Automated coherence measures fail to index thought disorder in individuals at risk for psychosis

Kasia Hitczenko[1], Henry R. Cowan[2], Vijay A. Mittal[2,3,4,5,6], Matthew Goldrick[1,6]
Departments of Linguistics[1], Psychology[2], and Psychiatry[3]
Institute for Policy Research[4], Medical Social Sciences[5],
Institute for Innovations in Developmental Sciences[6]
Northwestern University
kasia.hitczenko@northwestern.edu

Abstract

Thought disorder – linguistic disturbances including incoherence and derailment of topic – is seen in individuals both with and at risk for psychosis. Methods from computational linguistics have increasingly sought to quantify thought disorder to detect group differences between clinical populations and healthy controls. While previous work has been quite successful at these classification tasks, the lack of interpretability of the computational metrics has made it unclear whether they are in fact measuring thought disorder. In this paper, we dive into these measures to try to better understand what they reflect. While we find group differences between at-risk and healthy control populations, we also find that the measures mostly do not correlate with existing measures of thought disorder symptoms (what they are intended to measure), but rather correlate with surface properties of the speech (e.g., sentence length) and sociodemographic properties of the speaker (e.g., race). These results highlight the importance of considering interpretability front and center as the field continues to grow. Ethical use of computational measures like those studied here – especially in the high-stakes context of clinical care – requires us to devote substantial attention to potential biases in our measures.

1 Introduction

Individuals with psychosis exhibit language disturbances, often referred to as thought disorder. At the discourse level, this includes poverty of speech (low quantities of speech), poverty of speech content (vague, repetitive speech), as well as the focus of this work: incoherence and derailment (slow but steady loss of topic; e.g., '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.') (Andreasen, 1986; Bleuler, 1950; Kuperberg, 2010). These symptoms are used to diagnose psychotic disorders and are thought to have predictive clinical value (Andreasen, 1979, 1986; Andreasen and Grove, 1986; First, 1997; Roche et al., 2016; Wilcox et al., 2012). Similar, but attenuated, symptoms are observed in individuals who do not have psychosis, but who meet criteria for being at clinical high-risk for psychosis (CHR). In this population, the presence of these linguistic symptoms predicts later transition to psychosis (Bearden et al., 2011; Demjaha et al., 2017; Perkins et al., 2015).

However, despite the clinical value of these measures, these symptoms have generally been evaluated via self-report and/or overall clinician impressions, which may capture only the most extreme disturbances. Manual annotations of specific linguistic features may allow for more nuanced measures; however, they are time-intensive and infeasible to apply on a wide scale. As a result, these linguistic measures, despite their clinical value, have been underused in the field.

There is a growing body of literature trying to automatically quantify these linguistic differences using methods from computational linguistics, both in psychosis (Elvevåg et al., 2007; Iter et al., 2018; Just et al., 2019; Hitczenko et al., 2020) and CHR populations (Bedi et al., 2015; Corcoran et al., 2018; Gupta et al., 2018; Corcoran et al., 2020). This work has been quite successful, replicating group differences between patient and healthy populations and accurately categorizing individuals into appropriate groups. However, much of the focus of this work has been on separating groups, and there has been less of a focus on relating these metrics to symptoms. Work examining this relationship has sometimes found correlations between these computational metrics and relevant symptoms, but has often failed to find such relationships.

In order for these measures to be useful clinically, it is important to establish their construct validity:

Do they relate to relevant symptoms? Or, do their instead reflect other linguistic/demographic factors? Establishing trust for a system's predictions is particularly important in the clinical/medical setting where these systems could have substantial consequences (Ribeiro et al., 2016). This is especially true as the machine learning systems that these metrics rely on are known to exhibit potentially harmful biases in other domains (Bolukbasi et al., 2016; Caliskan et al., 2017; Koenecke et al., 2020).

In this paper, we dive into measures utilized in previous work to try to understand what they reflect. Following this work, we use a suite of models to quantify incoherence and derailment on speech produced by the CHR vs. HC groups (individuals who meet criteria for being at high-risk for psychosis vs. healthy controls). We examine group differences, finding significant differences using a subset of measures (at uncorrected $\alpha = .05$). We then critique these measures to determine if they reflect the target thought disorder symptoms – and fail to find specific correlations. Finally, we consider what these measures *do* reflect, finding that they partially reflect surface properties of the speech (sentence length) and sociodemographic properties of the speaker. These results highlight the need to consider the interpretability of these measures as the field continues to grow.

2 A Note on Terminology

Past work applying computational methods to study thought disorder in psychosis has used the words 'incoherence' or 'tangentiality' to describe their object of study, which has focused on the cohesion between sentences. However, this terminology is somewhat misaligned with the terminology discussed in the original thought disorder literature, which uses 'incoherence' to describe a lack of cohesion *within* sentences and 'tangentiality' for cases where participants give an off-topic response to a question (Andreasen, 1986). In this paper, we follow the naming conventions of past computational work in this area. We will refer to methods measuring the cohesion between neighboring sentences as 'coherence measures' and methods measuring how much a text drifts off topic as 'tangentiality measures'. However, it is very important to note that these methods better relate to derailment as defined in Andreasen (1986), as they measure how much a participant shifts topics between sentences.

	CHR	HC
Sociodemographics		
Age	21.0(2.3)	21.6(3.2)
Sex		
Female	47%	71%
Male	53%	29%
Education Level	14.4(2.1)	14.6(2.2)
Racial Identity		
First Nations	0%	2%
East Asian	9%	7%
Southeast Asian	0%	5%
South Asian	6%	2%
Black	37%	17%
Central/South American	11%	2%
West/Central Asia and ME	0%	2%
White	31%	51%
Interracial	6%	10%
Ethnicity		
Hispanic	23%	12%
Not Hispanic	77%	88%
WRAT Score	108(15)	118(13)
Speech Samples		
Sentence Length	29.2(6.5)	30.8(10.1)
Lexical Diversity	0.70(0.04)	0.71(0.03)
Response Length	295(169)	275(121)

Table 1: Summary of participant and speech sample measures. ME = Middle East.

3 Methods

3.1 Participants

Speech samples were obtained from 77 participants aged 16-30: (a) 36 who met criteria for being at clinical high-risk for psychosis, and (b) 41 healthy controls. Participants were recruited from the larger Chicago, Illinois area through newspaper, transit, and Craigslist ads, e-mail postings, flyers, and community professional referrals. The Structured Interview for Prodromal Syndromes (SIPS) was used to determine the CHR vs. HC status of the participant (Miller et al., 1999) and to assess symptomatology. The Structured Clinical Interview for the DSM (First, 1997) was used to rule out Axis I psychotic disorder diagnoses within both groups.

Written informed consent was obtained from all participants. Data collection took place in a research lab setting and was approved by the institutional review board at Northwestern University.

3.2 Participant Measures

We obtained self-reported demographic information from participants (including age, sex, education level, and racial identity). In addition, participants completed the Word Reading subtest of the fourth edition of the Wide Range Achievement Test (WRAT) (Wilkinson and Robertson, 2006), which is a measure of scholastic achievement, strongly associated with general intelligence (Johnstone et al., 1996). As described, symptom severity was measured using the SIPS clinical interview. Our analyses focused on the following symptom items: P5 (“disorganized communication”) (range 0-6), N5 (“ideational richness”) (0-6), and D2 (“bizarre thinking”) (0-6), in addition to the positive symptoms subscale total (0-30), the negative symptoms subscale total (0-36), and the disorganized symptoms subscale total (0-24) (see Miller et al. (1999), McGlashan et al. (2001), and Appendix A for more details about the SIPS).

3.3 Speech Measures

3.3.1 Speech Elicitation

Participants were prompted to describe (1) a challenge they had overcome, (2) a self-defining memory, (3) a turning-point memory, and (4) an unusual memory (see Appendix B for full prompts). Their responses were professionally transcribed. For the CHR group, responses were 275 words long on average (range: 111-835 words), while for the HC group, responses were 255 words long on average (range: 98-559 words). We analyze the first full uninterrupted response participants provided and remove the following filler words: *um, uh, you know, I mean, okay, so, actually basically, right, yeah* as in Iter et al. (2018) (see Appendix E for analyses with filler words included). We analyzed each participant’s four responses separately before averaging them to obtain a mean coherence and a mean tangentiality score for each individual.

3.3.2 Automated Coherence/Tangentiality Measures

We obtain a measure of **coherence**, using the same word embedding methods used in past work on both psychosis and CHR populations (Bedi et al., 2015; Corcoran et al., 2018). At a high-level, this measure represents how similar, on average, the adjacent sentences in each participant’s speech samples are to one another. If their sentences tend to be dissimilar to one another, then this is taken as evidence of incoherence.

To do this, we represent each word in the speech sample as a vector (using one of three pre-trained word embedding models e.g., word2vec), and combine the vectors of the words in a sentence (using one of 4 methods e.g., by averaging the word vectors) to obtain a vector for each sentence. We then calculate the cosine similarity between each pair of adjacent sentences, and average these, to obtain one coherence score per speech sample. We average across speech samples to obtain one overall score per participant.

We also obtain a measure of **tangentiality** as in Elvevåg et al. (2007) and Iter et al. (2018). At a high-level, this measure represents how quickly the topic of the speech sample changes. To do this, with sentence-level vectors in hand, we calculate the cosine similarity between the first sentence of a speech sample and each subsequent sentence (i.e., sentence 1 vs. sentence 2, sentence 1 vs. sentence 3, and so forth). We then fit a linear regression model to these values, treating the sentence number as the independent variable and the similarity score against the first sentence as the dependent variable. We use the slope of this line as the tangentiality measure. As with coherence, we obtain one measure for each speech sample, which we average within participants to obtain one overall tangentiality score per participant.

We follow Iter et al. (2018) in deciding which embedding models to use to obtain the sentence-level vector representations needed for these measures. We use either LSA (Landauer et al., 1998), GLoVE (Pennington et al., 2014), or word2vec (Mikolov et al., 2013) to obtain word-level vectors.[1] For sentence embedding methods, we simply average the vectors of all of the words in the sentence (**Mean(All)**), or use one of three methods that puts more weight on the content words of the sentence. **Mean(Content)** averages only the content word vectors of the sentence. **TF-IDF** divides each word’s embedding by its frequency (operationalized as the number of times it occurs in a large corpus, like Wikipedia), essentially calculating a weighted average where more frequent words (e.g.,‘the’) are given less weight (Lintean et al., 2010). **SIF** also computes a weighted average for each sentence, but then removes the projection of the first principal component of the singular value decomposition of the sentence embedding matrix, which removes “semantically meaningless directions” (Arora et al., 2017). Finally, we use sent2vec, which works similarly to word2vec but on the sentence level: it directly learns sentence representations that predict neighboring sentences (Pagliardini et al., 2017). Using these methods, we obtain one coherence score per participant for

[1] We focus on LSA, GLoVE, word2vec, and sent2vec in the main text to align with past work, but Appendix D shows that results are qualitatively similar for the more modern and contextualized ELMo and BERT embeddings.

each combination of sentence and word embedding models, plus one for sent2vec (13 total). We refer the reader to Corcoran et al. (2018), Iter et al. (2018), and Hitczenko et al. (2020) for more details on embedding models.

3.3.3 Other Speech Measures

In addition to automated coherence and tangentiality, we calculated the average sentence length (number of words per sentence) for each participant as well as a measure of each participant's lexical diversity. For lexical diversity, we used the moving average type-to-token ratio (MATTR) with a window of 50 words (Covington and McFall, 2010), which calculates the word type to word token ratio over each overlapping window of 50 words, and then averages them to obtain one overall measure of lexical diversity.

3.4 Analyses and Predictions

First, we ask whether there are group differences in coherence and tangentiality between the CHR and HC groups by running two sample t-tests as in past work. We expect to observe significant differences between the groups, with the HC group being more coherent and less tangential than the CHR group.

Second, we ask whether these automated scores correlate with item scores on the SIPS clinical interview related to disorganized speech or thought disorder, as well as with overall symptomatology measured by the SIPS. Where tested, past work has reported mixed findings, with some seeing correlations between automated measures and symptom severity (Just et al., 2019), but many not (Corcoran et al., 2018; Iter et al., 2018). As these automated measures are intended to measure thought disorder, we expect to find that worse symptom severity (i.e. higher symptom scores) is associated with worse coherence scores (i.e. lower coherence scores), especially for P5 (“disorganized communication”).

Finally, we ask whether these automated linguistic scores relate to other linguistic properties of the speech (i.e., sentence length and lexical diversity) as well as sociodemographic factors of the individuals speaking (i.e., scholastic achievement/general intelligence, education, race, etc.). We calculate correlations for continuous measures and compare groups for discrete measures.

4 Results

4.1 Question 1: Are there CHR vs HC group differences in coherence/tangentiality?

As shown in Table 2, we find significant differences in coherence between the CHR and HC groups in 3 out of 13 of the methods we report (see Appendix C.1 for difference plots). However, it is important to note that these differences may be spurious based on multiple comparisons; with a Bonferonni correction (α = .004), these differences no longer reach significance. In 6 out of the remaining 10 methods, the healthy controls have numerically, but non-significantly, greater coherence scores than the CHR group. In the remaining 4 methods, the groups show near identical scores.

For tangentiality, we do not find any significant differences in tangentiality between the CHR and HC groups (Table 3). As a result, we do not conduct additional analyses of this measure.

These results suggest that these automated measures of thought disorder are very sensitive to the particular method used to derive it. Notably, previous work has not found any particular method to be consistently successful in separating groups. One of the methods where we find a significant difference is also successful in Just et al. (2019), who find significant coherence differences using TF-IDF GLoVE and no significant differences in tangentiality. However, Iter et al. (2018) only found differences in coherence using SIF word2vec, while other papers (Bedi et al., 2015; Elvevåg et al., 2007; Corcoran et al., 2018) have found significant differences using LSA Mean(All).

Overall, while we do not find group differences in tangentiality, we do find the predicted group differences in coherence between CHR and HC in a subset of cases. However, more work needs to be done to understand whether these are meaningful effects and what they reflect. To this end, for the remainder of the paper, we ask whether these automated linguistic methods of coherence relate to symptoms or other linguistic/sociodemographic factors. For these analyses, we zoom in on the sentence/word embedding models that separate CHR from HC groups. We present GLoVE Mean(Content) analyses in the main text; all other analyses are presented in Appendix C.

Sentence	Word	CHR mean	HC mean	CHR sd	HC sd	T-stat	P-value
Mean	LSA	0.58	0.60	0.07	0.06	-1.12	0.13
(All)	word2vec	0.79	0.80	0.04	0.04	-0.94	0.18
	GLoVE	**0.92**	**0.93**	**0.02**	**0.02**	**-1.89**	**0.03**
Mean	LSA	0.31	0.30	0.07	0.06	0.77	0.77
(Content)	word2vec	0.63	0.65	0.05	0.06	-1.17	0.12
	GLoVE	**0.81**	**0.82**	**0.04**	**0.03**	**-1.74**	**0.04**
TF-IDF	LSA	0.42	0.44	0.07	0.07	-1.05	0.15
	word2vec	0.75	0.76	0.04	0.05	-0.71	0.24
	GLoVE	**0.87**	**0.89**	**0.03**	**0.02**	**-2.14**	**0.02**
SIF	LSA	0.08	0.08	0.09	0.07	0.23	0.59
	word2vec	0.03	0.02	0.06	0.06	0.96	0.83
	GLoVE	0.05	0.04	0.06	0.06	1.08	0.86
sent2vec	sent2vec	0.47	0.48	0.04	0.05	-1.21	0.11

Table 2: Coherence results. We see a significant difference between groups in 3/13 methods (in bold), though these differences are no longer significant using the Bonferroni correction for multiple comparisons (α = 0.004).

Sentence	Word	CHR mean	HC mean	CHR sd	HC sd	T-stat	P-value
Mean	LSA	-0.007	-0.018	0.03	0.05	1.18	0.88
(All)	word2vec	-0.007	-0.01	0.02	0.02	0.68	0.75
	GLoVE	-0.002	-0.004	0.01	0.01	0.94	0.83
Mean	LSA	-0.017	-0.013	0.04	0.03	-0.5	0.31
(Content)	word2vec	-0.013	-0.016	0.02	0.03	0.57	0.72
	GLoVE	-0.007	-0.01	0.02	0.02	0.76	0.78
TF-IDF	LSA	-0.011	-0.017	0.04	0.04	0.59	0.72
	word2vec	-0.008	-0.011	0.02	0.03	0.52	0.70
	GLoVE	-0.004	-0.006	0.01	0.02	0.8	0.79
SIF	LSA	-0.02	-0.029	0.07	0.07	0.59	0.72
	word2vec	-0.029	-0.039	0.05	0.08	0.66	0.74
	GLoVE	-0.034	-0.04	0.06	0.07	0.41	0.66
sent2vec	sent2vec	-0.013	-0.011	0.03	0.03	-0.34	0.37

Table 3: Tangentiality results. We observe no significant differences between the CHR vs. HC groups.

4.2 Question 2: Do automated coherence scores correlate with symptoms?

Do lower coherence scores (within the CHR group) relate to worse thought disorder? We examine this using symptoms in the SIPS that are related to thought disorder. As shown in Figure 1, we find generally poor correlations. The computational measures intended to measure thought disorder do not show any correlation with currently used clinical interviews measuring thought disorder in the CHR group. This result adds to a growing but mixed literature on the relationship between automated linguistic measures and the symptoms they are intended to measure.

Of past work that has reported correlations, Corcoran et al. (2018) and Iter et al. (2018) found no correlation between coherence scores and clinical interview symptoms, while Just et al. (2019) found their coherence measures did correlate negatively with symptom severity as measured by the Scale for the Assessment of Negative Symptoms (Andreasen, 1989). Bedi et al. (2015) included coherence in a canonical correlation identifying the maximal correlation between a linear combination of 3 linguistic features – coherence, maximal word phrase length, and number of determiners – and a linear combination of the positive and negative SIPS subscales. They found an overall positive correlation, but it's unclear what role coherence played in driving this correlation. Taken together, our results and previous results suggests that coherence scores are not reliably related to clinical measures of thought

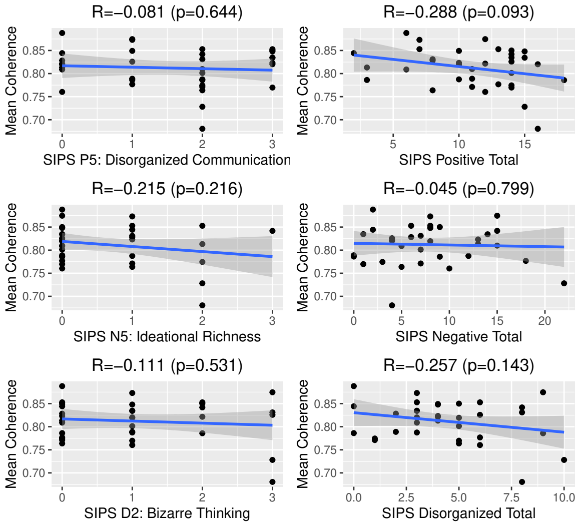

Figure 1: Correlation between mean coherence scores and relevant SIPS subitems and SIPS symptom totals. The lines show the estimated linear regression models and the shading shows 95% confidence intervals. Each point represents one participant.

disorders; however, a high-powered investigation is warranted.

4.3 Question 3: Do automated coherence scores correlate with linguistic features of speech samples or sociodemographic factors of the speaker?

If these measures are not capturing thought disorder symptoms, what are they measuring? To examine this issue, we examine the relationship of these computational measures to surface linguistic features of the speech samples and sociodemographic factors of the speakers. We focus on three features that show a significant relationship to this 'coherence' measure – sentence length, a measure of general intelligence, and racial identity of the speaker – and report non-significant correlations in Appendix C.

4.3.1 Sentence length

We find a significant positive correlation between average sentence length and automated measures of coherence: that is, longer sentences are measured as more coherent (*r* (75)=0.66; p<0.001) all else being equal (Figure 2).

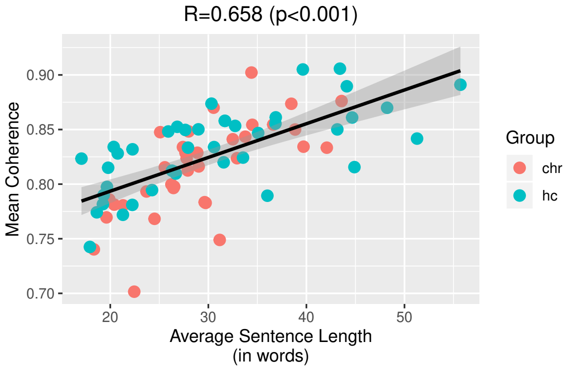

Figure 2: Correlation between mean coherence scores and average sentence length. The line show the estimated linear regression model and the shading shows 95% confidence intervals. Each point represents one participant, colored by CHR status.

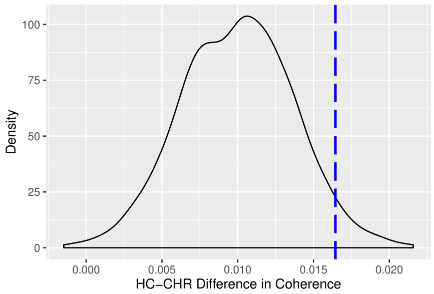

Figure 3: Length-only baseline distribution of HC-CHR differences in coherence (1000 samples). The vertical line shows the location of the true difference in this distribution.

This raises the possibility that the observed CHR-HC difference simply reflects differences in average sentence length (CHR mean: 29 words/sentence; HC mean: 31 words/sentence). To test for this possibility, we calculated the distribution of group differences predicted by a length-only baseline. In particular, we use a Monte Carlo method to compare the group differences in coherence scores against a surface-only baseline based on sentence length. We estimate this baseline by randomly replacing each word in our corpus – generating random word strings matching the length of our participants' productions. We then recalculate the group difference, providing an estimate of the difference in coherence scores predicted to occur by differences in sentence length alone. This procedure is repeated 1000 times to estimate the distribution of baseline differences. If the difference in coherence scores is based on the content of what participants are saying, then the observed difference should lie at the extreme tail of this baseline distribution.

As shown in Figure 3, only 3.9% of the runs had a more extreme HC-CHR difference than observed in the original participant data (shown with the blue dotted line), suggesting that there is something in the linguistic content that is contributing to the difference observed above and beyond the sentence lengths. However, we also note that the baseline difference is always greater than zero. Even though we completely randomized the content of the speech in both groups, the sentence length differences observed between the groups still resulted in greater coherence for the HC group, suggesting that sentence length plays a large role in the observed outcomes. Group differences can be obtained without considering any of the linguistic content spoken by participants. This is not a good property for this measure.

4.3.2 WRAT scores

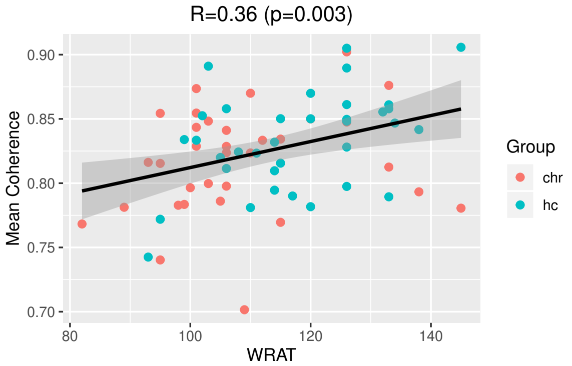

Figure 4: WRAT vs. Coherence Scores. The line shows the estimated linear regression model and the shading shows 95% confidence intervals. Each point represents one participant, colored by CHR status.

Next we observe in Figure 4 that higher coherence is associated with higher scores on the WRAT, a measure of scholastic achievement, associated with general intelligence ($r(75)$=0.36; p<0.001). Those with higher WRAT scores tend to produce more coherent speech (though it could also be that they tend to produce longer sentences). As with sentence length, we cannot make conclusions about causality here. However, this finding again reduces our confidence in the use of this computational measure as an index of thought disorder. Future work utilizing this coherence measure must control for the correlation with WRAT.

4.3.3 Race

Finally, as shown in Figure 5, coherence scores may be correlated with racial identity. In our sample, Black speakers' speech was measured as less 'coherent' than that of White speakers' (all else being equal). However, it is critical to note that these analyses were based on a small numbers of participants (including just 7 Black participants in the healthy control group), and this warrants a high-powered study directly investigating the relationship between coherence models and racial identity.

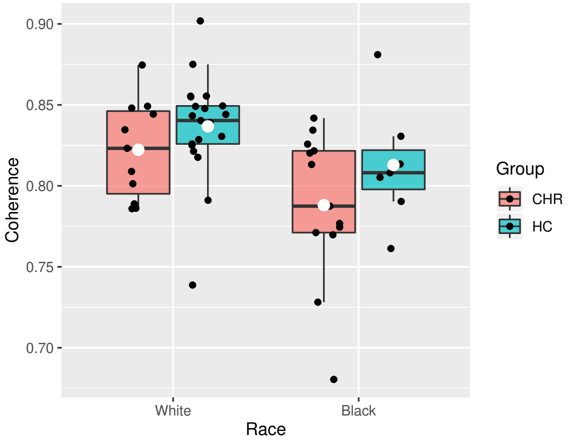

Figure 5: Coherence scores by race and clinical group. Each black point represents the mean coherence score of one individual grouped by their race (Black vs. White) and clinical status (HC vs. CHR). The four white dots represent the mean value for each group.

Nonetheless, this is a troubling finding that calls for a deeper dive into understanding what factors these computational measures are sensitive to before they can be used clinically. In particular, this result parallels other findings from the computational world - e.g., that ASR systems and computer vision systems work less well for Black individuals than White individuals (Koenecke et al., 2020; Buolamwini and Gebru, 2018). As the field develops, it is crucial to place analyses such as these front and center to ensure that this does not become another domain that perpetuates existing systemic biases.

4.3.4 Relationship between effects

In summary, we observed relationships between automated coherence scores and (1) average sentence length, (2) intelligence/achievement scores as measured by the WRAT, and (3) racial identity. To get a better understanding of these effects and their interrelationship, we fit a linear model predicting average coherence scores from average sentence length, WRAT score, and race. We found that coherence scores were significantly higher for participants with longer average sentences ($\hat{\beta} = 0.001$, p = 0.009), but found no other significant effects – suggesting that the relationships between coherence and racial identity as well as scholastic achievement reflected correlations of these factors with sentence length. Indeed, White speakers produced longer sentences than Black speakers (White mean: 32 words, Black mean: 26 words) and individuals with higher WRAT scores produced longer sentences and passages than those with lower WRAT scores ($r = 0.3$; p = 0.01).

Overall, the findings in this section make clear that there is more work to be done to ensure that group differences reported in this body of literature reflect the differences in thought disorder they are meant to reflect, especially given non-correlations with SIPS symptoms measuring thought disorder. Of all of the factors, including thought disorder symptoms, sentence length was the factor that most correlated with coherence scores. Our results not only suggest that these measures may not be measuring what we think they are, but that this could have harmful downstream consequences (e.g., predicting lower coherence scores for Black speakers than White speakers).

5 Discussion

We tested methods of quantifying coherence and tangentiality, applying them to speech samples produced by individuals at clinical high-risk for psychosis. We found group differences between the CHR and HC groups for a subset of the tested methods (3 out of 13, significant only at uncorrected $\alpha = .05$). Surprisingly, we did not find significant correlations with items from clinical interviews that measure thought disorder (i.e. what these measures are meant to capture). In order for these measures to be useful clinically, it is important to show construct validity – that the measures actually index what they are meant to, rather than other features of the speech/speaker. This is especially true as the methods we use here have been shown to exhibit potentially harmful biases in other work. To this end, our final exploratory analyses were designed to better understand what these measures *are* capturing. We found correlations with sentence length, WRAT scores, and race, which suggests that these methods partially reflect properties that these measures are *not* intended to measure. These results suggest that there is substantial and careful work that needs to be done for these methods to be useful clinically.

5.1 Group differences are sensitive to the methods used and vary across papers

Replicating past work, we find group differences in coherence between the CHR vs. HC groups. However, as in past work using multiple word/sentence embedding methods, we find this difference in a subset of cases, suggesting this finding is sensitive to the particular method used. We fail to find group differences in tangentiality between CHR vs. HC groups. While these results overlap with those of one paper (Just et al., 2019), they do not overlap with other work (Corcoran et al., 2018; Bedi et al., 2015; Elvevåg et al., 2007; Iter et al., 2018) (and there is substantial variation within these papers as well). We offer two possible factors underlying these diverging findings. First, each paper has made different methodological decisions. Research differs in: the kinds of speech samples collected (shorter vs. longer length, individuals with vs. at-risk for psychosis); the analysis methods (some researchers remove fillers but others do not); and modeling decisions (some compare similarity between sentences, while others compare similarity between windows of words of length N), and so forth. These differences could easily give rise to differences across studies. Second, the true effect size could be quite small to begin with, especially in the CHR group who displays attenuated symptoms, and we know there is substantial heterogeneity between individuals. Some healthy individuals show linguistic disturbances, while some individuals with psychosis do not show any or show disturbances of almost opposite nature (e.g., perseveration, staying fixed on a single topic) (Andreasen, 1979). The substantial heterogeneity and differing sample sizes observed could also give rise to substantial differences between studies.

Overall, while past work has highlighted successes in the important goal of establishing differences between groups, it is critical to acknowledge where this line of work has fallen short: small changes in the particular methods used can substantially change the outcome, and which methods are successful varies unpredictably between studies. Moving forward, it may be useful to better align the methodological, analytical, and modeling choices across studies to better understand what gives rise to these differences. Due to the heterogeneity observed, it may also be worth focusing less on group differences and more on symptoms and outcome measures. In addition, as these methods continue to develop, it may be easier to accurately and more transparently evaluate their performance, by testing them on speech samples that are known to contain vs. not contain the particular studied linguistic disturbances. This shift in focus may allow us to gain a better understanding of what these measures reflect and how they can be useful on an individual basis.

5.2 Lack of correlations with SIPS thought disorder symptoms

We did not find correlations with the SIPS items that are thought to measure disorganized language and thought disorder. We note that is possible that, with 36 CHR participants, we did not have sufficient power to detect existing correlations with SIPS symptoms. However, this null finding adds to a growing literature of inconsistent findings, with some past work finding correlations with thought disorder and/or other clinical symptoms, but other past work failing to find these same correlations. This underscores the importance of doing careful work to establish construct validity with automated measures. Rigorous testing is needed to verify that novel measures relate to the properties of speech and cognition that they are intended to index.

5.3 Coherence scores correlate with sentence length and speaker sociodemographics

Perhaps most troublingly, we find that the differences in coherence between groups partially reflect irrelevant surface properties of the speech and sociodemographic qualities of the speakers. In fact, the single factor that best correlated with these measures was the length of the sentence. On the one hand, this raises concern that we are not measuring what we think we are. On the other hand, due to the fact that other factors (e.g., racial identity, achievement and intelligence, as measured by the WRAT) correlate with differences in average sentence length, this could have downstream harmful consequences (e.g., rating Black speakers as less coherent than White speakers due to differences unrelated to coherence). Overall, these results provide evidence that there is substantial work to be done to understand what these measures reflect to a degree where they can be used clinically.

5.4 Ethics and Broader Impacts Statement

Ethical use of computational measures like those studied here – especially in the high-stakes context of clinical care – requires us to devote substantial attention to potential biases in our measures. To that end, we recommend that future researchers in this area conduct and report analyses examining relations to symptoms, as well as the linguistic and sociodemographic factors studied here. This will allow us to gain a better understanding of what these measures reflect, and make sure that they are developed to be equally useful for all. To this end, we provide all of our code to hopefully facilitate these crucial cross-study comparisons.[2]

5.5 Conclusion

Linguistic disturbances characterize psychosis, yet they have been understudied in the field, largely due to how time-intensive it is to obtain meaningful and reliable measures of them. Automated linguistic methods have the potential to transform the scale at which we can study and identify these linguistic disturbances. However, with this strength come some downsides that the field must address: these methods are less transparent and can be harder to interpret. Facing these challenges head-on will allow us to develop a stronger, more ethical practice in this important and promising area of research.

Acknowledgements

This work was supported by the National Institutes of Health (NIH R21 MH119677 and T32 NS047987) and the Canadian Institutes of Health Research (CIHR DFS-152268). We thank Rachel Ostrand and Ewan Dunbar for their helpful suggestions and feedback on this work.

References

Nancy C Andreasen. 1979. Thought, language, and communication disorders: II. Diagnostic significance. *Archives of General Psychiatry*, 36(12):1325–1330.

Nancy C Andreasen. 1986. Scale for the assessment of thought, language, and communication (TLC). *Schizophrenia Bulletin*, 12(3):473.

Nancy C Andreasen. 1989. The scale for the assessment of negative symptoms (SANS): Conceptual and theoretical foundations. *The British Journal of Psychiatry*, 155(S7):49–52.

Nancy C Andreasen and William M Grove. 1986. Thought, language, and communication in schizophrenia: Diagnosis and prognosis. *Schizophrenia Bulletin*, 12(3):348–359.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *5th International Conference on Learning Representations, ICLR 2017*.

[2] github.com/khitczenko/chr_coherence

Carrie E Bearden, Keng Nei Wu, Rochelle Caplan, and Tyrone D Cannon. 2011. Thought disorder and communication deviance as predictors of outcome in youth at clinical high risk for psychosis. *Journal of the American Academy of Child & Adolescent Psychiatry*, 50(7):669–680.

Gillinder Bedi, Facundo Carrillo, Guillermo A Cecchi, Diego Fernández Slezak, Mariano Sigman, Natália B Mota, Sidarta Ribeiro, Daniel C Javitt, Mauro Copelli, and Cheryl M Corcoran. 2015. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia*, 1(1):1–7.

Eugen Bleuler. 1950. Dementia praecox or the group of schizophrenias. *International Universities Press*.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 4356–4364.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Cheryl M Corcoran, Facundo Carrillo, Diego Fernández-Slezak, Gillinder Bedi, Casimir Klim, Daniel C Javitt, Carrie E Bearden, and Guillermo A Cecchi. 2018. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, 17(1):67–75.

Cheryl M Corcoran, Vijay A Mittal, Carrie E Bearden, Raquel E Gur, Kasia Hitczenko, Zarina Bilgrami, Aleksandar Savic, Guillermo A Cecchi, and Phillip Wolff. 2020. Language as a biomarker for psychosis: A natural language processing approach. *Schizophrenia Research*, 226:158–166.

Michael A Covington and Joe D McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.

Arsime Demjaha, Sara Weinstein, Daniel Stahl, Fern Day, Lucia Valmaggia, Grazia Rutigliano, Andrea De Micheli, Paolo Fusar-Poli, and Philip McGuire. 2017. Formal thought disorder in people at ultra-high risk of psychosis. *BJPsych Open*, 3(4):165–170.

Brita Elvevåg, Peter W Foltz, Daniel R Weinberger, and Terry E Goldberg. 2007. Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophrenia Research*, 93(1-3):304–316.

Michael B First. 1997. Structured Clinical Interview for DSM-IV Axis I disorders. *Biometrics Research Department*.

Tina Gupta, Susan J Hespos, William S Horton, and Vijay A Mittal. 2018. Automated analysis of written narratives reveals abnormalities in referential cohesion in youth at ultra high risk for psychosis. *Schizophrenia Research*, 192:82–88.

Kasia Hitczenko, Vijay A Mittal, and Matthew Goldrick. 2020. Understanding language abnormalities and associated clinical markers in psychosis: The promise of computational methods. *Schizophrenia Bulletin*.

Dan Iter, Jong Yoon, and Dan Jurafsky. 2018. Automatic detection of incoherent speech for diagnosing schizophrenia. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 136–146.

Brick Johnstone, Charles D Callahan, Cynthia J Kapila, and Dawn E Bouman. 1996. The comparability of the WRAT-R reading test and NAART as estimates of premorbid intelligence in neurologically impaired patients. *Archives of Clinical Neuropsychology*, 11(6):513–519.

Sandra Just, Erik Haegert, Nora Kořánová, Anna-Lena Bröcker, Ivan Nenchev, Jakob Funcke, Christiane Montag, and Manfred Stede. 2019. Coherence models in schizophrenia. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 126–136.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Gina R Kuperberg. 2010. Language in schizophrenia part 1: An introduction. *Language and Linguistics Compass*, 4(8):576–589.

Thomas K Landauer, Peter W Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284.

Mihai Lintean, Cristian Moldovan, Vasile Rus, and Danielle McNamara. 2010. The role of local and global weighting in assessing the semantic similarity of texts using latent semantic analysis. In *Twenty-Third International FLAIRS Conference*.

Thomas H McGlashan, Barbara C Walsh, Scott W Woods, J Addington, K Cadenhead, T Cannon, and E Walker. 2001. Structured Interview for Psychosis-risk Syndromes. *New Haven, CT: Yale School of Medicine*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tandy J Miller, Thomas H McGlashan, Scott W Woods, Kelly Stein, Naomi Driesen, Cheryl M Corcoran, Ralph Hoffman, and Larry Davidson. 1999. Symptom assessment in schizophrenic prodromal states. *Psychiatric Quarterly*, 70(4):273–287.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2017. Unsupervised learning of sentence embeddings using compositional n-gram features. *arXiv preprint arXiv:1703.02507*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GLoVE: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Diana O Perkins, Clark D Jeffries, Barbara A Cornblatt, Scott W Woods, Jean Addington, Carrie E Bearden, Kristin S Cadenhead, Tyrone D Cannon, Robert Heinssen, Daniel H Mathalon, et al. 2015. Severity of thought disorder predicts psychosis in persons at clinical high-risk. *Schizophrenia Research*, 169(1-3):169–177.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Eric Roche, John Lyne, Brian O'Donoghue, Ricardo Segurado, Caragh Behan, Laoise Renwick, Felicity Fanning, Kevin Madigan, and Mary Clarke. 2016. The prognostic value of formal thought disorder following first episode psychosis. *Schizophrenia Research*, 178(1-3):29–34.

James Wilcox, George Winokur, and Ming Tsuang. 2012. Predictive value of thought disorder in new-onset psychosis. *Comprehensive Psychiatry*, 53(6):674–678.

GS Wilkinson and GJ Robertson. 2006. Wide Range Achievement Test 4 Professional Manual. *Psychological Assessment Resources*.

A More Information on the Structured Interview for Psychosis-Risk Syndromes

The SIPS is a clinical interview administered by experienced clinicians that is used to classify individuals as being at clinical high-risk for psychosis. It consists of 19 symptoms that are grouped into four symptom classes: 5 positive (P) symptoms, 6 negative (N) symptoms, 4 disorganized (D) symptoms, and 4 general (G) symptoms. Patients are rated along each of the 19 individual symptoms (scores for each individual symptom range from 0, least severe, to 6, most severe). The scores on the individual symptoms within each of the four classes are totaled to get total positive (range 0-30), negative (range 0-36), disorganized (range 0-24), and general (range 0-24) symptom scores. Our analyses focus on items P5 ("Disorganized Communication"), N5 ("Ideational Richness"), and D2 ("Bizarre Thoughts"), as well as the positive, negative, and disorganized symptom totals, as described below. We refer readers to Miller et al. (1999) and McGlashan et al. (2001) for more information about the SIPS.

Positive Symptoms [0-30]: There are five positive symptoms: P1 (Unusual Thought Content/Delusional Ideas), P2 (Suspiciousness/Persecutory Ideas), P3 (Gradiose Ideas), P4 (Perceptual Abnormalities/Hallucinations), and P5 (Disorganized Communication).

P5- Disorganized Communication [0-6]: The types of inquiries used to establish the score include:

- Do people ever tell you that they can't understand you? Do people ever seem to have difficulty understanding you?
- Are you aware of any ongoing difficulties getting your point across, such as finding yourself rambling or going off track when you talk?
- Do you ever completely lose your train of thought or speech, like suddenly blanking out?

Negative Symptoms [0-36]: There are six negative symptoms: N1 (Social Anhedonia), N2 (Avolition), N3 (Expression of Emotion), N4 (Experience of Emotions and Self), N5 (Ideational Richness), and N6 (Occupational Functioning).

N5- Ideational Richness [0-6]: The types of inquiries used to establish the score include:

- Do you sometimes find it hard to understand what people are trying to tell you because you don't understand what they mean?
- Do people more and more use words that you don't understand?

Disorganized Symptoms [0-24]: There are four disorganized symptoms: D1 (Odd Behavior or Appearance), D2 (Bizarre Thinking), D3 (Trouble with Focus and Attention), and D4 (Impairment in Personal Hygiene). In our analyses, we use the total disorganized score (range: 0-24), as well as the D2 item (bizarre thinking).

D2- Bizarre Thinking [0-6]: The types of inquiries used to establish the score include:

- Do people ever say your ideas are unusual or that the way you think is strange or illogical?

General Symptoms [0-24]: We do not include these symptoms in our analyses, but there are four general symptoms: G1 (Sleep Disturbance), G2 (Dysphoric Mood), G3 (Motor Disturbances), and G4 (Impaired Tolerance to Normal Stress).

B Complete Question Prompts

- **Challenge**: Looking back over your life, what do you think is the single greatest challenge you have ever faced? Tell me the story of that challenge, what it is or was, how did the challenge or problem develop, and how did you address or deal with the challenge or problem?

- **Self-Defining**: A self-defining memory is a scene or an episode from your life that was very important for how you see yourself. This would be something that happened at least one year ago that you have thought about many times since it happened so that the memory of it is clear and familiar to you. This scene or episode helps you know who you are as a person. You might even tell this story to a friend if you wanted to help them understand you better. I'd like you to take a moment to think of a self-defining memory like this and then tell me the story of that memory and specifically what happened, when and where it happened, and who was involved?

- **Turning Point**: In most people's lives we experience episodes that change the direction of our lives or change how we see ourselves in some important way. We call those memories turning points. Looking back over your life, there may be a few key moments that stand out as turning points or episodes that marked an important change in you or your life story. I'd like you to identify a particular memory that you see as a turning point in your life and then tell me the story about that turning point: what happened, when and where it happened, and who was involved?

- **Unusual**: Next I'll ask you about an unusual experience that you might have had. Any unusual, strange or profound things that are hard to explain, for example, some coincidences, supernatural events, seeing visions of spirits, feeling like you're the center of attention, like you have special powers, or like one of your dreams had really happened. These experiences might be difficult to explain and might feel like the world is not as it seems or like your mind is playing tricks on you in some way. Take a moment to think of an unusual experience like this and then tell me the story of that experience: what happened, when and where it happened, and who was involved?

C Additional Analyses: Participant's Main Response with Fillers Removed

The main text reports results from running the participant's first main response, with fillers removed. This section provides additional analyses that were omitted from the main text, including correlations for all three models that were found to be significant (GLoVE TF-IDF, GLoVE Mean(All), and GLoVE Mean(Content)), as well as non-significant correlations (e.g., for age and education).

C.1 Group Differences

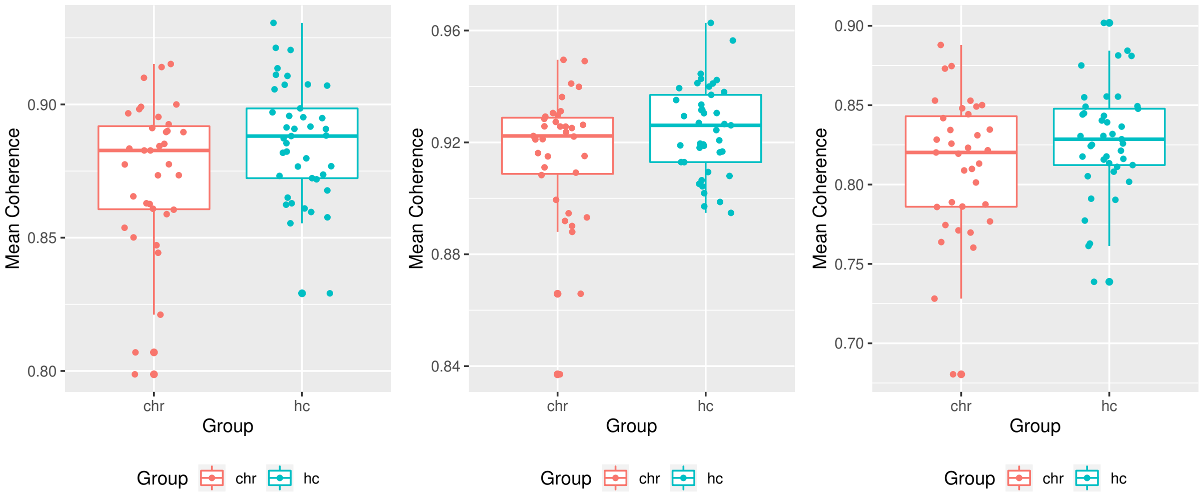

Figure 6: Coherence scores by group for each of the three methods that yield significant differences between the CHR and HC groups: GLoVE TF-IDF, GLoVE Mean(All), and GLoVE Mean(Content).

C.2 Correlations with thought disorder symptoms

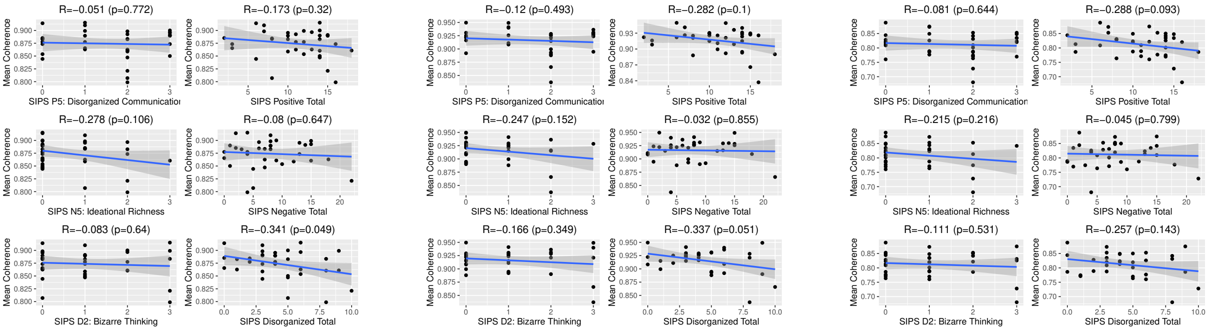

Figure 7: Correlations between coherence scores and SIPS symptoms for methods that yielded significant results (from left to right: GLoVE TF-IDF, GLoVE Mean(All), GLoVE Mean(Content)). Most correlations are not significant with one exception: GLoVE TF-IDF coherence scores correlate negatively with SIPS Total Disorganized Scores (r = -0.34, p = 0.049).

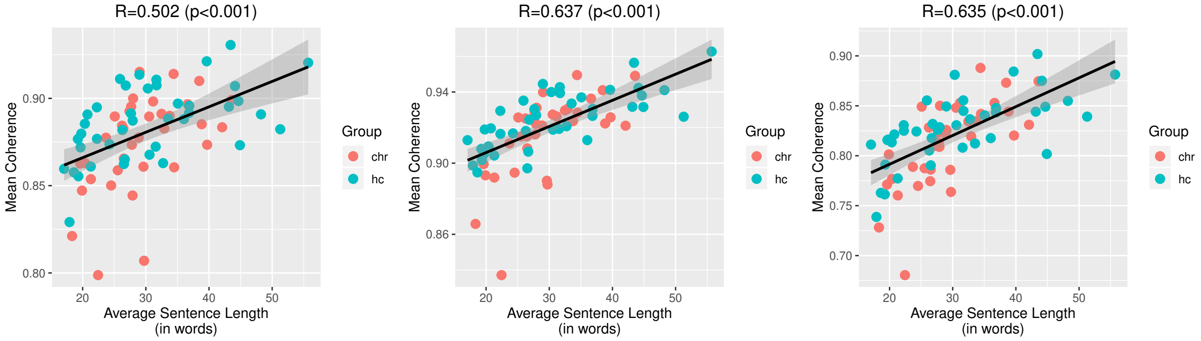

Figure 8: In all three cases (L-to-R: GLoVE TF-IDF, GLoVE Mean(All), GLoVE Mean(Content)), we observe significant positive correlations between average sentence length and average coherence with correlation coefficients ranging from 0.5 to 0.64.

C.3 Correlations with sentence length

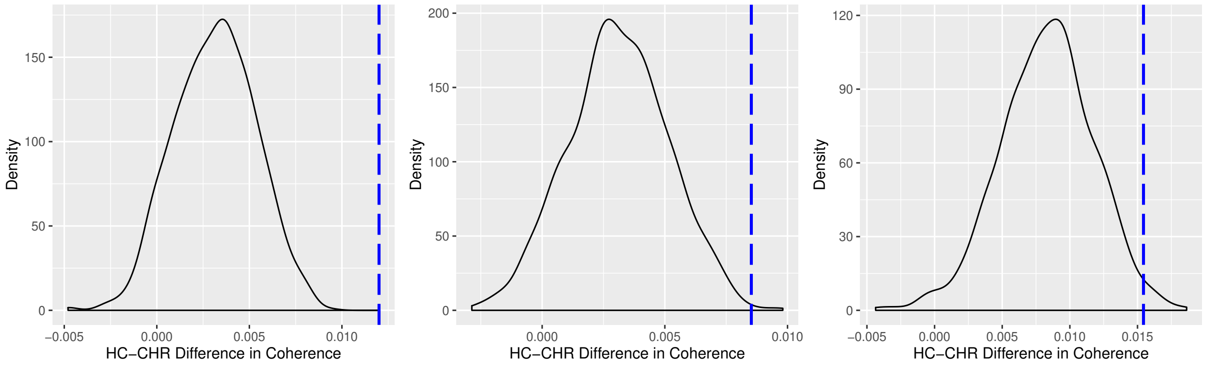

Figure 9: For each of the three significant methods (L-to-R: GLoVE TF-IDF, GLoVE Mean(All), GLoVE Mean(Content)), we randomly replace words and recalculate the coherence scores 1000 times. This graph shows the distribution of HC-CHR differences over these 1000 runs. For all three graphs, the vast majority of the differences are positive, meaning that the HC group scores as more coherent than the CHR group despite complete randomization of words. Nonetheless, the true difference (shown in the blue dotted line) is more extreme than most (GLoVE Mean(All), GLoVE Mean(Content)) or all (GLoVE TF-IDF) of the 1000 differences, suggesting that the coherence measures are partially based on the content of the speech.

C.4 Correlations with lexical diversity

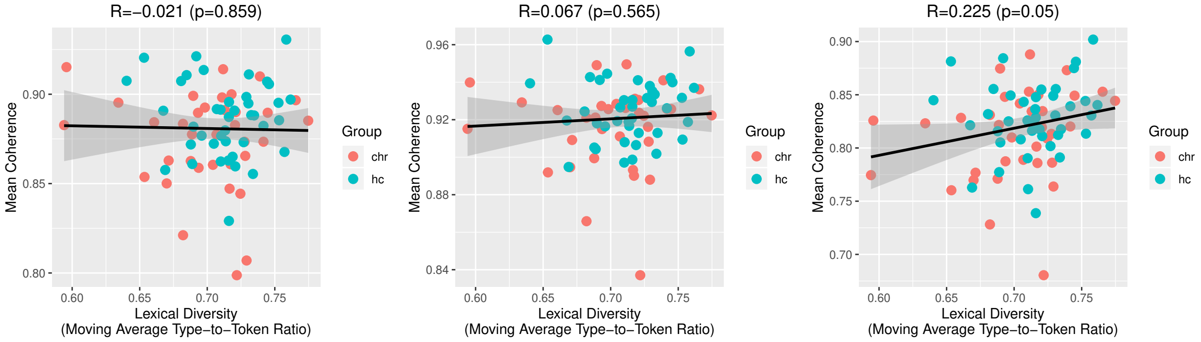

Figure 10: Left-to-right: GLoVE TF-IDF, GLoVE Mean(All), GLoVE Mean(Content). Lexical diversity, as measured by MATTR, does not correlate with coherence scores, though the correlation approaches significance for GLoVE Mean(Content), such that greater lexical diversity is associated with greater average coherence. As these automated measures calculate similarity between sentences, we might expect that repeating words would be associated with greater coherence scores. However, we do not observe this effect.

C.5 Correlations with Scholastic Achievement and Intelligence (WRAT)

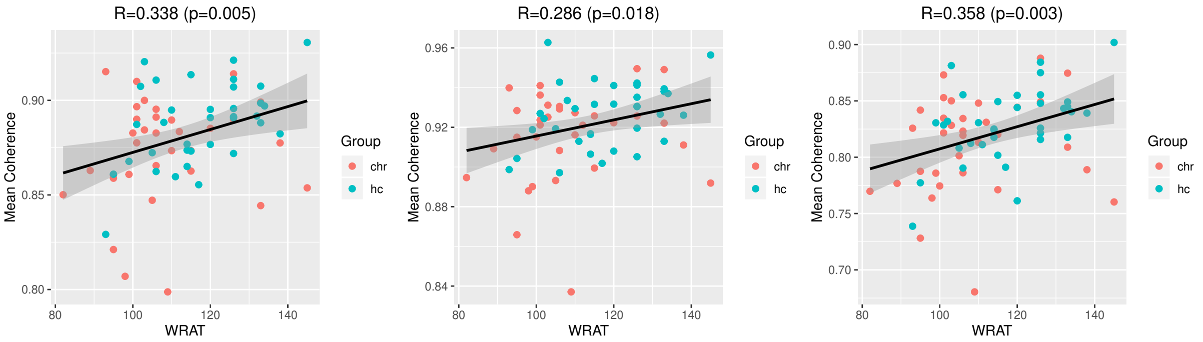

Figure 11: Left-to-right: GLoVE TF-IDF, GLoVE Mean(All), GLoVE Mean(Content). In all three methods, WRAT scores correlate positively with coherence scores, such that greater coherence is associated with higher WRAT scores. Correlation coefficients range from 0.29 to 0.36.

C.6 Race

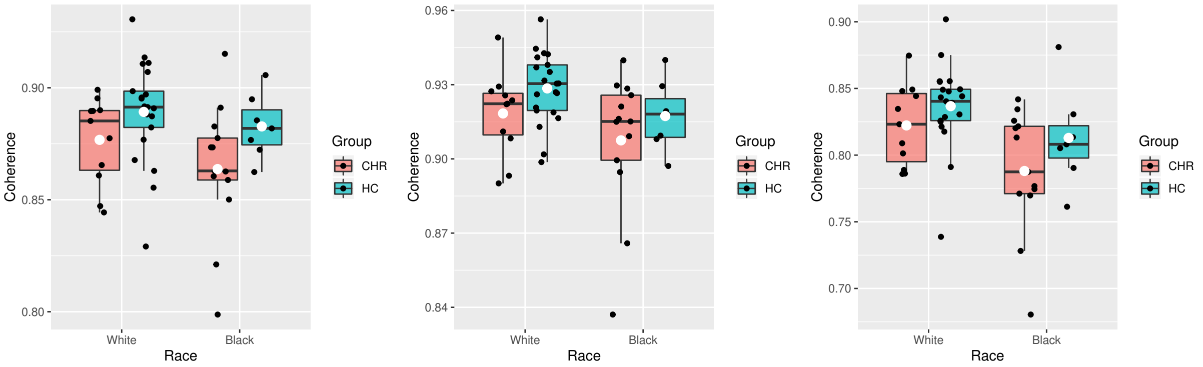

Figure 12: Left-to-right: GLoVE TF-IDF, GLoVE Mean(All), GLoVE Mean(Content). Coherence scores by CHR status (HC vs. CHR) and racial identity (Black vs. White). Across the three methods, these automated measures rate Black speakers as less coherent than White speakers.

C.7 Age

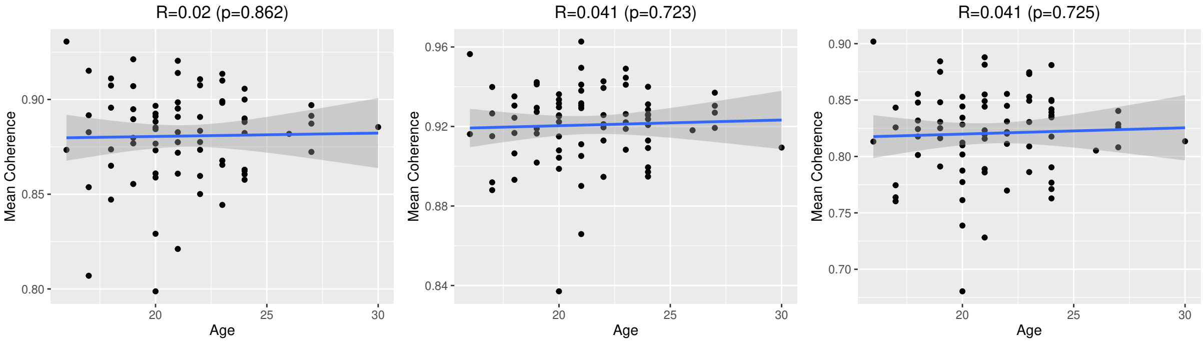

Figure 13: Left-to-right: GLoVE TF-IDF, GLoVE Mean(All), GLoVE Mean(Content). As expected, we find no correlation between age and coherence scores, although we note that this relationship has been observed in past work with older individuals scoring as more coherent (Corcoran et al., 2018).

C.8 Education

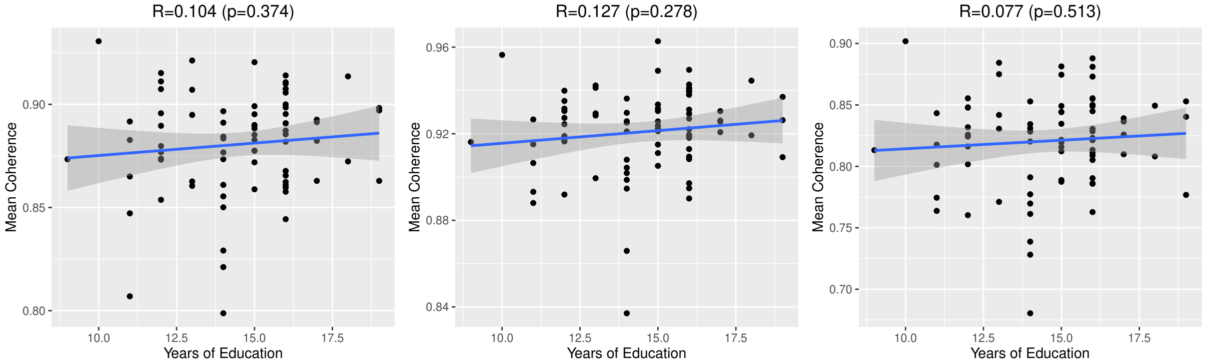

Figure 14: Left-to-right: GLoVE TF-IDF, GLoVE Mean(All), GLoVE Mean(Content). Finally, as expected, we find no correlation between level of education and coherence scores.

D Contextualized Word Embeddings (Fillers Removed)

To align with past work, the main text reports results using word2vec, GLoVE, and LSA embeddings. Here, we show similar results for the more modern, contextualized embeddings from BERT and ELMo.

The analyses for ELMo mirror those for word2vec, GLoVE, and LSA: once we have word embeddings from ELMo, we obtain sentence embeddings by averaging all of the words (Mean(All)) or just the content words (Mean(Content)) or using TF-IDF or SIF weights, which both essentially give more weight to more content-bearing words.

For BERT, however, we used a different approach, taking advantage of in-built features of the model. In particular, BERT embeddings are trained by giving the model two sentences and having the model predict whether or not one immediately followed the other (Next Sentence Prediction). That means that given a first sentence and a second sentence, we can obtain a score for how likely it is that the second sentence directly follows the first one. We used this to obtain coherence scores for each participant's speech sample, with the idea that more coherent passages will have adjacent sentences that are more predictive of one another. We obtained BERT embeddings for each word in the participant's speech. Then, directly from these embeddings, for each pair of adjacent sentences in a speech sample, we obtained the model's score for how likely it was that the second sentence followed the first sentence (BERT Next Sentence Prediction). We averaged these scores within speech samples to obtain one coherence score for each speech sample (which, in turn, were averaged to obtain one coherence score per participant).

Word	Sentence	CHR mean	HC mean	CHR sd	HC sd	T-stat	P-value
ELMo	Mean (All)	0.71	0.72	0.03	0.03	-0.86	0.20
	Mean (Content)	0.62	0.63	0.05	0.05	-0.68	0.25
	TF-IDF	0.69	0.70	0.03	0.04	-0.85	0.20
	SIF	0.02	0.01	0.05	0.05	0.98	0.84
BERT	n/a	0.977	0.983	0.05	0.05	-1.07	0.14

Table 4: Coherence results, using ELMo embeddings. We find no significant differences between groups.

Although we found no significant differences between groups, we checked whether these embeddings also exhibited the same crucial problem of being correlated with sentence length and found that they did (Figure 15). The effect is reduced using BERT, as many sentence pairs are predicted to be adjacent with scores approaching 1; however, we still observe a significant correlation between average sentence length and mean coherence, finding that participants who produce shorter sentences are relatively more likely to have lower coherence scores.

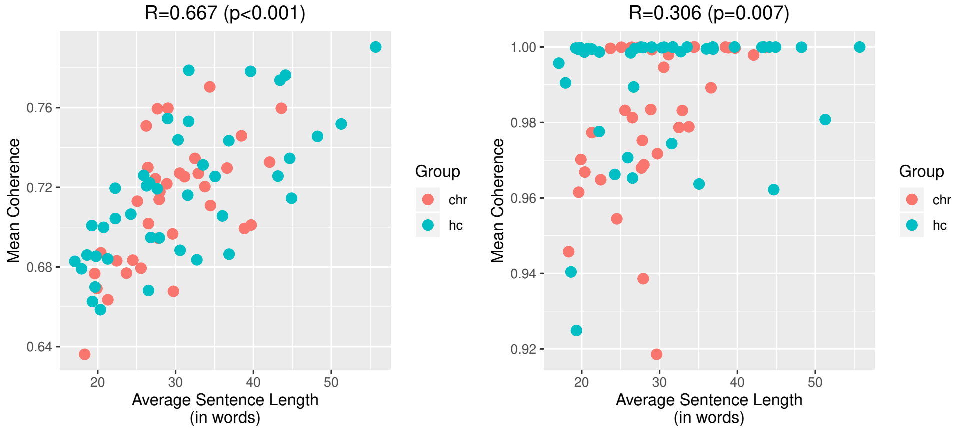

Figure 15: Left-to-right: ELMo (Mean(All)), BERT (Next Sentence Prediction). In both cases, we see a correlation between automated coherence scores and sentence length.

E Including Filler Words

In the main text, we report findings from analyzing the participants' first uninterrupted response removing filler words as in Iter et al. (2018). Here, we report results from the same speech samples, but with fillers included.

E.1 Group Differences

We test for group differences between the CHR and HC groups. As in the main text, we find significant differences in coherence for a subset of the methods used (here 2/13: GLoVE Mean(All) is no longer significant), but no significant differences in tangentiality. For the remainder of the analyses, we focus on the two methods that yielded significant differences between groups: coherence as measured by GLoVE TF-IDF and GLoVE Mean(Content).

Sentence	Word	CHR mean	HC mean	CHR sd	HC sd	T-stat	P-value
Mean (All)	LSA	0.57	0.59	0.08	0.07	-1.16	0.12
	word2vec	0.80	0.81	0.03	0.04	-0.98	0.17
	GLoVE	0.91	0.92	0.03	0.02	-1.55	0.06
Mean (Content)	LSA	0.32	0.31	0.07	0.07	0.43	0.66
	word2vec	0.64	0.66	0.05	0.06	-1.38	0.09
	GLoVE	**0.82**	**0.83**	**0.04**	**0.04**	**-1.81**	**0.04**
TF-IDF	LSA	0.42	0.44	0.07	0.08	-1.35	0.09
	word2vec	0.78	0.78	0.04	0.05	-0.38	0.35
	GLoVE	**0.88**	**0.89**	**0.03**	**0.03**	**-1.75**	**0.04**
SIF	LSA	0.10	0.10	0.09	0.07	0.18	0.57
	word2vec	0.05	0.04	0.05	0.06	1.37	0.91
	GLoVE	0.07	0.06	0.06	0.08	0.96	0.83
sent2vec	sent2vec	0.47	0.48	0.05	0.05	-1.37	0.09

Table 5: Coherence results. We see a significant difference between groups in 2/13 methods (GLoVE TF-IDF and GLoVE Mean(Content)), though these differences are no longer significant using the Bonferroni correction for multiple comparisons.

Sentence	Word	CHR mean	HC mean	CHR sd	HC sd	T-stat	P-value
Mean(All)	LSA	-0.015	-0.022	0.04	0.05	0.73	0.77
	word2vec	-0.006	-0.01	0.02	0.02	0.88	0.81
	GLoVE	-0.004	-0.004	0.02	0.01	0.1	0.54
SIF	LSA	-0.026	-0.027	0.06	0.06	0.07	0.53
	word2vec	-0.032	-0.038	0.08	0.07	0.34	0.63
	GLoVE	-0.038	-0.037	0.08	0.06	-0.02	0.49
TF-IDF	LSA	-0.016	-0.019	0.04	0.04	0.35	0.64
	word2vec	-0.005	-0.009	0.02	0.02	0.96	0.83
	GLoVE	-0.004	-0.006	0.01	0.01	0.5	0.69
Mean(Content)	LSA	-0.02	-0.014	0.04	0.03	-0.69	0.25
	word2vec	-0.011	-0.015	0.02	0.03	0.69	0.75
	GLoVE	-0.007	-0.009	0.02	0.02	0.37	0.64
sent2vec	sent2vec	-0.017	-0.011	0.03	0.03	-0.84	0.20

Table 6: Tangentiality results. As in the main text, we observe no significant differences between groups.

E.2 Correlations with thought disorder symptoms

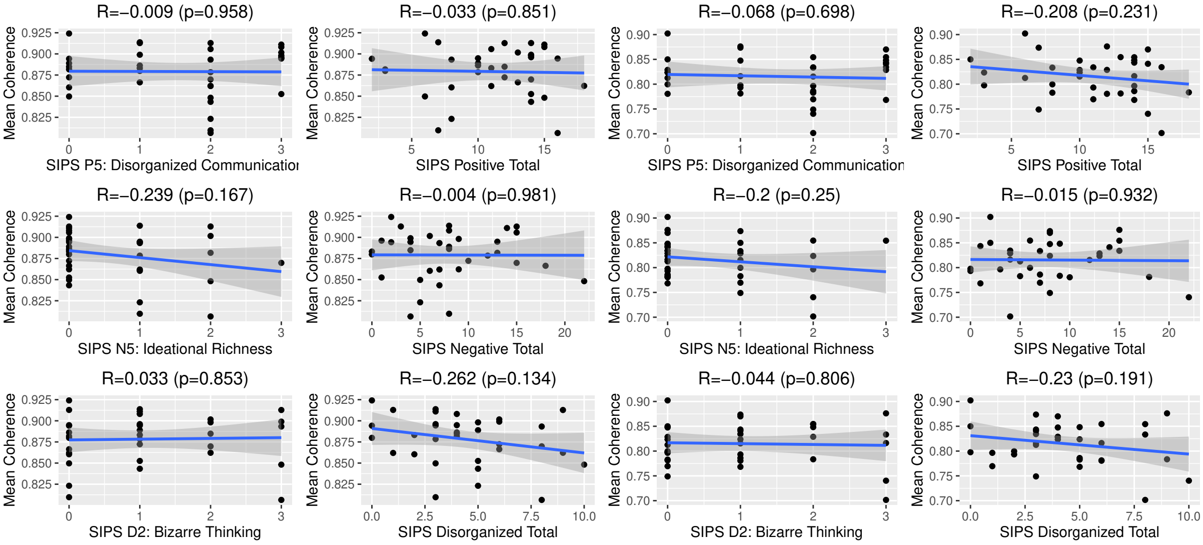

Figure 16: Correlations between coherence scores and SIPS symptoms for methods that yielded significant results (left two columns: GLoVE TF-IDF, right two columns: GLoVE Mean(Content)). As in the main text, we observe no significant correlations between SIPS symptoms and mean coherence.

E.3 Correlations with sentence length

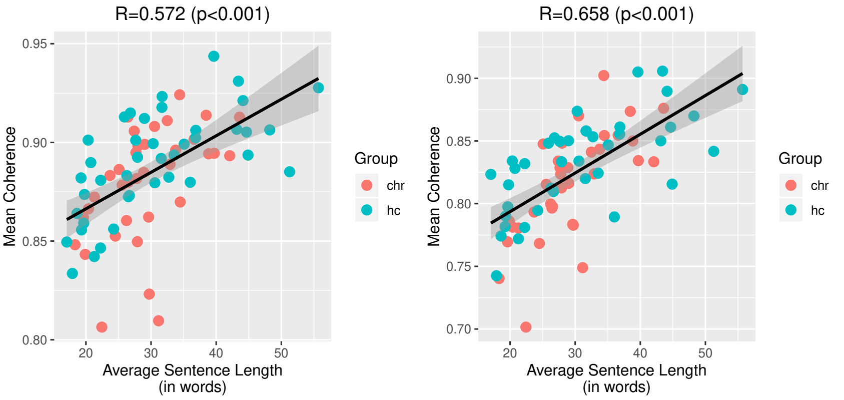

Figure 17: As in the main text, in both cases (L-to-R: GLoVE TF-IDF, GLoVE Mean(Content)), we observe significant positive correlations between average sentence length and average coherence with correlation coefficients.

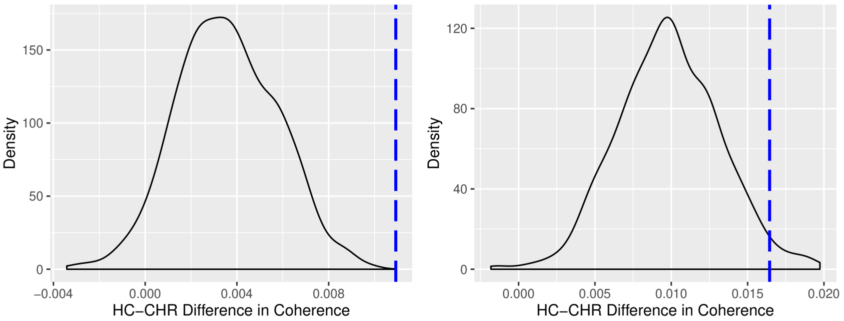

Figure 18: Left: GLoVE TF-IDF, Right: GLoVE Mean(Content). As in the case of removing fillers, in both graphs, the vast majority of the differences are positive, meaning that the HC group scores as more coherent than the CHR group despite complete randomization of words.

E.4 Correlations with lexical diversity

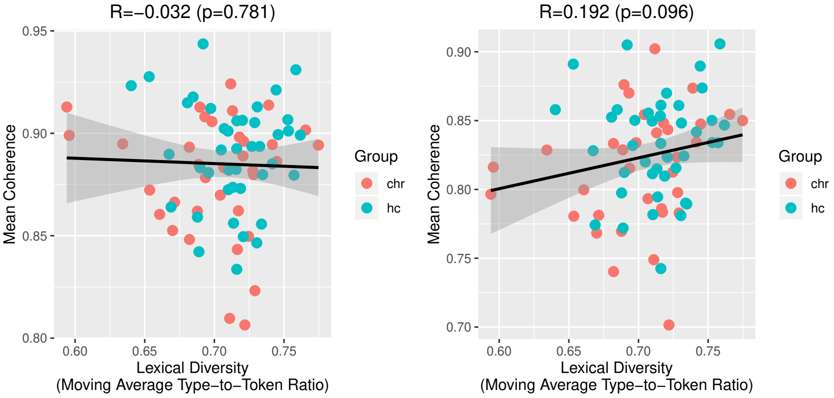

Figure 19: Left-to-right: GLoVE TF-IDF, GLoVE Mean(Content). As in the case of removing fillers, we find no significant correlation between lexical diversity, as measured by the MATTR, and mean coherence scores.

E.5 Correlations with Scholastic Achievement and Intelligence (WRAT)

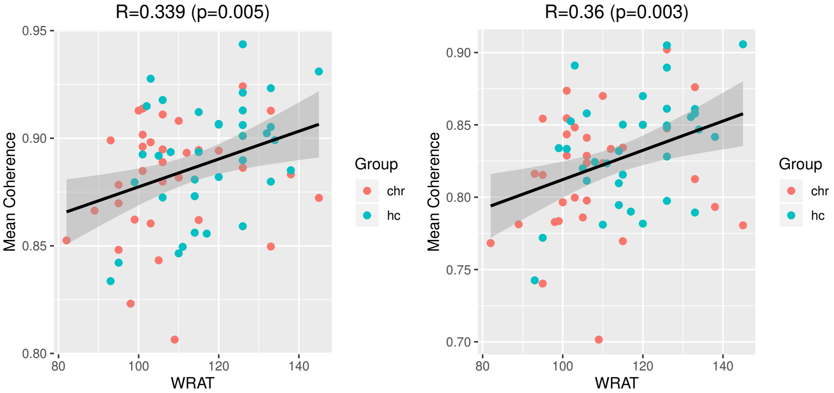

Figure 20: Left-to-right: GLoVE TF-IDF and GLoVE Mean(Content). As in the main text, WRAT scores correlate positively with coherence scores, such that greater coherence is associated with higher WRAT scores (a measure of achievement, associated with intelligence).

E.6 Race

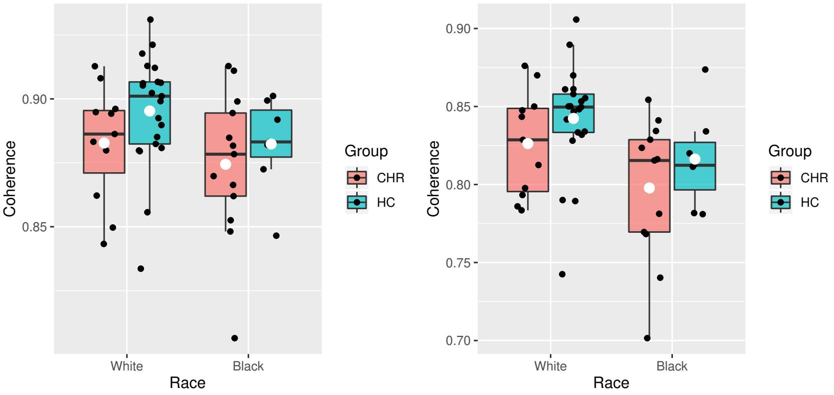

Figure 21: Left-to-right: GLoVE TF-IDF and GLoVE Mean(Content). As in the case of removing fillers, we find that both of these automated methods assign lower coherence scores to Black speakers than White speakers.

E.7 Age

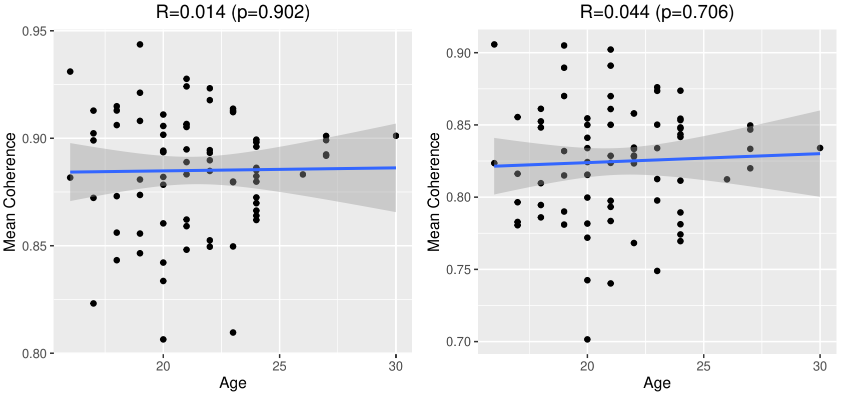

Figure 22: Left-to-right: GLoVE TF-IDF and GLoVE Mean(Content). Using both methods, we find no correlation between age and coherence scores.