

Exploratory Analyses

While the previous methods have studied well-documented language abnormalities in psychosis, investigators have also adopted a more exploratory approach to see whether individuals with psychotic disorders differ from controls in the topics they discuss and words they use, primarily focusing on social media language.[35,44–47] Some of these studies[45] have used Linguistic Inquiry and Word Count (LIWC),[48] which counts the proportion of words that fall within certain predefined categories (eg, negative or positive affect and anxiety). Others have used topic modeling,[35,43] which automatically discovers which topics participants discuss[93] without prespecifying them. Some of the most promising and consistent results suggest that individuals with psychosis use more function words (eg, "the" and "a"), first-person singular pronouns (eg, "I"), auxiliary verbs, negative emotion words, insight words, and health words, but show a decreased focus on leisure.[35,44–47] However, there has been substantial variability in findings, with sometimes opposing effects. For example, of 5 studies, 2 studies[45,46] found that controls used more first-person plural pronouns ("we") than the psychosis group, but another[47] found the opposite, and the remaining studies reported no difference between groups. In addition, the use of social media data means that these results cannot be linked with symptomatology.

Moving Forward

Across domains of language structure and use, computational methods have shown promise in being able to identify the linguistic properties that differentiate individuals with psychosis from healthy controls. But there are clear challenges that the field must address, especially given its high social impact. It can be difficult to evaluate how well these methods are measuring linguistic abnormalities due, in part, to an overreliance on categorization methods. Discrepancies in findings across studies undermine confidence that these methods are generalizable.

How can we move forward? Much of the work thus far has focused on the successes of these methods; an increased focus on when and why these methods fail will help refine our work. A great deal of research has been exploratory in nature; adopting a more hypothesis-driven approach that relates these automated measures to other known, relevant measures in psychosis will help ground these methods in the wider psychosis literature. Finally, we emphasize the importance of considering sociodemographic factors front and center when evaluating these models, especially in light of an extensive literature documenting that computational methods magnify biases. We discuss each of these in turn.

Difficulty Evaluating Performance

Overreliance on Categorization. Much of the past work has focused on developing functions that categorize patients as having (or developing) psychosis or not. While this is important work, overly focusing on categorization creates several interrelated issues. Given the dimensional aspect of these abnormalities—not all patients exhibit these abnormalities, some healthy individuals do, and some patients exhibit opposite patterns of impairment (eg, alogia vs pressured speech and derailment vs perseverance/repetition)—it is unclear how to evaluate classification accuracy. It is unlikely that one can classify based solely on speech/language, and the true target accuracy is likely to vary between studies. On the other hand, categorization functions are very likely to "overfit" the data—that is, learn and rely on spurious differences between the (necessarily limited size) psychosis and control groups that do not necessarily generalize to other datasets, an issue exacerbated by how heterogeneous the manifestations of psychosis are.[32] This could, in part, explain how some models have achieved 100% on one dataset, while being at chance on another. Finally, overly focusing on categorization makes it difficult to evaluate construct validity. Demonstrating that a measure can categorize individuals into 2 groups well does not reveal how and why the measure works, as well as what constructs it is tapping in to. Instead of simply focusing on classification accuracy, it may instead be more useful to (1) evaluate computational methods on speech samples that are known to contain (or not) particular linguistic abnormalities, (2) focus primarily on comparisons with symptoms, behavior, neurocognitive variables, and clinical ratings (less emphasized in past work), and (3) start to tackle questions about the sensitivity of these methods, how specific they are to psychotic disorders vs other illnesses, and what the time course of their predictive value is.

Increasing Comparability of Studies. Although many of the individual studies we report on show promising findings, these findings do not always align with one another. The studies we review have studied different and heterogeneous subgroups (eg, individuals with schizotypy, CHR youth, individuals with schizophrenia, schizoaffective disorder, mania, and individuals with or without thought disorder), in a variety of contexts (hospitals, research labs, and on the internet), using different kinds of prompts (written vs spoken, spontaneous speech vs read speech, and more or less personal questions) that elicited varied lengths of responses. In addition, these studies have made different modeling decisions (eg, have used different categorization techniques) and have studied different subsets of linguistic variables measured in different ways. Any of these differences could have contributed to the heterogeneity across studies; however, the discrepancies make it difficult to evaluate the generalizability of these methods.

This is why it is critical that computational studies make direct comparisons with past work. To facilitate this, studies should share their analyses so that replications are possible. Direct comparison of results can also be helpful for considering qualitatively different methods. For example, directly comparing word graphs vs vector-based coherence measures on the same sample would allow for a better understanding of what each of the methods is capturing and what their relative benefits are.

Where possible, standardizing elicitation methods for speech and written text, or explicitly considering differences between elicitation methods, would be helpful. It has become clear that different methods result in different speech sample lengths, which can add noise to automatic speech and language measures; this leads investigators to make different modeling decisions (eg, Bedi et al[26] vs Corcoran et al[28]), further exacerbating differences. Ideally, research would be done on larger samples of data collected specifically for the purpose of analyzing language[94]; barring this, computational models should be run across multiple datasets to ensure that the model is not overly sensitive to idiosyncratic properties of one dataset.[26,28–30]

Understanding Model Failures/Successes for Model Refinement

To improve modeling, there now needs to be a shift away from emphasizing the good performance of models toward more of a focus on where and why these models fail. This can be done by performing detailed error analyses of the systems. In particular, it would be helpful to examine the speech/language tasks that the model incorrectly marked as having high or low levels of a particular abnormality to identify classes of recurring patterns that the model does not handle well. This approach has successfully fueled innovation in analysis methods. Error analyses of particular language samples allowed Iter[29] et al to realize that methods from Elvevåg et al[23] and Bedi et al[26] performed poorly on text that is heavy with verbal fillers (eg, "uh," "like," and "I mean") and repetitions, and also to realize that sentence length was related to disorganization scores.

In addition to leading to refinements, qualitative analyses of errors can reveal the strengths of methods that might not otherwise have been appreciated. In trying to understand why the biobehavioral measures they studied did not mirror the large effects in clinical ratings, Cohen et al[92] were able to show a temporal resolution at which their measures did show larger effects. This revealed an additional potential benefit of automated methods—that they can capture differences at resolutions that clinical ratings cannot. Especially at the early stage of development, this type of analysis can help move the field in the right direction (and is currently being more emphasized in computational research for this very reason).

Adopting a Hypothesis-Driven Approach

Most of the research thus far has been data-driven and exploratory in nature. While this work has been promising, more focus on theoretical validity and direct connections between computational analyses and the broader psychosis literature could help address some of the issues outlined above. For example, acoustic differences that individuals with psychosis exhibit relative to controls could be due to documented motor difficulties,[4] differences in how individuals represent particular sounds,[95] cognitive difficulties,[1] aprosody,[86] and so forth. Each of these possibilities makes different predictions about (1) the symptoms, (2) behavioral task performance, and (3) neural abnormalities the changes in speech acoustics should be associated with. This can drive more targeted, well-controlled analyses that will yield more reliable performance with the small, heterogeneous samples that characterize this area of research. By expanding the types of questions being asked beyond categorization, hypothesis-driven work can also clearly improve our understanding of what these linguistic measures reflect.

Bias in Computational Methods

Sociodemographic factors, such as race, age, education, gender, as well as linguistic and geographical background, have been understudied in relation to automated methods in psychosis. On the one hand, an extensive literature has documented harmful bias in computational methods across domains,[96] including in some of the very methods described here: vector embeddings show biases based on race and gender,[97,98] automatic speech recognition systems show greater error rates for black speakers than white speakers,[99] and facial recognition software currently used is being recalled because of performance disparities.[100] It is critical to ensure that the models we describe are not plagued by similar biases.

There is some evidence that they may be. For example, Bedi et al[26] found an association with age, such that older individuals had more organized speech samples, but age has not been controlled for in most of the reported analyses, even when patients and controls are not matched on age.[26]Similarly, Mota et al[101] found an association between graph-based speech connectedness and education. In measuring flat affect, researchers have used acoustic cues such as formant values and pitch[51,52,86,90]; however, these acoustics are affected by a number of other factors, including vowel type, neighboring sounds, dialect, gender, and age[52,102–104]—factors which were not modeled in previous work. In fact, Cohen et al[91] found that when controlling for social factors and task type, all group differences disappeared. Controlling for potential social factor confounds is clearly a key area for development.

At the same time, it is important to recognize that speech and language measures must ultimately be evaluated in a social context, as what is considered "normal" (eg, a normal response length to a question) varies drastically by culture. Body language, gestures, and intonation can change how something is perceived, so these methods may ultimately need to be used in conjunction with such measures.[105–109] In addition, most of the models have been developed for English, and other languages may require different, tailored approaches to measuring the same constructs. Although these issues are by no means unique to automated approaches, models that gloss over cultural/contextual factors could magnify the problem, especially as one of the potential benefits of computational methods is that they can reach a wider range of individuals. The field must confront these issues early and consistently to ensure its benefits reach everyone.

These Issues Will Remain Even With Improved Measures

Computational linguistics is a rapidly developing field. Static word embeddings are being replaced with context-sensitive models (eg, BERT and ELMo). Automated speech analysis is yielding a more accurate measurement of a wider range of acoustic measures. Advances in related areas may allow for these methods to be used in conjunction with automated measures of body language, gesture, facial expressions, and so forth.[105–109] As we capitalize on these advances, we will still need to address the core issues we have identified here: compare performance across models, identify their strengths and areas for potential improvement, link model results to the broader psychosis literature (eg, through hypothesis-driven methods), and inspect models for bias.

Conclusions

Abnormalities in language production are characteristic of psychosis, present prior to disease onset, and can directly contribute to worse outcomes. Computational methods can be used to automatically detect these language abnormalities and have shown great promise in being able to classify and predict psychosis, sometimes outperforming clinical measures. These methods are particularly promising, as they are objective and cost-effective, meaning that they could be applied on a wide scale to reach and help individuals who might previously fall through the cracks. Much of the work to this point has understandably focused on demonstrating the successes of these methods. However, to best move the field forward, we argue that the field should now shift focus toward understanding when and why current models fail. Accomplishing this will require collaborations between psychosis researchers, linguists who understand the measures and language abnormalities, as well as computational researchers who can develop and refine these models to be appropriate for this area. By performing qualitative error analyses, testing the generalizability of these models, adopting a more hypothesis-driven approach where possible, and aligning results with decades of psychosis research, we can better adapt these methods to the psychosis domain, to ensure that these methods can be as beneficial for all as quickly as possible.

Funding

This was supported by the National Institutes of Health (R21MH119677 to M.G. and V.A.M.).

Acknowledgment

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

References

1. Heinrichs RW, Zakzanis KK. Neurocognitive deficit in schizophrenia: a quantitative review of the evidence. *Neuropsychology* 1998;12(3):426–445.
2. Green MF, Kern RS, Braff DL, Mintz J. Neurocognitive deficits and functional outcome in schizophrenia: are we measuring the "right stuff"? *Schizophr Bull.* 2000;26(1):119–136.
3. Andreasen NC, Paradiso S, O'Leary DS. "Cognitive dysmetria" as an integrative theory of schizophrenia: a dysfunction in cortical-subcortical-cerebellar circuitry? *Schizophr Bull.* 1998;24(2):203–218.
4. Middleton FA, Strick PL. Basal ganglia and cerebellar loops: motor and cognitive circuits. *Brain Res Brain Res Rev.* 2000;31(2-3):236–250.
5. van Harten PN, Walther S, Kent JS, Sponheim SR, Mittal VA. The clinical and prognostic value of motor abnormalities in psychosis, and the importance of instrumental assessment. *Neurosci Biobehav Rev.* 2017;80:476–487.
6. Mittal VA, Bernard JA, Northoff G. What can different motor circuits tell us about psychosis? An RDoC perspective. *Schizophr Bull.* 2017;43(5):949–955.
7. Covington MA, He C, Brown C, et al. Schizophrenia and the structure of language: the linguist's view. *Schizophr Res.* 2005;77(1):85–98.
8. Kuperberg GR. Language in schizophrenia part 1: an Introduction. *Lang Linguist Compass.* 2010;4(8):576–589.
9. Andreasen NC. Thought, language, and communication disorders. I. Clinical assessment, definition of terms, and evaluation of their reliability. *Arch Gen Psychiatry.* 1979;36(12):1315–1321.
10. Andreasen NC. Thought, language, and communication disorders. II. Diagnostic significance. *Arch Gen Psychiatry.* 1979;36(12):1325–1330.
11. Morice RD, Ingram JC. Language analysis in schizophrenia: diagnostic implications. *Aust N Z J Psychiatry.* 1982;16(2):11–21.
12. Morice RD, Igram JC. Language complexity and age of onset of schizophrenia. *Psychiatry Res.* 1983;9(3):233–242.
13. Fraser WI, King KM, Thomas P, Kendell RE. The diagnosis of schizophrenia by language analysis. *Br J Psychiatry.* 1986;148:275–278.
14. Kircher TT, Oh TM, Brammer MJ, McGuire PK. Neural correlates of syntax production in schizophrenia. *Br J Psychiatry.* 2005;186:209–214.
15. Obrębska M, Obrębski T. Lexical and grammatical analysis of schizophrenic patients' language: a preliminary report. *Psychol Lang Comm.* 2007;11(1):63–72.
16. Spoerri TH. Speaking voice of the schizophrenic patient. *Arch Gen Psychiatry.* 1966;14(6):581–585.
17. Alpert M, Rosen A, Welkowitz J, Sobin C, Borod JC. Vocal acoustic correlates of flat affect in schizophrenia: similarity to Parkinson's disease and right hemisphere disease and contrast with depression. *Br J Psychiatry.* 1989;154(S4):51–56.
18. Bearden CE, Wu KN, Caplan R, Cannon TD. 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(7):669–680.
19. 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(2-3):278–283.
20. Sichlinger L, Cibelli E, Goldrick M, Mittal VA. Clinical correlates of aberrant conversational turn-taking in youth at clinical high-risk for psychosis. *Schizophr Res.* 2019;204:419–420.

21. Corcoran CM, Mittal VA, Bearden CE, et al. Language as a biomarker for psychosis: a natural language processing approach. *Schizophr Res.* doi:10.1016/j.schres.2020.04.032.
22. Solomon M, Olsen E, Niendam T, et al. From lumping to splitting and back again: atypical social and language development in individuals with clinical-high-risk for psychosis, first episode schizophrenia, and autism spectrum disorders. *Schizophr Res.* 2011;131(1-3):146–151.
23. Elvevåg B, Foltz PW, Weinberger DR, Goldberg TE. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophr Res.* 2007;93(1-3):304–316.
24. Elvevåg B, Foltz PW, Rosenstein M, Delisi LE. An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. *J Neurolinguistics.* 2010;23(3):270–284.
25. Rapcan V, D'Arcy S, Yeap S, Afzal N, Thakore J, Reilly RB. Acoustic and temporal analysis of speech: a potential biomarker for schizophrenia. *Med Eng Phys.* 2010;32(9):1074–1079.
26. 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.
27. Martínez-Sánchez F, Muela-Martínez JA, Cortés-Soto P, et al. Can the acoustic analysis of expressive prosody discriminate schizophrenia? *Span J Psychol.* 2015;18:E86. doi:10.1017/sjp.2015.85.
28. Corcoran CM, Carrillo F, Fernández-Slezak D, et al. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry.* 2018;17(1):67–75.
29. Iter D, Yoon J, Jurafsky D. Automatic detection of incoherent speech for diagnosing schizophrenia. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic; June 5, 2018, New Orleans, LA: Association for Computational Linguistics; 2018: 136–146.
30. Just S, Haegert E, Kořanova N, et al. Coherence models in schizophrenia. In: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology; June 6, 2019, Minneapolis, MN: Association for Computational Linguistics; 2019: 126–136.
31. Gutiérrez ED, Cecchi G, Corcoran C, Corlett P. Using automated metaphor identification to aid in detection and prediction of first-episode schizophrenia. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing; September 7–11, 2017, Copenhagen, Denmark; 2017: 2923–2930.
32. Voleti R, Woolridge S, Liss JM, Milanovic M, Bowie CR, Berisha V. Objective assessment of social skills using automated language analysis for identification of schizophrenia and bipolar disorder. *INTERSPEECH.* 2019;
33. Gupta T, Hespos SJ, Horton WS, Mittal VA. Automated analysis of written narratives reveals abnormalities in referential cohesion in youth at ultra high risk for psychosis. *Schizophr Res.* 2018;192:82–88.
34. Minor KS, Willits JA, Marggraf MP, Jones MN, Lysaker PH. Measuring disorganized speech in schizophrenia: automated analysis explains variance in cognitive deficits beyond clinician-rated scales. *Psychol Med.* doi:10.1017/S0033291718001046.
35. Mitchell M, Hollingshead K, Coppersmith, G. Quantifying the language of schizophrenia in social media. In: Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality; Denver, CO. 2015: 11–20.
36. Mota NB, Vasconcelos NA, Lemos N, et al. Speech graphs provide a quantitative measure of thought disorder in psychosis. *PLoS One.* 2012;7(4):e34928.
37. Mota NB, Copelli M, Ribeiro S. Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance. *npj Schizophr.* 2017;3(1):1–10.
38. Rezaii N, Walker E, Wolff P. A machine learning approach to predicting psychosis using semantic density and latent content analysis. *npj Schizophr.* 2019;5(1):1–12.
39. Palaniyappan L, Mota NB, Oowise S, et al. Speech structure links the neural and socio-behavioural correlates of psychotic disorders. *Prog Neuropsychopharmacol Biol Psychiatry.* 2019;88:112–120.
40. Docherty NM. Communication disturbances in schizophrenia: a two-process formulation. *Compr Psychiatry.* 1995;36(3):182–186.
41. Docherty NM, DeRosa M, Andreasen NC. Communication disturbances in schizophrenia and mania. *Arch Gen Psychiatry.* 1996;53(4):358–364.
42. Andreasen NC. Scale for the assessment of thought, language, and communication (TLC). *Schizophr Bull.* 1986;12(3):473.
43. Kayi ES, Diab M, Pauselli L, Compton M, Coppersmith G. Predictive linguistic features of schizophrenia. In: Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017); August 3–4, Vancouver, Canada. 2017: 241–250.
44. Coppersmith G, Dredze M, Harman C, Hollingshead K. From ADHD to SAD: analyzing the language of mental health on Twitter through self-reported diagnoses. In: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality; June 5, 2015, Denver, CO; 2015: 1–10.
45. Zomick J, Levitan SI, Serper M. Linguistic analysis of schizophrenia in Reddit posts. In: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology; June 6, 2019, Minneapolis, MN: Association for Computational Linguistics; 2019: 74–83.
46. Lyons M, Aksayli ND, Brewer G. Mental distress and language use: linguistic analysis of discussion forum posts. *Comput Human Behav.* 2018;87:207–11.
47. Birnbaum ML, Ernala SK, Rizvi AF, De Choudhury M, Kane JM. A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. *J Med Internet Res.* 2017;19(8):e289.
48. Pennebaker JW, Francis ME, Booth RJ. *Linguistic inquiry and word count: LIWC 2001.* Mahway, NJ: Lawrence Erlbaum Associates. 2001;71.
49. Parola A, Simonsen A, Bliksted V, Fusaroli R. Voice patterns in schizophrenia: a systematic review and Bayesian meta-analysis. *Schizophr Res.* 2019;216:24–40.
50. Cohen AS, Alpert M, Nienow TM, Dinzeo TJ, Docherty NM. Computerized measurement of negative symptoms in schizophrenia. *J Psychiatr Res.* 2008;42(10):827–836.

51. Bernardini F, Lunden A, Covington M, et al. 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.* 2016;239:253–258.
52. Covington MA, Lunden SL, Cristofaro SL, et al. Phonetic measures of reduced tongue movement correlate with negative symptom severity in hospitalized patients with first-episode schizophrenia-spectrum disorders. *Schizophr Res.* 2012;142(1-3):93–95.
53. Arevian AC, Bone D, Malandrakis N, et al. Clinical state tracking in serious mental illness through computational analysis of speech. *PLoS One.* 2020;15(1):e0225695.
54. Harrow M, Marengo JT. Schizophrenic thought disorder at followup: its persistence and prognostic significance. *Schizophr Bull.* 1986;12(3):373–393.
55. Andreasen NC, Olsen S. Negative v positive schizophrenia. Definition and validation. *Arch Gen Psychiatry.* 1982;39(7):789–794.
56. Docherty NM, Cohen AS, Nienow TM, Dinzeo TJ, Dangelmaier RE. Stability of formal thought disorder and referential communication disturbances in schizophrenia. *J Abnorm Psychol.* 2003;112(3):469–475.
57. Kuperberg GR, McGuire PK, David AS. Sensitivity to linguistic anomalies in spoken sentences: a case study approach to understanding thought disorder in schizophrenia. *Psychol Med.* 2000;30(2):345–357.
58. Aloia MS, Gourovitch ML, Missar D, Pickar D, Weinberger DR, Goldberg TE. Cognitive substrates of thought disorder, II: specifying a candidate cognitive mechanism. *Am J Psychiatry.* 1998;155(12):1677–1684.
59. Kostova M, Passerieux C, Laurent JP, Hardy-Baylé MC. N400 anomalies in schizophrenia are correlated with the severity of formal thought disorder. *Schizophr Res.* 2005;78(2-3):285–291.
60. Moritz S, Andresen B, Domin F, et al. Increased automatic spreading activation in healthy subjects with elevated scores in a scale assessing schizophrenic language disturbances. *Psychol Med.* 1999;29(1):161–170.
61. Goldberg TE, Aloia MS, Gourovitch ML, Missar D, Pickar D, Weinberger DR. Cognitive substrates of thought disorder, I: the semantic system. *Am J Psychiatry.* 1998;155(12):1671–1676.
62. Harvey PD, Earle-Boyer EA, Levinson JC. Cognitive deficits and thought disorder: a retest study. *Schizophr Bull.* 1988;14(1):57–66.
63. Bagner DM, Melinder MR, Barch DM. Language comprehension and working memory language comprehension and working memory deficits in patients with schizophrenia. *Schizophr Res.* 2003;60(2-3):299–309.
64. Kuperberg GR, McGuire PK, David AS. Reduced sensitivity to linguistic context in schizophrenic thought disorder: evidence from on-line monitoring for words in linguistically anomalous sentences. *J Abnorm Psychol.* 1998;107(3):423–434.
65. Andrews S, Shelley AM, Ward PB, Fox A, Catts SV, McConaghy N. Event-related potential indices of semantic processing in schizophrenia. *Biol Psychiatry.* 1993;34(7):443–458.
66. Kuperberg G, Heckers S. Schizophrenia and cognitive function. *Curr Opin Neurobiol.* 2000;10(2):205–210.
67. Sans-Sansa B, McKenna PJ, Canales-Rodríguez EJ, et al. Association of formal thought disorder in schizophrenia with structural brain abnormalities in language-related cortical regions. *Schizophr Res.* 2013;146(1-3):308–313.
68. Palaniyappan L, Mahmood J, Balain V, Mougin O, Gowland PA, Liddle PF. Structural correlates of formal thought disorder in schizophrenia: an ultra-high field multivariate morphometry study. *Schizophr Res.* 2015;168(1-2):305–312.
69. Weinstein S, Werker JF, Vouloumanos A, Woodward TS, Ngan ET. Do you hear what I hear? Neural correlates of thought disorder during listening to speech in schizophrenia. *Schizophr Res.* 2006;86(1-3):130–137.
70. Kircher TT, Bulimore ET, Brammer MJ, et al. Differential activation of temporal cortex during sentence completion in schizophrenic patients with and without formal thought disorder. *Schizophr Res.* 2001;50(1-2):27–40.
71. Kircher TT, Liddle PF, Brammer MJ, Williams SC, Murray RM, McGuire PK. Reversed lateralization of temporal activation during speech production in thought disordered patients with schizophrenia. *Psychol Med.* 2002;32(3):439–449.
72. Wilcox J, Winokur G, Tsuang M. Predictive value of thought disorder in new-onset psychosis. *Compr Psychiatry.* 2012;53(6):674–678.
73. Gooding DC, Ott SL, Roberts SA, Erlenmeyer-Kimling L. Thought disorder in mid-childhood as a predictor of adulthood diagnostic outcome: findings from the New York High-Risk Project. *Psychol Med.* 2013;43(5):1003–1012.
74. Wilcox J, Briones D, Quadri S, Tsuang M. Prognostic implications of paranoia and thought disorder in new onset psychosis. *Compr Psychiatry.* 2014;55(4):813–817.
75. Roche E, Creed L, MacMahon D, Brennan D, Clarke M. The epidemiology and associated phenomenology of formal thought disorder: a systematic review. *Schizophr Bull.* 2015;41(4):951–962.
76. Docherty N, Schnur M, Harvey PD. Reference performance and positive and negative thought disorder: a follow-up study of manics and schizophrenics. *J Abnorm Psychol.* 1988;97(4):437–442.
77. Harvey PD, Docherty NM, Serper MR, Rasmussen M. Cognitive deficits and thought disorder: II. An 8-month followup study. *Schizophr Bull.* 1990;16(1):147–156.
78. Tan EJ, Thomas N, Rossell SL. Speech disturbances and quality of life in schizophrenia: differential impacts on functioning and life satisfaction. *Compr Psychiatry.* 2014;55(3):693–698.
79. Nagels A, Fährmann P, Stratmann M, et al. Distinct neuropsychological correlates in positive and negative formal thought disorder syndromes: the thought and language disorder scale in endogenous psychoses. *Neuropsychobiology* 2016;73(3):139–147.
80. Knight RA, Roff JD, Barrnett J, Moss JL. Concurrent and predictive validity of thought disorder and affectivity: a 22-year follow-up of acute schizophrenics. *J Abnorm Psychol.* 1979;88(1):1–12.
81. Gur RE, Kohler CG, Ragland JD, et al. Flat affect in schizophrenia: relation to emotion processing and neurocognitive measures. *Schizophr Bull.* 2006;32(2):279–287.
82. Evensen J, Røssberg JI, Barder H, et al. Flat affect and social functioning: a 10 year follow-up study of first episode psychosis patients. *Schizophr Res.* 2012;139(1-3):99–104.
83. Lepage M, Sergerie K, Benoit A, Czechowska Y, Dickie E, Armony JL. Emotional face processing and flat affect in schizophrenia: functional and structural neural correlates. *Psychol Med.* 2011;41(9):1833–1844.

84. Andreasen NC, Grove WM. Thought, language, and communication in schizophrenia: diagnosis and prognosis. *Schizophr Bull.* 1986;12(3):348–359.

85. Mota NB, Furtado R, Maia PP, Copelli M, Ribeiro S. Graph analysis of dream reports is especially informative about psychosis. *Sci Rep.* 2014;4:3691.

86. Compton MT, Lunden A, Cleary SD, et al. The aprosody of schizophrenia: computationally derived acoustic phonetic underpinnings of monotone speech. *Schizophr Res.* 2018;197:392–399.

87. Thomas P. Syntactic complexity and negative symptoms in first onset schizophrenia. *Cogn Neuropsychiatry.* 1996;1(3):191–200.

88. Marcus M. *Building a Large Annotated Corpus of English: The Penn Treebank*. Fort Belvoir, VA: Defense Technical Information Center; 1993.

89. Toutanova K, Klein D, Manning CD, Singer Y. Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology – NAACL '03; Edmonton, Canada; 2003: 173–180.

90. Wörtwein T, Baltrušaitis T, Laksana E, et al. Computational analysis of acoustic descriptors in psychotic patients. In: Proceedings of Interspeech 2017; August 20–24, 2017, Stockholm, Sweden.

91. Cohen AS, Mitchell KR, Docherty NM, Horan WP. Vocal expression in schizophrenia: less than meets the ear. *J Abnorm Psychol.* 2016;125(2):299–309. doi:10.1037/abn0000136.

92. Cohen AS, Schwartz E, Le TP, et al. Digital phenotyping of negative symptoms: the relationship to clinician ratings. *Schizophrenia Bulletin.* 2021;47(1):44–53.

93. Blei DM. Latent Dirichlet allocation. *J Mach Learn Res.* 2003;3(1):993–1022.

94. Foltz PW, Rosenstein M, Elvevåg B. Detecting clinically significant events through automated language analysis: quo imus? *npj Schizophr.* 2016;2:15054.

95. Cienfuegos A, March L, Shelley AM, Javitt DC. Impaired categorical perception of synthetic speech sounds in schizophrenia. *Biol Psychiatry.* 1999;45(1):82–88.

96. Blodgett SL, Barocas S, Daumé H III, Wallach H. *Language (Technology) is Power: a Critical Survey of "Bias" in NLP*. arXiv. Preprint posted online May 29, 2020. doi:arXiv:2005.14050v2.

97. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. *Science* 2017;356(6334):183–186.

98. Garg N, Schiebinger L, Jurafsky D, Zou J. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc Natl Acad Sci U S A.* 2018;115(16):E3635–E3644.

99. Koenecke A, Nam A, Lake E, et al. Racial disparities in automated speech recognition. *Proc Natl Acad Sci U S A.* 2020;117(14):7684–7689.

100. Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency 2018; February 23–24, 2018, New York, NY; 77–91.

101. Mota NB, Sigman M, Cecchi G, Copelli M, Ribeiro S. The maturation of speech structure in psychosis is resistant to formal education. *npj Schizophr.* 2018;4(1):25.

102. Hillenbrand JM, Clark MJ. The role of f0 and formant frequencies in distinguishing the voices of men and women. *Atten Percept Psychophys.* 2009;71(5):1150–1166.

103. Hillenbrand J, Gayvert RT. Vowel classification based on fundamental frequency and formant frequencies. *J Speech Hear Res.* 1993;36(4):694–700.

104. Hillenbrand JM, Clark MJ, Nearey TM. Effects of consonant environment on vowel formant patterns. *J Acoust Soc Am.* 2001;109(2):748–763.

105. Alvino C, Kohler C, Barrett F, Gur RE, Gur RC, Verma R. Computerized measurement of facial expression of emotions in schizophrenia. *J Neurosci Methods.* 2007;163(2):350–361.

106. Kupper Z, Ramseyer F, Hoffmann H, Kalbermatten S, Tschacher W. Video-based quantification of body movement during social interaction indicates the severity of negative symptoms in patients with schizophrenia. *Schizophr Res.* 2010;121(1-3):90–100.

107. Wang P, Barrett F, Martin E, et al. Automated video-based facial expression analysis of neuropsychiatric disorders. *J Neurosci Methods.* 2008;168(1):224–238.

108. Cohen AS, Cowan T, Le TP, et al. Ambulatory digital phenotyping of blunted affect and alogia using objective facial and vocal analysis: proof of concept. *Schizophr Res.* 2020;220:141–146.

109. Gupta T, Haase CM, Strauss GP, Cohen AS, Mittal VA. Alterations in facial expressivity in youth at clinical high-risk for psychosis. *J Abnorm Psychol.* 2019;128(4):341–351.