Published in final edited form as:
J Biomed Inform. 2022 February ; 126: 103998. doi:10.1016/j.jbi.2022.103998.

Fully Automated Detection of Formal Thought Disorder with Time-series Augmented Representations for Detection of Incoherent Speech (TARDIS)

Weizhe Xu, BS[1], **Weichen Wang, MS**[3], **Jake Portanova, BA, BS**[1], **Ayesha Chander, MRes**[2], **Andrew Campbell, PhD**[3], **Serguei Pakhomov, PhD**[4], **Dror Ben-Zeev, PhD**[2], **Trevor Cohen, MBChB, PhD**[1,2]

[1]Biomedical Informatics and Medical Education, University of Washington, Seattle, WA

[2]Behavioral Research in Technology (BRiTE) Center, Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA

[3]Department of Computer Science, Dartmouth College, Hanover, NH

[4]Pharmaceutical Care and Health Systems, University of Minnesota, MN

Abstract

Formal thought disorder (ThD) is a clinical sign of schizophrenia amongst other serious mental health conditions. ThD can be recognized by observing incoherent speech - speech in which it is difficult to perceive connections between successive utterances and lacks a clear global theme. Automated assessment of the coherence of speech in patients with schizophrenia has been an active area of research for over a decade, in an effort to develop an objective and reliable instrument through which to quantify ThD. However, this work has largely been conducted in controlled settings using structured interviews and depended upon manual transcription services to render audio recordings amenable to computational analysis. In this paper, we present an evaluation of such automated methods in the context of a fully automated system using Automated Speech Recognition (ASR) in place of a manual transcription service, with "audio diaries" collected in naturalistic settings from participants experiencing Auditory Verbal Hallucinations (AVH). We show that performance lost due to ASR errors can often be restored through

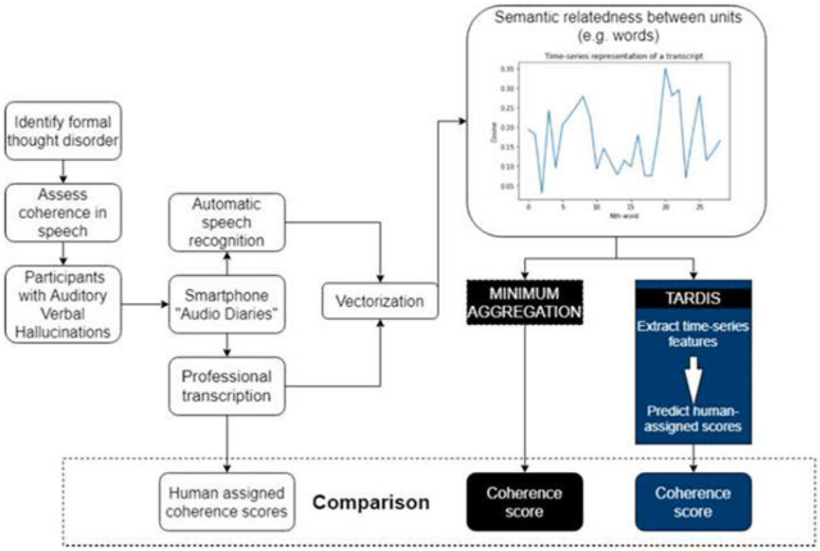
Weizhe Xu: study conceptualization, design of experiments, and coherence software development with input from **Trevor Cohen** and **Serguei Pakhomov.**
Dror Ben-Zeev and **Ayesha Chander**: data collection and curation.
Andrew Campbell and **Weichen Wang**: data collection and provided additional metadata.
Weichen Wang: additional metadata interpretation
Serguei Pakhomov: designing automated speech recognition (ASR) pipeline.
Jake Portanova: label creation for the dataset.
Weizhe Xu wrote the initial draft of the manuscript, with input from **Trevor Cohen** and **Serguei Pakhomov**. All authors read and reviewed the manuscript, providing edits and suggestions for improvement where appropriate.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Disclosure:
Dr. Ben-Zeev has an intervention content licensing agreement with Pear Therapeutics and financial interests in FOCUS technology and in Merlin LLC. He has consulted for Trusst Health, eQuility, Otsuka Pharmaceuticals, and Deep Valley Labs.

the application of Time-Series Augmented Representations for Detection of Incoherent Speech (TARDIS), a novel approach that involves treating the sequence of coherence scores from a transcript as a time-series, providing features for machine learning. With ASR, TARDIS improves average AUC across coherence metrics for detection of severe ThD by 0.09; average correlation with human-labeled derailment scores by 0.10; and average correlation between coherence estimates from manual and ASR-derived transcripts by 0.29. In addition, TARDIS improves the agreement between coherence estimates from manual transcripts and human judgment and correlation with self-reported estimates of AVH symptom severity. As such, TARDIS eliminates a fundamental barrier to the deployment of automated methods to detect linguistic indicators of ThD to monitor and improve clinical care in serious mental illness.

Graphical Abstract


Identify formal thought disorder
Assess coherence in speech
Participants with Auditory Verbal Hallucinations
Automatic speech recognition
Smartphone "Audio Diaries"
Professional transcription
Vectorization
Semantic relatedness between units (e.g. words)
Time-series representation of a transcript
0.35
0.30
0.25
0.20
0.15
0.10
0.05
Cosine
0
5
10
15
20
25
Nth word
MINIMUM AGGREGATION
TARDIS
Extract time-series features
Predict human-assigned scores
Human assigned coherence scores
Comparison
Coherence score
Coherence score

1 Introduction

Coherent speech is apparent when individual utterances are logically connected and relate to a global theme. Speech lacking such coherence is clinically indicative of a range of serious mental illnesses, notably schizophrenia, where it is considered to be a manifestation of an underlying formal thought disorder (ThD) [1]. Consider for example this excerpt from the speech of a patient with schizophrenia, presented by Andreasen and Grove [1]:

> 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, m-my folks always thought they should, get pop but the best thing to get, is motor oil, and money.

While connections between sentences are perceivable at times, there are lapses in which it is difficult to understand how elements of the discourse relate to one another - such as why one might need soap to buy gasoline - or to a central theme. In psychiatry, lack of clear associative connections in spoken language is referred to as *derailment, loose* *associations,* or *flight of ideas*, with standardized instruments further distinguishing the total absence of connectivity between ideas as a distinct construct [1-2]. The notion that coherence might be automatically estimated using methods of distributional similarity was proposed by Foltz et al. [3], who derived a measure of text coherence by estimating the semantic relatedness between successive text segments using Latent Semantic Analysis (LSA) [4], for the purpose of studying the relationship between narrative coherence and text comprehension. The utility of this approach as a diagnostic instrument was evaluated by Elvevag, Foltz, and colleagues [5], who showed significant differences in automated coherence metrics between patients with schizophrenia and healthy controls, and between patients that were clinically rated as having higher degrees of ThD and other patients. Coherence scores from this method were subsequently incorporated into a predictive model that distinguished between patients and their healthy relatives with an accuracy of approximately 86% in cross-validation experiments [6]. In more recent work, LSA-based coherence metrics were incorporated as features in a classifier that predicted the *onset* of psychosis in a small sample (n=34) of youths experiencing prodromal symptoms with perfect accuracy in participant-level leave one out cross-validation experiments [7]. A variant of this approach was subsequently shown to retain 83% accuracy in predicting psychosis onset when evaluated using a larger data set from another site [8]. Neural word embeddings, as an alternative to LSA, have also been explored in automated coherence analysis, showing promising results in predicting clinical ratings of ThD [9,10]. Most recently, differences between participants with schizophrenia spectrum disorders and healthy controls were found in coherence estimates using sentence embeddings from a pretrained deep learning architecture, Bidirectional Embedding Representations from Transformers (BERT) [11,12].

While these studies provide support for the validity of semantic distance-based measures of coherence as indicators of ThD, they also raise questions about barriers to the deployment of these measures in practice. Work in this area has depended upon manual transcriptions of lengthy clinical interviews (e.g., > 1 hour in duration in Bedi, et al [7]), or data gathered in the context of structured tasks (such as story recall in Elvevag et al. [5]) conducted in controlled settings by interviewers with specialized training. The use of manual transcribers presents logistical challenges to automated speech assessment in serious mental illness, as this requires transferring potentially sensitive information to a third-party transcription service and would result in delays in response to changes in the clinical state if applied for the purpose of real-time monitoring. Extended structured interviews provide granular information but would place excessive demands on both staff and patients if applied with the frequency prerequisite for early detection of exacerbation in clinical symptoms.

Recent research suggests pathways through which to negotiate these challenges to automated, speech-based monitoring of symptoms in serious mental illness. The pervasiveness of smartphone technology presents the opportunity for real-time, real-place granular capture of speech data. Individuals with mental illness are more likely to own a smartphone than a computer [13], with survey estimates as high as two-thirds [14]. In our own recent work, we have demonstrated that individuals with schizophrenia are able to use smartphones for illness monitoring and relapse detection for extended periods (i.e., up to a year [15-17]); that people with active psychotic symptoms are willing to use their personally owned smartphones to record "audio diaries" describing them; and that automated analysis of manual transcriptions of these audio diaries approximates human judgment in the identification of thought disorder (e.g., semantic incoherence) [18]. This work showed that meaningful linguistic markers indicating thought disorder can be reliably extracted from transcripts of short (three minutes or less) smartphone-derived audio recordings of spontaneous speech captured in naturalistic settings in response to an open-ended prompt. This raises the question as to whether similar alignment with human judgment can be achieved in the context of Automated Speech Recognition (ASR), which would eliminate an important logistical barrier to deployment.

There is reason to believe this may be the case - in a recent study, Holmlund and colleagues [19] demonstrated that automated estimates of performance on story recall tasks using manual and ASR-derived transcripts of smartphone-derived audio recordings were highly correlated with one another, and comparably correlated with human ratings [19]. While this work did not concern formal thought disorder, the method used to derive automated recall scores depends upon estimates of semantic relatedness derived from vector representations of words similar to those that underlie previous work on automated coherence estimates. The authors argue that the robustness of automated scoring performance in the context of ASR errors is in part attributable to the mapping between variant forms of words on account of distributional similarity, reducing the dependence on perfectly accurate word recognition such that even recognition of a word fragment may be sufficient for meaningful estimation of the relatedness between text passages. In subsequent work [20], the automated story recall scores were incorporated amongst a battery of smartphone-delivered neuropsychological tests, further underscoring the potential of smartphone technology for scalable deployment of neuropsychological assessments.

In the current work, we assess the robustness of automated estimates of coherence [18] to errors introduced during the process of ASR. We also devise a novel representational approach for these coherence estimates called Time-series Augmented Representations for Detection of Incoherent Speech (TARDIS) and compare this with the typical approach of aggregating coherence estimates across transcripts: taking the minimum coherence score. This approach predominates in prior work on automated estimates of coherence in the context of thought disorganization [5-8], underlies key validation studies in this area [5,7], and has been shown to outperform a range of other aggregate statistics in its agreement with human judgment and utility as a predictive feature of the onset of psychosis in high-risk individuals [7,18]. However, as it is based on a point estimate only, we hypothesized that it would be vulnerable to ASR errors. Using an in-house ASR system based on Baidu's Deep Speech 2 architecture [21], we generated automated transcriptions of 275 "Audio Diary" recordings of participants describing their experiences of Auditory Verbal Hallucinations (AVH) – another prominent symptom of psychotic-spectrum disorders such as schizophrenia - and compared the concordance of automated estimates of coherence derived from both these and professional transcriptions to the judgment of human annotators. To ensure the generalizability of our method, we employed an additional 2000+ unannotated "Audio Diary" recordings. We assessed the correlation between coherence estimates derived from manual and automated transcriptions and the strength of association between the resulting estimates and baseline estimates of the severity of related psychotic symptoms using a validated self-report scale. We hypothesized that ASR-based metrics would (1) largely retain their alignment with the human judgment of coherence; (2) correlate well with corresponding assessments from professional transcriptions; (3) retain their association with the severity of related psychotic symptoms. In addition, we hypothesized loss in performance with ASR in all three of these evaluations could be partly remediated using TARDIS – an alternative to the typical approach of using the lowest evaluation of coherence between semantic units (words, phrases, or sentences) of a transcript as a sole feature. This approach seemed to us particularly vulnerable to spuriously low coherence estimates introduced by ASR errors.

2 Method

2.1 Data sets

Data used in this work were provided by 384 participants experiencing AVH, of which 295 had significant clinical histories, including inpatient care (50%) and partial hospitalization (33%). Participants were drawn from 41 U.S. states, with the majority (approximately 80%) of participants recruited online. The participant pool was diverse, with approximately 20% of participants identifying as Black or African American, and approximately 15% identifying as Hispanic or Latino. Together, these participants contributed a total of 27,731 Ecological Momentary Assessment (EMA) self-reports and 4809 Audio Diary recordings, with 3040 of these – recordings of duration 30 seconds or longer – professionally transcribed [22]. From these data, we derived two datasets, one annotated with human-assigned estimates of incoherence (the labeled dataset) and the other without human annotation (the unlabeled set).

Figure 1 provides an overview of the data sets and how they were constructed. The partial overlap between the “Full” (all participants completing the study) and “Labeled” (from participants with data available in October of 2019) source sets can be explained by data in the smaller set, which was gathered earlier, from participants who did not complete the study in its entirety.

Labeled dataset: We used data collected by extracting a sample of up to three smartphone-derived Audio Diary recordings of duration thirty seconds or more per participant from the data collected up to date Oct 2019, resulting in a set of 310 transcripts of 142 participants describing their auditory hallucinations. These transcripts were manually annotated by two raters for the construct of “derailment” as defined in the Thought and Language Disorder Scale (TALD) [2]. Scores were assigned by 2 annotators independently and ranged from 0-4; with 0 indicating no evidence of derailment, 1-2 indicating mild to moderate derailment, 3 indicating severe derailment, and 4 indicating the text was incomprehensible. After independent reassessment of any discrepancies of 2 or more TALD units, agreement by quadratically weighted Kappa score was 0.71 [18]. For a small number of recordings, neither the ASR system nor the human transcribers were able to produce meaningful transcripts. In these cases, the human transcribers noted background noise, and the ASR system did not produce output. After removing these recordings and restricting to only those transcripts from which all coherence metrics produced a score (for example, sentence-based metrics require more than one sentence to be recognized), we arrived at a set of 275 paired (manually transcribed and ASR) labeled transcripts from 134 participants.

Unlabeled dataset: While most of the full set of 3033 recordings with transcriptions has not been annotated for derailment, this set nonetheless provides additional data for evaluation purposes. While the annotated set is not a subset of this set (because it includes data from participants that were enrolled earlier in the study but did not complete it), there was some overlap between the sets with 247 recordings occurring in both sets. After removing these files from the larger set and those in which the ASR system did not produce any output, we retained a total of 2359 unlabeled transcripts from 235 participants. Because it is without human annotation for coherence, this set was used to compare coherence scores derived from automated transcripts with either comparable scores from manual transcripts, or clinical rating scales. Specifically, the unlabeled dataset was used to (1) evaluate the correlation between ASR- and manual transcript-derived coherence scores; and (2) assess the relationship between these scores and scores from a validated self-report instrument for the assessment of the severity of other psychotic symptoms.

2.2 Automatic speech recognition

We trained an ASR system based on Baidu’s Deep Speech 2 architecture [21] implemented in PyTorch[1], and consisting of 3 convolutional neural network (CNN) layers, followed by 5 bidirectional recurrent neural network (RNN) layers with gated recurrent units (GRU), a single lookahead convolution layer followed by a fully connected layer and a single softmax layer. The system was trained using the Connectionist Temporal Classification (CTC) loss function [23]. In addition to the default greedy search decoding over the hypotheses produced by the softmax layer, the system’s implementation also can use a beam search decoder with a standard n-gram language model. We used default hyperparameters: the size of the RNN layers was set to 800 GRU units; starting learning rate was set to 0.0003 with the annealing parameter set to 1.1 and momentum of 0.9. Audio signal processing consisted of transforming the audio from the time to the frequency domain via Short-time Fourier transform as implemented by the Python librosa[2] library. The signal was sampled in frames of 20 milliseconds overlapping by 10 milliseconds. The resulting input vectors to the first CNN layer of the Deep Speech 2 network consisted of 160 values representing the power spectrum of each frame.

A collection of speech corpora available from the Linguistic Data Consortium were used as training data. These corpora include the Wall Street Journal (WSJ: LDC93S6A, LDC94S13B), Resource Management (RM - LDC93S3A), TIMIT (LDC93S1), FFMTIMIT (LDC96S32), DCIEM/HCRC (LDC96S38), USC-SFI MALACH corpus (LDC2019S11), Switchboard-1 (LDC97S62), and Fisher (LDC2004S13, LDC2005S13). In addition to these corpora, we used the following publicly available data: TalkBank (CMU, ISL, SBCSAE collections) [24], Common Voice (CV: Version 1.0)) corpus [3], Voxforge corpus [4], TED-

[1] Baidu Deep Speech: https://github.com/SeanNaren/deepspeech.pytorch
[2] Librosa: https://librosa.org/
[3] Common voice: http://voice.mozilla.org
[4] Voxforge: http://www.voxforge.org/

LIUM corpus (Release 2) [25], LibriSpeech [26], Flicker8K [27], CSTR VCTK corpus [28], and the Spoken Wikipedia Corpus (SWC-English [29]). Audio samples from all these data sources were split into pieces shorter than 25 seconds in duration. The total size of the resulting corpus was approximately 4,991 hours of audio (2,000 hours contributed by the Fisher corpus alone). Finally, we also used in-house audio data from various prior studies that were conducted at the University of Minnesota consisting of story recall, verbal fluency, and spontaneous narrative tasks. Apart from the Fisher and Switchboard corpora, all other data were recorded at a minimum of 16 kHz sampling frequency. The Fisher and Switchboard corpora contain narrow-band telephone conversations sampled at 8 KHz. All data were either down-sampled or upsampled and converted using the SoX toolkit[5] to a single channel 16-bit 16 kHz PCM WAVE format.

Beam-search decoding was used to produce raw ASR transcripts with a 4-gram language model constructed with the SRILM Toolkit [30] from the English language portion of the 1 Billion words text corpus[6] model with Kneser-Ney smoothing [31].

2.3 Post-processing of transcripts

2.3.1 Repunctuation—The raw output of the ASR pipeline is a sequence of words, without capitalization or punctuation. However, punctuation is necessary for the phrase- and sentence-level segmentation, which is required for certain coherence metrics. We therefore used the punctuator model of Tilk et al [32] to add punctuation to the transcriptions. This model uses a bidirectional recurrent neural network with an attention mechanism, trained on English TED talks (2.1M words). We used a publicly available pretrained model[7] to add punctuation marks such as commas, periods, and question marks to our ASR output. After repunctuation, we capitalized the first letter of each sentence (start of a line or following a period) and standalone “i” characters, to further improve transcript quality.

2.3.2 Segmentation—Automated estimates of coherence leveraging distributional similarity are estimated by comparing the semantic relatedness between units of text, where a unit might be an individual word, phrase, or sentence. Before coherence analysis, transcripts must be tokenized into such semantic units. Tokenization is a necessary process to break down the document into basic units (word/phrase/sentence), and in the case of larger units, further tokenization is required to construct semantic vectors for further analysis by averaging the vectors of the words they contain. We first removed the “stop-words”, words that do not carry semantic content (such as “a”, “an”, and “the”), using a commonly used list of stop-words provided by the NLTK toolkit [33]. Then we tokenized the transcripts into semantic units at three different levels of granularity: words, noun phrases, and sentences. The words and sentences were tokenized using the NLTK word and sentence tokenizer, respectively, and the noun phrases were tokenized using the noun-phrase tokenizer from the Spacy package [34].

[5] Sox toolkit: http://sox.sourceforge.net

[6] SRILM toolkit: https://github.com/ciprian-chelba/1-billion-word-language-modeling-benchmark

[7] Punctuator: https://github.com/ottokart/punctuator2

2.4 Assessment of coherence

In this section, we will describe our pipeline for automated estimation of coherence subsequent to the tokenization of incoming text into semantic units at the word, phrase, or sentence level. Once this segmentation is accomplished, the main questions to consider are: (1) how is semantic relatedness between units of text measured? (2) upon which units are these measurements based (e.g., sequential units, gapped units); and (3) how are these measurements aggregated across a transcript? We will commence by describing how semantic vector representations of words are used to calculate the relatedness between semantic units.

2.4.1 Skip-gram semantic vectors—Vector representations of words learned from large unlabeled corpora have a long track record of application in both automated natural language processing tasks, and cognitive models of lexical semantics [35-37]. Neural word embedding [38] is a widely used approach to generating semantic word vectors, on account of its ability to scale comfortably to large corpora. For the current research, we used publicly available pre-trained vectors derived using the FastText [39] package[8] consisting of 2-million-word vectors trained on a corpus derived from Common Crawl[9]. Individual words are represented by their vectors, and larger units (phrases or sentences) are represented as the normalized superposition of the vectors of the words they contain. For example, the noun phrase “bank account” can be represented by an embedding that is the normalized sum of the embedding of “bank” and the embedding of “account”. The same approach can be applied to each sentence. With some sentence-level variants (henceforth denoted with “IDF”), this superposition is weighted by the inverse document frequency of the terms concerned, such that relatively infrequent (and hence more informative) terms will carry more weight. The relatedness between any pair of semantic units is calculated as the cosine of the angle between the vectors that represent them.

2.4.2 Contextual semantic vectors—In addition to skip-gram semantic vector embeddings, we experimented with contextual semantic vector embeddings using the Bidirectional Encoder Representations from Transformers (BERT) model [12]. Coherence estimates based on BERT-derived sentence embeddings have recently been shown to differ between patients with schizophrenia spectrum disorders and healthy controls [11], and we wished to evaluate their utility as a means to model coherence in our data with and without TARDIS. BERT models are trained to predict ‘masked’ words within sentences and to predict whether one observed sentence follows another. This is accomplished using an attention mechanism, through which the contextual representation of a word is informed by the representations of other words in its vicinity. This context-specific representation differs from the single (global) vector representation of a word that underlies the distributional semantic vector representations we have discussed previously. We derived contextual embedding from the BERT model at token, phrase, and sentence levels. The token level embeddings were derived as the sum of the last four layers of the hidden state output for each input token [12]. The phrase embeddings were generated as the sum of the embeddings

[8] FastText: https://fasttext.cc/docs/en/english-vectors.html

[9] Common crawl: https://commoncrawl.org/2017/06

from the individual token components. At the sentence level, we experimented with a range of approaches: the second-to-last layer of hidden state output [12], the CLS token output (a special token prepended to the sequence that is typically used to generate sentence representations for the purpose of text categorization), the sum of the token embeddings that form the sentence, and the sentence embeddings from sentence-BERT [40]. Sentence-BERT uses Siamese and triplet network structures to derive semantically meaningful sentence embeddings with advantages in performance over prior methods in sentence similarity tasks [40]. We used the pretrained "BERT-base-uncased" model to derive most of the embedding variants and the publicly available "all-MiniLM-L6-v2" model to derive the embeddings from the sentence-BERT implementation [40]-.

2.4.3 Cosine calculations—*Sequential* estimates of coherence are based on measurement of the similarity between terms that are juxtaposed in sequence underlie most automated estimates of formal thought disorder. Sequential estimates have been validated in numerous prior studies [5,7]. However, in recent work, we have shown that *global* estimates of coherence, based on the similarity between terms in a text and their centroid (or vector average) can align better with annotator assessment of coherence than their sequential counterparts [18]. Motivated by these results, we calculated both sequential and centroid-based estimates of coherence for the current study. With centroid-based methods, we included both *static* and *cumulative* variants, where the former measure the relatedness between each term in a transcript and their centroid, and the latter measure the relatedness between each term and the centroid for all terms encountered up to the point in the sequence of the term under consideration. Where sequential approaches estimate the relatedness between ideas that are stated in proximity, global estimates measure the relatedness between each stated idea and the central topic of a body of text. The relatedness was measured in terms of cosine similarity such that each transcript was represented by a series of cosine values.

2.4.4 TARDIS: Time-series augmentation—Once the cosine similarities between all the units of interest in a text have been calculated, these are typically aggregated to provide a coherence estimate to serve as a single feature for analysis, or for downstream machine learning. The minimum cosine score is a commonly used aggregation function for this purpose [6-7], and this proved effective in our recent work also [18]. While this is convenient to calculate and has been effective as a basis for machine learning models [7-8], we were concerned that a single point estimate (such as the minimum cosine between successive phrases) or some other individual summary statistic (such as the mean cosine across all semantic units) may be vulnerable to occasional extremely low cosine values from errors arising during automated transcription. We were also concerned that simply taking an aggregation function would discard much information about the pattern of the cosine values as a transcript progresses, especially given that a decrease in coherence throughout an utterance was observed in participants judged to have high degrees of formal thought disorder in previous work [5]. Therefore, we developed TARDIS - a novel approach to representing coherence estimates from an entire transcript, which involves treating the cosine values for words that occur in sequence as time-series data, and further analyzing the coherence through time series feature extraction for machine-learning-based prediction of human-assigned scores.

We extracted key features from time-series data using the TSFRESH software package [41]. The TSFRESH package acquired features in the following main categories: (1) features from summary statistics (e.g. min, max, number of peaks) (2) additional characteristics of the sample distribution (e.g. binned energy, data symmetry) (3) features derived from observed dynamics (e.g. mean autocorrelation, the Fast Fourier Transformation coefficient). We removed length-dependent features, in order to avoid developing models that consider the volume rather than the coherence of language produced. We then used these features to train a support vector machine regression (SVR) model to predict annotator-assigned derailment scores. Additional statistically-based feature selection provided by TSFRESH [41] was performed before model training, but only with word-level semantic units because larger semantic units did not produce sufficient individual cosine values for this to be effective due to the limited length of our transcripts. The Scikit Learn package[10] implementation of the SVR model was used. We chose the radial basis function (RBF) kernel and kept other hyperparameters as their default values. We used leave-one-out (LOO) cross-validation to generate a regression result for each transcript (such that each transcript's prediction score was the output of the model trained using the other 274 transcripts). The SVR model predictions served as a final coherence assessment in terms of derailment.

2.5 Summary of the analytic pipeline

The coherence analytic pipeline described in sections 2.2-2.4 is be summarized in Figure 2, which demonstrates the path from an audio recording to estimation of a coherence score, through either TARDIS or the minimum aggregation function.

2.6 Experiments

2.6.1 Alignment between ASR-derived coherence metrics and human-labeled derailment scores—The goal of this experiment was to evaluate the extent to which errors introduced by ASR transcripts negatively influence the agreement between system- and human-assigned derailment scores. A secondary goal was to evaluate the extent to which using time-series features could recover lost performance. We measured the performance of each metric in two ways. To assess *overall agreement* with the average human-assigned score, we calculated the Spearman correlation between this average and each of our automated coherence estimates. To assess the ability to detect *severe cases* of incoherence, we rank-ordered transcripts by their automatically assigned coherence scores and calculated the area under the receiver operating characteristic curve (ROC AUC) using the labeled derailment scores to identify transcripts corresponding to severe levels of disorganization according to the TALD. Specifically, positive class labels for AUC calculation were affixed to transcripts with derailment scores of 3 or more. To evaluate the impact of ASR on the coherence metrics, we compare the performance of ASR-derived and professional transcription-derived coherence scores estimated using the minimum aggregation function (i.e. the lowest coherence score across a transcript). To evaluate the extent to which the

[10] Scikit learn: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

time-series method, TARDIS, restores performance, we compare the performance of ASR-derived coherence scores generated by the minimum aggregation function and the time-series methods. TARDIS in this experiment was evaluated using a leave-one-participant-out cross-validation procedure (i.e. for each transcript, train on all other transcripts and store the predicted score for this held-out test case), due to the limited sample size. In contrast, in the experiments that follow – which do not use the annotated set for evaluation purposes - the entire set of 275 annotated transcripts was used to train both TARDIS-based models and those trained on entity grid feature vectors that provide a point of comparison (see section 2.6.4).

2.6.2 Alignment of ASR-derived and professional transcription derived coherence metrics—In this experiment, we aimed to assess the correlation between the ASR-derived coherence metrics and professional transcription derived coherence metrics across a larger set of unannotated recordings, with a high correlation suggesting that few errors were introduced by the ASR process. Once again, we examined whether the time-series method could improve this correlation, by comparing it with the minimum aggregation function method. Correlation between ASR- and transcription-derived coherence scores was measured using Spearman Rho correlation. For this component, we used the 2359 unlabeled recordings to make the comparison.

In addition, we assessed the relationship between ASR accuracy and correlation between scores assigned to professional and corresponding ASR-derived transcripts of the same recordings, with the hypothesis that TARDIS would enhance the robustness of this correlation to ASR error, which was measured in word error rate (WER). This metric calculates the number of substitutions, deletions, and insertions divided by the number of words in the manual transcript.

2.6.3 TARDIS enhancement of coherence metrics derived from manual transcriptions.—In this experiment, we evaluated the potential for TARDIS to improve performance in the context of professionally transcribed recordings. Performance was measured as the Spearman Rho correlation with average annotator score, and the ROC AUC for detection of transcripts with average annotator scores >=3. The dataset concerned was the 275 annotated transcripts, and the time-series method was evaluated in a leave-one-out cross-validation configuration. The performance characteristics of this times-series method and the minimum aggregation method were calculated and compared.

2.6.4 Comparison of TARDIS metrics with Entity Grid coherence metrics.—Although they have not to our knowledge been used to model thought disorganization previously, we include Entity Grid coherence scores as an additional point of comparison. The Entity Grid method is a well-established approach to measuring textual coherence that operates by capturing the local syntactic transitions of entities – how they shift from one semantic role to another across sentences [42]. These role transitions (e.g. subject-to-object) are quantified to generate feature vectors for machine learning models, providing a syntax-informed point of comparison for the feature vectors emerging from TARDIS. We used the feature vectors created from entity grids to train an SVM regressor to serve as a baseline comparison to the TARDIS feature set. The entity grids and features were generated using the text-to-entity grid package[11].

2.6.5 Correlation with HPSVQ—The Hamilton Program for Schizophrenia Voices Questionnaire (HPSVQ) [43-44] is a validated self-report instrument for AVH. While this questionnaire does not measure the severity of other manifestations of psychotic episodes – such as thought disorganization – we nonetheless hypothesized that the HSPVQ total score, which indicates the *severity* of this symptom, would partly correlate with the severity of thought disorganization as estimated by coherence assessment because these aspects of psychosis are frequently observed together [45]. We further hypothesized that the correlation with the overall score would decrease with ASR on account of transcription errors and that time-series featurization may restore some of this correlation.

The HPSVQ was collected once when participants signed up for the study. We used the transcript with the lowest coherence score to represent the coherence score for each participant. Each of the coherence metrics (time-series vs. minimum) generated from either manual (professionally transcribed) or ASR-derived transcripts were compared for their correlation with the summary score of the HSPVQ.

3 Results

3.1.1 Alignment of ASR-derived coherence metrics and human-labeled derailment scores (annotated set) using skip-gram vectors:

Figure 3 provides a side-by-side comparison of the time-series method and minimum aggregation method across transcripts. Each 3-bar column represents a different coherence metric with different combinations of semantic units and computation methods. Within each column, the 3 bars represent evaluation scores for the minimum coherence method from manual transcripts, the minimum coherence method from ASR, and the TARDIS method from ASR (left to right, respectively). We can derive 2 main observations (1) For the minimum coherence method, performance usually drops when switching from manual transcript (first of three bars) to ASR transcript (second of three bars). (2) The TARDIS method (third of three bars) improves the performance of almost all the coherence metrics when using ASR. When considering AUC in the context of ASR, the time-series method produces the highest value of 0.744 and improves the average AUC across all the coherence metrics from 0.623 to 0.691. With respect to Spearman Rho with ASR, the time-series method achieves the highest performance with Rho=0.456, improving the average across all coherence metrics from 0.287 to 0.396.

3.1.2 Alignment of ASR-derived coherence metrics and human-labeled derailment scores (annotated set) using contextual vectors derived from BERT:

The contextual vectors derived from the BERT model were applied using both minimum coherence (manual and automated transcripts) and TARDIS (automated transcripts only). As shown in Figure 4, the results are similar to those obtained with skip-gram word

[11] Text to entity grid: https://github.com/MMesgar/text_to_entity_grid

embeddings. With ASR, TARDIS using BERT-derived vectors outperformed BERT-derived minimum coherence for most metrics in terms of ROC-AUC (the rightmost red bars are higher than the middle blue bars in each metric group). We also observed a similar pattern of performance drop with minimum aggregation when switching from manual transcript to ASR transcript (observed from the leftmost orange bars higher than the middle blue bars in each metric group). Additionally, we observed systematic improvements when moving from minimum coherence aggregation to TARDIS across most metrics in terms of Spearman Rho correlation. This shows the robustness of the advantage of time-series representations across different embedding approaches.

Table 1 shows a comparison between performance with skip-gram and contextual vectors at each semantic level. In terms of ROC-AUC, we did not find an improved maximum AUC with BERT for all the coherence metrics. However, we did find improvements within certain semantic units - specifically at the word and sentence levels. With respect to Spearman Rho correlation, performance was generally better with neural word embeddings. The only exception occurred when using contextual vectors with the sentence-level metrics. However, this did result in the best overall Spearman Rho correlation of 0.465 (as compared with 0.456 with skip-gram embeddings).

3.2.1 Alignment of ASR-transcript-derived coherence metrics and manual-transcript derived coherence metrics (unannotated set):

We explored the correlation between the coherence scores generated from ASR transcripts and manual transcripts using TARDIS and the standard aggregation approach of taking the minimum value for a transcript, with minimum aggregation used for manual transcripts and either TARDIS or minimum aggregation used with ASR. Higher correlation indicates relative robustness to transcription errors introduced by ASR. The results of this analysis are shown in Table 2.

These results demonstrate that the time-series method consistently improved the correlation between coherence scores from ASR-derived transcripts and those from professionally transcribed recordings, with a mean increase from 0.429 to 0.720 across all coherence metrics.

3.2.2 Impact of ASR error on coherence metrics:

ASR transcription error is a key reason for drops in performance in coherence evaluations. For example, an instance of “Craigslist ad” in a manual transcript was transcribed as “Craigslist dad” in an automated transcript, altering the meaning of the phrase. This section demonstrates evaluations of coherence metrics in the context of similar ASR errors measured by word-error-rate (WER) and character-error-rate (CER) metrics. As might be anticipated given the difficulties inherent in transcribing recordings captured in naturalistic settings, performance was closer to that documented with Deep Speech 2 with noisy speech (WER 21.59-42.55) than with standard evaluation sets (WER 3.10-12.73) [21], with a mean WER of 0.36 and CER of 0.2 across the transcripts used in our studies. Figure 5 shows the correlation between average coherence scores across all metrics derived from ASR and manual transcripts plotted against different ranges of ASR WERs (bins divided at each quantile). Higher correlation indicates higher similarity and potentially less performance loss between the ASR and manual transcripts. The results indicate that coherence metrics suffer correlation loss linearly as the ASR error rate increases up to an error rate of approximately 0.5, and that correlation declines precipitously after this point. TARDIS is more resistant to ASR error because it has a higher correlation at all error rates.

3.3 TARDIS improvement on manual-transcript-derived coherence scores—

Having observed a strong recovery in performance with ASR-derived transcripts when using time-series features, we proceeded to evaluate how this featurization approach affects performance in the context of professionally transcribed recordings.

The results of these experiments are shown in Figure 6, which shows a comparison between the time-series method and the original method of taking the minimum coherence across a transcript on manual transcripts (from professionally transcribed recordings). In terms of the AUC, we can observe that TARDIS improves the majority of metrics with both skip-gram and BERT-derived embeddings. However, it does not improve the best-performing metrics, including the sentence-based metrics and the phrase-based centroid metrics with skip-gram vectors. When considering Spearman correlation there is a clear advantage for TARDIS across all metrics (with both FastText and BERT embeddings), with an increase of maximum value from 0.525 to 0. 601. The overall performance of TARDIS indicates a general improvement in the alignment between coherence scores and human judgment, but not necessarily in the ability to identify severe cases (mean TALD >= 3).

3.4 Comparison between TARDIS metrics and the Entity Grid representation.—

Table 3 shows the SVM model prediction performance using TARDIS feature set and the entity grid feature set on automated and manual transcripts. The entity grid features led to a promising performance with both automated and manual transcripts. However, they did not outperform the best TARDIS metrics in any of the cases.

3.5 Alignment of the coherence scores to the HPSVQ [44] clinical scale.—

In this section, we evaluated the Spearman Rho correlation between various coherence metrics (the minimum coherence estimated for an individual) and the HPSVQ total score.

Table 4 indicates that some correlations exist between the coherence metrics and AVH severity measured by HPSVQ. Although not strongly so, self-reported AVH symptoms are shown to correlate with the automatically measured coherence in speech. In addition, the TARDIS method amplifies the mean correlations with the HPSVQ total scores in the context of both ASR and manual transcripts. The contextual embeddings also demonstrated potential in their correlation with HPSVQ scale because the strongest correlations in each experimental set up came from a metric with BERT-derived embeddings.

4. Discussion:

4.1 Key findings

In this study we evaluated a fully automated approach to quantify coherence from speech samples collected in naturalistic environments. Our results show that our novel featurization approach, TARDIS, effectively compensates for ASR errors when estimating coherence and improves performance with most coherence metrics with manual transcripts.

In comparison with professional manual transcriptions, ASR may introduce transcription errors. Our results show that these errors do impair the performance of coherence metrics to some degree. This impairment is most pronounced in the case of sentence-based metrics. This may be explained by that relatively few coherence measurements between units occur at this level of analysis. The effects of transcription errors on metrics of coherence are demonstrated through their loss of alignment with human-assigned derailment scores, and decreased correlation between coherence scores generated from manual and automated transcripts of the same set of recordings.

We found that these losses can be largely recovered by representing the full spectrum of coherence information generated from a transcript as a time series. This provides an alternative to the predominant approach of using the point of least coherence between elements of a transcript as a sole feature [7]. We do so by using an approach we call Time-series Augmented Representations for Detection of Incoherent Speech (TARDIS). TARDIS does not rely on a single extreme coherence value exclusively. As such, it is robust to such values being introduced by sporadic machine transcription errors. In addition, TARDIS leverages features derived from the trajectory of coherence estimates over the course of a transcript. This is in accordance with the seminal finding that automated estimates of coherence may decrease precipitously as speech progresses in the setting of severe thought disorder [5]. The recovery of performance with TARDIS is evident in our findings. Most of the coherence metrics applied to ASR transcripts show improved association with human-assigned scores with this approach. In many cases, performance recovers to that obtained with manual transcripts. Surprisingly, TARDIS performance with automated transcripts even surpasses the performance of the 'minimum coherence' approach with *manual* transcripts for some coherence metrics, suggesting that the benefits of time-series featurization on this task extend beyond their robustness to ASR error. Recovery with TARDIS is further supported by the considerably higher correlation between coherence scores derived from ASR (with TARDIS) and scores derived from manual transcripts (with minimum coherence as an aggregation function). This indicates a reduction in divergence between coherence assessments of ASR and manual transcripts when TARDIS is applied to automated transcripts.

The hypothesis that TARDIS may have benefits beyond robustness to ASR error is also supported by our subsequent findings. TARDIS also improves the alignment between human-assigned and automated estimates of coherence with manual transcripts. This is most evident in the Spearman Rho correlation with human-assigned derailment scores, where all coherence metrics show improvement with time-series featurization. This correlation is indicative of alignment with human annotators across the full spectrum of coherence levels in our dataset. However, the ability of models to identify cases of severe thought disorder (as indicated by a TALD derailment score >= 3) may provide a better estimate of their clinical utility in the context of smartphone-based continuous monitoring efforts to identify relapse events. TARDIS improves the majority of metrics' ability to identify such severe cases. However, in the context of these manual transcripts, the minimum aggregation approach has the highest overall AUC scores. These are obtained when it is applied to sentence-based coherence metrics. One explanation for this finding is that at the sentence level the number of time-series data points available for analysis is limited because the sentence is the largest semantic unit considered. Another is that with manual transcripts the extremely low scores captured by the minimum aggregation function are likely to indicate legitimate severe cases, rather than being artifacts of ASR error. However, in the context of ASR, sentence embeddings derived from word vectors (weighted or unweighted) suffer from a decline in performance with higher ASR error rate [47]. Thus, the sentence-based coherence metrics are severely limited by ASR error especially when only using the minimum value to represent the transcript. The TARDIS method we presented in this study did not improve the sentence embeddings themselves under the influence of ASR error. However, it did improve the performance of the downstream coherence evaluation task by incorporating more information about the transcript and limiting the impact of any individual error. In this way, the TARDIS method improves the performance of sentence-based coherence metrics considerably when ASR transcripts are used.

This robustness of TARDIS-based approaches to ASR error is illustrated by a series of coherence scores extracted from one of the annotated transcripts in our set (Figure 7). This is a time-series representation of a transcript when using the sequential word coherence metric, such that the score indicates the cosine of the angle between vectors representing sequential words. When using the minimum aggregation method, the minimum value 0.03 was directly taken as the coherence score for the entire transcript but most of the cosine values are well above this. The normalized human-assigned coherence for this transcript was 0.875 (indicating a high degree of coherence), the TARDIS coherence was 0.787 (also indicating a coherent transcript) but the minimum coherence score was 0.536 (all values were normalized (min-max scaled between 0-1) across all transcripts) and represent coherence instead of derailment). As such, it is readily apparent why TARDIS produces a better estimation of coherence in this case.

We also demonstrated the robustness of TARDIS-based metrics across a different set of word embeddings (FastText and BERT embeddings). The TARDIS metrics outperform the minimum coherence metrics using both skip-gram embeddings (FastText) and contextual embeddings (BERT). This observation shows that time-series features represent useful information for the task of estimating coherence, irrespective of whether the global or context-specific meaning of words is considered. In addition, when comparing skip-gram and contextual word embeddings, we found some potential advantages for including contextual information when estimating coherence. This is a novel finding - recent work using BERT embeddings did not include a comparison with established skip-gram semantics-based approaches. The best Spearman Rho correlation with human judgment using automated transcripts was achieved with BERT-derived embeddings. BERT-derived embeddings also improved ROC-AUC performance with word and sentence level metrics. Thus, applying contextual embeddings to the task of quantifying coherence for ThD may be a fruitful direction for future research.

Interestingly, TARDIS-based coherence scores also on average correspond better with clinical assessment of *other* features of psychosis, namely AVH. Despite disorganized thinking being a separate construct from AVH, we find a modest but significant Spearman Rho correlation between the lowest coherence score for the transcript from an individual, and their scores on the HPSVQ (which measures the severity of AVH) collected at baseline. This correlation is stronger with TARDIS with both manual and automated transcripts. This suggests an association between the severity of the AVH symptoms and the coherence of speech. This finding is consistent with previous research showing that AVH tends to cooccur with thought disorder [45], potentially because incoherence of covert (i.e. ‘internal’) speech may influence discourse processing and manifest as poorly organized overt speech [48].

Surprisingly, despite the general loss of performance when transitioning from manual to ASR-derived transcripts, some phrase-based coherence metrics did not lose performance in certain evaluations. One possible explanation for this observation is that the noun-phrase extractor extracts different amounts of data from manual and ASR transcripts. On account of ASR and repunctuation errors, more phrases are extracted from manual transcripts than from their ASR-derived counterparts. For example, a common ASR error involves the omission of a spoken word from a transcript, which would reduce the number of noun phrases extracted if this word were a noun. For the current experiments, the average number of noun phrases extracted from the manual transcripts was 20.5 whereas with the ASR transcripts this was reduced to 13.7 (P<.001). Thus, with ASR the unit of analysis is larger than with manual transcripts. This may have a smoothing effect, such that the effects of unrelated smaller phrases that would be ‘semantic outliers’ with manual transcripts are diluted within the larger phrases extracted from automated ones.

The entity grid approach, which has not to our knowledge been applied to model thought disorganization previously, incorporates structural elements by quantifying transitions between syntactic roles across sentences [42]. This approach yielded promising performance on the task of quantifying coherence in our AVH data. Although TARDIS metrics generally achieved better performance, the entity grid approach offers new insights into the usefulness of syntactic features for the detection of ThD. Prior work has prioritized semantics over syntax, perhaps because syntactic structures are thought to be preserved in schizophrenia even in the presence of thought disorganization [49]. Our findings suggest that the entity grid can be used when modeling thought disorganization. This may be explained by the fact that the entity grid measures the saliency of the entities in text [42]. Despite speech in ThD exhibiting correct syntactic structure, entity saliency is still an important feature to consider. While it takes syntax into account, the entity grid method is not intended to measure the correctness of syntactic structure. Rather, it uses this structure to identify salient entities in text. It is the transitions of the syntactic roles of these entities across sentences that are used to generate features from which to estimate coherence. This, and the performance of entity grid features in our evaluations, suggest that syntax-aware models offer potential as an alternative and likely complementary approach to established methods.

4.2 Implications

This study presents a new method (TARDIS) for quantifying coherence of speech and demonstrates its utility in the context of a novel pipeline for the automated assessment of thought disorder and AVH severity, without the need for manual transcription. The main implications of this study can be summarized as follows:

4.2.1 Coherence estimates using speech samples from naturalistic settings:

Traditionally the data needed to assess thought disorders are collected in laboratory settings, and elicited through structured interviews [7], or tasks such as story recall [5]. As a result, the participant pool size is limited by available resources and logistical constraints, and data are limited to brief snapshots that may miss clinically important fluctuations in symptom severity in between clinic visits. Our results demonstrate the feasibility of measuring coherence using speech samples gathered in a naturalistic environment using smartphone technology, with data captured in real-time. By eliminating logistical and resource-related barriers, it allows for the recruitment of larger participant pools in a domain where small proof-of-principle studies have predominated to date [50].

4.2.2 Automated speech analysis pipeline:

From speech recordings to transcripts to coherence scores, the process of quantifying coherence can be completed without manual input. Prior efforts [5,7] require manual transcription from speech data, a process that imposes further logistical constraints on the analysis process, such as time-to-transcription and transfer of sensitive data to a third party. With the introduction of the novel TARDIS method, the performance of coherence assessment using ASR transcripts improves, approximating performance with manual transcripts in most cases. This establishes the feasibility of the application of a fully automated pipeline from data collection to coherence assessment eliminating a crucial logistical barrier to real-time monitoring of patients for exacerbations of their psychotic symptoms. As an additional benefit, this pipeline enhances patient privacy. The fully automated ASR-based coherence measures can be estimated without sending any sensitive data to a third-party service.

4.3 Limitations and future work

One limitation of this study is the limited interpretability of the time-series features. The features extracted from the TSFRESH package are well-established as ways to represent information carried in time-series data [51]. Upon evaluation of feature importance to our Support Vector Regression models using Shapley [52] additive explanations, the most influential time-series features were the angle component of the Fourier coefficient [53], and the standard error and r-value of the regression line. However, in some cases, it is difficult to interpret how individual time-series features contribute to a particular aspect of coherence scoring. Some straightforward features such as the “minimum” can be readily interpreted as the worst possible coherence score for the entire document (often used as the sole feature in prior work), but others such as the Fourier coefficient are derived from multiple components which makes it difficult to isolate the utility of the information they carry. This is particularly important in the context of our current dataset, as we have observed a moderate correlation between human-assigned derailment scores and word count. That is to say, those transcripts viewed as less coherent by our annotators were generally longer than those viewed as more coherent, which is consistent with both the tendency of coherence to decrease as speech progresses [6] and the likelihood of incoherence being easier to observe with more data available. Nonetheless, we would not wish our incoherence models to react to sequence length independently, and with the minimum aggregation function, we can be certain this is not the case. With TARDIS, we deliberately removed features that were obviously length-dependent from consideration, such as the sum, time-series length, and the number of peaks. However, we cannot exclude the possibility that some of the remaining features are affected by word count in more subtle ways, and as such it is not clear how much of the additional correlation with human scores obtained with TARDIS may be a secondary effect of sensitivity to transcript length. While we note that an SVM regression model trained with word count as a sole feature has a correlation with human judgment considerably below that of the average of the TARDIS models for each coherence score (ASR: 0.227 vs. 0.396; manual: 0.329 vs. 0.543) future work with an evaluation set that is balanced with respect to word count would be required to resolve this conclusively. Similarly, an important direction for future work will be to isolate individual time series features and consider their importance to predictive models. By doing so, we will gain a better understanding of the contribution of each feature toward quantifying coherence and draw closer toward interpretability.

Another limitation concerns the use of only one construct, derailment, for the assignment of manual coherence scores. While this is arguably the construct most conducive to modeling in the context of short recordings of spontaneous speech without a directive prompt (as compared with, for example modeling *tangentiality* as a divergence from such a prompt), it is of interest for future work to determine the extent to which ASR errors affect other digital speech-based diagnostic approaches such as the use of speech graphs [54].

We have also yet to evaluate the influence of regional differences in language use and dialectical differences on the performance of our models. Recent research suggests these are important concerns for both speech recognition and NLP [55-57], but we have yet to establish the extent to which measures of coherence are affected. In future work, ASR accuracy may also be further improved by adapting the ASR model to individual participants via an enrollment process in which participants are asked to read a short passage with the resulting audio used to finetune the ASR model.

Another limitation is that this study did not evaluate neural models that have developed from the entity grid approach. We performed entity grid analysis on our data as an alternative feature representation for comparison to TARDIS features in our regression models. However, entity grids are based on discrete symbols while distributed representations, which permit models to draw associations between words with related meanings, are by now a well-established means to model semantics in ThD. Recent work has developed variants of the entity grid approaches that involve neural networks [58-60]. These, along with other neural network approaches that involve syntactic structure input [61] are known as neural coherence models. These models offer the advantage of considering both distributional semantics and syntactic structure from texts. While such models have yet to be evaluated for their utility as a means to model coherence in ThD, the performance of the entity grid model in our experiments suggests their potential as a direction for future work.

5 Conclusions

In the context of the task of quantifying formal thought disorder in participants experiencing AVH, this study considers the use of smartphone-based data collection in conjunction with ASR in speech data collection as a means to mitigate logistical constraints on the application of these methods for monitoring of symptoms between clinic visits. Our findings show that the robustness of coherence metrics to ASR-induced transcription errors is enhanced by our novel representation approach, TARDIS, which improves the alignment of automated assessments of derailment with human judgment. As such, our methods show potential as a way to enhance existing coherence metrics and pave the way toward fully automated detection of disorganized thinking in naturalistic settings with implications for research and practice. Our implementation of TARDIS and the underlying coherence metrics used in this work is publicly available on Github[12].

Acknowledgement

This research was supported by the National Institute of Mental Health grant R01MH112641, its supplement 3R01MH112641-03S2, and the National Institute of Aging grant R21AG069792.

References:

[1]. Andreasen NC, Grove WM, Thought, language, and communication in schizophrenia: diagnosis and prognosis., Schizophr. Bull (1986) pg.352. 10.1093/schbul/12.3.348.

[2]. Kircher T, Krug A, Stratmann M, Ghazi S, Schales C, Frauenheim M, Turner L, Fährmann P, Hornig T, Katzev M, Grosvald M, Müller-Isberner R, Nagels A, A rating scale for the assessment of objective and subjective formal thought and language disorder (TALD), Schizophr. Res (2014). 10.1016/j.schres.2014.10.024.

[3]. Foltz PW, Kintsch W, Landauer TK, The measurement of textual coherence with latent semantic analysis, Discourse Process. (1998). 10.1080/01638539809545029.

[4]. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R, Indexing by latent semantic analysis, J. Am. Soc. Inf. Sci (1990) pg.7–15. 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

[5]. Elvevåg B, Foltz PW, Weinberger DR, Goldberg TE, Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia, Schizophr. Res (2007). 10.1016/j.schres.2007.03.001.

[6]. 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). 10.1016/j.jneuroling.2009.05.002.

[7]. Bedi G, Carrillo F, Cecchi GA, Slezak DF, Sigman M, Mota NB, Ribeiro S, Javitt DC, Copelli M, Corcoran CM, Automated analysis of free speech predicts psychosis onset in high-risk youths, Npj Schizophr. (2015). 10.1038/npjschz.2015.30.

[8]. Corcoran CM, Carrillo F, Fernández-Slezak D, Bedi G, Klim C, Javitt DC, Bearden CE, Cecchi GA, Prediction of psychosis across protocols and risk cohorts using automated language analysis, World Psychiatry. (2018). 10.1002/wps.20491.

[9]. Just S, Haegert E, Kořánová N, Bröcker A-L, Nenchev I, Funcke J, Montag C, Stede M, Coherence models in schizophrenia, (2019). 10.18653/v1/w19-3015.

[10]. Just SA, Haegert E, Kořánová N, Bröcker AL, Nenchev I, Funcke J, Heinz A, Bermpohl F, Stede M, Montag C, Modeling Incoherent Discourse in Non-Affective Psychosis, Front. Psychiatry (2020). 10.3389/fpsyt.2020.00846.

[12] TARDIS: https://github.com/LinguisticAnomalies/Coherence

[11]. Tang SX, Kriz R, Cho S, Park SJ, Harowitz J, Gur RE, Bhati MT, Wolf DH, Sedoc J, Liberman MY, Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders, Npj Schizophr. (2021). 10.1038/s41537-021-00154-3.

[12]. Devlin J, Chang MW, Lee K, Toutanova K, BERT: Pre-training of deep bidirectional transformers for language understanding, in: NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf. (2019). 10.18653/v1/N19-1423

[13]. Aschbrenner KA, Naslund JA, Grinley T, Bienvenida JCM, Bartels SJ, Brunette M, A Survey of Online and Mobile Technology Use at Peer Support Agencies, Psychiatr. Q (2018). 10.1007/s11126-017-9561-4.

[14]. Torous J, Friedman R, Keshavan M, Smartphone Ownership and Interest in Mobile Applications to Monitor Symptoms of Mental Health Conditions, JMIR MHealth UHealth. (2014). 10.2196/mhealth.2994.

[15]. Varadan V, Mittal P, Vaske CJ, Benz SC, The integration of biological pathway knowledge in cancer genomics: A review of existing computational approaches, IEEE Signal Process. Mag (2012). 10.1109/MSP.2011.943037.

[16]. Buck B, Hallgren KA, Scherer E, Brian R, Wang R, Wang W, Campbell A, Choudhury T, Hauser M, Kane JM, Ben-Zeev D, Capturing behavioral indicators of persecutory ideation using mobile technology, J. Psychiatr. Res (2019). 10.1016/j.jpsychires.2019.06.002.

[17]. Buck B, Scherer E, Brian R, Wang R, Wang W, Campbell A, Choudhury T, Hauser M, Kane JM, Ben-Zeev D, Relationships between smartphone social behavior and relapse in schizophrenia: A preliminary report, Schizophr. Res (2019). 10.1016/j.schres.2019.03.014.

[18]. Xu W, Portanova J, Chander A, Ben-Zeev D, Cohen T, The Centroid Cannot Hold: Comparing Sequential and Global Estimates of Coherence as Indicators of Formal Thought Disorder, AMIA … Annu. Symp. Proceedings. AMIA Symp (2020).

[19]. Holmlund TB, Chandler C, Foltz PW, Cohen AS, Cheng J, Bernstein JC, Rosenfeld EP, Elvevåg B, Applying speech technologies to assess verbal memory in patients with serious mental illness, Npj Digit. Med (2020). 10.1038/s41746-020-0241-7.

[20]. Chandler C, Foltz PW, Cohen AS, Holmlund TB, Cheng J, Bernstein JC, Rosenfeld EP, Elvevåg B, Machine learning for ambulatory applications of neuropsychological testing, Intell. Med (2020). 10.1016/j.ibmed.2020.100006.

[21]. Amodei D, Ananthanarayanan S, Anubhai R, Bai J, Battenberg E, Case C, Casper J, Catanzaro B, Cheng Q, Chen G, Chen J, Chen J, Chen Z, Chrzanowski M, Coates A, Diamos G, Ding K, Du N, Elsen E, Engel J, Fang W, Fan L, Fougner C, Gao L, Gong C, Hannun AN, Han T, Johannes LV, Jiang B, Ju C, Jun B, Legresley P, Lin L, Liu J, Liu Y, Li W, Li X, Ma D, Narang S, Ng A, Ozair S, Peng Y, Prenger R, Qian S, Quan Z, Raiman J, Rao V, Satheesh S, Seetapun D, Sengupta S, Srinet K, Sriram A, Tang H, Tang L, Wang C, Wang J, Wang K, Wang Y, Wang Z, Wang Z, Wu S, Wei L, Xiao B, Xie W, Xie Y, Yogatama D, Yuan B, Zhan J, Zhu Z, Deep speech 2: End-to-end speech recognition in English and Mandarin, in: 33rd Int. Conf. Mach. Learn. ICML 2016. (2016).

[22]. Ben-Zeev D, Buck B, Chander A, Brian R, Wang W, Atkins D, Brenner CJ, Cohen T, Campbell A, Munson J, Mobile RDoC: Using Smartphones to Understand the Relationship Between Auditory Verbal Hallucinations and Need for Care, Schizophr. Bull. Open (2020). 10.1093/schizbullopen/sgaa060.

[23]. Graves A, Fernández S, Gomez F, Schmidhuber J, Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks, in: ACM Int. Conf. Proceeding Ser. (2006). 10.1145/1143844.1143891.

[24]. MacWhinney B, Wagner J, Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository., Gesprachsforsch. Online-Zeitschrift Zur Verbalen Interaktion (2010).

[25]. Rousseau A, Deléglise P, Estève Y, Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks, in: Proc. 9th Int. Conf. Lang. Resour. Eval. Lr. 2014. (2014).

[26]. Panayotov V, Chen G, Povey D, Khudanpur S, Librispeech: An ASR corpus based on public domain audio books, in: ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc. (2015). 10.1109/ICASSP.2015.7178964.

[27]. Hodosh M, Young P, Hockenmaier J, Framing image description as a ranking task: Data, models and evaluation metrics, J. Artif. Intell. Res (2013). 10.1613/jair.3994.

[28]. Veaux C, Yamagishi J, MacDonald K, CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit, Cent. Speech Technol. Res (2016).

[29]. Köhn A, Stegen F, Baumann T, Mining the spoken Wikipedia for speech data and beyond, in: Proc. 10th Int. Conf. Lang. Resour. Eval. Lr. 2016. (2016).

[30]. Stolcke A, SRILM - An extensible language modeling toolkit, in: 7th Int. Conf. Spok. Lang. Process. ICSLP 2002. (2002).

[31]. Ney H, Essen U, Kneser R, On structuring probabilistic dependences in stochastic language modelling, Comput. Speech Lang (1994). 10.1006/csla.1994.1001.

[32]. Tilk O, Alumäe T, Bidirectional recurrent neural network with attention mechanism for punctuation restoration, in: Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH. (2016). 10.21437/Interspeech.2016-1517.

[33]. Bird S, Bird S, Loper E, NLTK : The natural language toolkit NLTK : The Natural Language Toolkit, Proc. ACL-02 Work. Eff. Tools Methodol. Teach. Nat. Lang. Process. Comput. Linguist 1. (2016).

[34]. Honnibal M, Montani I, Van Landeghem S, Boyd A, spaCy: Industrial-strength Natural Language Processing in Python, (2020). 10.5281/zenodo.1212303.

[35]. Cohen T, Widdows D, Empirical distributional semantics: Methods and biomedical applications,J. Biomed. Inform (2009). 10.1016/j.jbi.2009.02.002.

[36]. Turney PD, Pantel P, From frequency to meaning: Vector space models of semantics, J. Artif. Intell. Res (2010). 10.1613/jair.2934.

[37]. Handbook of Latent Semantic Analysis, Part1. (2007). 10.4324/9780203936399.

[38]. Mikolov T, Chen K, Corrado G, Dean J, Efficient estimation of word representations in vector space, in: 1st Int. Conf. Learn. Represent. ICLR 2013 - Work. Track Proc. (2013).

[39]. Joulin A, Grave E, Bojanowski P, Mikolov T, Bag of tricks for efficient text classification, in: 15th Conf. Eur. Chapter Assoc. Comput. Linguist. EACL 2017 - Proc. Conf. (2017). 10.18653/v1/e17-2068.

[40]. Reimers N, Gurevych I, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf., 2020. 10.18653/v1/d19-1410.

[41]. Christ M, Braun N, Neuffer J, Kempa-Liehr AW, Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package), Neurocomputing. (2018). 10.1016/j.neucom.2018.03.067.

[42]. Barzilay R, Lapata M, Modeling local coherence: An entity-based approach, Comput. Linguist (2008). 10.1162/coli.2008.34.1.1.

[43]. Kim SH, Jung HY, Hwang SS, Chang JS, Kim Y, Ahn YM, Kim YS, The usefulness of a self-report questionnaire measuring auditory verbal hallucinations, Prog. Neuro-Psychopharmacology Biol. Psychiatry (2010). 10.1016/j.pnpbp.2010.05.005.

[44]. Van Lieshout RJ, Goldberg JO, Quantifying self-reports of auditory verbal hallucinations in persons with psychosis, Can. J. Behav. Sci (2007). 10.1037/cjbs2007006.

[45]. Sommer IE, Derwort AMC, Daalman K, de Weijer AD, Liddle PF, Boks MPM, Formal thought disorder in non-clinical individuals with auditory verbal hallucinations, Schizophr. Res (2010). 10.1016/j.schres.2010.01.024.

[46]. Fisher RA, Statistical Methods for Research Workers, In: Kotz S, Johnson NL (eds) Breakthroughs in Statistics pg.66–70. Springer Series in Statistics (Perspectives in Statistics). Springer, New York, NY (1992). 10.1007/978-1-4612-4380-9_6.

[47]. Voleti R, Liss JM, Berisha V, Investigating the Effects of Word Substitution Errors on Sentence Embeddings, in: ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., 2019. 10.1109/ICASSP.2019.8683367.

[48]. Hoffman RE, Verbal hallucinations and language production processes in schizophrenia, Behav. Brain Sci (1986). 10.1017/S0140525X00046781.

[49]. Covington MA, He C, Brown C, Naçi L, McClain JT, Fjordbak BS, Semple J, Brown J, Schizophrenia and the structure of language: The linguist's view, Schizophr. Res (2005). 10.1016/j.schres.2005.01.016.

[50]. Corcoran CM, Mittal VA, Bearden CE, Gur RE, Hitczenko K, Bilgrami Z, Savic A, Cecchi GA, Wolff P, Language as a biomarker for psychosis: A natural language processing approach, Schizophr. Res (2020). 10.1016/j.schres.2020.04.032.

[51]. Christ M, Kempa-Liehr AW, Feindt M, Distributed and parallel time series feature extraction for industrial big data applications, CoRR. abs/1610.0 (2016). http://arxiv.org/abs/1610.07717.

[52]. Lundberg SM, Lee SI, A unified approach to interpreting model predictions, in: Adv. Neural Inf. Process. Syst (2017).

[53]. Bocher M, Introduction to the Theory of Fourier's Series, Ann. Math (1906). 10.2307/1967238.

[54]. Mota NB, Vasconcelos NAP, Lemos N, Pieretti AC, Kinouchi O, Cecchi GA, Copelli M, Ribeiro S, Speech graphs provide a quantitative measure of thought disorder in psychosis, PLoS One. (2012). 10.1371/journal.pone.0034928.

[55]. Tan S, Joty S, Varshney L, Kan M-Y, Mind Your Inflections! Improving NLP for Non-Standard Englishes with Base-Inflection Encoding, (2020). 10.18653/v1/2020.emnlp-main.455.

[56]. Tan S, Joty S, Kan M-Y, Socher R, It's Morphin' Time! Combating Linguistic Discrimination with Inflectional Perturbations, (2020). 10.18653/v1/2020.acl-main.263.

[57]. Koenecke A, Nam A, Lake E, Nudell J, Quartey M, Mengesha Z, Toups C, Rickford JR, Jurafsky D, Goel S, Racial disparities in automated speech recognition, Proc. Natl. Acad. Sci. U. S. A (2020). 10.1073/pnas.1915768117.

[58]. Nguyen DT, Joty S, A neural local coherence model, in: ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. Long Pap. (2017). 10.18653/v1/P17-1121.

[59]. Mohiuddin T, Joty S, Nguyen DT, Coherence modeling of asynchronous conversations: A neural entity grid approach, in: ACL 2018 - 56th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. Long Pap. (2018). 10.18653/v1/p18-1052.

[60]. Moon HC, Mohiuddin T, Joty S, Chi X, A unified neural coherence model, in: EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf. (2020). 10.18653/v1/d19-1231.

[61]. Jeon S, Strube M, Centering-based Neural Coherence Modeling with Hierarchical Discourse Segments, (2020). 10.18653/v1/2020.emnlp-main.604.

Highlights:

- Quantifying coherence in speech identifies formal thought disorder automatically
- Manual transcription constrains research and practice applications
- Standard coherence estimates are vulnerable to automated transcription errors
- TARDIS - our novel method for estimating coherence - is robust to such errors
- TARDIS applies to both contextual and skip-gram semantic embeddings
- TARDIS better aligns with coherence estimates from professional transcripts
- This facilitates scalable, privacy-preserving automated coherence estimation

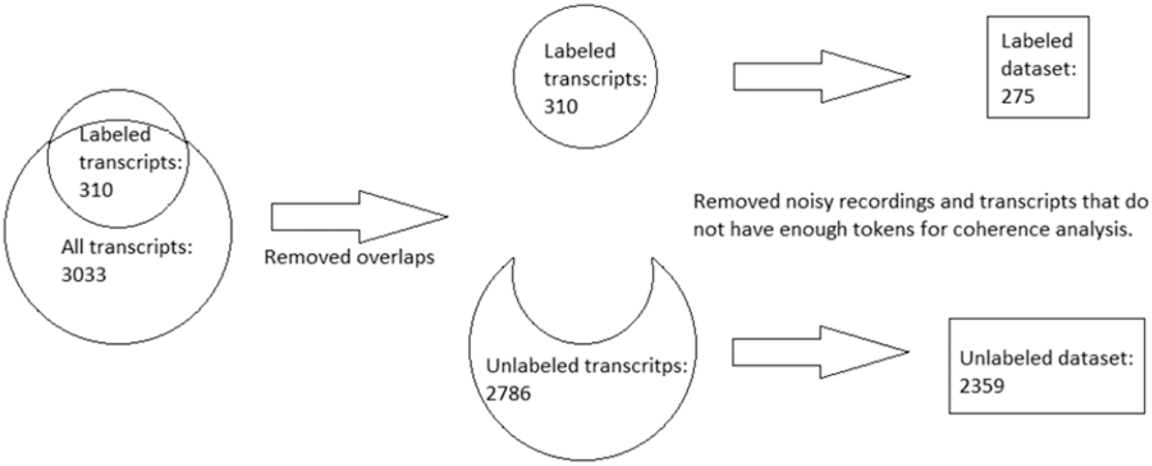


Figure 1:

Data aggregation and processing.

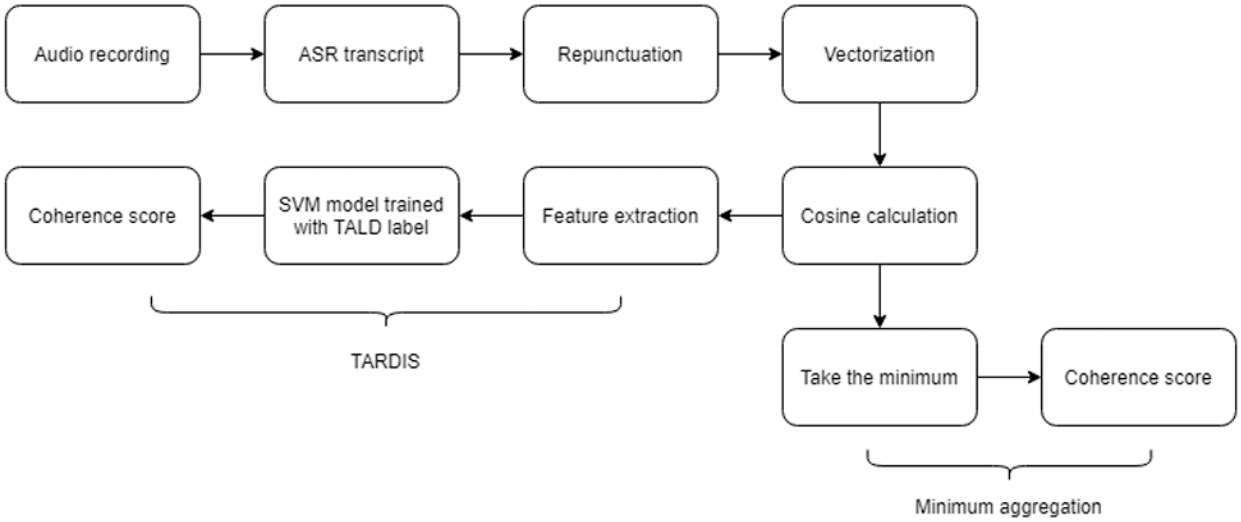


Figure 2:
Summary of coherence analytical pipeline.

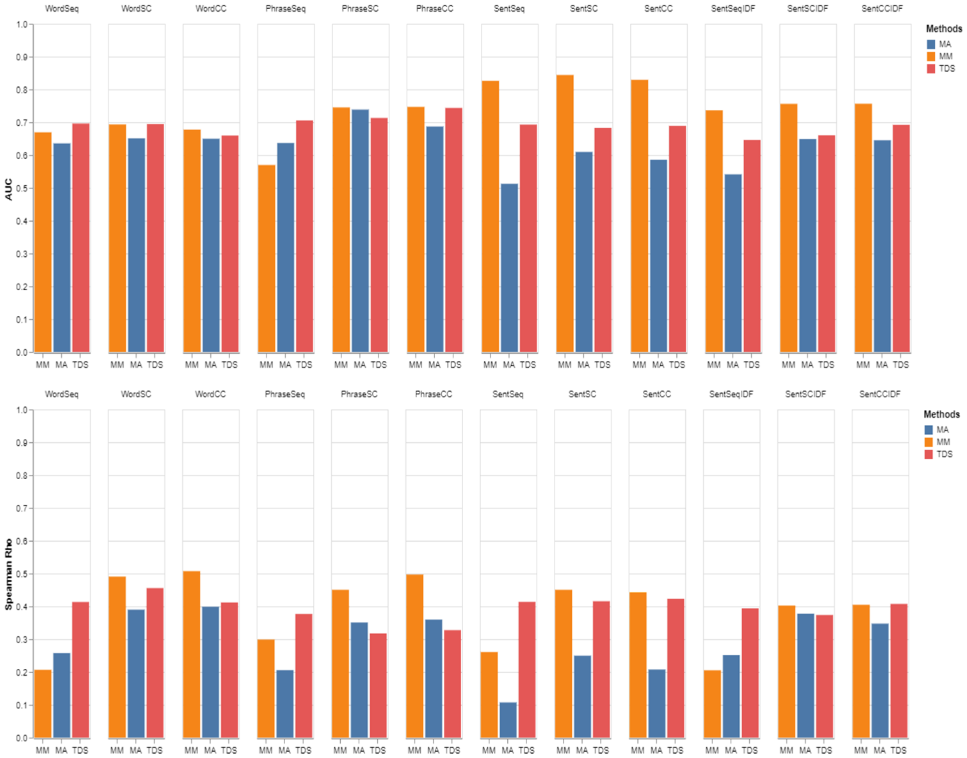


Figure 3:

The performance comparison among manual minimum coherence (reference metric), ASR minimum coherence, and ASR time-series coherence. (a) Top: Evaluation by AUC, (b) Bottom: Evaluation by Spearman Rho. **MM** = minimum coherence with manual transcripts. **MA** = minimum coherence with automated transcripts. **TDS** = time-series based coherence with automated transcripts. **SC** = static centroid. **CC** = cumulative centroid. **IDF** = inverse document frequency.

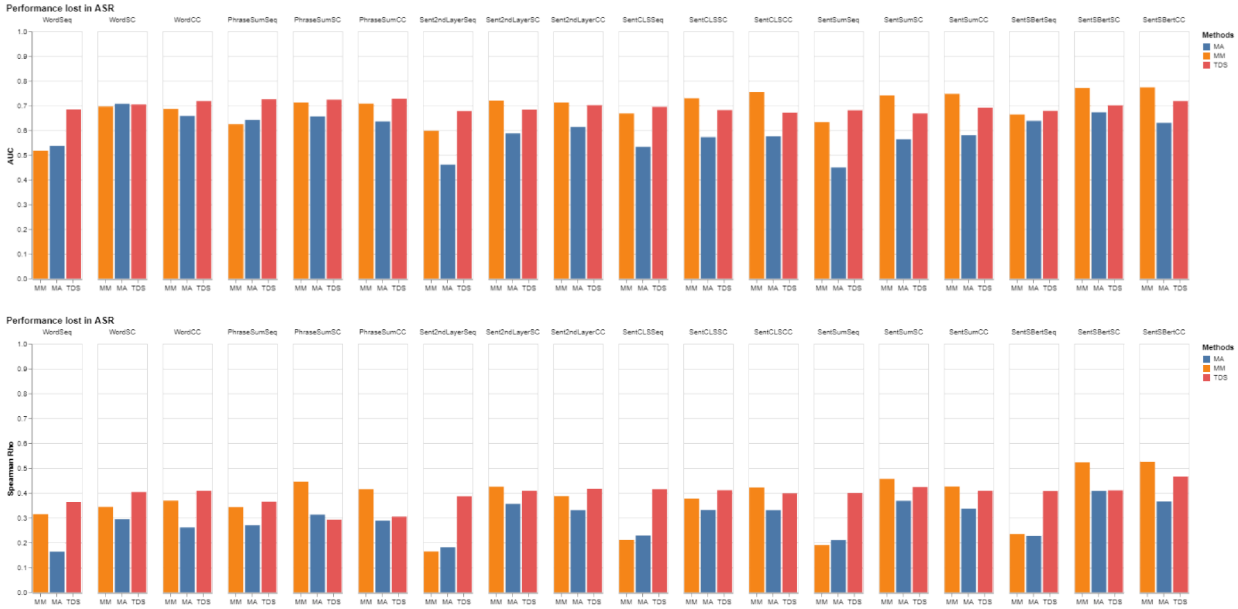


Figure 4:

The performance comparison among manual minimum coherence (reference metric), ASR minimum coherence, and ASR time-series coherence with BERT-derived contextual vectors. (a) Top: Evaluation by AUC, (b) Bottom: Evaluation by Spearman Rho. **MM** = minimum coherence with manual transcripts. **MA** = minimum coherence with automated transcripts. **TDS** = time-series based coherence with automated transcripts. **SC** = static centroid. **CC** = cumulative centroid. **CLS** = embeddings from the CLS token. **Sum** = obtained from the sum of individual word vectors. **2ndLayer** = 2nd to last layer of BERT hidden state output. **SBert** = vectors from the sentence-BERT package [40].

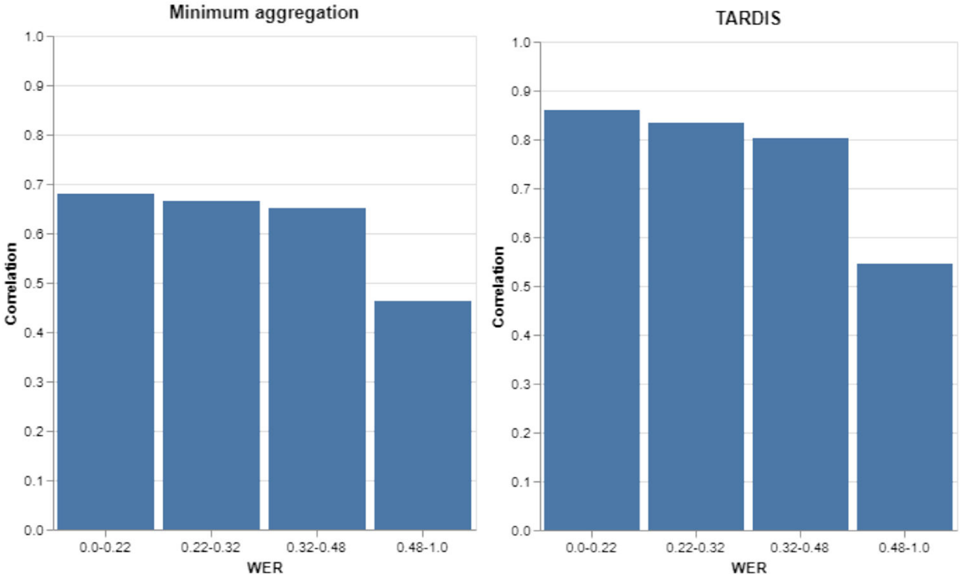


Figure 5:

Correlations between average coherence scores derived from ASR and manual transcripts. Each bin corresponds to one quartile from the distribution of coherence scores for each transcript. (A) The left figure shows averages for coherence metrics with minimum aggregation. (B) The right figure shows averages for TARDIS-derived coherence metrics.

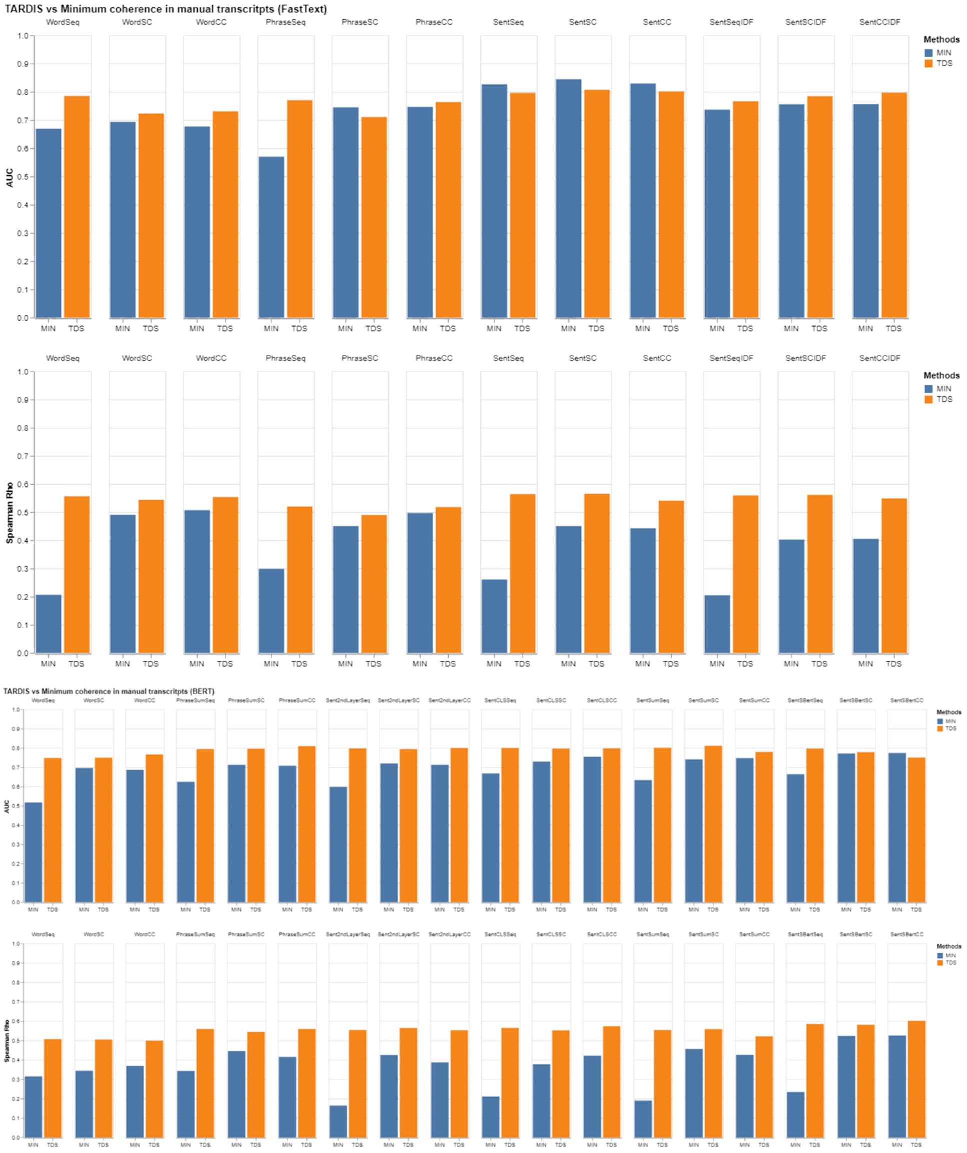


Figure 6:

Comparison of TARDIS and Minimum coherence performance using FastText (top) and BERT embeddings (bottom) with manual transcripts. **MIN**= minimum coherence with manual transcripts. **TDS** = time-series based coherence with automated transcripts. **SC** = static centroid. **CC** = cumulative centroid. **Sum** = obtained from the sum of individual word vectors. **2ndLayer** = 2nd to last layer of BERT hidden state output. **SBERT** = vectors from the sentence-BERT package [40].

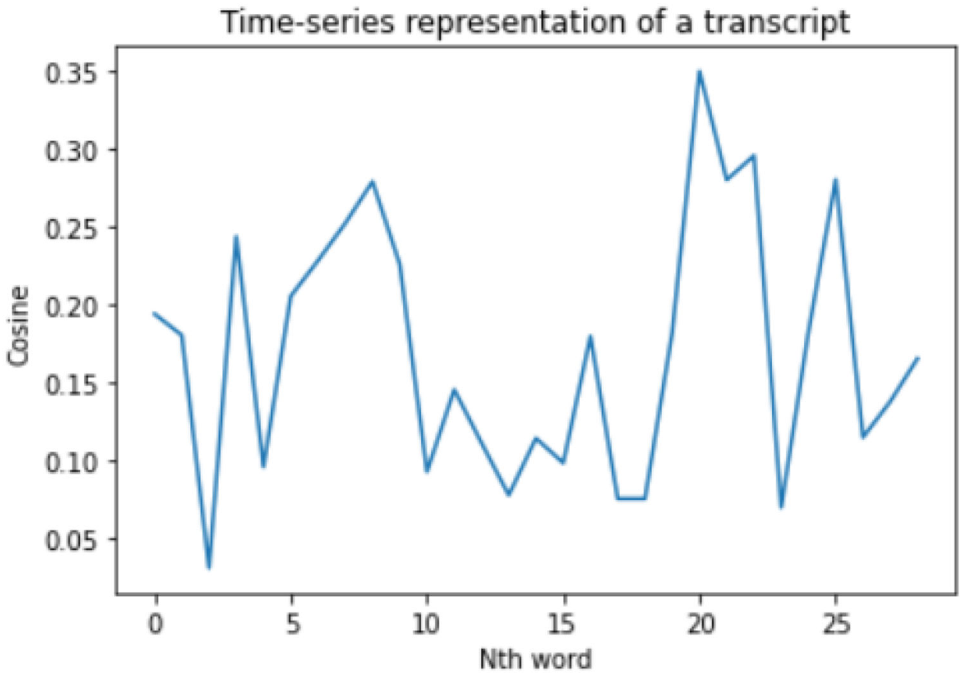


Figure 7:

Time-series representation of an illustrative transcript. The y-axis represents cosine values computed between adjacent word vectors while the x-axis represents the order in which the words appear in the text. The TARDIS takes into consideration many characteristics of this plot such as the mean, max, or regression line slope, not just the minimum, which is the only data point considered by standard approaches.

Table 1:

Comparison of different embeddings with the best performing metric on each semantic level (Using TARDIS on ASR transcripts). Best results in **boldface**. (**IDF** = inverse document frequency, **SBERT** = vectors from the sentence-BERT package [40].)

	Word		**Phrase**		**Sentence**	
	Skip-gram embeddings (FastText)	**Contextual embeddings (BERT)**	**Skip-gram embeddings (FastText)**	**Contextual embeddings (BERT)**	**Skip-gram embeddings (FastText)**	**Contextual embeddings (BERT)**
Best ROC-AUC	0.696	**0.718**	**0.744**	0.728	0.698	**0.718**
Best Metric	Sequential	Cumulative Centroid	Cumulative Centroid	Cumulative Centroid	IDF Cumulative Centroid	SBERT Cumulative Centroid
Best Spearman Rho	**0.456**	0.409	**0.377**	0.364	0.414	**0.465**
Best Metric	Static Centroid	Cumulative Centroid	Sequential	Sequential	Cumulative Centroid	SBERT Cumulative Centroid

Table 2:

Spearman Rho correlations between manually transcribed and ASR transcript derived coherence scores. **IDF** = inverse document frequency, **BERT2ndLayer** = 2nd to last layer of BERT hidden state output, **BERTCLS** = embeddings from the CLS token, **BERTSum** = obtained from the sum of individual word vectors, **SBERT** = vectors from the sentence-BERT package [40].

Metrics	**TARDIS**	**Minimum**
Word Sequence (FastText)	**0.698**	0.453
Word Centroid (FastText)	**0.740**	0.504
Word Cumulative Centroid (FastText)	**0.738**	0.603
Phrase Sequence (FastText)	**0.764**	0.445
Phrase Centroid (FastText)	**0.772**	0.689
Phrase Cumulative Centroid (FastText)	**0.767**	0.723
Sentence Sequence	**0.695**	0.186
Sentence Centroid	**0.692**	0.408
Sentence Cumulative Centroid	**0.684**	0.398
Sentence IDF Sequence	**0.694**	0.252
Sentence IDF Centroid	**0.679**	0.454
Sentence IDF Cumulative Centroid	**0.669**	0.461
Word Sequence (BERT)	**0.757**	0.386
Word Centroid (BERT)	**0.770**	0.431
Word Cumulative Centroid (BERT)	**0.791**	0.421
Phrase Sequence (BERT)	**0.756**	0.337
Phrase Centroid (BERT)	**0.765**	0.567
Phrase Cumulative Centroid (BERT)	**0.789**	0.601
Sentence BERT2ndLayer Sequence	**0.717**	0.192
Sentence BERT2ndLayer Centroid	**0.704**	0.405
Sentence BERT2ndLayer Cumulative Centroid	**0.701**	0.404
Sentence BERTCLS Sequence	**0.692**	0.201
Sentence BERTCLS Centroid	**0.689**	0.341
Sentence BERTCLS Cumulative Centroid	**0.691**	0.372
Sentence BERTSum Sequence	**0.709**	0.196
Sentence BERTSum Centroid	**0.703**	0.422
Sentence BERTSum Cumulative Centroid	**0.701**	0.423
Sentence SBERT Sequence	**0.679**	0.312
Sentence SBERT Centroid	**0.699**	0.635
Sentence SBERT Cumulative Centroid	**0.706**	0.639
Mean	**0.720**	0.429

Table 3:

Performance comparison across the best performing TARDIS metrics and the entity grid metric. (**SBERT** = vectors from the sentence-BERT package [40])

	Auto transcripts		**Manual transcripts**	
	AUC-ROC	**Spearman Rho**	**AUC-ROC**	**Spearman Rho**
Entity grid	0.733	0.457	0.767	0.438
TARDIS	**0.744**	**0.465**	**0.811**	**0.601**
TARDIS metric	Phrase Cumulative Centroid (FastText)	Sentence SBERT Cumulative Centroid	Sentence SBERT Cumulative Centroid	Word Sequential (BERT)

Table 4:

Mean and max Spearman Rho correlations between coherence scores and HPSVQ total score. The mean and max were aggregated across all coherence metrics and the metrics that produced the max correlation were also included in the table. The p-values for mean correlation were calculated using Fisher's combined probability test [46]. (**BERTCLS** = embeddings from the CLS token, **SBERT** = vectors from the sentence-BERT package [40].)

	Minimum (ASR)	**TARDIS (ASR)**	**Minimum (Manual)**	**TARDIS (Manual)**
Mean correlation	0.181 (P<.001)	**0.203** (P<.001)	0.191 (P<.001)	**0.237** (P<.001)
Max correlation	**0.240** (P<.001)	0.237 (P<.001)	0.271 (P<.001)	**0.275** (P<.001)
Max correlation metric	Word Static Centroid (BERT)	Phrase Sequence (BERT)	Sentence SBERT Static Centroid	Sentence BERTCLS Sequence