Comprehensive assessment and quantification of incoherent speech using natural language processing

Weizhe Xu

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Trevor A. Cohen, Chair

Serguei Pakhomov

Patrick Heagerty

Eric Horvitz

Program Authorized to Offer Degree:
Biomedical Informatics and Medical Education
University of Washington

University of Washington

Abstract

Comprehensive assessment and quantification of incoherent speech using natural language processing

Weizhe Xu

Chair of the Supervisory Committee:

Trevor A. Cohen

Department of Biomedical Informatics and Medical Education

Coherence is a linguistic feature that is defined as the orderly and interconnected flow of ideas. The disruption of coherence is a linguistic anomaly that is commonly observed in a group of psychiatric disorders known as schizophrenia spectrum disorders (SSD), where disorganized thoughts manifest as incoherent speech. While early detection of symptoms can potentially lead to better outcomes, manual assessment of symptom severity can be time-consuming and require specialized expertise. Therefore, symptom evaluation through automated coherence assessment methods is desired.

However, gaps remain in prior research on this area, namely 1) most prior work focuses on the estimation of local coherence (coherent transitions between adjacent semantic units) via computation of cosine values between vector representations of sequential semantic units. The estimation of global coherence (sustaining a theme or topic throughout a narrative) has received much less attention; 2) the impact of automated speech recognition (ASR) errors receives little attention. Prior work mainly focused on using manual transcript data; 3) there is limited exploration on using language model perplexity to assess coherence, especially given the recent advancement of large language models (LLM).

This work bridges the gaps through the following contributions: 1) Two new global coherence assessment methods were developed based on centroids of embeddings (vector representation of semantic space). We found that the global coherence methods align better with human judgment than local coherence methods. 2) A time-series feature extraction pipeline is used to replace the aggregation step in coherence assessment pipelines. We found that by using this method, coherence evaluation process is resistant to the impact of ASR errors in the text input. 3) Two sentence-level perplexity-based coherence methods were developed, and we revealed that combining perplexity features with traditional coherence scores (proximity features because they are based on cosine similarity) resulting in better prediction models than using proximity or perplexity features alone. 4) The innovations and classical approaches were combined into the Comprehensive Coherence Calculator (CCC), a software package that can perform comprehensive coherence analysis with a myriad of configurations. With these contributions, fully automated coherence assessment pipeline can be established to offer patients easy monitoring at home, clinicians necessary information to provide better care and researchers an objective quantitative basis for the study of semantic coherence.

Table of contents

Acknowledgements

I would like to offer my deepest gratitude to my advisor Dr. Trevor Cohen for his guidance throughout my doctoral studies. He is knowledgeable, supportive, and always responsive to any questions I may have. I am also thankful to my reading committee members Dr. Serguei Pakhomov, Dr. Patrick Heagerty and Dr. Eric Horvitz for their insightful feedback and valuable suggestions throughout the journey and graduate school representative Dr. David Beck for holding my work up to UW standard.

I extend my thanks to all members of the Halo study, led by Dr. Dror Ben-Zeev and Dr. Trevor Cohen. Thanks for building a warm and positive working environment and supporting my doctoral work. I am also grateful for BIME faculty and staff. The classes I had were the most memorable and they are always ready to help me overcome any challenges.

Finally, I thank my family for their unconditional love and always having my back whatever happens. Also, special thanks to my good friends Yifan, Qifei, Changye and Feng for bracing many challenges of graduate school together.

Dedicated to my mentors, who guided me throughout this journey.

List of Figures

List of Tables

Chapter 1: Introduction and overview

1.1 Schizophrenia and benefits of early detection

Schizophrenia is a serious mental illness that affects 4.6 per 1000 persons globally (Saha et al., 2005). It is associated with severe deterioration of quality of life (Bobes et al., 2007), neurocognitive function (Heinrichs & Zakzanis, 1998) and social cognition (Green et al., 2015). Without intervention, the condition has a high rate of fatality through the means of suicide (Pompili et al., 2007).

Early detection is key to mitigating the effects and severity of symptoms of schizophrenia. Longer duration of untreated psychosis is directly associated with negative outcomes while reduced duration of untreated psychosis can lead to benefits such as reduced suicide rate (McGorry et al., 2008; Nordentoft et al., 2009).

1.2 Thought disorder and incoherent speech

To achieve early detection, one method is to evaluate the manifestation of thought disorder (TD), an important diagnostic feature of schizophrenia (Andreasen & Grove, 1986; Andreasen & TUCKER, 1991; Ludwig & Othmer, 1977). TD can manifest as abnormalities in speech patterns, ranging from loose association of content (derailment) to completely incomprehensible speech (Andreasen, 1986). To capture these patterns, clinical evaluation scales have been developed. A widely accepted scale is the Thought, Language and Communication scale (TLC) (Andreasen, 1986), which touches on many aspects of speech abnormalities such as poverty of speech, pressure of speech and derailment. A newer and more comprehensive scale is the Thought and Language Disorder (TALD) scale (Kircher et al., 2014), which includes 30 items based on

review of the literature since early 20th century and a sizable clinical trial, which demonstrated the scale's ability to portray various dimensions of thought disorder. Aside from these, self-report scales have also been developed (Barrera et al., 2008; Liddle et al., 2002). However, these are limited by their capacity to capture objective symptoms or subtle changes in thought process. On the other hand, more comprehensive clinical scales such as the Positive and Negative Syndrome Scale (PANSS) (Kay et al., 1987) require specialized training and expertise to administer and are time-consuming. To address these issues, automated analysis methods can be employed to detect abnormalities in speech samples. This dissertation focuses on the detection of incoherent speech or derailment, a major aspect of TD, through automated natural language processing (NLP) methods. Note that while the "derailment" of the train of thought can manifest over the span of several sentences or clauses, incoherent speech can appear as irregularities within sentences and therefore is a broader concept.

Incoherent speech or derailment can be interpreted as loose association of contents or conceptual disorganization. An example of incoherent speech can be found in Elyn Sak's autobiographical account of her own experience of a psychotic spectrum condition, The Center Cannot Hold:

> *The memo materials have been infiltrated. They are jumping around. I used to be good at the broad jump, because I'm tall. I fall. People put things in and then say it's my fault. I used to be God, but I got demoted.*

While the sentences are syntactically correct, it is hard to perceive the overall meaning because the topics transition drastically from sentence to sentence with little or no apparent connection. Different degrees of this type of incoherent speech are referred to as “derailment”, “loose associations”, or “flight of ideas” in psychiatry.

1.3 Automated assessment of incoherent speech: an overview

This section provides an overview of existing methods of automated assessment of incoherent speech through Table 1.1. Each of the categories is elaborated in Chapter 2 of this document.

Table 1.1: Overview of documented coherence assessment methods

Main category	Key technology	Description
Sequential cosine similarity (Proximity)	LSA vectors (Bedi et al., 2015; Britton & Gülgöz, 1991; Corcoran et al., 2018; Elvevåg et al., 2007, 2010; Foltz et al., 1998)	Captures semantic relatedness through co-occurrence.
	Static word embeddings (Iter et al., 2018; S. Just et al., 2019)	Semantic word embeddings derived from neural networks via self-supervised learning. Context is not considered.
	Contextual word embeddings (Figueroa-Barra et al., 2022; Tang et al., 2021)	Semantic word embeddings derived from neural networks via self-supervised learning. Context is considered.

Perplexity	Language model (GPT) (Fradkin et al., 2023; Sharpe et al., 2024)	Perplexity or probability derived from a language model to represent coherence scores.
Speech graph	Graph properties (Mota et al., 2012)	Texts mapped to graphs and use graph properties to assess coherence.
Syntactical alteration	Entity grid (Barzilay & Lapata, 2008; Mohiuddin et al., 2018)	Captures the frequency of syntactical role transition of entities in sentences.

While the current literature offers a solid foundation for automated coherence assessment, there is room for improvement for each category of methods. For example, the first main category, which employs various versions of distributional representations within a semantic space, computes the consecutive cosine similarities between such representations of units of text that occur in sequence. As such, the coherence assessment only captures the transition between consecutive semantic units. Thus, it only captures coherence on a local level (local coherence) while offering no insight into the ability to sustain a topic throughout the speech (global coherence).

Additionally, aside from speech graphs and entity grids, the other methods all produce a sequence of scores related to each semantic unit. To represent an overall coherence score for an entire speech transcript, these scores are usually aggregated. However, doing so might cause the overall score to be affected by extreme values depending on the chosen aggregation function,

especially when using transcripts generated from automated speech recognition (ASR) programs because machine transcription errors could cause erratic changes in speech content.

Last but not least, there is limited research on using different categories of methods in combination with each other. While each of the methods is attempting to capture some aspects of coherence, they are all limited in their capability to provide a comprehensive picture of coherence. As such, it may be beneficial to explore potential combinations of different coherence assessment approaches to improve evaluation quality.

1.4 Hypotheses

Based on the above discussion, I propose the following hypotheses to address each of the concerns.

Concern 1: Sequential coherence methods only capture local coherence which may offer insufficient insight into coherence assessment.

Hypothesis 1: Automated assessments of global coherence can identify linguistic manifestations of formal thought disorder as or more effectively than established sequential measures.

Concern 2: Aggregating unit coherence scores makes the total score vulnerable to the influence of extreme values, especially when using ASR transcripts.

Hypothesis 2: Coherence assessments incorporating time-series analysis will be more robust to speech recognition errors than those based on minimum aggregation.

Concern 3: There is limited exploration for combining coherence assessment methods to achieve better scoring quality

Hypothesis 3: Perplexity from LLM complements proximity methods, providing additional information to improve alignment with human appraisal of disorganized thinking.

1.5 Specific aims

I evaluated these hypotheses with the following aims:

Aim 1: Expand the scope of coherence assessment by developing global coherence methods

I developed novel methods to evaluate global coherence, by comparing units of text to an average representation of the document from which they are derived using state-of-the-art NLP technologies. I compared expert- and model-assigned coherence scores for speech samples across a psychosis-related dataset for the purpose of evaluating H1.

Aim 2: Reinforce against ASR errors

I developed and evaluated coherence methods that are robust against ASR errors by treating coherence measurements as a time-series, rather than relying on point estimates (such as the minimum) as in previous work. To evaluate the new methods, I examined the level of correspondence to human labeled coherence scores under different word error rates from ASR algorithms and compare the results produced by the new time-series method to the results produced by the aggregation method to evaluate H2.

Aim 3: Harness the power of large language models (LLMs)

I developed perplexity-based coherence evaluation methods using the state-of-the-art LLMs, further expanding the boundaries of current perplexity-based methods. I combined scores from both perplexity and proximity methods to create a more comprehensive model of coherence. The scores were compared to human annotations, and the performance were compared to baseline performance scores to evaluate H3.

1.6 Road map

This document features 8 chapters. Chapter 2 provides a literature review on current coherence assessment methods, summarizing the studies performed and related results and discoveries, while also pointing out the limitations. Chapter 3 introduces the main innovations in this work: the global coherence evaluation methods, the time-series augmentation, and the combination of proximity- and perplexity-based methods. This chapter also provides an overview of Chapter 4-6, which go over each of the innovations and their evaluation in detail. Chapter 7 introduces a software package, the Comprehensive Coherence Calculator (CCC) as a practical product as a result of the innovations from chapter 4-6. Chapter 8 discusses the contributions, assessment of the hypotheses, potential applications, and future work.

Chapter 2: Background and related work

2.1 Initial evidence that distributional semantics can be used to evaluate coherence

The earliest attempt at assessing semantic coherence using an automated method was Latent Semantic Analysis (LSA), which was presented in Landauer et al's work (Landauer et al., 1998) and subsequently applied to evaluate semantic coherence (Foltz et al., 1998). LSA is a fully automated mathematical technique for extracting and inferring semantics from contextual usage of words in passages of discourse (Landauer et al., 1998). It is designed to uncover the hidden (latent) semantic structure in a collection of texts using linear algebra to derive reduced-dimensional representations of words that reflect semantic similarities.

The first step of LSA analysis is to have the texts represented by a matrix, where each row stands for a unique word and each column stands for a text passage. Each cell contains the frequency the word appears in the passage represented by the columns. This frequency can be raw count or term frequency – inverse document frequency (TF-IDF), which offers better representation because it is less skewed by common words that have low information (such as "a", "the", and "is") because such words are weighed down as they appear in nearly all documents. Next, LSA applies singular value decomposition (SVD), which is a mathematical process that decomposes the singular matrix into the product of three other matrices. The three matrices that the SVD process yields represent word vectors, document vectors and diagonal matrix of singular values, which are sorted from the largest to smallest and the smallest values are pruned to reduce dimensionality of the final vectors, only keeping the top K dimensions. After multiplying the pruned matrices together, we obtain a dense representation of words in the form of vectors,

capable of capturing the semantic relations between words. Subsequently, the LSA vectors are used in cosine similarity computations between sequential semantic units to find incoherence in texts (for semantic units larger than words (i.e. sentences), the vectors are usually represented by the average of word vector components).

Stepping back from the abstract mathematical processes, Landauer et al (Landauer et al., 1998) provides a vivid example of the relationships that LSA is capturing. Consider the sentence:

John is Bob's father and Mary is Ann's mother.

And now suppose we have the second piece of the puzzle:

Mary is Bob's mother.

We can infer from the two sentences that Bob and Ann are probably brothers and sisters, and John and Mary are husband and wife even though the relations are not explicitly stated. The LSA representation of words is of a similar nature. By capturing co-occurrence of words in large number of different passages, indirect associations can be inferred and therefore provide valuable information about the words' semantic similarities with other words.

The LSA technique has been used in many studies to evaluate semantic coherence. Foltz el al re-evaluated two previous coherence evaluation studies using LSA (Britton & Gülgöz, 1991; McNamara et al., 1996), with the former showing that LSA derived scores strongly correlated with readers' comprehension performance and the latter showing that LSA outperformed a

baseline word-overlap approach in predicting low-knowledge readers' comprehension. Later, Elvevåg et al (Elvevåg et al., 2007) compared speech samples from schizophrenia patients and healthy controls completing narrative and discourse tasks. The scores for the speech samples were computed using cosine similarities in an LSA space between adjacent sentences and summarized using aggregation functions (such as minimum or mean). The patient group showed reduced semantic coherence LSA scores, which also correlated with clinical ratings of thought disorder. The same team took a step further in a later study (Elvevåg et al., 2010) that included first-degree relatives of patients as participants and found that LSA scores could detect subtle coherence differences not only in patients but also in their relatives, indicating LSA coherence scores are potentially endophenotypic marker of schizophrenia. Further evidence from Holshausen et al (Holshausen et al., 2014) who used a similar LSA method (cosine similarity between adjacent words), indicated that lower coherence is associated with disconnected speech and the length of the LSA word vectors represents a level of word unusualness, which was significantly associated with word fluency task scores.

In contrast with these studies, which directly examined relationships between LSA scores and measurements of coherence or thought disorder, later studies used LSA derived scores as a means to predict the risk of psychosis. Bedi et al (Bedi et al., 2015) used cosine similarities between adjacent phrases represented by LSA vectors as a feature in a classifier model, which correctly predicted the onset of psychosis in 5 of the patients (individuals with prodromal symptoms) (total of 34) with 100% accuracy in a leave-one-out cross-validation analysis. In a follow-up study conducted by Corcoran et al (Corcoran et al., 2018), LSA coherence scores, the variance of the coherence scores, and the use of possessive pronouns were used as features in a

machine learning model to predict the onset of psychosis. This model achieved 83% prediction accuracy when using texts collected under the same prompt protocol (intra-protocol), 79% prediction accuracy when a different narrative-based prompt protocol (cross-protocol), and 72% accuracy when discriminating speeches of psychosis patients from those of healthy controls. These results established the utility of LSA coherence scores as a useful feature in predictive modeling of schizophrenia spectrum disorder (SSD).

2.2 The emergence of neural word embeddings

With the recent advancement in neural network based machine learning technologies, other representations of word semantics in vector space (called word embeddings) emerged through application of self-supervised learning. Mikolov et al (Mikolov, Chen, et al., 2013) proposed two foundational methods to derive word embedding: the continuous bag-of-words model (CBOW) and the continuous skip-gram model, both of which involve training of a shallow neural network that consists of input, projection, and output layers.

In the CBOW model, the input layer represents words that occur before and after a target word and the neural network is trained to predict the target word as the output. The words in the input layer are encoded as “one-hot” vectors of size V, where V is the size of the vocabulary. The input layer is projected using a projection layer that has dimensionality of N x D using a shared dimension matrix, where N is the number of words to be used in contexts and D is the dimension of embeddings. A softmax function is used to compute scores that determine the most likely words in the output layer.

In the skip-gram model, the model follows a similar set up to the CBOW model but it uses a single input word to predict the context of the word both before and after within a certain range. Figure 2.1 shows the two model architectures from Mikolov et al (Mikolov, Chen, et al., 2013).

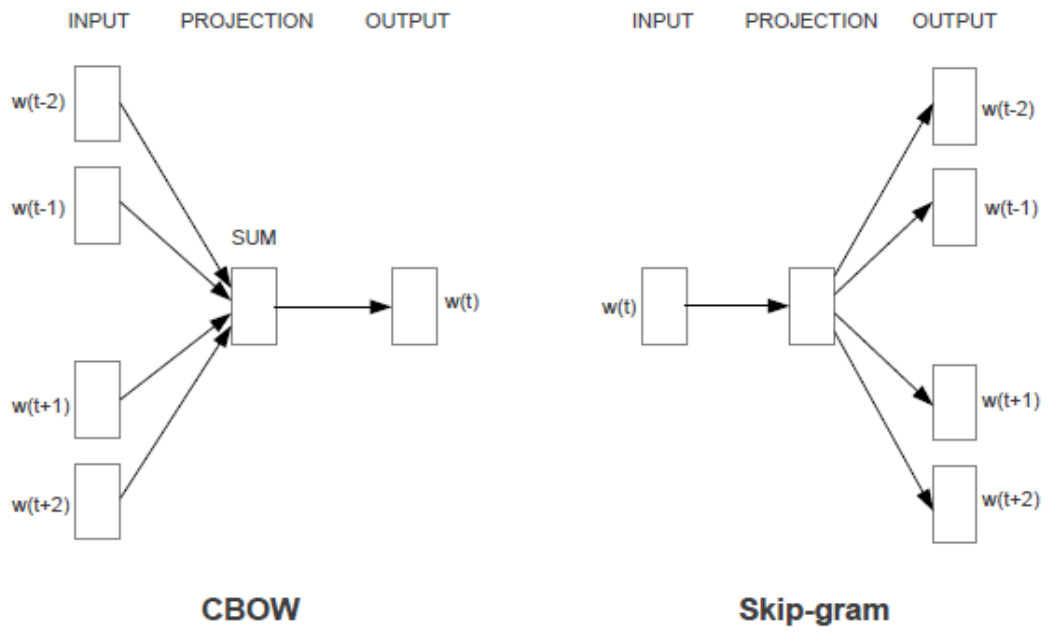


Figure 2.1: Model architecture for continuous bag-of-words and skip-gram (Mikolov, Chen, et al., 2013).

The training of these neural networks involves processing each word in a training corpus, running predictions of either the middle word (CBOW) or the context (skip-gram), and updating the weights based on the gradient of the loss function using back propagation. Eventually, the word embeddings can be derived from the input layer weights after training. These embeddings capture the semantic representations of words because each time a word is encountered in the

corpus, the model weights are updated in regards to its context and over time they will represent how the word is used throughout the corpus and therefore can serve as an effective semantic representation. Because the training process is based on co-occurrence of words, but no additional annotations are provided other than the training texts, this is considered a self-supervised learning process.

These models provide the foundation of a commonly used word embedding software package known as Word2Vec[1]. Later, Bojanowski et al (Bojanowski et al., 2016) further expanded the potential of this approach by including subword information to improve the out-of-vocabulary performance and created the FastText package[2]. Pennington et al (Pennington et al., 2014) combined the strengths of LSA and Word2Vec to develop global vectors (GloVe), which uses a global co-occurrence matrix (like LSA) but instead of using SVD, it uses stochastic gradient decent (SGD) to optimize a loss function (like Word2Vec), which aims to bring the dot product of two embeddings to predict the log of their co-occurrence count.

Neural embeddings have also been used in studies of semantic coherence. Iter el al (Iter et al., 2018), used neural word embeddings in a coherence study with speech samples collected from schizophrenia patients and healthy controls. The patient group had lower coherence scores compared to the control group, whether using LSA vectors or neural word embeddings. However, neural word embeddings demonstrated better performance than LSA vectors in many

[1] Word2Vec software package URL: https://github.com/tmikolov/word2vec
[2] Fasttext software package URL: https://fasttext.cc/

cases. Just el al (S. Just et al., 2019) also employed a similar setup with a different dataset. One configuration with GloVe embeddings achieved significant difference between patient and control groups.

2.3 Bringing context to word embeddings

Although neural word embeddings of this nature provide improved representations of semantics in vector space, the embeddings are static for each word, imposing the limitation that the representation is not context aware. For example, in the phrases *river bank* and *national bank*, the word *bank* has different meanings based on its context, but it has the same static neural word embedding. To represent semantics more accurately in their context, context-aware embeddings were created. The foundational method is through Embeddings from Language Models (ELMo) (Peters et al., 2018), which uses bidirectional long short-term memory (LSTM) neural network weights as word embeddings. However, with the continued advancement of graphical processing units (GPUs), the LSTM architecture was soon outperformed by transformers, which can take full advantage of parallel computing capabilities offered by GPUs. As a result, the most used contextual embeddings come from the transformer architecture through Bidirectional Encoder Representation from Transformers (BERT) (Devlin et al., 2019). Unlike earlier models such as Word2Vec or GloVe, which assign each lexical item a single, context-independent vector, these deep neural networks construct embeddings dynamically based on the surrounding linguistic environment. These contextual embeddings arise from a sequence of transformations beginning with tokenization and composite input embeddings and culminating in deep bidirectional self-attention across stacked transformer layers (Vaswani et al., 2017).

One crucial mechanism that mediates contextual embedding is that BERT augments token embeddings with two additional forms of information to create its initial input representation: positional embeddings and segment embeddings. Positional embeddings encode the sequential index of each token, enabling the otherwise order-agnostic Transformer architecture to model word order. Segment embeddings differentiate tokens belonging to sentence A from those belonging to sentence B, a requirement introduced by BERT's Next Sentence Prediction (NSP) pretraining objective. The final input embedding for token i is thus constructed as the element-wise sum of the token embedding, the positional embedding for position i, and the segment embedding reflecting sentence membership. This composite representation becomes the input to the first Transformer layer.

Another important technique is self-attention. Self-attention is the mechanism through which BERT learns context-dependent meaning. For each token representation entering a Transformer layer, three learned linear projections derive the corresponding query, key, and value vectors. Attention scores are computed via scaled dot products between a token's query and every other token's key. After normalization through a softmax function, these scores yield attention weights, which serve as coefficients in a weighted sum of value vectors. The result is a new representation for each token that directly incorporates information from the rest of the sequence, including long-distance dependencies that earlier RNN-based architectures struggled to capture.

Multi-head attention extends this process by performing attention calculations in parallel across multiple subspaces. Each head can focus on different linguistic relationships, such as coreference, syntactic dependencies, or semantic relatedness. The outputs of all heads are concatenated and subsequently transformed through a linear projection, enabling each encoder layer to integrate diverse relational cues. The resulting intermediate representations are then refined through a feed-forward network, which introduces additional nonlinearity and capacity. Stacking many such layers yields embeddings that are deeply conditioned on both local and global context.

Devlin et al (Devlin et al., 2019) provided two training tasks for BERT architecture: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). MLM provides token-level supervision by randomly masking approximately 15% of input tokens and tasking the model with predicting the original identities of the masked items. This prediction relies on the contextualized representations produced by the uppermost Transformer layer. The associated cross-entropy loss backpropagates through the entire model, updating not only the projection layers and Transformer weights but also the token, positional, and segment embedding parameters. Consequently, even the base embeddings evolve to reflect the statistics of natural language. NSP supplies a complementary sentence-level signal. During pretraining, the model receives pairs of sentences and must classify whether the second sentence naturally follows the first in the corpus. BERT uses the embedding of the special [CLS] token—a learned representation of the entire input sequence—to make this decision. The NSP loss further shapes the contextual embedding space by embedding discourse-level relationships into the model's parameters.

Aside from pulling embeddings from BERT hidden states or the [CLS] token, additional methods were derived from BERT. One of these is Sentence-BERT (Reimers & Gurevych, 2020). Sentence-BERT (SBERT) is a modification of BERT designed to produce high-quality sentence-level embeddings, something vanilla BERT is not optimized for. While BERT generates contextual token embeddings, they are not optimized for sentence representation or semantic similarity. SBERT solves this by placing a pretrained BERT into a Siamese (a network that contains two identical encoders) or triplet network (a network with multiple inputs), allowing it to encode each sentence independently and enabling fast cosine-similarity comparisons. It adds a pooling layer (usually mean pooling) to convert token embeddings into a fixed-length sentence vector, and it re-trains the model on semantic similarity tasks so that semantically similar sentences map to nearby points in embedding space. The result is a model that produces robust, efficient, and semantically meaningful sentence embeddings suitable for search, clustering, and retrieval. These embeddings can be further improved by replacing SBERT's supervised classification-style training with a powerful contrastive learning objective that produces cleaner and more meaningful embedding geometry, such as the SimCSE (Gao et al., 2021) and DiffCSE (Chuang et al., 2022) embeddings.

The usage of contextual embeddings was also seen in coherence evaluation studies. Tang et al (Tang et al., 2021), used BERT derived sentence embeddings and computed distance metrics between interviewer prompts and participant responses. They achieved 0.91 AUC distinguishing schizophrenia spectrum disorder patients (SSD, n=20) from healthy controls (n=11). Figueroa-Barra et al (Figueroa-Barra et al., 2022), built a classification model that achieved 85.9%

accuracy in distinguishing schizophrenia groups from healthy controls. The most important feature from the classification model is a semantic coherence feature that is based on contextual embeddings with cosine similarity between 5-word interval units.

2.4 Methods outside of the proximity paradigm

The methods discussed so far were all based on computing distances or similarities between vector representations of semantic units (proximity methods). These methods measure semantic transitions directly and therefore offer a direct estimation of coherence. However, other methods have been proven useful in evaluating severity of thought disorder or schizophrenia spectrum disorder (SSD) and may be evaluating coherence indirectly. These methods are discussed subsequently.

2.4.1 From proximity to perplexity

With the advancement of large language models, there are reasons to believe that performance could be further improved through the use of Generative Pre-trained Transformer (GPT) family language models. Notably, BERT's bidirectional encoder is inconsistent with the unidirectional process of language generation during speech. In addition, both its size and volume of training data are considerably smaller than contemporary transformer models.

GPT and BERT differ primarily in architecture, attention direction, and purpose: GPT is a decoder-only, causal (left-to-right) transformer trained with an autoregressive objective to predict

the next token, making it ideal for text generation. In contrast, BERT is an encoder-only, bidirectional transformer trained with masked language modeling (and originally next-sentence prediction), allowing it to use context from both sides of each token. Figure 2.2 shows the difference between encoder and decoder transformer components.

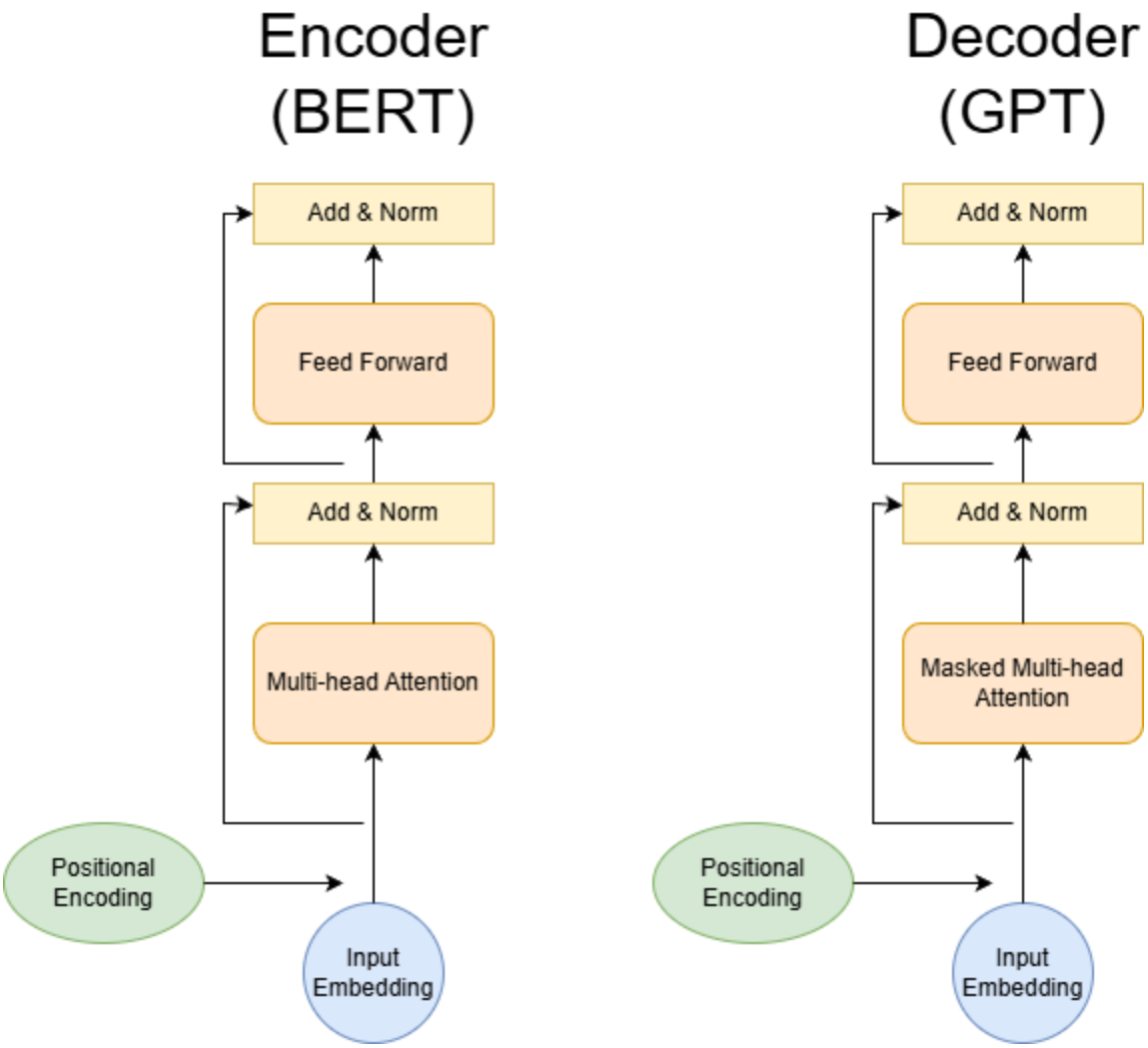


Figure 2.2: Comparison of encoder and decoder transformer architectures. Note that they are very similar except that for the decoder, the attentions are masked so that the model can only focus on the prior context.

As a result, encoder-based models (BERT) produce useful representations of semantics in vector space due to tokens in full view of each other to maximize the understanding of their contexts, whereas decoder-based models (GPT) excel at next token prediction task.

For the task of evaluating coherence, we have discussed that coherence estimation with encoder models can be used to generate useful embeddings for representing semantics; but how are decoder models used to estimate coherence? This is a less explored territory than encoder models. Decoder models, which focus on next token generation, do not offer straightforward application for coherence estimation and only became well known recently. However, recent evidence from certain research efforts indicates that decoder models can be used to evaluate coherence in certain ways. Fradkin et al (Fradkin et al., 2023), employed a generative language model (GPT-2) to generate simulated thought disorder speech samples by perturbing the model's temperature (a hyperparameter that controls the creativity of the language model's output) and memory span. Results indicate an association between these perturbations to GPTs output, and human ratings of the semantic coherence of this output. Additionally, work described in a preprint by Sharpe et al. (Sharpe et al., 2024) found that the speech of participants with SSD was more difficult for LLMs to predict than that from healthy controls, with more marked differences when additional context was provided to the LLMs by increasing the size of the sliding window within which the perplexity was measured. This finding suggested a previously identified difficulty (Kuperberg, 2010b) for SSD patients to integrate distal context in their speech and could inform the mechanism through which decoder generative model are used to evaluate coherence.

2.4.2 Speech graphs

A graph is data structure consisting of nodes and edges, where edges are connections between nodes. Mota el al (Mota et al., 2012) proposed that certain qualities of texts represented by a graph could be useful to evaluate thought disorder. To represent texts as a graph, each canonical element in the text (lexeme (e.g. woke up)) is considered a node and a directed edge is formed between the nodes if they are adjacent in the speech. For example, consider the sentence "*I walked into a place, and I found my grandma. I hugged her strongly and I woke up."* It can be represented as the graph shown in Figure 2.3 (Mota et al., 2012).

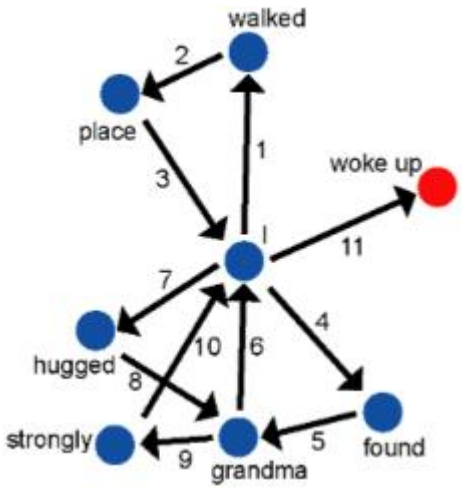


Figure 2.3: Graph representation of the text "*I walked into a place, and I found my grandma. I hugged her strongly and I woke up." (Reproduced from (Mota et al., 2012)*

After representing the text as a graph, Mota et al suggested some graph properties could be used to assess clinical conditions. These were general features: number of nodes (N) and number of edges (E); connectivity-related: including number of nodes on the largest connected component (LCC), number of nodes on the largest strongly connected component (LSC), and average total

degree (ATD); recurrence-related: including number of parallel edges (PE), and number of loops with one, two or three nodes (L1, L2, L3) (Mota et al., 2012). In addition to these local graph measures that concern each node and its neighbors, global graph measures could also be taken such as graph diameter, density, and average shortest path (ASP). In the context of this study (Mota et al., 2012), which included data about descriptions of dreaming and waking moments, topic deviations could also be represented by the number of waking moment related nodes and edges, but this information may not be available outside the scope of this dataset. The study found that graphs from the schizophrenia group tend to have fewer edges, smaller LCC and LSC, indicating a more fragmented and less connected graph.

2.5 Syntactic coherence methods

There are many computational methods that evaluate an aspect of text coherence, which focuses more on the structural coherence of texts rather than the semantic shifts. These methods focus on detecting syntactic abnormalities instead of semantic shifts. In thought disorder, there is little evidence that syntactical structure is heavily affected. So, these methods to evaluate text coherence are out of scope of this work. However, they are briefly discussed in this chapter to demonstrate the distinction from coherence methods mentioned in previous sections.

2.5.1 Entity grid

The original entity grid model, introduced by Barzilay and Lapata (Barzilay & Lapata, 2008), represents a document as a table whose rows correspond to sentences and whose columns correspond to discourse entities, with each cell indicating the syntactic role of an entity in a

given sentence (such as subject, object, other, or absent). Coherence is then quantified by examining the probability of transitions between these roles as one moves down each entity's column - for example, how often an entity remains in the subject position across adjacent sentences or shifts from subject to object or to being omitted. To evaluate coherence, the model constructs positive examples (unaltered documents) and negative examples (sentence-permuted versions), extracts transition patterns from their grids, and trains a ranking SVM to score original texts higher than their scrambled counterparts. Across tasks such as sentence ordering and summary coherence evaluation, this entity-transition framework consistently outperformed lexical or repetition-based baselines, establishing the entity grid as the standard approach to modeling local text coherence.

2.5.2 Neural coherence model

Building on the original entity grid idea, a line of work develops neural models that operate directly on entity grids or their graph-based variants to learn richer coherence representations. Instead of hand-engineered transition features, models such as Nguyen and Joty's neural entity grid (Nguyen & Joty, 2017) embed the grid cells and pass them through convolutional or recurrent layers that automatically learn which patterns of entity transitions signal coherence, improving performance on sentence ordering and summary coherence tasks. Extensions such as Mohiuddin and Joty's conversational neural entity grid (Mohiuddin et al., 2018) adapt this idea to asynchronous dialogue, augmenting the grid with conversation structure and using it for thread reconstruction and conversation coherence assessment. Across these models, the entity grid provides an inductive bias toward referential structure, while neural architectures learn nuanced

patterns that go beyond simple role-transition probabilities, typically outperforming classic grid models and serving as strong baselines for modern syntactic coherence evaluation.

2.6 Gaps of prior work

2.6.1 Dataset limitations

Most prior work mentioned relied on long-form interview datasets that were gathered in controlled experimental settings. While evaluation under controlled conditions using standardized tasks or extended interviews has advanced the science of automated measures of coherence considerably, the translational impact of these methods is contingent upon their being readily deployable under naturalistic conditions, such that they can capture fluctuation in symptom severity without the need for additional clinic visits.

Additionally, some foundational work in this area had small sample sizes. For example, Bedi el al (Bedi et al., 2015) achieved 100% model accuracy only when using leave-one-out cross-validation on a dataset with a total sample of 34. Smaller sample size may lead to undercoverage bias, causing the model to overfit to the characteristics of the sample population. This may cause inaccurate evaluation metrics when the model is applied to a larger population.

2.6.2 Proximity methods based exclusively on local coherence

Most proximity methods discussed in Chapter 2.1-2.3 use the cosine similarities between juxtaposed semantic units exclusively. This includes both comparison of the subsequent units

(termed "first order" coherence) and comparison of gapped units with an intervening unit in between ("second order" coherence). For example, in the 3-word "w1 w2 w3" sequence, word-based variants using first order coherence will compare w1 : w2 and w2 : w3 for coherence calculation, whereas the second order coherence calculation will compare w1 : w3. This method may capture local coherence characteristics but it does not consider coherence globally and global coherence - the ability to sustain a topic throughout spoken discourse - is an important consideration for normal speech capabilities (Ellis et al., 2016).

Some recent work (Burke et al., 2023) evaluated a form of global coherence, but it involved a comparison between the speech sample and a reference sample (such as the interview prompt). While it may prove to be effective in the proposed setting, the reference may not be available outside of experimental settings. Also, global coherence should be considered as an innate property of participant speech and therefore the evaluation method ideally should not rely on outside information to assess it.

2.6.3 Static aggregation techniques

Coherence scores for an entire transcript were primarily derived from an aggregation of cosine similarities between juxtaposed units in prior work, such as the mean or minimum aggregation. However, when used in conjunction with an automated speech recognition (ASR) model, the coherence assessment pipeline with an aggregation component can be sensitive to extreme values introduced by ASR errors, especially when using recordings captured under noisy conditions. This issue has not been addressed in prior studies because only high quality professionally

transcribed text data were used. However, 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. Therefore, methods to mitigate the sensitivity to ASR errors caused by static aggregation are needed.

2.6.4 Limited exploration on LLM perplexity

Most prior work on semantic coherence used the proximity-based methods discussed in Chapter 2.1-2.3. However, few studies have explored perplexity-based methods for coherence assessment, which seems a missed opportunity, especially when considering the current generation of large language models (LLMs – with "large" typically referring to models with a billion or more parameters at the time of this writing) that have demonstrated unprecedented capabilities in many NLP tasks (Bubeck et al., 2023). Even with the studies that examined applications of LLMs (Sharpe et al., 2024), none to date have evaluated the perplexity and proximity methods in concord, causing a fragmentation in coherence assessment research.

2.6.5 Lack of comprehensive analysis

Each study mentioned in this chapter provides a facet of the collection of coherence assessment methods, but none have systematically examined multiple types of methods (such as local vs global coherence and perplexity vs proximity coherence) in the context of the same dataset. A comprehensive analysis could be instrumental in understanding which aspect of the disease

manifestation each coherence method is most sensitive to and in revealing potential useful combinations of complementing coherence methods.

Chapter 3 Key innovations and overview of coherence analysis pipeline

3.1 Innovative concepts

3.1.1 Using centroids as global coherence evaluation methods

In this work we propose two novel coherence evaluation methods that aim to evaluate global coherence by employing vector centroids: the static and cumulative centroid. For the static centroid, we compute each vector's similarity to the mean vector, or centroid, of the semantic vectors for a transcript. Similarity is calculated as the cosine of the angle between two vectors - one representing the centroid and the other representing a semantic unit (word, phrase, sentence) - with the centroid calculated as the vector average of the individual semantic vectors for each unit in the transcript. The idea underlying this approach is that the dispersion of units from the centroid gives a measure of the extent to which they diverge in meaning from the central topic of a transcript. As this central topic may evolve as speech proceeds, we also developed and evaluated a cumulative centroid coherence metric, where centroid of a document changes as more vectors are considered in sequence. The cumulative centroid is therefore sensitive to the position of each semantic unit within the document, and measures whether what has been said is consistent with what was said previously.

To demonstrate the difference between the novel centroids methods and the well-established first order (sequential) and second order (gap) coherence methods, consider a paragraph consisting of three sentences, each represented by a form of embedding V1, V2, and V3. Figure 3.1 (Xu et al., 2020) shows a side-by-side comparison of how each coherence method computes the coherence scores from the embeddings. Note that while sequential and gap coherence focus on embeddings'

relationships with their neighbor embeddings, the static and cumulative centroid focus on the similarities between each embedding and the centroid embedding, effectively representing how far each semantic unit strays away from the central topic of the input text and therefore achieving an estimation of global coherence.

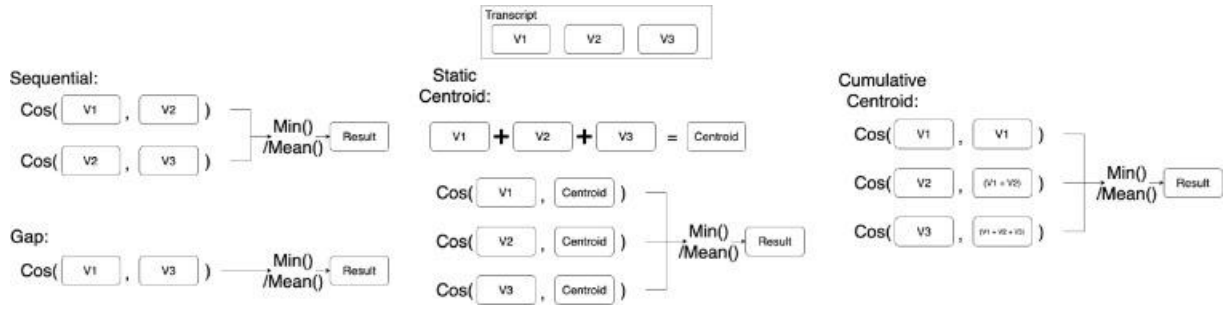


Figure 3.1: Vector computation comparison: suppose a transcript is tokenized into three units, which are then represented by vectors V1, V2, and V3. The coherence metrics will be calculated as shown above (Xu et al., 2020).

3.1.2 Time-series augmentation

In this work, we devised a novel representational approach for these coherence estimates called Time-series Augmented Representations for Detection of Incoherent Speech (TARDIS) and used this as an alternative to the typical approach of aggregating coherence estimates across transcripts, which is based on a point estimate only.

When cosine similarities between embeddings are calculated, whether to represent local coherence (between embeddings of adjacent units) or global coherence (between individual embeddings and the centroid embedding), it can be considered that a series of "coherence

scores" are estimated. However, a single final coherence score is still needed to represent the coherence level of the whole transcript. Instead of using the minimum aggregation, the series of coherence scores are treated as a time series, from which a set of features can be derived to train a downstream machine learning model to predict human-assigned coherence scores. The key features are extracted from time-series data using the TSFRESH software package (Christ et al., 2018a). The TSFRESH package acquires 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). Length-dependent features were removed, in order to avoid developing models that consider the volume rather than the coherence of language produced. Then these features are used to train a support vector machine regression (SVR) model to predict annotator-assigned derailment scores (Xu et al., 2022).

With this approach, the information from semantic shifts represented by cosine similarities between embeddings is captured in the format time-series features. Instead of a single feature derived from an aggregation, this multi-feature approach allows a machine learning model to learn the usefulness of each feature and therefore mitigates the impact of extreme values imposed on the arbitrary single feature.

3.1.3 Sentence level perplexity coherence methods and its combined usage with proximity methods

In this work, we developed a new sentence-level perplexity-based estimate of thought disorganization (as reflected by diminished semantic coherence) based on the perplexity of LLMs, which measures the level of "surprise" of a LLM after observing a sentence given its context. The perplexities are derived from an LLM's probability output when predicting the next token. However, perplexity is intuitively inversely correlated with probability because the higher the probability the lower the level of surprise from the language model. Based on the work of Sap, Jafarpour et al (Sap et al., 2020), we use the negative log likelihood (NLL) as a surrogate for perplexity, which is monotonically associated with NLL. The NLL of a sentence S given context C can be calculated using the following equation:

$$NLL(S|C) = -\frac{1}{|S|}\sum_{i=0}^{|S|} \log\left(P(T_i|C, T_{0\ldots i-1})\right)$$

where *P* is the probability output of the language model and $T_{0\ldots i}$ are the tokens in sentence *S*.

Following the work of Sap, Jafarpour et al (Sap et al., 2020), we adapted the "chain model", where perplexities are used with a cumulative context (all history sentences), and the "bag model", where perplexities are used with a central static context (a summary), as probabilistic analogues of semantic distance-based coherence measures. The chain and bag models are components of a measure of linearity developed in this work, which was used to quantify narrative flow to distinguish between real and imagined stories. In the current work, we focused on the chain and bag model components' utilities as LLM-based measures of sequential and global coherence, respectively. For the chain model, the perplexities of each of the subsequent sentences were computed by including all previous sentences and the initial context as the

context C. For the bag model, the perplexities of each of the sentences were computed individually with a constant initial context (a summary of the text). Figure 3.2 demonstrates how the chain and bag models handle contexts in perplexity calculation. The bag model is an estimate of global coherence as each sentence is assessed in the context of text representing a central theme, similar to its counterpart in proximity metrics (static centroid). The chain model, resembling the design of the cumulative centroid in proximity metrics, measures an evolving global coherence by including contextual information up until the point of measurement.

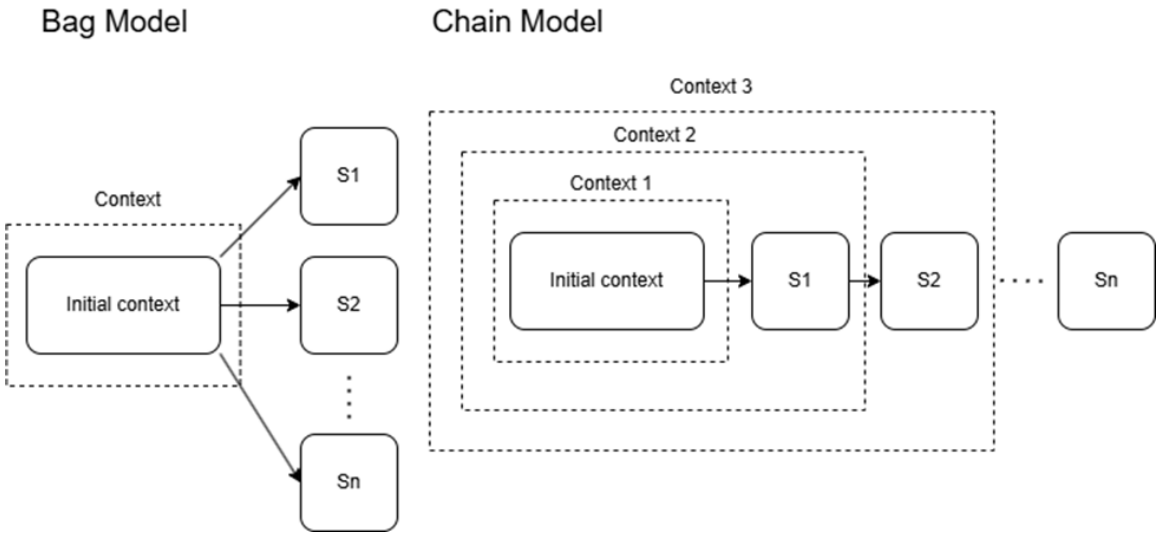


Figure 3.2: The difference between chain and bag models in choosing contexts for perplexity calculations at each sentence S1-Sn. Arrows indicate that the preceding segments inform the prediction of those that follow (Xu et al., 2025).

Another point of innovation is that we developed and evaluated methods that combine coherence scores from the perplexity and proximity-based methods because we hypothesized that these

methods measure a different aspect of coherence and therefore can provide a more accurate coherence assessment when used together. Specifically, proximity features should be measuring the semantic similarity between two text units while the perplexity of a LLM measures the level of "surprise" of the LLM given a prior context. As such, the perplexity feature reacts not only to semantic similarity, but also to continuity, which indicates how likely a sentence is to occur after a prior sentence regardless of their semantic similarity. This contributes additional and useful information to the proximity features and could potentially lead to better assessment of coherence.

3.1.4 Comprehensive Coherence Calculator (CCC)

With multiple novel modifications to established coherence methods at different stages of the analysis pipeline, it can be difficult to keep track of the configurations of the coherence scores. Therefore, the CCC is developed in this work to integrate the three major categories of in-scope coherence methods discussed in Chapter 2, namely, proximity methods, perplexity methods, and speech graph methods. The CCC implements methods derived from aforementioned innovative concepts and also classical approaches, allowing users to generate coherence scores from a variety of different configurations and enabling comprehensive analysis of coherence within the same frame of reference.

3.2 Overview of coherence assessment pipeline

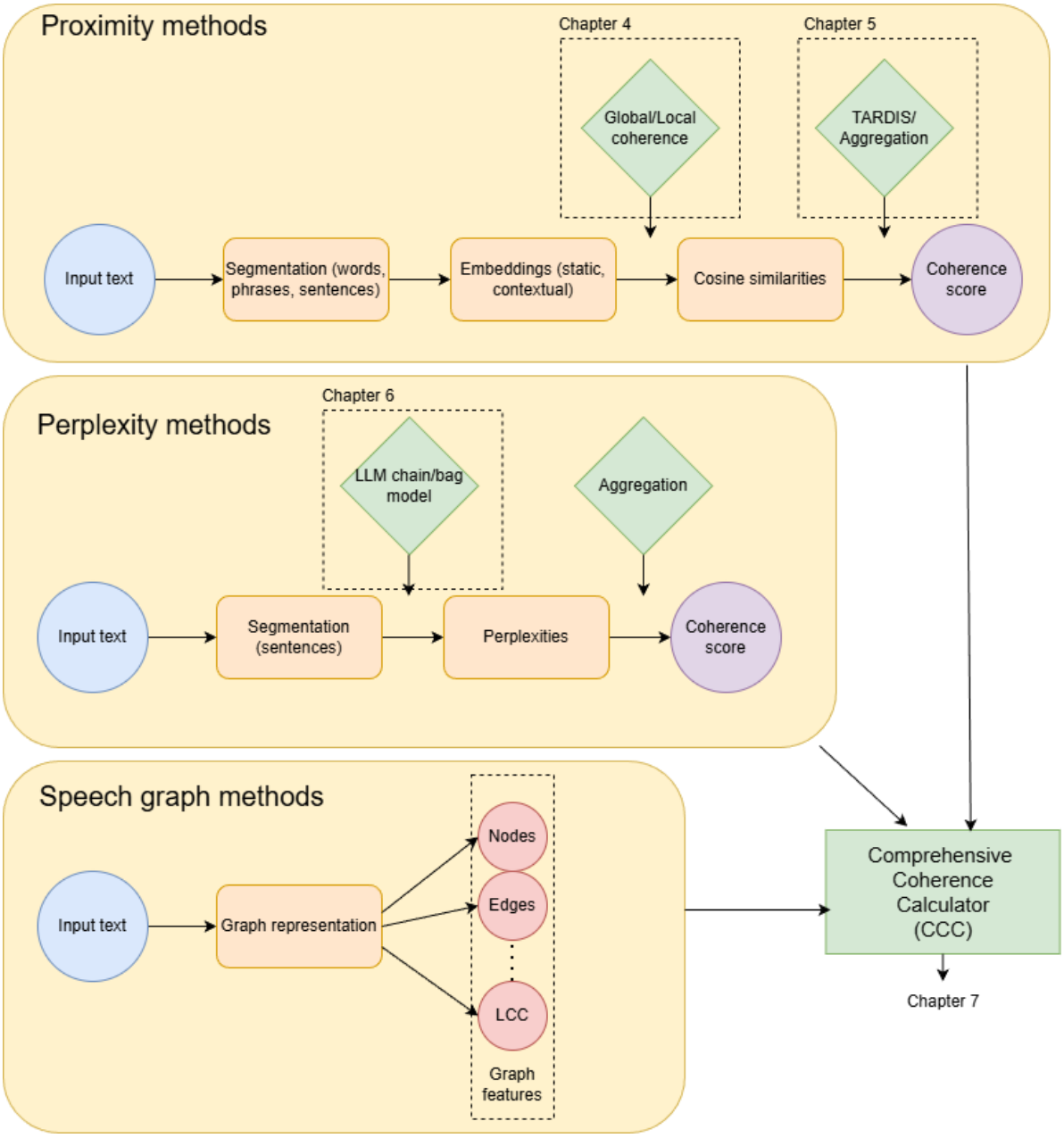


Figure 3.3: Overview of coherence assessment pipeline and chapter topics

Figure 3.3 provides an overview of the coherence assessment methods and how each of the subsequent chapters fit into the big picture. Chapter 4 focuses on the innovation of global coherence methods and their comparison with local coherence methods. Chapter 5 goes into detail about TARDIS and its comparison with aggregation in the context of both ASR transcripts

and manual transcripts. Chapter 6 describes the evaluation details of the perplexity-based coherence methods, and it's combined application with proximity-based coherence methods. Together, with a reimplementation of the speech graph method, are included in the CCC, which is discussed in Chapter 7.

Chapter 4: The Centroid Cannot Hold: Comparing Sequential and Global Estimates of Coherence as Indicators of Formal Thought Disorder

Parts of this chapter are drawn from a previously published paper

Xu, W., Portanova, J., Chander, A., Ben-Zeev, D., & Cohen, T. (2020a). The Centroid Cannot Hold: Comparing Sequential and Global Estimates of Coherence as Indicators of Formal Thought Disorder. AMIA ... Annual Symposium Proceedings. AMIA Symposium.

4.1 Introduction

The chapter title "The Centroid Cannot Hold" takes inspiration from the book "The Center Cannot Hold: My Journey Through Madness" by Elyn Saks, which vividly describes her experiences with schizophrenia. This work was motivated by the thought that the "centroid" (global coherence) methods could deter of thought disorder using speech samples from patients so that those who need care can receive it sooner. As discussed in Chapter 1 that coherence, or the lack there of, is an important indicator symptom severity in thought disorder and more broadly, schizophrenia spectrum disorder. With various embedding strategies discussed in chapters 2.1-2.3 and the use of cosine similarities between the embeddings, automated coherence evaluation methods were developed, allowing quantification of coherence from text data and forming the foundational coherence evaluation pipeline (proximity-based methods). However, there are gaps in research on proximity-based coherence methods. In this chapter, we focus on addressing the issue of limited data sample size (Section 2.6.1) and the lack of global coherence evaluation methods (Section 2.6.2).

In this work, we aim to address these issues, while presenting and evaluating a novel approach to quantify global speech coherence. To overcome the limitation of participant pool size and address the need to estimate coherence in naturalistic settings, we evaluate our methods in the context of speech samples from a smartphone application that collects "audio diaries" describing participant experiences of auditory verbal hallucinations (AVH) recorded in naturalistic environments. AVH, like TD, are an important diagnostic consideration in schizophrenia and may be caused by the disordered monitoring of inner speech. The sample collection process did not involve structured questions, and the recordings were limited to three minutes in duration, which enhances the scalability of the data collection procedure. However distributional semantics-based estimates of coherence have yet to be validated in the context of relatively unstructured speech samples of this length. The speech samples used in the current study were collected using this smartphone application from more than 150 participants, with the potential to scale to much larger participant pools. As such, validation of coherence metrics in the context of these data would provide a solution to the scalability limitations inherent in laboratory-based evaluations using lengthy interviews or structured instruments that require specialized expertise to use. To analyze transcribed speech samples for their coherence, we evaluated a range of established and novel distributional semantics-based approaches that collectively not only compare consecutive semantic units, but also compute a mean vector (the centroid) to estimate global coherence. We also explored the utility of different semantic units by conducting a comprehensive analysis comparing them. The coherence scores generated by each approach were compared to human annotations for validation and comparative evaluation.

4.2 Method

4.2.1 Participants

Data were obtained from a study of participant experience of AVH, which uses a smartphone platform to capture a range of ecological momentary assessment and sensor-derived variables (Ben-Zeev et al., 2020). The study was approved by the institutional review boards at the University of Washington and Dartmouth College. Participants experiencing AVH were recruited via both in-person and online means. Informed consents from participants were obtained through a rigorous procedure involving triple confirmations from a screening questionnaire. All participants were asked to install a mobile application, which had the capability of recording and uploading audio diaries, and were prompted to describe their experiences of AVH, as well as anything else they would like to share or think it would be helpful for the research team to know, with prompts for audio diaries following the collection of other data, and the option to record an entry directly on demand. Although no monetary incentives were offered for the audio diary component, most participants submitted their recorded audio diaries. We used data collected up to October 18th 2019, consisting of 1868 recordings from 202 users. As short recordings seldom contained interpretable language, we restricted this set to recordings of length 30 seconds or more (maximum three minutes), leaving 909 recordings from 154 users. We randomly sampled up to three recordings per user, leaving 355 recordings which were professionally transcribed. After manual inspection, we retained 310 transcripts with interpretable content, covering 142 participants (Table 4.1).

Table 4.1: Characteristics of participant pool.

Gender	**Number**	**Percentage**	**Age**	***Number***	***Percentage***
Male	56	39.4%	19-29	24	16.9%
Female	82	57.8%	30-39	52	36.7%
Transgender (MTF)	3	2.1%	40-49	37	26.1%
Transgender (FTM)	1	0.7%	>=50	29	20.4%

4.2.2 Transcripts

Each transcript was labeled by two human annotators with a score between 0 and 4 to indicate the degree of derailment, which is an indicator of TD, and was selected as a construct for the current study because it does not concern deviation from the topic of a question, and audio diary prompts were open-ended in nature. Annotation was guided by the definitions and training materials for the Thought and Language Disorder (TALD) rating scale (Kircher et al., 2014), a validated instrument for the assessment of TD. A score of 0 indicates that derailment is not present. A score of 4 indicates that speech is incomprehensible. Scores from 1 to 3 represent intermediate degrees of derailment, in which the connections between sentences grow less recognizable as the score increases. The raters each rated all transcripts. Any transcripts with a disagreement of two or more units on the scale (n=22) were re-evaluated independently, to reach a quadratically-weighted Kappa of 0.71. Note that quadratically-weighted Kappa scores penalize larger differences between scale categories more than smaller ones, which we deemed appropriate given the subtle distinctions between neighboring TALD categories. The average

score of the two raters for each transcript was calculated to be used for further analysis. Table 4.2 shows the number of transcripts by average rater score.

Table 4.2: Transcripts by mean rater score. Line ruled between 2.5 and 3 indicates categorization threshold.

Score (x)	**Number**	**Percentage**	**TALD category** (paraphrased and abridged)
0	35	11.29%	*not present:* no derailment
0.5	62	20.00%	
1	93	30.00%	*doubtful:* connections still obvious
1.5	53	17.10%	
2	24	7.74%	*moderate:* sometimes disconnected from prior speech
2.5	25	8.06%	
3	8	2.58%	*severe:* no meaningful connection between ideas
3.5	8	2.58%	
4	2	0.65%	*extreme:* interview is incomprehensible
Total	**310**	**100.00%**	

4.2.3 Preprocessing:

Transcripts were pre-processed by stop-word removal and term tokenization. Stop-words are filler words with little or no semantic content (such as "a", "the", and "on"). These words were defined by the stopword list distributed with the natural language tool kit (NLTK)(Bird et al., 2016), an open-source tool, and their occurrences were removed from the transcripts to reduce noise. Tokenization is a process of extracting semantic units from documents so that they can be represented by vectors for similarity comparison. For example, a word tokenizer extracts individual words from a document while maintaining their sequential order. In this study, because we were interested in conducting a comprehensive analysis of various semantic units, we tokenized the transcripts into three different units: words, noun phrases and sentences. The word and sentence tokenizations were performed using the NLTK word and sentence tokenizers. The noun-phrase tokenization involved a different tool, Textblob (Loria et al., 2014), which is also publicly available.

4.2.4 Semantic vectors (word embeddings):

Most previous work modeling coherence using distributional semantics has employed LSA to generate semantic vectors for words. LSA creates vectors based on the distributional statistics of words in a corpus (usually the Touchstone Applied Science Associates (TASA) corpus, which was used in previous studies of coherence in schizophrenia (Bedi et al., 2015; Corcoran et al., 2018; Elvevåg et al., 2007, 2010)). However, *neural word embeddings,* distributed representations of words derived from neural networks trained to predict words in proximity to an observed word (such as the popular *skipgram* and *continuous bag of words* architecture

(Mikolov, Chen, et al., 2013) embodied in the widely used word2vec[3] and FastText[4] software packages), have been shown to outperform matrix decomposition-based approaches like LSA (Baroni et al., 2014), especially on novel tasks involving solving proportional analogy problems using geometric operators (Mikolov, Yih, et al., 2013). While some of these improvements in performance have subsequently been shown to be contingent upon the selection of task-specific optimized hyperparameters (Levy et al., 2015), it remains true that the efficient algorithms used to train neural embeddings allow for training on much larger corpora in a relatively short time. Thus, in this study, neural embedding was used as a technique to generate the vector space for automated analysis of the transcripts. Publicly available FastText pre-trained word embeddings[5] were selected for this study. These vectors were trained on a large corpus derived from Common Crawl[6], comprised of approximately 600 billion word-level tokens, as compared with approximately 12 million in the TASA corpus, and without the use of subword embeddings.

While we used the aforementioned FastText-derived space for the majority of our experiments, we used four additional vector spaces - two Wikipedia-derived, and two trained on the TASA corpus - to evaluate two methodological variants applied in prior studies. Firstly, the utility of lemmatization of words was evaluated as this has been used in prior work on coherence(Bedi et al., 2015). Lemmatization refers to the process of reducing various forms of a word to a canonical form, for example converting "does", "did", "doing" and "done" to "do".

[3] Word2Vec software package URL: https://github.com/tmikolov/word2vec
[4] Fasttext software package URL: https://fasttext.cc/
[5] Fasttextpretrained vectors. URL: https://fasttext.cc/docs/en/english-vectors.html
[6] Common crawl corpus. URL: https://commoncrawl.org/

Lemmatization has been used as a normalization procedure to accommodate morphological variants in distributional semantics (Turney & Pantel, 2010).

For the purpose of comparison, we trained neural word embeddings using the open source Gensim[7] implementation of the skipgram-with-negative-sampling algorithm (Mikolov, Sutskever, et al., 2013) (which is also a component of word2vec and FastText), to generate a vector space from lemmatized and non-lemmatized versions of a Wikipedia-derived corpus. We generated a 100-dimensional vector space without imposing frequency thresholds (i.e. including all terms), using a window size of 5, a subsampling threshold of 10^{-3} and five iterations of training across the corpus. The transcripts were also lemmatized when using the vectors trained on the lemmatized corpus. In addition, we trained word vectors on the TASA corpus using both LSA and neural word embeddings, both using Semantic Vectors[8] which implements a number of distributional semantics algorithms in a manner conducive to comparative evaluation (e.g. with consistent pre-processing). Both spaces were 300-dimensional. With LSA we used log-entropy weighting of terms. Neural embeddings were trained using the skipgram-with negative sampling algorithm with five negative samples per observed term, a subsampling threshold of 10^{-3}, and ten iterations of training across the corpus. For both models we excluded terms that occurred fewer than five times or more than 15,000 times. The latter constraint approximates a stopword list, which in our experience is important for the quality of LSA vectors in particular.

4.2.5 Semantic units:

[7] Gensim software package. URL: https://radimrehurek.com/gensim/models/word2vec.html
[8] Semantic vectors software package URL: https://github.com/semanticvectors/semanticvectors

One goal of the current work was to conduct a comprehensive analysis of the utility of modeling coherence using differently sized semantic units. The semantic units considered in this study were words, noun phrases, and sentences. Words as a unit are straightforward, in that individual embeddings can be retrieved directly from the lookup table of a vector space. For noun phrases, once extracted from documents using Textblob, vectors were calculated by summing the vectors of individual words that composed a phrase. Similarly, a sentence vector was also calculated by summing of vectors representing component words. We explored one additional sentence vector variant by multiplying each component word vector by the relevant word's inverse document frequency (IDF) (Jones, 1972). The IDF of a word is derived from the total number of documents N, and the number of documents that contain the word of interest n as log^{N}. The higher the IDF, the rarer the word, and it has been argued on theoretical grounds that IDF is an optimal measure of a word's importance for information retrieval (Papineni, 2001). The IDF of each word was obtained from distributional statistics derived from the Wikipedia corpus, and used to scale each corresponding word vector before summation. In all cases, the resulting vectors were normalized to unit length.

4.2.6 Centroid-derived metrics:

We implemented the two centroid coherence metrics discussed in Section 3.1.1.

4.2.7 Aggregation:

The sequential, gap, centroid and cumulative centroid metrics were then applied to words (all metrics), noun phrases and sentences (sequential and centroid metrics only) with and without

IDF weighting for a total of 13 coherence metrics. Each metric produces a series of similarity calculations, one for each comparison it makes. For example, the sequential word-level metric produces a cosine value for each pair of neighboring words, and the centroid-based metrics produce a cosine value for the comparison between each independent unit and the centroid. Thus, the output for every metric was an array of cosine values. We then calculated the *minimum* and *mean* value of the array to evaluate their utility as transcript-level coherence scores. Our motivation for doing so was that previous studies suggested the minimum and mean of the cosine array were effective in representing coherence (Bedi et al., 2015).

4.2.8 Evaluation:

For each of the metrics, an area under the curve (AUC) of a receiver operating characteristic (ROC) curve was calculated, using 1 - the coherence score of a transcript *t* (1 — *coherence(t))* as an estimate of incoherence, and comparison against average human annotations with derailment score >=3 labeled 1, and derailment score < 3 labeled 0. Our choice of this threshold was motivated by the immediate clinical implications of severe to extreme degrees of TD, and the likely utility of a downstream application that could detect deterioration to this point. The coherence metrics' performance was further evaluated by computing their Spearman Rho correlation coefficient with the average of the scores assigned by human annotators. The Spearman Rho correlation is a ranked-based correlation metric that evaluates the monotonic relationship between two continuous or ordinal variables. Consequently, the Spearman Rho does not require that the two variables under consideration change together in a linear fashion, which makes it a suitable metric to evaluate the relationship between automatically generated coherence

scores and human ratings. We measured the correlation between the average human rating and 1 — *coherence(t)* for each method.

4.3 Results

4.3.1 Aggregation:

The *mean* and *minimum* aggregation methods were evaluated by comparing the number of coherence metrics that performed best in terms of ROC curve AUC and Spearman correlation with each method. With ROC curve AUC, nine out of thirteen coherence metrics performed better when summarized by the *minimum* method. The 4 exceptions were the *centroid* metric at phrase level and the *cumulative centroid* metric at word, phrase, and weighted sentence levels. With Spearman correlation, the *mean* performed better than the *minimum* for only two of thirteen coherence metrics: *centroid* and *cumulative centroid* both at phrase level. Because of the generally better performance of the *minimum,* we report results with this approach to aggregation for the remainder of the paper.

4.3.2 Coherence metrics:

The results of our experiments comparing coherence metrics are shown in Table 4.3. Across both metrics (AUC and Spearman RhO) and all unit types, the best-performing metric is always one of the centroid variants, with the *cumulative centroid* ($CTRD_{cuml}$) predominating in two of the eight configurations, the *static centroid (*$CTRD_{stat}$) variant predominating in three, and these two metrics tied for best performance in the remaining three. The sentence level

sequential *(SEQ)* model performs well with respect to AUC, but relatively poorly when considering correlation, suggesting that it is effective in identifying severe TD, but less well equipped to identify subtler manifestations of this condition. IDF weighting did not improve sentence vector performance.

Table 4.3: ROC Curve AUC (left) and Spearman Rho (right) for each of the metrics. **Boldface** indicates best performance across models, and underscored text indicates best performance across unit types. *SEQ :* sequential, *GAP* :gapped, $CTRD_{stat}$ and $CTRD_{cumi}$ *:* static and cumulative variants of the centroid respectively.

AUC						Spearman Rho				
	SEQ	***GAP***	$CTRD_{stat}$	$CTRD_{cumi}$			***SEQ***	***GAP***	$CTRD_{stat}$	$CTRD_{cumi}$
Word	0.67	0.55	**0.70**	0.68		**Word**	0.21	0.26	0.50	**0.51**
Noun-phrase	0.69	**-**	**0.78**	0.77		**Noun-phrase**	0.38	**-**	0.49	**0.50**
Sentence	0.83	**-**	***0.84***	0.83		**Sentence**	0.26	**-**	**0.44**	**0.44**
Sentence with IDF	0.74	**-**	**0.76**	**0.76**		**Sentence with IDF**	0.21	**-**	**0.41**	**0.41**

4.3.3 Lemmatization:

Out of thirteen coherence metrics, only two performed better with the vectors trained on the lemmatized Wikipedia corpus: the *centroid* and *cumulative centroid* at phrase level. The performance of the remaining eleven metrics with vectors trained on the original unlemmatized Wikipedia corpus was better in terms of the AUC of the ROC curve. The Spearman Rho coefficient revealed a different ratio of eight to five (instead of 11:2), but the original Wikipedia-trained vectors still predominated over those trained on a lemmatized corpus. The five exceptions were *sequential* at phrase level, *centroid* at phrase, sentence and weighted sentence level, and *cumulative centroid* at weighted sentence level. Overall, lemmatization did not improve performance on the task of quantifying coherence.

4.3.4 Distributional models:

To evaluate the influence of the underlying method of distributional semantics, LSA (a matrix-decomposition-based method) and skipgram-with-negative sampling (SGNS) (a neural network based method) were compared. Both models were trained on the TASA corpus. Of thirteen coherence metrics, nine performed better when implemented with LSA vectors in terms of ROC AUC. The four exceptions were *sequential* and *cumulative centroid* at word level and *centroid* and *cumulative centroid* at phrase level. Similar results were observed with Spearman correlation, aside from that instead of *sequential* at word level, it was *sequential* at phrase level where SGNS performed better. These results show LSA outperforming neural embeddings when trained on a relatively small corpus, a finding consistent with previous research (Zipitria et al., 2006). However, we note that the performance of either TASA-trained

space in the majority of metrics was exceeded by the performance of models using neural embeddings trained on Common Crawl. Thus, while LSA appears to offer advantages when restricted to smaller corpora, the capacity of neural embedding models to scale to much larger corpora appears advantageous for automated estimates of coherence.

4.4 Discussion

In this chapter we present a comprehensive study of automated methods of measuring speech coherence in the context of transcripts of short (<3 minutes) responses to an open-ended prompt. When evaluated for their agreement with human annotators, our results show strong performance for two novel coherence metrics: the *centroid* and *cumulative centroid.* When considering the consistency with speech coherence level rated by humans, the two novel metrics outperformed the established *sequential* and *gap* metrics of coherence in terms of both their ability to detect severe cases (as estimated by the AUC of the ROC curve) and their correlation with average annotator scores across all categories of severity (as estimated by the Spearman Rho coefficient). This observation holds true for all semantic units considered in this study: words, noun phrases, sentences and IDF-weighted sentences. For detection of severe cases, the best performing metric of coherence is the *centroid,* while the *cumulative centroid* performed slightly better with respect to overall correlation. The *centroid* measure attained an AUC of the ROC curve above 0.7 for all semantic unit types with some above 0.8. For overall correlation, this metric attained a Spearman Rho coefficient larger than 0.4, with in some cases larger than 0.5. While not presented in detail for succinctness, we note that the centroid-based methods performed best across all of the vector spaces generated during the course of this research, whether derived from TASA, Wikipedia or Common Crawl, and irrespective of whether neural embeddings or LSA were used. These

findings suggest that in the context of short unstructured speech samples, coherence metrics using distributional similarity perform better when modeling global coherence with a centroid vector.

When considering different semantic units, using a sentence as a unit performed best on the task of identifying severe cases, with AUC values above 0.8. For overall correlation, using a word as a unit led to best performance with *centroid* metrics, while noun phrase units performed best with the *sequential* metric. The disparity between these findings may be due to the nature of the tasks concerned. The AUC measures the ability of the model to predict positive cases at relatively low false positive rate, with "positive" in our case indicating severe derailment. Thus, the AUC measure focuses on the coherence metric's ability to identify severely incoherent speech. On the other hand, the Spearman Rho coefficient is a rank-based correlation measure that takes into consideration all coherence categories. It does not require the imposition of a dichotomous classification threshold like the AUC and thus, it measures the overall prediction quality of the coherence metrics. Therefore, the sentence semantic unit appears best for identifying severely incoherent cases, and word or phrase units appear best used to model subtler distinctions in coherence.

To focus on detection of manifestations of severe TD, we set a threshold at an average human rating of three to calculate the AUC. However, detection of milder degrees of TD is also of interest, and previous work, albeit with a different rating scale based on clinical observation rather than text, has employed a threshold of two on a five point scale to identify TD. To verify

the consistency of our findings at a different threshold level, we also computed the AUC of the ROC curve with threshold of two, with a more than threefold increase in the number of positive examples. Our main findings were consistent at this threshold: the *centroid* measures still outperformed the *sequential* and *gap* measures with every type of semantic unit. Of note, a difference is that the *cumulative centroid* measures now have higher AUCs than their static counterpart (best AUC of 0.78 at phrase level), which is consistent with this metric's better performance for correlation across all levels of severity. Experiments with aggregates in this study suggest the *minimum* of a set of similarity values performs better than the *mean* in most cases. The few exceptions are from the *centroid* metrics (best AUC of 0.83 for *cumulative centroid* at phrase level), indicating the *minimum* aggregate does not impair performance of *sequential* or *gap* coherence metrics.

We also examined the effects of lemmatization of the corpus used to train word embeddings (as well as the transcripts themselves). Our findings with word embedding methods suggest lemmatization does not improve the performance of most coherence methods, which is consistent with previous work evaluating the utility of lemmatization for automated grading of summaries for coherence (Zipitria et al., 2006). There is some degree of disparity between the AUC and Spearman Rho when considering the ratios of number of best performing measures (unlemmatized:lemmatized - 11:2 for AUC, 8:5 for Spearman Rho). However, this is likely due to the different aspects of performance these metrics focus on, as discussed earlier. As coherence metrics without lemmatization generally outperform their lemmatized counterparts, the use of lemmatization is not recommended.

In addition, the comparison between LSA and neural word embeddings trained on the TASA corpus found LSA vectors performed better with most coherence measures. This finding suggests that neural embeddings are not inherently better than LSA for automated estimates of coherence. LSA remains an robust alternative to generate word vectors, especially with a small corpus, which is consistent with prior work comparing these models[44]. The main advantage of neural embeddings over LSA appears to be attributable to the availability of neural embeddings trained on larger and more comprehensive corpora. This advantage was substantial with the best-performing sentence-level *centroid* metrics where improvements in performance of in AUC were observed with the vectors trained on Common Crawl as compared with those trained on the TASA corpus.

4.5 Conclusion

In the work described in this chapter, we compared the performance of novel *centroid*-based estimates of speech coherence to established *sequential* measures in the context of transcribed recordings of responses to an open-ended prompt. The novel methods agreed better with human annotation both for detection of severe cases, and in terms of overall coherence. In addition, we evaluated a number of methodological alternatives, providing guidance for future efforts toward automated detection of linguistic manifestations of disordered thinking.

Chapter 5: Time-series Augmented Representations for Detection of Incoherent Speech (TARDIS)

Parts of this chapter are drawn from a previously published paper

Xu, W., Wang, W., Portanova, J., Chander, A., Campbell, A., Pakhomov, S., Ben-Zeev, D., & Cohen, T. (2022). Fully automated detection of formal thought disorder with Time-series Augmented Representations for Detection of Incoherent Speech (TARDIS). Journal of Biomedical Informatics, 126. https://doi.org/10.1016/j.jbi.2022.103998

5.1 Introduction

In this chapter, we continue to explore methods that improve proximity-based coherence evaluation pipelines. In particular, we address the limitation in the current work that coherence scores are largely dependent on static aggregation techniques of cosine similarities (discussed in chapter 2.7.3). The reason to pursue an alternative to aggregation is that aggregated values are vulnerable to extreme values caused by undesirable errors such as those produced by an ASR pipeline, which could be necessary because 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. Alternatively, the use of manual transcription service could lead to logistic issues and privacy concerns because audio recordings need to be sent to a third-party transcription vender.

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 (Aschbrenner et al., 2018), with survey estimates as high as two-thirds (Torous et al., 2014). Our previous work (Xu et al., 2020) 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 (Holmlund et al., 2020) 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 (Holmlund et al., 2020). 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 (Chandler et al., 2020), 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 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. Using an in-house ASR system based on Baidu's Deep Speech 2 architecture (Amodei et al., 2016), 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.

5.2 Method

5.2.1 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 EMA self-reports and 4809 Audio Diary recordings, with 3040 of these – recordings of duration 30 seconds or longer – professionally transcribed (Ben-Zeev et al., 2020). 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 5.1 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.

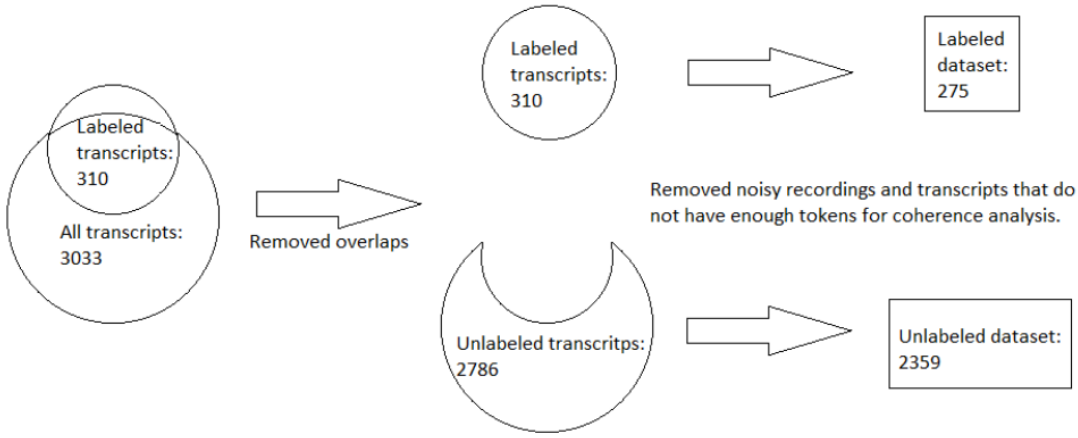


Figure 5.1: Data aggregation and processing.

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) (Kircher et al., 2014). 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 (Ben-Zeev et al., 2020). 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.

5.2.2 Automatic speech recognition

We trained an ASR system based on Baidu's Deep Speech 2 architecture (Amodei et al., 2016) implemented in PyTorch[9], 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 (Graves et al., 2006). 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 librosa2 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

[9] https://github.com/SeanNaren/deepspeech.pytorch

[2]https://librosa.org/

the following publicly available data: TalkBank (CMU, ISL, SBCSAE collections) (MacWhinney & Wagner, 2010), Common Voice (CV: Version 1.0)) corpus [10], Voxforge corpus [11], TED-LIUM corpus (Release 2) (Rousseau et al., 2014), LibriSpeech (Panayotov et al., 2015), Flicker8K (Hodosh et al., 2013), CSTR VCTK corpus (Veaux et al., 2016), and the Spoken Wikipedia Corpus (SWC-English (Köhn et al., 2016)). 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[12] 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 (Stolcke, 2002) from the English language portion of the 1 Billion words text corpus[13] model with Kneser-Ney smoothing (Ney et al., 1994).

5.2.3 Post-processing of transcripts

[10] http://voice.mozilla.org
[11] http://www.voxforge.org/
[12] http://sox.sourceforge.net
[13] https://github.com/ciprian-chelba/1-billion-word-language-modeling-benchmark

5.2.3.1 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 (Tilk & Alumäe, 2016) 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 pre-trained model[14] 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.

5.2.3.2 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 (Bird et al., 2016). Then we tokenized the transcripts into semantic units at three different levels

[14] https://github.com/ottokart/punctuator2

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 (Honnibal et al., 2020).

5.2.4 Assessment of coherence

In this section, we 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.

5.2.4.1 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 (Cohen & Widdows, 2009; “Handbook of Latent Semantic Analysis,” 2007; Turney & Pantel, 2010). Neural word embedding (Mikolov, Chen, et al., 2013) 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 (Joulin et al., 2017) package[15] consisting of 2 million word vectors trained on a corpus derived from Common Crawl[16]. 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.

5.2.4.2 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 (Devlin et al., 2019). Coherence estimates based on BERT-derived sentence embeddings have recently been shown to differ between patients with schizophrenia spectrum disorders and healthy controls (Tang et al., 2021), 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

[15] https://fasttext.cc/docs/en/english-vectors.html
[16] https://commoncrawl.org/2017/06

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 tokens (Devlin et al., 2019). The phrase embeddings were generated as the sum of the embeddings 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 (Devlin et al., 2019), the CLS token output (a special token appended 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 (Reimers & Gurevych, 2020). 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 (Reimers & Gurevych, 2020). 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 (Reimers & Gurevych, 2020).

5.2.4.3 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 (Elvevåg et al., 2007, 2010). 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 (Xu et al., 2020). 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.

5.2.4.4 TARDIS: Time-series augmentation

Following the extraction of cosine values, we applied the TARDIS feature extraction pipeline as described in chapter 3.1.2. Specifically for this dataset, additional statistically-based feature selection provided by TSFRESH (Christ et al., 2018b) 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.

For the downstream machine learning model, the Scikit Learn package[17] 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

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

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.

5.2.5 Summary of analytic pipeline

The coherence analytic pipeline described in this chapter is summarized in Figure 5.2, which demonstrates the path from an audio recording to estimation of a coherence score, through either TARDIS or the minimum aggregation function.

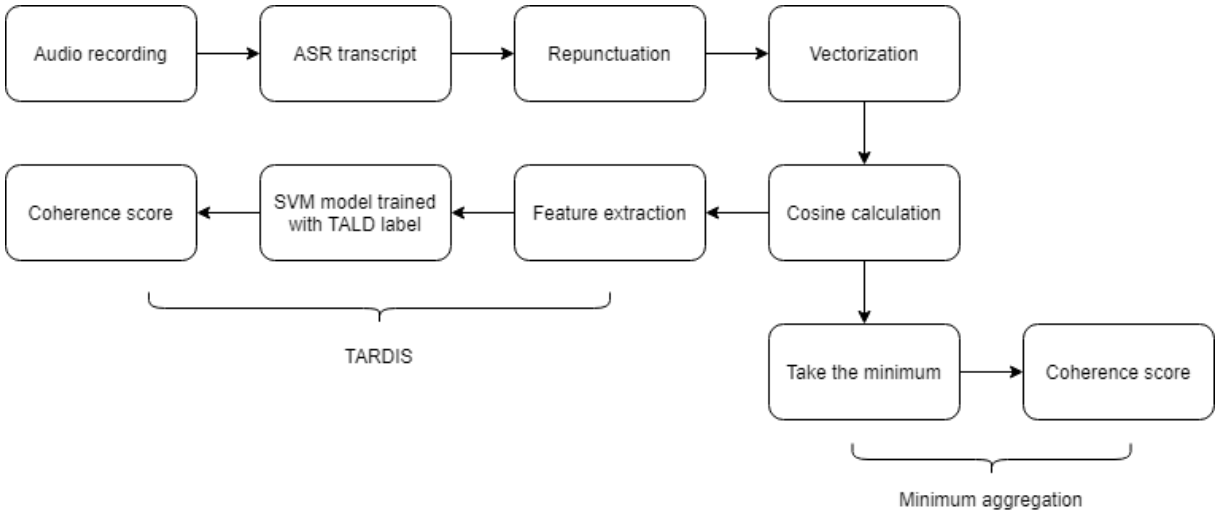


Figure 5.2: Summary of coherence analytical pipeline.

5.2.6 Experiments

5.2.6.1 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 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.

5.2.6.2 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 the 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.

5.2.6.3 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.

5.2.6.4 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 (Barzilay & Lapata, 2008). 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[18].

5.2.6.6 Correlation with HPSVQ

The Hamilton Program for Schizophrenia Voices Questionnaire (HPSVQ) (Kim et al., 2010)-(Van Lieshout & Goldberg, 2007) 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 (Sommer et al., 2010). We further hypothesized that the correlation

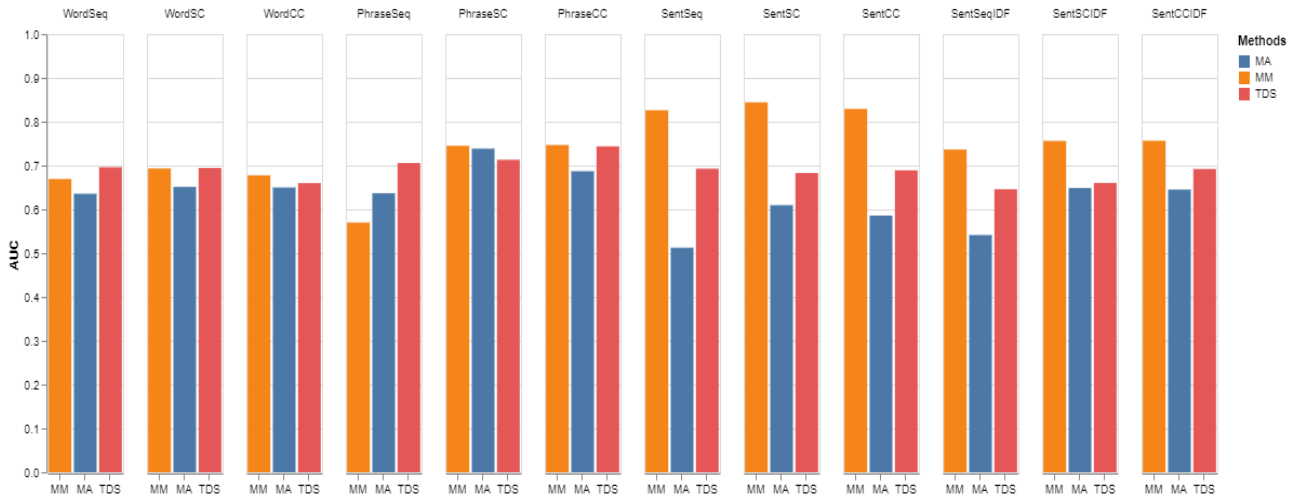
[18] Text to entity grid: https://github.com/MMesgar/text_to_entity_grid

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. Consequently, 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.

5.3 Results

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



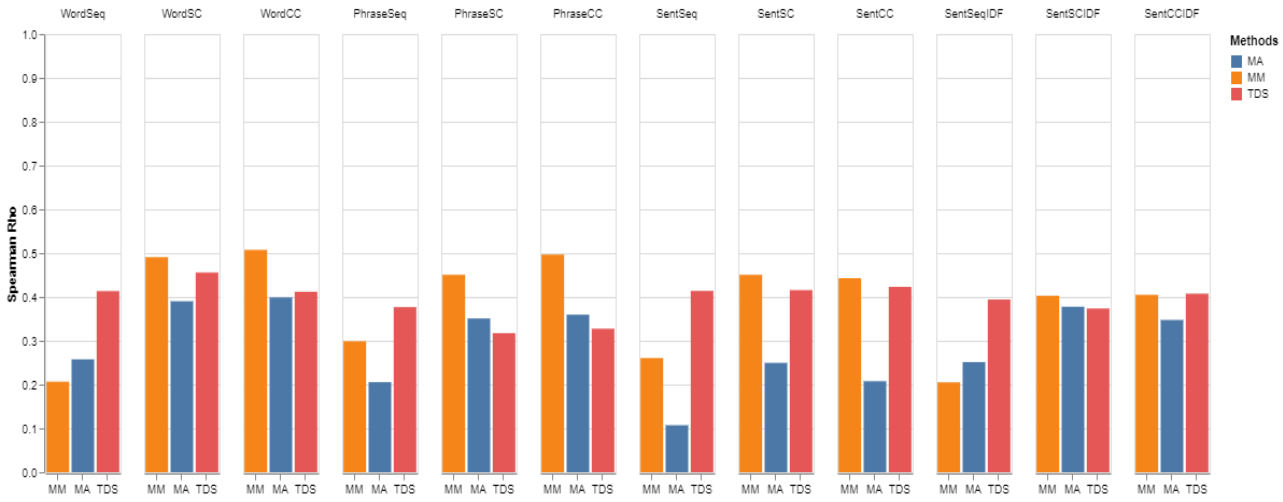


Figure 5.3: The performance comparison among manual minimum coherence (reference metric), ASR minimum coherence, and ASR time-series coherence. (a) Top: Evaluation by AUC, (b) Bottom: Evaluation by Spearman Rho. **MM** = minimum coherence with manual transcripts. **MA** = minimum coherence with automated transcripts. **TDS** = time-series based coherence with automated transcripts. **SC** = static centroid. **CC** = cumulative centroid. **IDF** = inverse document frequency.

Figure 5.3 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.

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

Figure 5.4: 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 (Reimers & Gurevych, 2020).

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 5.4, the results are similar to those obtained with skip-gram word 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 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.

	Word		Phrase		Sentence	
	Skip-gram embedding	Contextual embeddings (BERT)	Skip-gram embedding	Contextual embeddings (BERT)	Skip-gram embedding	Contextual embeddings (BERT)

	s (FastText)		s (FastText)		s (FastText)	
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 5.1: 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 (Reimers & Gurevych, 2020).)

Table 5.1 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 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). Because the contextual embedding coherence metrics did not show a systematic improvement on all fronts, we limited our experiments with the unannotated set to skip-gram embedding coherence metrics as a default.

5.3.2.1 Alignment of ASR-transcript-derived coherence metrics and manual-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. Higher correlation indicates relative robustness to transcription errors introduced by ASR. The results of this analysis are shown in Table 5.2.

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 5.2: 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 (Reimers & Gurevych, 2020).)

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.

5.3.2.2 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) (Amodei et al., 2016), with a mean WER of 0.36 and CER of 0.2 across the transcripts used in our studies. Figure 5.5 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 errors because it has a higher correlation at all error rates.

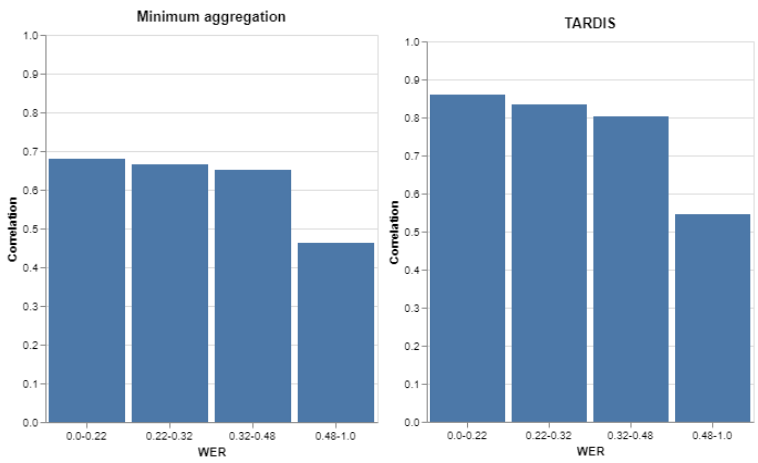
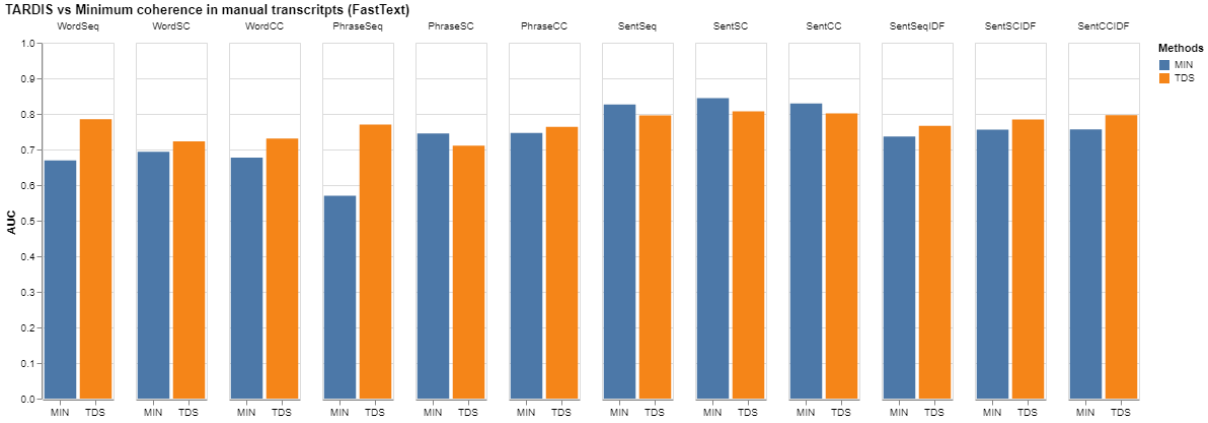
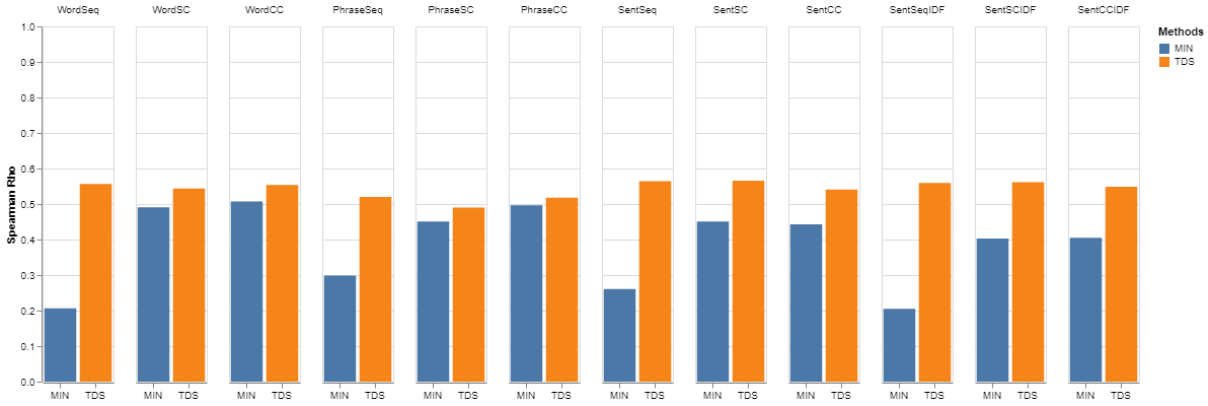


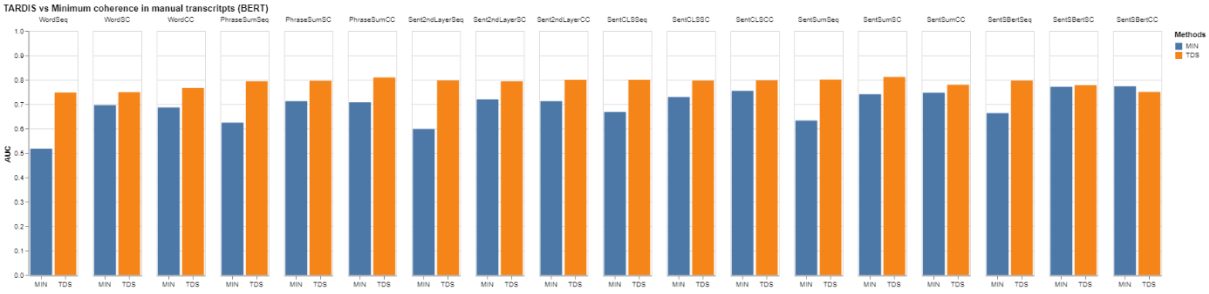
Figure 5.5: Correlations between average coherence scores derived from ASR and manual transcripts. Each bin corresponds to one quartile from the distribution of coherence scores for each transcript. (A) The left figure is plotted using traditional coherence metrics. (B) The right figure is plotted using TARDIS coherence metrics.

5.3.3 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.


TARDIS vs Minimum coherence in manual transcritpts (FastText)
WordSeq
WordSC
WordCC
PhraseSeq
PhraseSC
PhraseCC
SentSeq
SentSC
SentCC
SentSeqIDF
SentSCIDF
SentCCIDF
AUC
Methods
MIN
TDS


WordSeq
WordSC
WordCC
PhraseSeq
PhraseSC
PhraseCC
SentSeq
SentSC
SentCC
SentSeqIDF
SentSCIDF
SentCCIDF
Spearman Rho
Methods
MIN
TDS


TARDIS vs Minimum coherence in manual transcritpts (BERT)
AUC
Methods
MIN
TDS

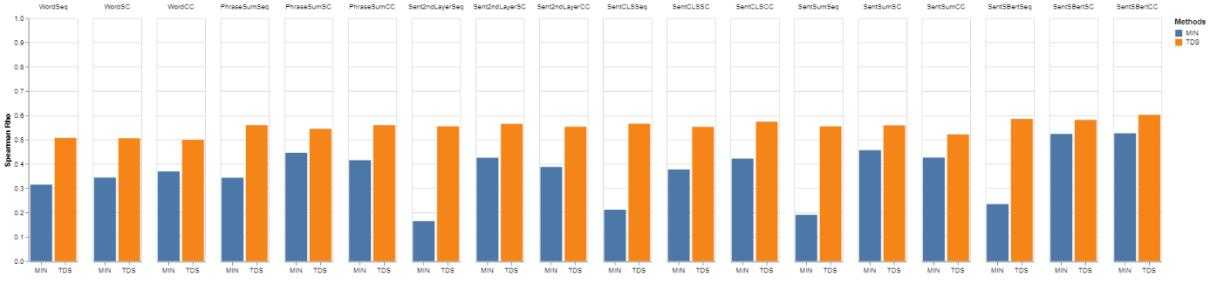

Spearman Rho
Methods
MIN
TDS

Figure 5.6: 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 (Reimers & Gurevych, 2020).

The results of these experiments are shown in Figure 5.6, 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).

5.3.4 Comparison between TARDIS metrics and the Entity Grid representation.

Table 5.3 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 all cases.

	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 5.3: Performance comparison across the best performing TARDIS metrics and the entity grid metric. (**SBERT** = vectors from the sentence-BERT package (Reimers & Gurevych, 2020))

5.3.5 Alignment of the coherence scores to the HPSVQ (Van Lieshout & Goldberg, 2007) clinical scale.

We evaluated the Spearman Rho correlation between various coherence metrics (the minimum coherence estimated for an individual) to the HPSVQ total score.

	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

Table 5.4: 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 (Fisher, 1992) to leverage the p-values of each individual correlation. (**BERTCLS** = embeddings from the CLS token, **SBERT** = vectors from the sentence-BERT package (Reimers & Gurevych, 2020).)

Table 5.4 indicates that some correlations exist between the coherence metrics and AVH severity measured by HPSVQ. Although not strong, the 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 measuring the correlation with HPSVQ scale because the max correlations in each experimental set up came from a metric with BERT-derived embeddings.

5.4 Discussion:

5.4.1 Key findings

In this chapter we presented an evaluation of 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 the fact 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 (Bedi et al., 2015). 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 (Elvevåg et al., 2007). 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 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 (Voleti et al., 2019). 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. Therefore, 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 5.7). 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 preliminary 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 between 0-1 and represent coherence instead of derailment). As such, it is readily apparent why TARDIS produces a better estimation of coherence in this case.

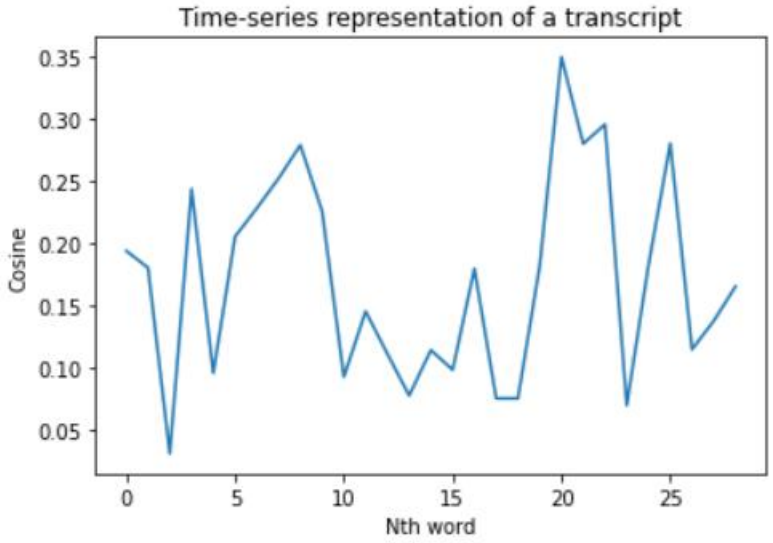


Figure 5.7: Time-series representation of an illustrative transcript. The y-axis represents cosine values computed between adjacent word vectors while the x-axis represents the order in which

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.

We also demonstrated the robustness of TARDIS-based metrics across different sets 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 (Sommer et al., 2010), potentially because incoherence of covert (i.e. 'internal') speech may influence discourse processing and manifest as poorly organized overt speech (Hoffman, 1986).

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 (Barzilay & Lapata, 2008). 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 (Covington et al., 2005). 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 (Barzilay & Lapata, 2008). 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 as features 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.

5.5 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.

Chapter 6: Perplexity and proximity: Large language model perplexity complements semantic distance metrics for the detection of incoherent speech

Parts of this chapter are drawn from a previously published paper

Xu, W., Pakhomov, S., Heagerty, P., Horvitz, E., Bradley, E. R., Woolley, J., Campbell, A., Cohen, A., Ben-Zeev, D., & Cohen, T. (2025). Perplexity and proximity: Large language model perplexity complements semantic distance metrics for the detection of incoherent speech. Journal of Biomedical Informatics, 170, 104899. https://doi.org/10.1016/J.JBI.2025.104899

6.1 Introduction

In this chapter, we expand our horizon beyond proximity-based coherence metrics and incorporate perplexity from decoder family models to coherence assessment methods. This aims to address the gap in current research landscape that there is limited exploration of applying perplexity on the task of coherence assessment (discussed in chapter 2.7.4). Prior studies did not explore the potential of sentence-level LLM-based scores and focused primarily on token level scores. Research on proximity metrics (Bedi et al., 2015; Xu et al., 2020, 2022) shows that sentence-level scores align well with human annotations, suggesting that perplexity methods at this level of granularity would be worth exploring also. Additionally, none of the prior work includes a comparison between LLM- and proximity-based measures of semantic coherence.

In this work, we address this gap in literature by developing a novel sentence-level perplexity-based metric and applying perplexity-and proximity-based measures in concord. Our work is inspired by the work of Sap, Jafarpour and colleagues (Sap et al., 2020, 2022). This work showed that recalled and imagined stories can be differentiated by a metric called linearity,

which is derived from language model *perplexity*, which measures the extent to which a pre-trained model finds previously unseen text unpredictable. We hypothesized that incoherent speech could also be distinguished by an LLM's perplexities because LLMs, trained on large amounts of presumably coherent language, should be "surprised" when presented with incoherent speech samples, resulting in higher perplexity measurements. We also hypothesized that LLM metrics (*perplexity* metrics) would be complementary to the classic semantic distance metrics (*proximity* metrics) because of their differences in conception. To evaluate this hypothesis, we used perplexity metrics derived from the LLaMA3 (Team & Meta, 2024) LLM, which has 70B parameters, and proximity metrics from the comprehensive coherence calculator (CCC)[19] (Xu et al., 2022), which implements a range of proximity-based metrics using both global and contextual embeddings. We trained predictive models using proximity features, perplexity features and a combination of the two feature sets on a dataset collected from people experiencing auditory verbal hallucinations (AVH) (with annotated derailment severity) (Ben-Zeev et al., 2020) and tested the resulting models on a dataset derived from clinical interviews with schizophrenia patients (with annotated thought disorder severity) (Bradley et al., 2024).

6.2 Method

6.2.1 Datasets:

6.2.1.1AVH dataset:

The AVH dataset was obtained from a prior study (Ben-Zeev et al., 2020) that collected speech samples using a smartphone application from n=384 participants who experienced hallucinations. Participants experiencing AVH were recruited via both in-person and online

[19] https://github.com/LinguisticAnomalies/Coherence

means. A clinical diagnosis was not a requirement for participation, though some participants (n=151) self-reported diagnoses of SSDs. Informed consent from participants was obtained through a rigorous procedure involving triple confirmations from a screening questionnaire. All participants were asked to install a mobile application, which had the capability of recording and uploading audio diaries, and were prompted to describe their experiences of AVH, as well as anything else they would like to share or think it would be helpful for the research team to know. Prompts for audio diaries followed the collection of other data, and the participants also had the option to record an entry directly on demand. Although no monetary incentives were offered for the audio diary component, most participants submitted their recorded audio diaries. We used data collected up to October 18th, 2019, consisting of 1868 recordings from 202 participants. From these recordings, we selected 310 transcripts that satisfy the following criteria: 1. They are longer than 30 seconds. 2. Each participant has at most 3 samples. 3. They have coherence labels from expert annotators. Two annotators labeled the transcripts for their degree of incoherence based on the Thought and Language Disorder (TALD) scale (Kircher et al., 2014), using the construct of derailment which was selected on account of its fit with the open-ended nature of the audio diary task, where a coherent flow of topics is more relevant than adherence to an initial topic or question. The TALD score ranges from 0-4 and represents greater incoherence as the score increases. The inter-rater agreement between annotators was 0.71, as measured by quadratically-weighted Kappa. This set contains samples that have a mean TALD score of 1.21 with a standard deviation of 0.87. As in previous research (Xu et al., 2020, 2022), severe cases of thought disorganization were defined by scores of 3 or higher. This threshold was used for the purpose of computing the area under the receiver operating characteristic curve (ROC-AUC) in our experiments. We selected this dataset for this study because 1) people with AVH experiences

may suffer from varying degrees of thought disorder, which ensures a diverse degree of semantic coherence in the sample pool, and 2) we have evaluated traditional proximity-based coherence measures using this set previously (Xu et al., 2020, 2022).

6.2.1.2 Clinical interview dataset:

This dataset was obtained from clinical interviews of 39 male outpatients diagnosed with schizophrenia spectrum disorders (SSDs) participating in a study of oxytocin conducted at the University of California, San Francisco. (UCSF)[20] (Bradley et al., 2024). The interviews are semi-structured and the resulting clinical assessments by trained raters are provided in the form of a composite score combining the *conceptual disorganization* item (ranging from 1-7 with increasing severity) (Kay et al., 1987) from the Positive and Negative Syndrome Scale (PANSS) and the *incoherent speech* item (ranging from 0-5 with increasing severity) from the Comprehensive Assessment of Symptom and History (CASH) (Andreasen et al., 1992). The composite score was derived as the sum of the two individual items. We collected 39 manually transcribed interviews with corresponding composite scores between 2 and 8 (in the range of 0-12), with a mean of 3.36 and standard deviation of 1.80. We included this dataset as an additional evaluation of the proposed method's ability to detect incoherent speech because 1) it is structured differently (dialogue vs. monologue) and gathered from a different population, mediating evaluation of the model's capacity for generalization; 2) it permits evaluation at degrees of thought disorganization typical of outpatients with SSDs, arguably the population in which automated coherence assessment would be of greatest clinical value as a means to anticipate

[20] The referenced work used 37 participants because 2 of the participants did not complete all the tasks required. Because we only need the interview transcripts, we included all 39 in this work.

relapse events; and 3) the scores reflect direct assessments of patients by clinicians, rather than judgments arising from the examination of transcribed speech, providing an indication of whether models trained on the latter assessments can generalize to the former.

6.2.2 Proximity metrics

The proximity metrics were based on the best performing 3 coherence metrics from our previous work (Xu et al., 2020), namely the sentence-level sequential, static centroid, and cumulative centroid coherence metrics. These methods perform a semantic distance evaluation between consecutive sentence embeddings (sequential) or between each sentence embedding and a central vector of the transcript, derived as the vector average (or centroid) of the vector representations of the sentences it contains. With the static method, all sentences were used to estimate the centroid. With the cumulative method, only those that preceded the sentence under consideration were used. The sequential coherence metric estimates the local coherence whereas the centroid coherence metrics estimate global coherence. We implemented the 3 metrics with both static embeddings and contextual embeddings, bringing the total number of proximity metrics to 6. The static embeddings were trained using FastText (Bojanowski et al., 2016), and publicly released as a set of 2-million-word vectors[21] trained on Common Crawl[22]. The contextual embeddings were obtained using Sentence-BERT (Reimers & Gurevych, 2020), a model specifically trained on semantic search tasks to produce meaningful sentence embeddings for cosine similarity calculations. Specifically, we used the "all-MiniLM-L6-v2" Sentence-BERT

[21] https://fasttext.cc/docs/en/english-vectors.html
[22] https://commoncrawl.org/

model[23]. For simplicity, we show the best performing embedding version of the metric in the result section. These proximity metrics are well-established tools for evaluating semantic coherence as a proxy for organized thinking and provide strong baselines for this study. Because previous work has shown that neural word embeddings generally have better performance or greater utility in tasks that involve cosine similarity calculations than LSA embeddings (Fang et al., 2016; Miani et al., 2022), we did not include LSA derived word embeddings in this analysis.

6.2.3 Perplexity metrics

The core concept of perplexity metrics is discussed in detail in chapter 3.1.3. In this section, we provide additional implementation details for the experiments used in this chapter. For the AVH data, we used a summary from BART, a leading transformer-based summarization model (Lewis et al., 2019) as the initial context. For the clinical interview data, we used the first sentence of the transcript as the initial context due to the difficulty of obtaining good summaries of data in long-form conversational formats. Because of the conversational nature of this dataset, we also used a different pre-processing strategy to represent sentences as semantic units: the conversations contain short and less meaningful phrases such as “Yes.” or “Of course.”, which are nonetheless categorized as independent sentences by the sentence tokenizer. After initial experiments revealed this limitation, we elected to count any sentence with 4 words or less as a part of the next sentence, to preserve the narrative without introducing semantic gaps by treating phrases with little semantic content as independent sentences. Following this procedure, we used concatenated participant speech for both the proximity and perplexity methods to avoid

[23] https://sbert.net/docs/sentence_transformer/pretrained_models.html

considering the interviewer's speech. The LLM we used to obtain the perplexities was a locally hosted instruction-tuned[24] LLaMA-3.0 model (70 billion parameters), with 4-bit quantized weights obtained with Exllama2[25] , which we used for accelerated inference. After obtaining a series of perplexity values for the sentences in a transcript, we used the best performing aggregation function (mean, median, max, min, 10th and 90th percentiles) to represent the perplexity coherence feature for the input transcript for downstream analysis in the experiments. A detailed discussion of the process of choosing an aggregation function can be found in the *Experiments* sub-section of this manuscript.

6.2.4 Experiments

6.2.4.1 Overview

Figure 6.1 provides an overview of the experiments in this study.

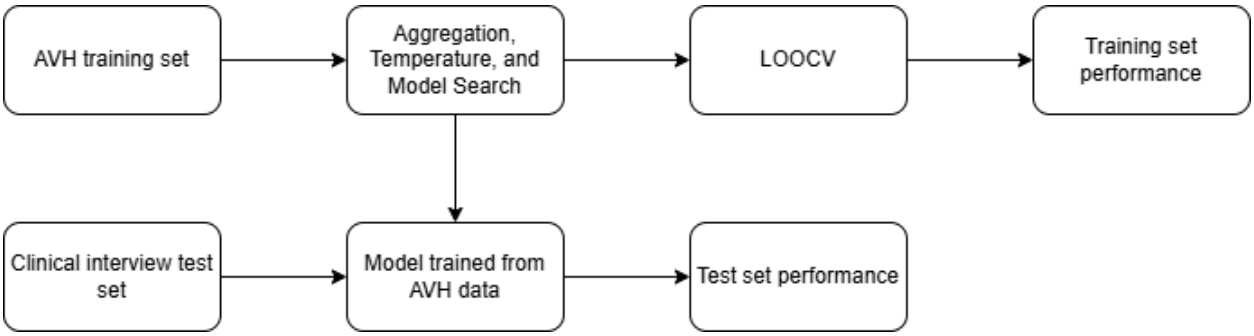


Figure 6.1: An overview of experimental design. LOOCV = leave one out cross-validation.

[24] While instruction-tuning is not necessary to estimate perplexity, we found it did not degrade performance in preliminary experiments and consequently elected to use the instruction-tuned variant of the model hosted internally as a shared service for use by members of the lab.
[25] https://huggingface.co/turboderp/Llama-3-70B-exl2

6.2.4.2 Aggregation, temperature, and model choice

Both the perplexity and proximity metrics produce scores on a sentence-to-sentence scale. So, we examined several aggregation strategies to collect these sentence-level scores as a single feature for the transcript. These include minimum, maximum, median, mean, and the 10^{th} and 90^{th} percentiles. The temperature setting of an LLM is typically applied during text generation, where a higher temperature will encourage the model to sample more broadly from probable next tokens. However, it can also be applied at the point at which perplexity is measured, such that tokens further down the ranked list of predictions have higher probabilities as temperature rises. We evaluated temperature settings of LLaMA3 ranging from 0.1-2.1 with an increment of 0.1 and then from 2.1-9.1 with an increment of 1. To determine the best temperature, we examined the Spearman rank correlation of model loss (i.e. NLL for the transcript) with the annotations in the AVH training set and selected the best temperature based on the correlation score. Using this temperature, we tested different aggregation strategies for the series of NLL values generated from the bag and chain models. We also evaluated different regression models, specifically linear regression, support vector machine regression, and gradient boosting regression to probe for both linear and non-linear relationships between the features (sets of proximity-based and perplexity-based coherence measures) and human ratings of incoherence (TALD derailment or the composite measure). We used leave-one-out (LOO) cross validation to obtain predicted scores from the regression models and computed Spearman rank correlation with human annotations. We chose the best performing model on the training set for the following tests of the utility of different feature sets. Because the training set has a defined criterion for severe cases (TALD score >=3), we also included an evaluation of Area Under the Curve of the Receiver Operating

Characteristic curve (ROC-AUC) for this dataset as an additional evaluation of performance. However, the selection of the best configuration is based on Spearman rank correlation because it is the metric used for the test set due to the lack of a defined threshold in the clinical evaluations accompanying this set.

6.2.4.3 Performance on training set

To evaluate the perplexity and proximity features, we trained models using the following types of feature sets: single proximity features, single perplexity features, and dual-feature combinations of any single feature so that we can observe the effects of combined proximity and perplexity features. The model performance was evaluated by calculating the Spearman rank correlation between the LOO predictions for each transcript and the corresponding human annotations. Through this design, we can determine 1) how the proximity and perplexity feature set alone performs on this dataset and 2) if the combination feature sets including both perplexity and proximity-based coherence estimates outperform the single feature models, or the dual feature models trained using two proximity or perplexity features only.

6.2.4.4 Performance on test set

We tested the performance of our proposed method on the clinical interview dataset by training LR models using the entirety of the AVH dataset. We followed the same feature configuration to test how each feature set performs on the test set. Because the scores used in the clinical interview dataset do not have a well-defined threshold for severe cases, we used the Spearman rank correlation as the evaluation score.

6.2.4.5 Adversarial paraphrase testing

Adversarial paraphrase testing provides a way to test the robustness of a model's interpretation of semantics. However, sentence-level paraphrasing may still introduce differences by perturbing the stylistic continuity of a passage. Therefore, we used adversarial paraphrasing to evaluate how each of the best-performing perplexity and proximity components behave under such perturbations. To do so, we paraphrased the AVH dataset using the FLAN-T5 paraphraser,[26] a model trained to generate paraphrase samples on a sentence-by-sentence basis. We then performed the same LOO evaluation on the paraphrased samples using single feature set regression models to observe how each feature reacts to the paraphrase.

6.3 Results

6.3.1Aggregation method and temperature

We found that a temperature of **1.1** produced the best correlations in the training set across aggregation techniques. **Figure 6.2** shows model NLLs' correlation with human judgment under different temperature settings when used with median aggregation.

[26] https://huggingface.co/alykassem/FLAN-T5-Paraphraser

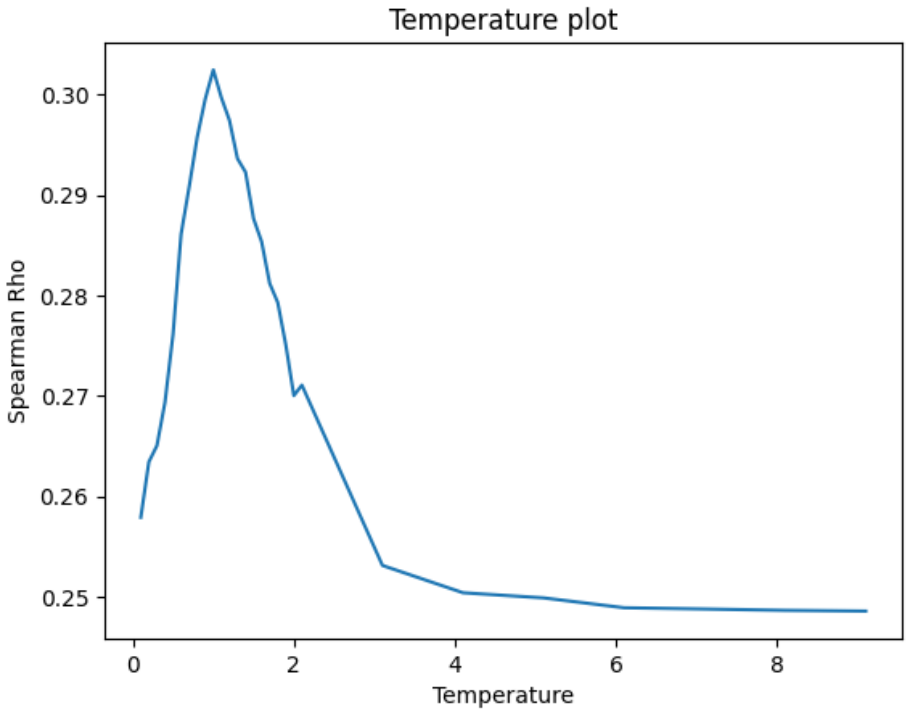


Figure 6.2: Spearman rank correlation between model NLL and human annotation at different temperatures on the training set. Temperature 1.1 produces the best performance.

With temperature 1.1, we also evaluated the Spearman rank correlation with the annotated score for each aggregation strategy as shown in **Table 6.1.** We found that the minimum aggregation was the most effective aggregation strategy for the sequential proximity metrics, which is consistent with previous findings (Bedi et al., 2015; Xu et al., 2020). However, when prioritizing overall alignment (Spearman rank correlations) over the ability to identify severe cases (ROC-AUC), we found that the mean aggregation was better or maintained equal performance to minimum aggregation when using the centroid coherence metrics. Similarly, we found that the median aggregation performed best for the bag model and that the mean aggregation worked best for the chain model. The median aggregation function was proposed to address the

disproportionately high perplexities observed in early sentences when little context was available and was chosen by selecting the aggregation function that produces the best Spearman rank correlation and ROC-AUC with the human annotated scores in the AVH training set. Of note, when considering the best aggregation function, the best-performing *perplexity* metric was the chain model (r=0.47, perplexity), which is also the best performing method in terms of AUC (0.83) and the best-performing *proximity* metric was the cumulative centroid (r=-0.57, proximity).

Table 6.1: Spearman rank correlation (ROC-AUC in parenthesis) of different aggregation strategies for each individual metric with the annotated scores in the AVH dataset (ROC-AUC calculated with the threshold of TALD >= 3). The best Spearman Rho value is highlighted in bold in each category. Note that the transcript level perplexity does not need to be aggregated so it is not shown in this table.

Proximity metrics (*higher* values indicate *more* coherence)				Perplexity metrics (*higher* values indicate *less* coherence)	
	Sequential (static embedding)	Static Centroid (contextual embedding)	Cumulative Centroid (contextual embedding)	Chain Model	Bag Model
Minimum	**-0.27 (0.81)**	-0.51 (0.73)	-0.53 (0.77)	0.01 (0.51)	-0.01 (0.55)
Mean	0.00 (0.66)	**-0.53 (0.79)**	**-0.57 (0.80)**	**0.47 (0.83)**	0.36 (0.73)

Median	-0.03 (0.65)	-0.51 (0.78)	-0.53 (0.77)	0.42 (0.81)	**0.39 (0.76)**
Maximum	0.23 (0.56)	-0.37 (0.76)	0.01 (0.56)	0.44 (0.79)	0.33 (0.63)
90^{th} Percentile	-0.12 (0.52)	-0.44 (0.79)	-0.55 (0.80)	0.39 (0.76)	0.28 (0.67)
10^{th} Percentile	-0.15 (0.76)	-0.50 (0.73)	-0.53 (0.77)	0.27 (0.66)	0.30 (0.66)

6.3.2 Model selection

We also evaluated the performance of linear and non-linear regression models on the training set. We included the model predictions' correlation with annotated scores using each metric as a single feature training data as shown in **Table 6.2**. The predictions were obtained using LOO cross validation, with a prediction for the held-out transcript from each LOO iteration. Our results indicate that linear regression is the best modeling approach for this task. Non-linear models did not offer improvement on the LOO cross validation performance, perhaps because they overfit to these single-feature inputs.

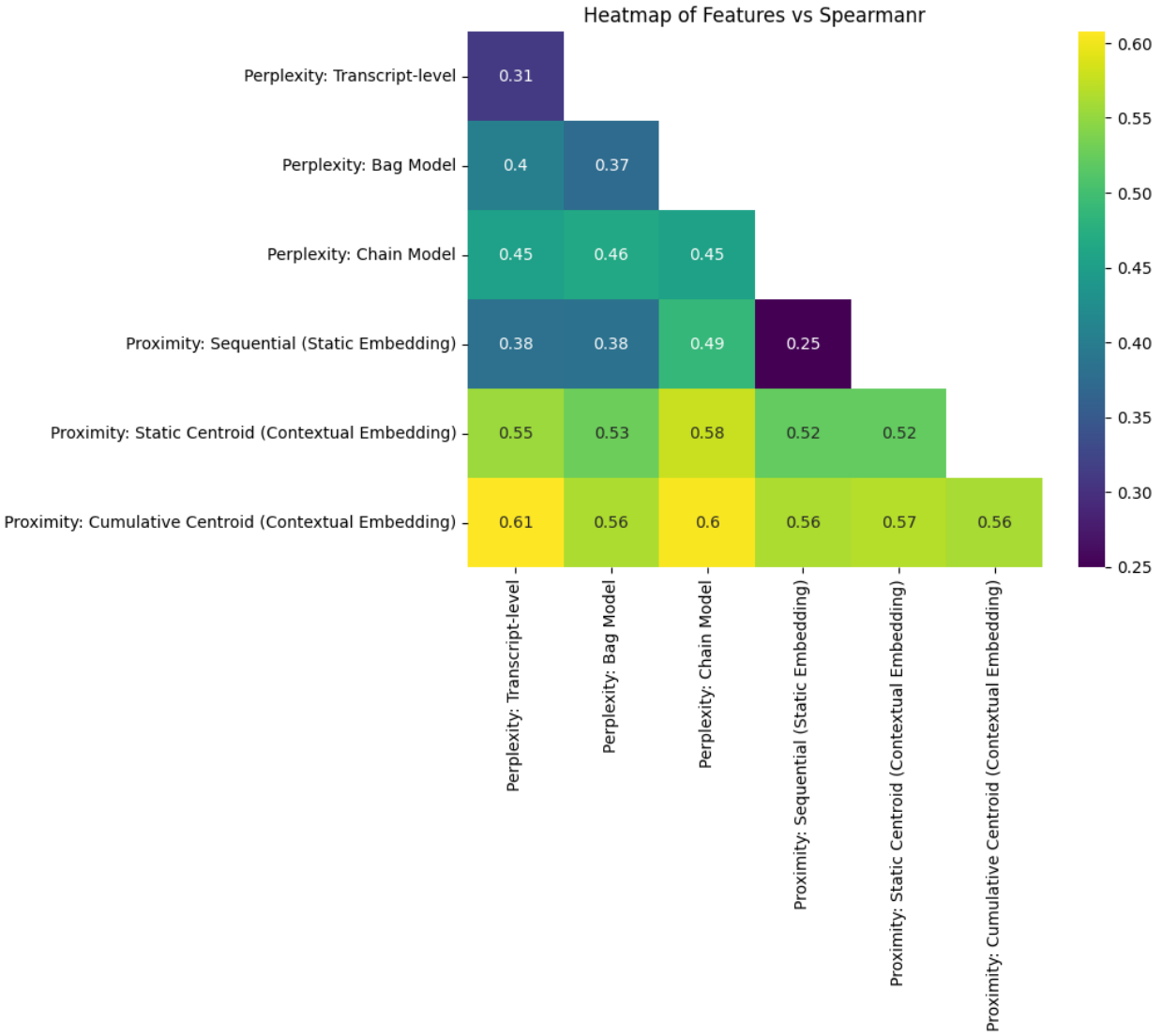
Table 6.2: Spearman rank correlation between regression scores and human annotations (ROC-AUC in parenthesis) using single feature with the best aggregation across LR (linear regression), SVM (support vector machine), and GBR (gradient boosting regression).

	Proximity metrics	Perplexity metric

	Sequential (static embedding)	Static Centroid (contextual embedding)	Cumulative Centroid (contextual embedding)	Chain Model	Bag Model	Transcript-level Perplexity
LR	**0.23 (0.81)**	**0.51 (0.76)**	**0.56 (0.79)**	**0.45 (0.79)**	**0.36 (0.71)**	**0.31 (0.81)**
SVM	0.21 (0.76)	0.51 (0.76)	0.54 (0.79)	0.43 (0.78)	0.31 (0.72)	0.28 (0.77)
GBR	0.09 (0.65)	0.40 (0.80)	0.45 (0.81)	0.35 (0.70)	0.25 (0.64)	0.21 (0.72)

6.3.3 Training set LOO performance

We used LOO cross validation to generate regression predictions using either 1 or 2 feature models. The performance was evaluated using Spearman rank correlation between the predicted scores and human annotations. The results are shown in **Figure 6.3**. We found that the dual feature combinations generally show improvements over the performance of individual feature models, especially when combining a **perplexity metric** and a **proximity metric**. The best performing feature set is a combination of transcript-level perplexity and cumulative centroid features, with a correlation of 0.61 to human annotated scores. This feature set also performed well for identification of severe cases, with an ROC-AUC of 0.88. However, on this task it was outperformed by the combination between the chain model (perplexity) and sequential measure (proximity), with an ROC-AUC of 0.89, both of which performed well (though not as well, with ROC-AUC of 0.79 and 0.81 respectively) for identification of severe cases as individual features.


Heatmap of Features vs Spearmanr
Perplexity: Transcript-level
Perplexity: Bag Model
Perplexity: Chain Model
Proximity: Sequential (Static Embedding)
Proximity: Static Centroid (Contextual Embedding)
Proximity: Cumulative Centroid (Contextual Embedding)
0.31
0.4 0.37
0.45 0.46 0.45
0.38 0.38 0.49 0.25
0.55 0.53 0.58 0.52 0.52
0.61 0.56 0.6 0.56 0.57 0.56
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25

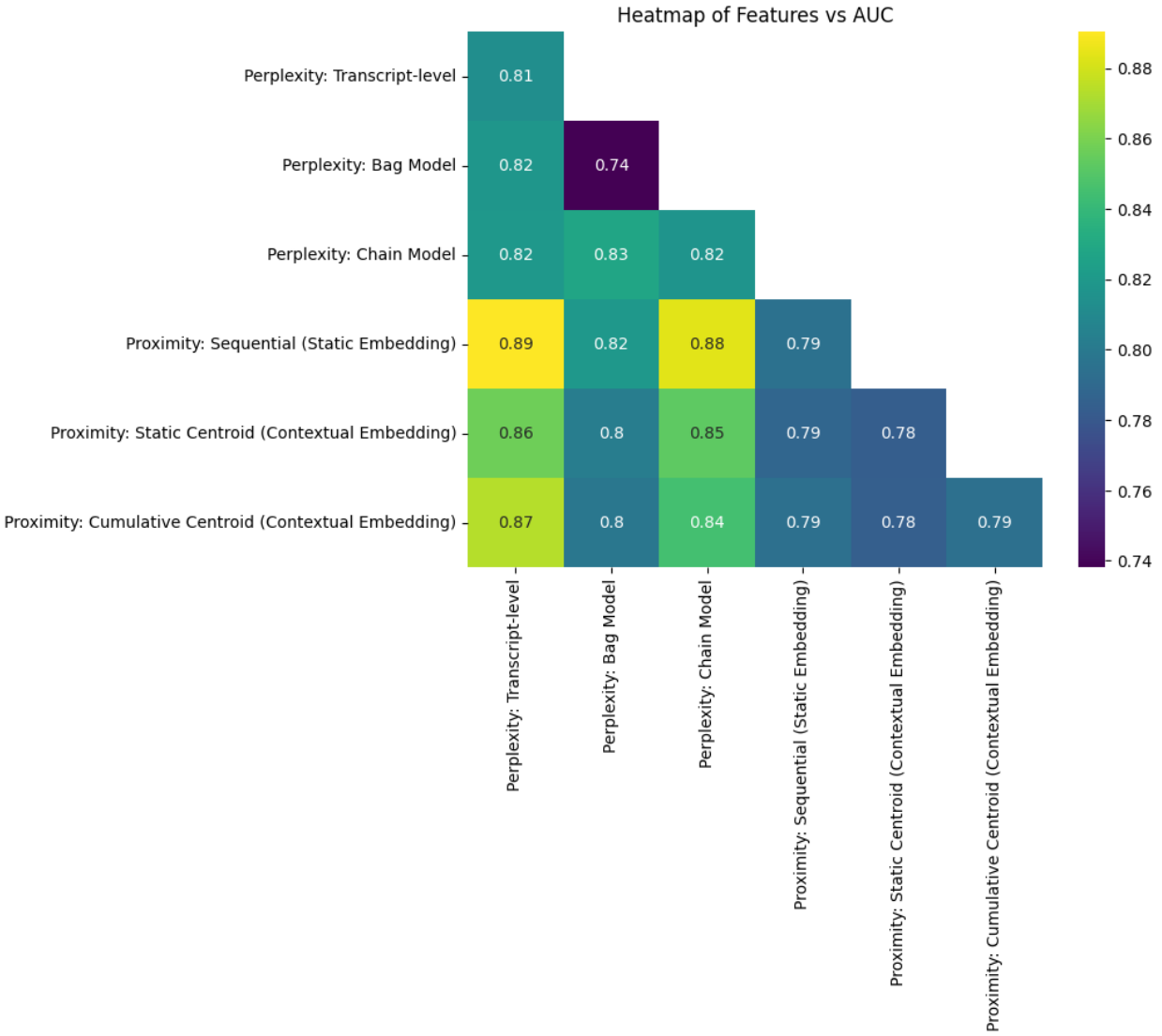


Figure 6.3: Training set cross-validation Spearman rank correlation (TOP) and ROC-AUC scores (BOTTOM) between model predictions and human annotations on the AVH dataset, including single feature models (in the diagonal of the heatmap) and 2 feature models of all combinations. Brighter shading indicates better scores.

6.3.4 Test set performance

We applied the models trained on the AVH dataset to the clinical interview dataset for further evaluation. The results are shown in **Figure 6.4**. A similar pattern can be observed to that seen in the LOO cross validation result in the AVH dataset: combining a perplexity metric and proximity metric as a dual feature set produced the best results. In the test set, the sequential proximity model achieved a 0.52 correlation with the human annotations. However, this performance is further enhanced by the combination with the chain model feature resulting in the best performance of **0.54** correlation with human annotations. The linear regression model coefficients for the best performing model (sequential + chain) are -0.29 for the proximity feature and 0.41 for the perplexity feature, suggesting approximately equal contribution. Of note, all of the feature combinations that included proximity: sequential have correlations that are highly significant ($p<0.01$) and all of those combinations that include proximity: static centroid are significant ($p<0.05$) except the combination with cumulative centroid.

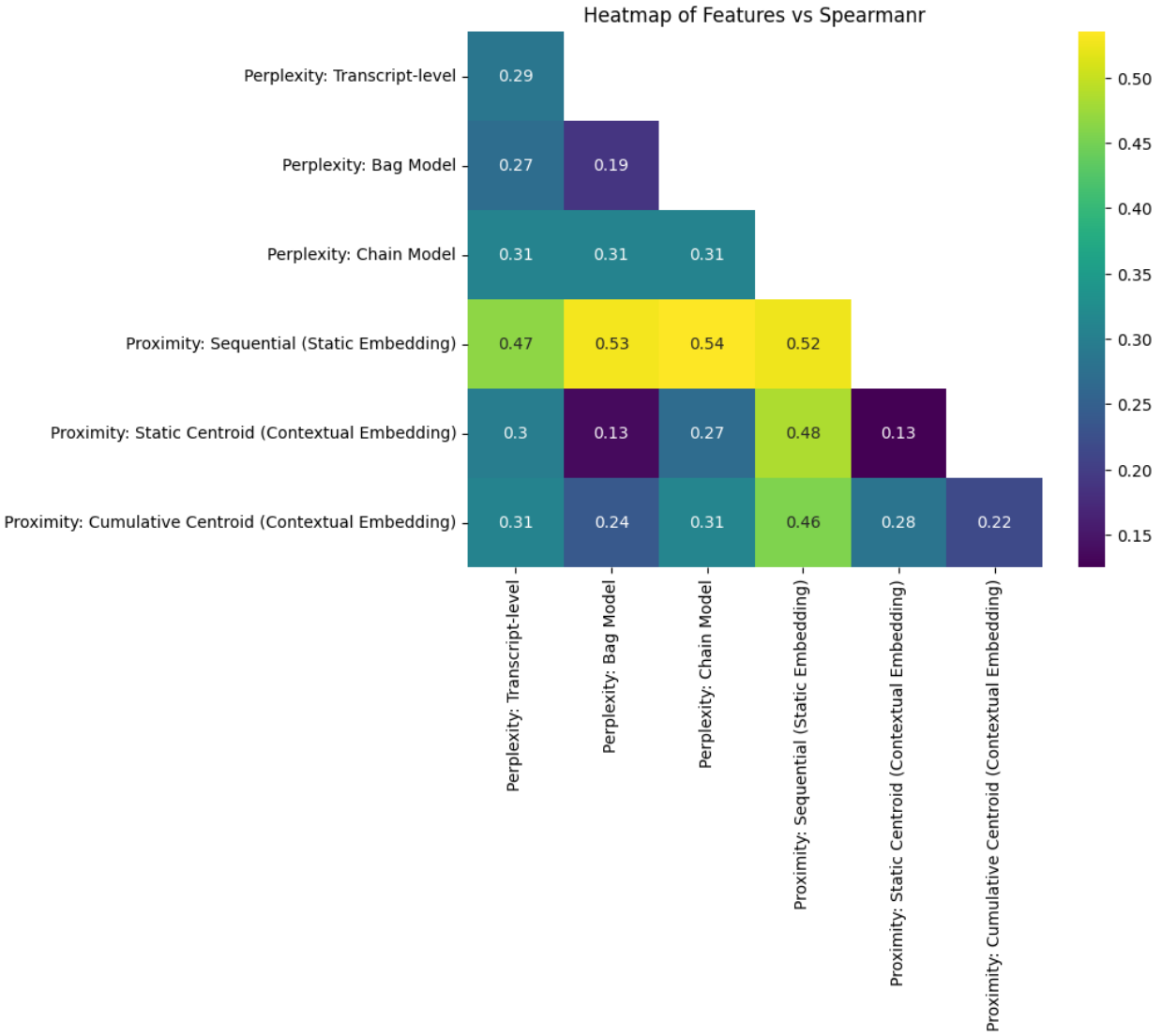


Figure 6.4: Test set Spearman rank correlation between model predictions and human annotations, including single feature models (in diagonal of the heatmap) and dual feature models of all combinations.

6.3.5 Adversarial paraphrase test

We performed the LOO evaluation on the paraphrased AHV dataset. The results are shown in Table 6.3. The perplexity metrics are most affected by this perturbation, whereas the proximity

metrics maintain performance on the paraphrased data. This was to some extent expected because the paraphrase was performed on individual sentences. While the semantics of each sentence might be similar to the original transcripts, the connectivity between each sentence could be lost in the process as the paraphrase model does not have access to previous text as context. We posit that perplexity features are more sensitive to such continuity within the transcripts, and disturbances in the flow between sentences likely obscure a key signal that the feature is responding, to, and lead to high perplexity in general. Here are some examples of the original and paraphrased transcripts to demonstrate this disturbance.

Original transcript:

"I'm a really nice person and I don't judge people at face value. There's some truth that, physically, more symmetrical faces equate to better behavior in society and higher intelligence. Especially attractive people are more likely to behave well in society, and I am a very pretty girl, very pretty."

The corresponding paraphrased transcript is shown below:

"I don't judge people by their appearance. Symmetry is a good thing in life and it can help you to be more intelligent. A pretty girl is more likely to behave well in society and i have a pretty profile."

The paraphrased transcript reads as disjointed sentences that do not follow the natural flow of speech, and the second sentence distorts the meaning of the original transcript which does not

suggest that symmetry can improve intelligence. However, this distinction is unlikely to be picked up by proximity models that solely rely on word emebedding similarities.

Original transcript:

“Hello.”

Paraphrased transcript:

“Hello and welcome to the etsy store.”

This differentiation is caused by the hallucination of the paraphrase model. This can result in unpredictable effects based on the hallucinated content but it is likely to disrupt the continuity between the sentences and thus cause inaccurate perplexity assessment.

Table 6.3: LOO performance (Spearman rank correlation with human annotation) comparison between the original transcripts and paraphrased transcripts using single feature linear regression models.

	Original transcript performance	Paraphrased transcript performance
Perplexity	0.31	0.12
Perplexity: Bag Model	0.38	0.29
Perplexity: Chain Model	0.46	0.11

Proximity: Sequential	0.25	0.37
Proximity: Static Centroid	0.50	0.47
Proximity: Cumulative Centroid	0.56	0.56

6.4 Discussion

6.4.1Key findings

In this study, we explored the utility of probabilistic approaches to automatically estimate deficits in semantic coherence using LLM perplexity, including a novel sentence-level approach. The scores produced by both transcript-level LLM perplexity and the new approach correlate with human judgments of disorganized thinking across two different datasets as an individual feature. While the best-performing proximity metrics from prior work exhibited stronger performance as individual features, combining perplexity and proximity metrics in a multi-variate model yielded the best overall performance in terms of correlation with human judgement on both datasets used in the study. This suggests that perplexity metrics and proximity metrics are complementary and can be combined into a composite feature set to improve the measurement of semantic coherence.

6.4.2 The unique contribution of perplexity

From the results, we observe that the combination of LLaMA3-derived perplexity feature and FastText derived proximity features attain the highest correlation with human judgement in both

training set LOO cross validation and test set evaluation scenarios. We initially attributed this high performance to LLaMA3's better representation of text semantics. However, further experiments using LLaMA3-derived word embeddings to calculate proximity features showed that this is not the case. Our conclusion stems from two observations: 1). Using the training set, we found that the proximity metrics using LLaMA3-derived embeddings have higher correlation with other proximity metrics than the perplexity metrics, which are also derived from LLaMA3. For example, the values from the cumulative centroid using LLaMA3 embeddings have a correlation of 0.79 with the cumulative centroid using Sentence-BERT embeddings but it only has -0.31 and -0.49 correlations with the chain and bag models respectively. This indicates that proximity metrics using LLaMA3-derived embeddings behave more similarly to proximity features derived from other embeddings than to LLaMA3-derived perplexity features. This observation suggests that the improved performance is due to LLM perplexity's unique contributions rather than better representation of text in embeddings from an LLM.

6.4.3 The complementary nature of the two feature sets

Like previously established distributional representations of language (Cohen & Widdows, 2009; Turney & Pantel, 2010), neural word embeddings are vector representation of semantics derived such that words with similar meanings will have vectors oriented in a similar direction (Mikolov, Chen, et al., 2013). It follows from this that proximity features should be measuring the semantic similarity between two text units. In contrast, the perplexity of a LLM measures the level of "surprise" of the LLM given a prior context. As such, the perplexity feature reacts not only to semantic similarity, but also to *continuity*, which indicates how likely a sentence is to occur after a prior sentence regardless of their semantic similarity. This contributes additional, and

apparently useful, information to the downstream regression model. This suggests that the perplexity features respond to both similarity and continuity, rather than selectively, to similarity alone. Both similarity and continuity are aspects of coherent speech. Therefore, adding the perplexity feature contributes additional information that is not available with the proximity feature alone.

We show some specific snippets of transcripts to gain insight into what the perplexity and proximity features are responding to. We analyze these transcripts through the lens of the chain model (perplexity) and the sequential model (proximity) because they are easier to interpret with sentence-to-sentence transitions. For example, consider the following snippet of a transcript:

"I don't know if it's real or somebody's messing with me or something. I know I heard it. It's not in my head."

In this example, because "I know I heard it." and "It's not in my head." have low semantic relatedness, the proximity sequential method provides a coherence score of 0.58, which is relatively low for cosine similarity between sentence embeddings (mean value 0.70). However, the perplexity score for the last sentence is 0.91, which is also low for perplexities (mean value 2.07), suggesting a coherent transition from the previous context (which, in the case of the chain model, includes earlier sentences also). This illustrates how the perplexity feature can accurately capture normal transitions between sentences that have been misrepresented as incoherent by the proximity feature due to semantic dissimilarity.

On the other hand, the perplexity feature may be insensitive to sudden transitions of topic that indicate the derailment of the train of thought. In an incoherent transcript:

"Well, the pedophile finally got me up out of the out of the devil's out of the devil's trap. I'm way uptown now. I got away from out of Satan's grasp. I'm up here by public XXX public market." (location name redacted to protect privacy

In this example, the speaker shifts from describing a metaphorical escape from the devil to describing their location. The perplexity score of the last sentence is 0.40, which is very low and indicates a coherent transition. In contrast, the proximity feature is very sensitive to the semantic shift from the diabolical to a marketplace, so the score for the last sentence is 0.47, which indicates low coherence. Therefore, proximity features can sometimes detect incoherent transitions that may not be apparent to perplexity features. These examples illustrate the complementary nature of these approaches.

Reflecting upon specific mechanisms thought to underlie the language differences in SSD, we theorize that the concepts of semantic similarity and continuity may correspond to differences in the subject's semantic and working memory respectively. The theory of semantic memory assumes that in our mind, words and concepts are linked in a network based on their association (Anderson, 1983; Kuperberg, 2010a). In accordance with this proposal, language differences in SSD's related to semantic memory may manifest as the "loosening of association" we observe in affected speech samples (Kuperberg, 2010a). As the distance between word vectors (proximity

metrics) represents semantic similarity, it fits the needs of an automated tool to assess differences related to semantic memory (the proposal that text-derived proximity metrics provide a computational model of semantic memory dates back at least as far back as Lund, 1995 (Lund et al., 1995)). On the other hand, working memory and executive functions can also be affected in SSDs (Kerns, 2007; Kuperberg, 2010a). The impairment of working memory can be reflected as the inability to recognize the constraints imposed by the context in previous text (Kuperberg, 2010b). Sharpe et al (Sharpe et al., 2024) measured the NLL values of a GPT3 model with varying length of contexts and discovered that the difference between control and schizophrenia group widens as longer contexts are used in language models, suggesting that the impairment of working memory in the schizophrenia group might have caused deviation from distant context. As such, because the perplexity metrics are based on the next token probabilities of a LLM, we speculate that it may serve as an automated tool to assess aspects of incoherence related to working memory differences in SSDs.

6.5 Conclusion

In this chapter, we leveraged features from an LLM to enhance the detection of incoherent speech. We found that features derived from LLM's perplexity can complement features derived from the proximity of word embeddings. The improved correlation with human judgment exhibited with combined models in both training and test sets indicates an important role for LLM-derived metrics in the assessment of speech for indicators of thought disorganization, with the potential to improve automated diagnosis and monitoring of patients with schizophrenia-spectrum disorders.

Chapter 7: Comprehensive Coherence Calculator (CCC)

7.1 Overview

The CCC is an open source freely available software tool[27] that implements the innovative methods discussed in previous chapters and incorporates foundational methods in the field to allow for comprehensive analysis of semantic coherence. The CCC includes proximity, perplexity and speech graph coherence pipelines with various configurations (refer to Figure 6 in Section 3.2). The tool accepts text data and outputs numerical scores to indicate the level of coherence of the text input. The users are able to compare coherence scores obtained through different methods and may use them in any combination or with other features in downstream analysis. This tool is available to be installed on a Python environment with Pip[28], a standard Python package installer and with all dependencies configured.

7.2 Implementation details

7.2.1 Proximity pipeline

For the proximity pipeline, the CCC supports segmentation into semantic units of various sizes, including tokens (for BERT model), words, noun phrases, sentences, and custom separations of arbitrary segments, which has proven helpful in studying coherence in conjunction with natural pauses in speech (Chen et al., 2025).

[27] CCC: https://github.com/LinguisticAnomalies/Coherence
[28] Pip: https://pypi.org/project/pip/

For embeddings, the CCC includes FastText[29] as a native static embedding and supports any word embedding files that can be read via the SemVecPy[30] framework. Larger embeddings such as noun phrases and sentences are computed as the normalized sum of word embedding vectors. For contextual embeddings, the CCC supports the use of embeddings that are derived from the base BERT model. The different types of contextual embeddings are consistent with those that are discussed in Section 5.2.4.2.

For cosine value calculations, the CCC includes sequential, static centroid, and cumulative centroid methods for both local and global coherence assessments, which are implemented based on the descriptions in Section 3.1.1. After the cosine values are generated, the CCC offers both the TARDIS and aggregation strategies to generate the final coherence scores. Because the TARDIS method uses machine learning models to process time-series features, the CCC offers the original SVM model trained using the AVH dataset labeled with TALD (chapter 5.2.1) as an innate regressor model for coherence scoring. However, given the limited diversity of the original training set, the results may not generalize well to data that have a drastically different format from the AVH set. To mitigate this issue, the CCC also supports training of custom TARDIS models provided users have their own text data and coherence labels for model training, after which the users can apply the newly trained TARDIS model on a held-out set for coherence evaluation.

[29] Fasttextpretrained vectors. URL: https://fasttext.cc/docs/en/english-vectors.html
[30] Semantic vectors software package URL: https://github.com/semanticvectors/semanticvectors

7.2.2 Perplexity pipeline

For the perplexity pipeline, the CCC implements both token and sentence level perplexity methods. For the token level approach, the CCC implements the methods used in Li et al's work (Li et al., 2025), where a sliding window is used to iterate over all the input tokens and a negative log loss (NLL) is computed for each window similar to how it is calculated in Section 3.1.3, except the sentences are replaced by sliding windows and the context is limited to the token prior to the window. A maximum aggregation is taken to represent the coherence scores after the NLLs are calculated.

For the sentence level perplexity methods, the CCC implements the chain and bag models discussed in chapter 3.1.3. To facilitate the distribution of this package, the CCC implements the perplexity methods using a relatively small LLM (Pythia 1B model[31]). While this model has been shown to perform well with sliding window perplexity (Li et al., 2025), it limits the performance of sentence-level bag and chain models comparing to using a larger LLM such as a LLaMA 70B model. However, because such models used in the original experiments were hosted on a local server with considerable GPU resources, they are infeasible for inclusion in a light-weight installable Python package. At the point of this writing, the CCC only uses the smaller LLM for both perplexity-based methods but in the future, an option could be added for users to use their own model checkpoints for a larger LLM if they so choose.

[31] Pythia model: https://huggingface.co/EleutherAI/pythia-1b

7.2.3 Speech graph pipeline

The CCC also incorporates the speech graph method and its features as coherence assessment scores. Mota el al (Mota et al., 2012)'s work on speech graphs is discussed in Section 2.5.2. However, because the original manuscript did not provide an open-source code base to guide the implementation, we sought the more recent work from the same group (Carrillo et al., 2016), and implemented the speech graph based on their published codebase[32]. The CCC currently supports the following graph features: number of nodes, number of edges, number of parallel edges, number of strongly connected components, largest strongly connected component (LSC), loop of length 1 (L1), density, average degree, and standard deviation of degree. Of note, the L1 feature could potentially be a valuable supplement to proximity-based coherence scores because it measures the repetition of the same word, which may cause undesirable inflation of cosine similarity values in proximity methods, and therefore offers an opportunity to correct this bias for proximity-based coherence scores in downstream analysis.

[32] Speech graph: https://github.com/guillermodoghel/speechgraph

Chapter 8: Discussion and Conclusion

8.1 Evaluation of hypotheses

Hypothesis 1: Automated assessments of global coherence can identify linguistic manifestations of formal thought disorder as or more effectively than established sequential measures.

In this work, we developed two novel global coherence assessment methods and applied them to a naturally collected speech dataset with derailment labels in the form of TALD scores. Based on experimental results discussed in Chapter 4, we found that the global coherence assessment methods aligned better with human judgement on both AUC and Spearman Rho metrics when compared to traditional sequential measures. This also suggests that measurement of global coherence does not necessarily need an external reference text as a comparison to be useful. Therefore, our results support this hypothesis.

Hypothesis 2: Coherence assessments incorporating time-series analysis will be more robust to automatic speech recognition (ASR) errors than those based on minimum aggregation.

To verify this hypothesis, we implemented the TARDIS framework to take advantage of time-series features and to use as an alternative to cosine similarity aggregation. It was applied to a dataset containing both manual transcripts and ASR transcripts with high word error rate. Based on the experimental results discussed in Chapter 5, we found that the TARDIS can prominently mitigate the negative impacts introduced by high ASR errors, achieving similar performance to

manual transcripts with even better performance in some cases. Therefore, our results provide strong support for this hypothesis.

Hypothesis 3: Perplexity from LLM complements proximity methods, providing additional information to improve alignment with human appraisal of disorganized thinking

In this work, we developed two new perplexity-based coherence methods that include the use of state of the art LLM. From the experiments discussed in Chapter 6, we found that they did not outperform proximity-based coherence methods when used alone, but when used in conjunction with proximity-based methods as a combined feature set for a regression model, the performance was better than using either proximity or perplexity features alone. These results provide robust support for this hypothesis.

8.2 Contributions

Given the gaps in coherence research, this dissertation aims to bridge the gaps through the following contributions.

8.2.1 New methodologies that align better with human judgement

The new methods include global coherence metrics that measure global coherence as an innate feature of participant speech, a time-series approach as an alternative to static aggregation, and perplexity-based methods derived from LLMs.

The new global coherence metrics could fill the gap by the fact that established proximity coherence methods are focused on local coherence. The new global coherence methods developed in this work are independent from an interview prompt or any outside reference text, providing a solid foundation for a data collection pipeline that does not require extensive interview protocols.

My novel time-series augmented coherence score calculation improves upon static aggregation by incorporating more information from the sequence of cosine similarities, allowing coherence assessment methods to be more robust to ASR errors. This makes deploying a coherence assessment pipeline after an ASR pipeline more feasible, circumventing the costly manual transcription step.

Last but not least, the perplexity-based methods provide new exploration on the frontier of LLM applications, bringing more insights into the decoder modeling approach as it pertains to coherence assessment.

8.2.2 Comprehensive view of coherence evaluation in the context of schizophrenia research

Following the development of new coherence evaluation methods, this dissertation also implements a variety of coherence methods (old and new) including proximity, perplexity and speech graphs-based methods into the form of the Comprehensive Coherence Calculator (CCC)

software package, which can be deployed on any text data. This work also provides comprehensive analysis in the context of human annotated data that were gathered outside of controlled experimental settings. This comprehensive analysis illuminates the behaviors of different coherence methods, including comparison between global and local coherence methods, time-series and aggregation strategies, and combining perplexity and proximity-based strategies. It also reveals effects of nuanced configurations for each method category, such as effects of lemmatization for word embeddings, and effects of LLM temperature for perplexity-based coherence scores.

8.3 Generalizability of results

The main results derived from this work are three-fold: 1) the usefulness of the vector centroid as the basis for an effective representation of global coherence 2) time-series augmentation to substitute static aggregation can mitigate the impact of ASR errors, and 3) perplexity-based coherence method can complement proximity-based method and provide more accurate assessment when used together. Since the publication of the results, there are various levels of evidence that suggest these methods can generalize well to datasets beyond those that they were developed from.

For the centroid global coherence method, Just el al (S. A. Just et al., 2023), performed a study that used speech samples from patients with a diagnosis of schizophrenia or schizoaffective disorder that were obtained at two measurement points, 6 months apart. The dataset is a completely different set from the AVH dataset discussed in Chapter 4 and includes key clinical

metrics such as scores from the Positive and Negative Syndrome Scale (PANSS). These authors essentially implemented the static centroid method as an estimate for global coherence (although they did not directly credit this work in the manuscript). From their results, we can observe that the global coherence method outperformed local coherence in many evaluations including an assessment of correlation with the key indicators of thought disorder: PANSS positive symptom and PANSS disorganized symptom. This external evidence suggests that the centroid method can serve as a good indication of global coherence, which may outperform local coherence method when considering coherence assessment for thought disorganization.

For the TARDIS method, although we could not find external research that directly studies the effects of ASR errors on downstream coherence pipeline performance, recent work from Chen et al (Chen et al., 2025) included the TARDIS method in an analysis pipeline and found high association with human judgement of thought disorganization on two additional datasets from PsychosisBank (Tan et al., 2023), a global effort to study thought, language and communication anomalies in psychosis. One is the TOPSY dataset, which is labeled with the Thought and Language Index (TLI) and the other is the PsyCL dataset, which is labeled with the Thought Language and Communication (TLC) scale. While the effectiveness of TARDIS on these additional datasets does not directly show the mitigation of ASR effect errors, it still indicates that using the time-series information on cosine similarities is an effective way to estimate coherence and can outperform static aggregations in some situations.

For the perplexity-based method, because this work was recently published, we did not identify external evidence of generalizability. However, the experiments discussed in Chapter 6 indicate that a regressor trained using combined perplexity and proximity feature performed the best not only on a leave-one-out cross validation test on the training set (AVH set) but also performed the best on a heldout test set (clinical interview set) that has different formats and annotation scales. This is evidence that the complementary nature of proximity and perplexity metrics is a generalizable quality.

8.4 Applications

An end-to-end coherence assessment pipeline: This work demonstrates that using a fully automated pipeline, from speech recordings to transcripts to coherence scores, the process of quantifying coherence can be completed without manual effort. Such a pipeline can be deployed in a smart-phone application that monitors symptom improvements after treatments by simply collecting speech samples in natural settings or serve as a clinical decision support tool that provides additional information on the patients' symptom severity as they visit the clinic.

Facilitating further research in SSD or other conditions: Although incoherent speech is a prominent symptom of SSD, it can appear in other conditions too. One example is Alzheimer's disease (AD): manual analysis evidence (Dijkstra et al., 2004) suggests that global coherence can be impaired in speech samples from AD patients. In fact, we have preliminary evidence that suggests applying the TARDIS pipeline to a free form conversation dataset (Davis & Pope, 2011) consisting of AD patient and healthy controls can result in high accuracy (0.847) in the dementia

classification task, higher than that of a finetuned BERT text classifier (0.772). Further analysis shows that models using global coherence features drastically outperform models using local coherence features, confirming the evidence from prior research (Dijkstra et al., 2004). Because the topic is not in the scope of this dissertation, it was not pursued further. However, this suggests that the evidence and tools developed in this work can be useful in facilitating further valuable research in SSD or other conditions like AD.

8.5 Limitations

8.5.1 Only evaluating one diagnostic construct

A limitation of this study is that only one construct (i.e. derailment) was used to focus the development efforts of coherence methods. Although it is considered a prominent symptom, there are a number of other constructs that have yet to be modeled. These constructs include tangentiality, logorrhoea and poverty of speech, which if accurately recognized would result in a more granular automated system for the characterization of linguistic manifestations of disordered thinking. In addition, paralinguistic characteristics from the audio were not considered in this work. Audio features such as pause duration (Chen et al., 2025), may provide useful indicators of symptom severity.

8.5.2 Coherence methods reacting to input length

This limitation concerns the possibility that the longer one speaks, the more chances one can derail from a topic. This apparent incoherence may be observed regardless of the speaker's

mental status. Proximity-based coherence methods with minimum aggregation strategies are especially vulnerable to this, because longer texts have more opportunities to produce low cosine values. Coincidentally, many incoherent responses in our dataset are also longer. However, using word count alone did not predict human annotations as well as using a coherence method. One way to circumvent this limitation is to strictly control the input length and should be an important consideration during the data collection process for future studies.

8.5.3 Unclear mechanism for complementary nature of proximity and perplexity features

The specific linguistic differences that perplexity and proximity features are measuring are not determined in this study. We interpreted our findings (Section 6.4) in relation to theories concerning the role of semantic and working memory in linguistic anomalies in SSD but we have not investigated these relations empirically.

8.5.4 Repetitions interfering with coherence methods

Repetitions of the same word can be a characteristic of disorganized thinking. However, in the context of proximity-based coherence methods, repetitions would result in high cosine similarities and therefore could undesirably elevate coherence scores (if not using minimum aggregation). For perplexity-based coherence methods, repetition may not be a significant obstacle. This seems counter-intuitive because a base decoder language model may have reduced perplexity when the same word is repeated multiple times as the element of "surprise" fades. However, an instruction-tuned LLM like the one used in experiments discussed in Chapter 6 that is designed to provide useful response should have high perplexity when nonsensical repetitions

are encountered. As previously discussed, one potential way to address this concern is to use the loop feature from speech graphs to adjust the coherence scores so that the factor of bad repetitions is considered in the overall coherence assessment process.

8.6 Future work

Many new research directions can stem from this work. One direction is to integrate coherence as a single feature in a more comprehensive view of SSD, integrating other symptom assessments such as delusions, persecutory ideations, emotional withdrawal and so forth. This effort is likely to result in an analytical pipeline that can better align with clinical scales.

Another direction is to include audio-related features to coherence analysis or analysis of other symptoms. Audio cues such as loudness, pitch or pause duration can be important indicators of a person's mental state, and including these in addition to the text-based coherence features may prove useful.

With the release of CCC software package, research on combinations of coherence methods may provide further insight into how each method works or produces better assessment. For example, a perturbation test on distal context that simulates a deficit in working memory could help further understand the relationship of proximity and perplexity methods to semantic and working memory. Another example might be a study of an integrated metric of the loop graph feature and proximity coherence to reduce the impact of repetition.

Other research directions include using the coherence assessment pipelines outside the bounds of SSD evaluation. One example is using coherence scores as a component to evaluate Alzheimer's disease, which can also be associated with loss of coherent speech. Text comprehension can be another topic of interest. The earliest automated coherence assessment study (Foltz et al., 1998) involved using text comprehension modifications as a means of producing texts with varying levels of coherence and found good association with LSA-based coherence scores. Our recent work (Cohen et al., 2025) that involved perplexity coherence metrics also indicates that interventions designed to improve comprehensibility also improve coherence. Exploring the possibilities of using coherence scores to evaluate text comprehensibility can be an interesting future direction.

8.7 Conclusion

In this work, we examined the development of computational methods for automated coherence assessment in the context of evaluating the severity of schizophrenia spectrum disorders (SSD). We addressed gaps in prior research by designing and evaluating novel global coherence assessment methods, using time-series augmentation to reinforce against ASR errors, and developing LLM-based perplexity coherence methods to combine with the traditional proximity-based methods for better assessment. Finally, we incorporated both new and classic methods into a comprehensive coherence calculator (CCC) software package for the benefit of research community. With these contributions, fully automated accurate coherence assessment pipeline

becomes a possibility and as such, SSD patients can enjoy stress-free evaluations at home and clinicians can receive additional information about patients' symptoms to achieve better care.

References:

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., … Zhu, Z. (2016). Deep speech 2: End-to-end speech recognition in English and Mandarin. *33rd International Conference on Machine Learning, ICML 2016*.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *22*(3), 261–295. https://doi.org/10.1016/S0022-5371(83)90201-3

Andreasen, N. C. (1986). Scale for the assessment of thought, language, and communication (TLC). *Schizophrenia Bulletin*, *12*(3). https://doi.org/10.1093/schbul/12.3.473

Andreasen, N. C., Flaum, M., & Arndt, S. (1992). The Comprehensive Assessment of Symptoms and History (CASH): An Instrument for Assessing Diagnosis and Psychopathology. *Archives of General Psychiatry*, *49*(8). https://doi.org/10.1001/archpsyc.1992.01820080023004

Andreasen, N. C., & Grove, W. M. (1986). Thought, language, and communication in schizophrenia: diagnosis and prognosis. *Schizophrenia Bulletin*. https://doi.org/10.1093/schbul/12.3.348

Andreasen, N. C., & TUCKER, G. J. (1991). Introductory Textbook of Psychiatry. *American Journal of Psychiatry*, *148*(5). https://doi.org/10.1176/ajp.148.5.670

Aschbrenner, K. A., Naslund, J. A., Grinley, T., Bienvenida, J. C. M., Bartels, S. J., & Brunette, M. (2018). A Survey of Online and Mobile Technology Use at Peer Support Agencies. *Psychiatric Quarterly*. https://doi.org/10.1007/s11126-017-9561-4

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, *1*. https://doi.org/10.3115/v1/p14-1023

Barrera, A., McKenna, P. J., & Berrios, G. E. (2008). Two new scales of formal thought disorder in schizophrenia. *Psychiatry Research*, *157*(1–3). https://doi.org/10.1016/j.psychres.2006.09.017

Barzilay, R., & Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*. https://doi.org/10.1162/coli.2008.34.1.1

Bedi, G., Carrillo, F., Cecchi, G. A., Slezak, D. F., Sigman, M., Mota, N. B., Ribeiro, S., Javitt, D. C., Copelli, M., & Corcoran, C. M. (2015). Automated analysis of free speech predicts

psychosis onset in high-risk youths. *Npj Schizophrenia*. https://doi.org/10.1038/npjschz.2015.30

Ben-Zeev, D., Buck, B., Chander, A., Brian, R., Wang, W., Atkins, D., Brenner, C. J., Cohen, T., Campbell, A., & Munson, J. (2020). Mobile RDoC: Using Smartphones to Understand the Relationship Between Auditory Verbal Hallucinations and Need for Care. *Schizophrenia Bulletin Open*. https://doi.org/10.1093/schizbullopen/sgaa060

Bird, S., Bird, S., & Loper, E. (2016). NLTK : The natural language toolkit NLTK : The Natural Language Toolkit. *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics-Volume 1*.

Bobes, J., Garcia-Portilla, M. P., Bascaran, M. T., Saiz, P. A., & Bousoño, M. (2007). Quality of life in schizophrenic patients. In *Dialogues in Clinical Neuroscience* (Vol. 9, Issue 2). https://doi.org/10.5294/aqui.2011.11.1.5

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051

Bradley, E. R., Portanova, J., Woolley, J. D., Buck, B., Painter, I. S., Hankin, M., Xu, W., & Cohen, T. (2024). Quantifying abnormal emotion processing: A novel computational assessment method and application in schizophrenia. *Psychiatry Research*, *336*, 115893. https://doi.org/10.1016/J.PSYCHRES.2024.115893

Britton, B. K., & Gülgöz, S. (1991). Using Kintsch's Computational Model to Improve Instructional Text: Effects of Repairing Inference Calls on Recall and Cognitive Structures. *Journal of Educational Psychology*, *83*(3), 329–345. https://doi.org/10.1037/0022-0663.83.3.329

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4*.

Burke, E., Gunstad, J., & Hamrick, P. (2023). Comparing global and local semantic coherence of spontaneous speech in persons with Alzheimer's disease and healthy controls. *Applied Corpus Linguistics*, *3*(3), 100064. https://doi.org/10.1016/J.ACORP.2023.100064

Carrillo, F., Mota, N., Copelli, M., Ribeiro, S., Sigman, M., Cecchi, G., & Slezak, D. F. (2016). Automated Speech Analysis for Psychosis Evaluation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9444 LNAI*, 31–39. https://doi.org/10.1007/978-3-319-45174-9_4

Chandler, C., Foltz, P. W., Cohen, A. S., Holmlund, T. B., Cheng, J., Bernstein, J. C., Rosenfeld, E. P., & Elvevåg, B. (2020). Machine learning for ambulatory applications of neuropsychological testing. *Intelligence-Based Medicine*. https://doi.org/10.1016/j.ibmed.2020.100006

Chen, F., Xu, W., Li, C., Pakhomov, S., Cohen, A., Bhola, S., Yin, S., Tang, S. X., Mackinley, M., Palaniyappan, L., Ben-Zeev, D., & Cohen, T. (2025). *Reading Between the Lines: Combining Pause Dynamics and Semantic Coherence for Automated Assessment of Thought Disorder*. https://arxiv.org/pdf/2507.13551

Christ, M., Braun, N., Neuffer, J., & Kempa-Liehr, A. W. (2018a). Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). *Neurocomputing*. https://doi.org/10.1016/j.neucom.2018.03.067

Christ, M., Braun, N., Neuffer, J., & Kempa-Liehr, A. W. (2018b). Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). *Neurocomputing*. https://doi.org/10.1016/j.neucom.2018.03.067

Chuang, Y.-S., Dangovski, R., Luo, H., Zhang, Y., Chang, S., Soljačić, M., Li, S.-W., Yih, W., Kim, Y., & Glass, J. (2022). *DiffCSE: Difference-based Contrastive Learning for Sentence Embeddings*. arXiv. https://doi.org/10.48550/ARXIV.2204.10298

Cohen, T., & Widdows, D. (2009). Empirical distributional semantics: Methods and biomedical applications. In *Journal of Biomedical Informatics*. https://doi.org/10.1016/j.jbi.2009.02.002

Cohen, T., Xu, W., Guo, Y., Pakhomov, S., & Leroy, G. (2025). Coherence and comprehensibility: Large language models predict lay understanding of health-related content. *Journal of Biomedical Informatics*, *161*. https://doi.org/10.1016/j.jbi.2024.104758

Corcoran, C. M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D. C., Bearden, C. E., & Cecchi, G. A. (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*. https://doi.org/10.1002/wps.20491

Covington, M. A., He, C., Brown, C., Naçi, L., McClain, J. T., Fjordbak, B. S., Semple, J., & Brown, J. (2005). Schizophrenia and the structure of language: The linguist's view. *Schizophrenia Research*. https://doi.org/10.1016/j.schres.2005.01.016

Davis, B. H., & Pope, C. (2011). Finding a balance: The Carolinas Conversation Collection. *Corpus Linguistics and Linguistic Theory*, *7*(1). https://doi.org/10.1515/CLLT.2011.007

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*.

Dijkstra, K., Bourgeois, M. S., Allen, R. S., & Burgio, L. D. (2004). Conversational coherence: discourse analysis of older adults with and without dementia. *Journal of Neurolinguistics*, *17*(4), 263–283. https://doi.org/10.1016/S0911-6044(03)00048-4

Ellis, C., Henderson, A., Wright, H. H., & Rogalski, Y. (2016). Global coherence during discourse production in adults: a review of the literature. In *International Journal of Language and Communication Disorders* (Vol. 51, Issue 4). https://doi.org/10.1111/1460-6984.12213

Elvevåg, B., Foltz, P. W., Rosenstein, M., & DeLisi, L. E. (2010). An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. *Journal of Neurolinguistics*. https://doi.org/10.1016/j.jneuroling.2009.05.002

Elvevåg, B., Foltz, P. W., Weinberger, D. R., & Goldberg, T. E. (2007). Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophrenia Research*. https://doi.org/10.1016/j.schres.2007.03.001

Fang, A., Macdonald, C., Ounis, I., & Habel, P. (2016). Using word embedding to evaluate the coherence of topics from twitter data. *SIGIR 2016 - Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1057–1060. https://doi.org/10.1145/2911451.2914729;WGROUP:STRING:ACM

Figueroa-Barra, A., Del Aguila, D., Cerda, M., Gaspar, P. A., Terissi, L. D., Durán, M., & Valderrama, C. (2022). Automatic language analysis identifies and predicts schizophrenia in first-episode of psychosis. *Schizophrenia 2022 8:1*, *8*(1), 53-. https://doi.org/10.1038/s41537-022-00259-3

Fisher, R. A. (1992). *Statistical Methods for Research Workers*. https://doi.org/10.1007/978-1-4612-4380-9_6

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*. https://doi.org/10.1080/01638539809545029

Fradkin, I., Nour, M. M., & Dolan, R. J. (2023). Theory-Driven Analysis of Natural Language Processing Measures of Thought Disorder Using Generative Language Modeling. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *8*(10), 1013–1023. https://doi.org/10.1016/J.BPSC.2023.05.005

Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple Contrastive Learning of Sentence Embeddings. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*. https://doi.org/10.18653/v1/2021.emnlp-main.552

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/1143844.1143891

Green, M. F., Horan, W. P., & Lee, J. (2015). Social cognition in schizophrenia. *Nature Reviews Neuroscience*, *16*(10), 620–631. https://doi.org/10.1038/NRN4005;TECHMETA

Handbook of Latent Semantic Analysis. (2007). In *Handbook of Latent Semantic Analysis*. https://doi.org/10.4324/9780203936399

Heinrichs, R. W., & Zakzanis, K. K. (1998). Neurocognitive deficit in schizophrenia: A quantitative review of the evidence. *Neuropsychology*, *12*(3), 426–445. https://doi.org/10.1037/0894-4105.12.3.426

Hodosh, M., Young, P., & Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*. https://doi.org/10.1613/jair.3994

Hoffman, R. E. (1986). Verbal hallucinations and language production processes in schizophrenia. *Behavioral and Brain Sciences*. https://doi.org/10.1017/S0140525X00046781

Holmlund, T. B., Chandler, C., Foltz, P. W., Cohen, A. S., Cheng, J., Bernstein, J. C., Rosenfeld, E. P., & Elvevåg, B. (2020). Applying speech technologies to assess verbal memory in patients with serious mental illness. *Npj Digital Medicine*. https://doi.org/10.1038/s41746-020-0241-7

Holshausen, K., Harvey, P. D., Elvevåg, B., Foltz, P. W., & Bowie, C. R. (2014). Latent semantic variables are associated with formal thought disorder and adaptive behavior in older inpatients with schizophrenia. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *55*(1), 88–96. https://doi.org/10.1016/J.CORTEX.2013.02.006

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python*. Zenodo. https://doi.org/10.5281/zenodo.1212303

Iter, D., Yoon, J. H., & Jurafsky, D. (2018). Automatic Detection of Incoherent Speech for Diagnosing Schizophrenia. *Proceedings of the 5th Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, CLPsych 2018 at the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies…*, 136–146. https://doi.org/10.18653/V1/W18-0615

Jones, K. S. (1972). A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL. *Journal of Documentation*, *28*(1), 11–21. https://doi.org/10.1108/EB026526

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*. https://doi.org/10.18653/v1/e17-2068

Just, S. A., Bröcker, A. L., Ryazanskaya, G., Nenchev, I., Schneider, M., Bermpohl, F., Heinz, A., & Montag, C. (2023). Validation of natural language processing methods capturing semantic incoherence in the speech of patients with non-affective psychosis. *Frontiers in Psychiatry*, *14*, 1208856. https://doi.org/10.3389/FPSYT.2023.1208856/BIBTEX

Just, S., Haegert, E., Kořánová, N., Bröcker, A.-L., Nenchev, I., Funcke, J., Montag, C., & Stede, M. (2019). *Coherence models in schizophrenia*. https://doi.org/10.18653/v1/w19-3015

Kay, S. R., Fiszbein, A., & Opler, L. A. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin*, *13*(2). https://doi.org/10.1093/schbul/13.2.261

Kerns, J. G. (2007). Verbal communication impairments and cognitive control components in people with schizophrenia. *Journal of Abnormal Psychology*, *116*(2), 279.

Kim, S. H., Jung, H. Y., Hwang, S. S., Chang, J. S., Kim, Y., Ahn, Y. M., & Kim, Y. S. (2010). The usefulness of a self-report questionnaire measuring auditory verbal hallucinations. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*. https://doi.org/10.1016/j.pnpbp.2010.05.005

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. (2014). A rating scale for the assessment of objective and subjective formal thought and language disorder (TALD). *Schizophrenia Research*. https://doi.org/10.1016/j.schres.2014.10.024

Köhn, A., Stegen, F., & Baumann, T. (2016). Mining the spoken Wikipedia for speech data and beyond. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*.

Kuperberg, G. R. (2010a). Language in schizophrenia Part 1: an Introduction. *Language and Linguistics Compass*, *4*(8), 576. https://doi.org/10.1111/J.1749-818X.2010.00216.X

Kuperberg, G. R. (2010b). Language in schizophrenia Part 2: What can psycholinguistics bring to the study of schizophrenia…and vice versa? *Language and Linguistics Compass*, *4*(8), 590. https://doi.org/10.1111/J.1749-818X.2010.00217.X

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 211–225. https://doi.org/10.1162/TACL_A_00134

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural

Language Generation, Translation, and Comprehension. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

Li, C., Xu, W., Pakhomov, S., Bradley, E., Ben-Zeev, D., & Cohen, T. (2025). *Bigger But Not Better: Small Neural Language Models Outperform Large Language Models in Detection of Thought Disorder*. https://arxiv.org/pdf/2503.20103

Liddle, P. F., Ngan, E. T. C., Caissie, S. L., Anderson, C. M., Bates, A. T., Quested, D. J., White, R., & Weg, R. (2002). Thought and Language Index: an instrument for assessing thought and language in schizophrenia. *The British Journal of Psychiatry*, *181*(4), 326–330. https://doi.org/10.1192/BJP.181.4.326

Loria, S., Keen, P., Honnibal, M., Yankovsky, R., Karesh, D., Dempsey, E., & others. (2014). Textblob: simplified text processing. *Secondary TextBlob: Simplified Text Processing*, *3*, 2014.

Ludwig, A. M., & Othmer, E. (1977). The medical basis of psychiatry. *American Journal of Psychiatry*, *134*(10). https://doi.org/10.1176/ajp.134.10.1087

Lund, A., Burgess, K., Atchley, C., Ann, R., Lund, K., Burgess, C., & Ann Atchley, R. (1995). UC Merced Proceedings of the Annual Meeting of the Cognitive Science Society Title Semantic and Associative Priming in High-Dimensional Semantic Space Publication Date Semantic and Associative Priming in High-Dimensional Semantic Space. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *17*(0), 17.

MacWhinney, B., & Wagner, J. (2010). Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository. *Gesprachsforschung : Online-Zeitschrift Zur Verbalen Interaktion*.

McGorry, P. D., Killackey, E., & Yung, A. (2008). Early intervention in psychosis: concepts, evidence and future directions. *World Psychiatry*, *7*(3), 148. https://doi.org/10.1002/J.2051-5545.2008.TB00182.X

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*(1), 1–43. https://doi.org/10.1207/S1532690XCI1401_1;WGROUP:STRING:PUBLICATION

Miani, A., Hills, T., & Bangerter, A. (2022). Interconnectedness and (in)coherence as a signature of conspiracy worldviews. *Science Advances*, *8*(43), 3668. https://doi.org/10.1126/SCIADV.ABQ3668/SUPPL_FILE/SCIADV.ABQ3668_SM.PDF

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*. https://arxiv.org/pdf/1310.4546

Mikolov, T., Yih, W., & Zweig, G. (2013). *Linguistic Regularities in Continuous Space Word Representations* (pp. 746–751). Association for Computational Linguistics. https://aclanthology.org/N13-1090/

Mohiuddin, T., Joty, S., & Nguyen, D. T. (2018). Coherence modeling of asynchronous conversations: A neural entity grid approach. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*. https://doi.org/10.18653/v1/p18-1052

Mota, N. B., Vasconcelos, N. A. P., Lemos, N., Pieretti, A. C., Kinouchi, O., Cecchi, G. A., Copelli, M., & Ribeiro, S. (2012). Speech graphs provide a quantitative measure of thought disorder in psychosis. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0034928

Ney, H., Essen, U., & Kneser, R. (1994). On structuring probabilistic dependences in stochastic language modelling. *Computer Speech and Language*. https://doi.org/10.1006/csla.1994.1001

Nguyen, D. T., & Joty, S. (2017). A neural local coherence model. *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*. https://doi.org/10.18653/v1/P17-1121

Nordentoft, M., Jeppesen, P., Petersen, L., Bertelsen, M., & Thorup, A. (2009). The rationale for early intervention in schizophrenia and related disorders. *Early Intervention in Psychiatry*, *3 Suppl 1*(SUPPL. 1). https://doi.org/10.1111/J.1751-7893.2009.00123.X

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. https://doi.org/10.1109/ICASSP.2015.7178964

Papineni, K. (2001). *Why Inverse Document Frequency?* https://aclanthology.org/N01-1004/

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 1532–1543. https://doi.org/10.3115/V1/D14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*. https://doi.org/10.18653/v1/n18-1202

Pompili, M., Amador, X. F., Girardi, P., Harkavy-Friedman, J., Harrow, M., Kaplan, K., Krausz, M., Lester, D., Meltzer, H. Y., Modestin, J., Montross, L. P., Bo Mortensen, P., Munk-Jørgensen, P., Nielsen, J., Nordentoft, M., Saarinen, P. I., Zisook, S., Wilson, S. T., & Tatarelli, R. (2007). Suicide risk in schizophrenia: learning from the past to change the future. *Annals of General Psychiatry*, *6*, 10. https://doi.org/10.1186/1744-859X-6-10

Reimers, N., & Gurevych, I. (2020). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. https://doi.org/10.18653/v1/d19-1410

Rousseau, A., Deléglise, P., & Estève, Y. (2014). Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks. *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*.

Saha, S., Chant, D., Welham, J., & McGrath, J. (2005). A systematic review of the prevalence of schizophrenia. In *PLoS Medicine* (Vol. 2, Issue 5). https://doi.org/10.1371/journal.pmed.0020141

Sap, M., Horvitz, E., Choi, Y., Smith, N. A., & Pennebaker, J. W. (2020). Recollection versus Imagination: Exploring Human Memory and Cognition via Neural Language Models. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1970–1978. https://doi.org/10.18653/V1/2020.ACL-MAIN.178

Sap, M., Jafarpour, A., Choi, Y., Smith, N. A., Pennebaker, J. W., & Horvitz, E. (2022). Quantifying the narrative flow of imagined versus autobiographical stories. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(45). https://doi.org/10.1073/pnas.2211715119

Sharpe, V., Mackinley, M., Nour Eddine, S., Wang, L., Palaniyappan, L., & Kuperberg, G. (2024). *GPT-3 reveals selective insensitivity to global vs. local linguistic context in speech produced by treatment-naive patients with positive thought disorder*. https://doi.org/10.1101/2024.07.08.602512

Sommer, I. E., Derwort, A. M. C., Daalman, K., de Weijer, A. D., Liddle, P. F., & Boks, M. P. M. (2010). Formal thought disorder in non-clinical individuals with auditory verbal hallucinations. *Schizophrenia Research*. https://doi.org/10.1016/j.schres.2010.01.024

Stolcke, A. (2002). SRILM - An extensible language modeling toolkit. *7th International Conference on Spoken Language Processing, ICSLP 2002*.

Tan, E. J., Sommer, I. E. C., & Palaniyappan, L. (2023). Language and Psychosis: Tightening the Association. *Schizophrenia Bulletin*, *49*(Suppl_2), S83–S85. https://doi.org/10.1093/SCHBUL/SBAC211

Tang, S. X., Kriz, R., Cho, S., Park, S. J., Harowitz, J., Gur, R. E., Bhati, M. T., Wolf, D. H., Sedoc, J., & Liberman, M. Y. (2021). Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. *Npj Schizophrenia*. https://doi.org/10.1038/s41537-021-00154-3

Team, L., & Meta, A. @. (2024). *The Llama 3 Herd of Models*.

Tilk, O., & Alumäe, T. (2016). Bidirectional recurrent neural network with attention mechanism for punctuation restoration. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. https://doi.org/10.21437/Interspeech.2016-1517

Torous, J., Friedman, R., & Keshavan, M. (2014). Smartphone Ownership and Interest in Mobile Applications to Monitor Symptoms of Mental Health Conditions. *JMIR MHealth and UHealth*. https://doi.org/10.2196/mhealth.2994

Turney, P. D., & Pantel, P. (2010a). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*. https://doi.org/10.1613/jair.2934

Turney, P. D., & Pantel, P. (2010b). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*. https://doi.org/10.1613/jair.2934

Van Lieshout, R. J., & Goldberg, J. O. (2007). Quantifying self-reports of auditory verbal hallucinations in persons with psychosis. *Canadian Journal of Behavioural Science*. https://doi.org/10.1037/cjbs2007006

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention Is All You Need*. 1. https://arxiv.org/pdf/1706.03762

Veaux, C., Yamagishi, J., & MacDonald, K. (2016). CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit. *The Centre for Speech Technology Research (CSTR)*.

Voleti, R., Liss, J. M., & Berisha, V. (2019). Investigating the Effects of Word Substitution Errors on Sentence Embeddings. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. https://doi.org/10.1109/ICASSP.2019.8683367

Xu, W., Pakhomov, S., Heagerty, P., Horvitz, E., Bradley, E. R., Woolley, J., Campbell, A., Cohen, A., Ben-Zeev, D., & Cohen, T. (2025). Perplexity and proximity: Large language

model perplexity complements semantic distance metrics for the detection of incoherent speech. *Journal of Biomedical Informatics*, *170*, 104899. https://doi.org/10.1016/J.JBI.2025.104899

Xu, W., Portanova, J., Chander, A., Ben-Zeev, D., & Cohen, T. (2020a). The Centroid Cannot Hold: Comparing Sequential and Global Estimates of Coherence as Indicators of Formal Thought Disorder. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*.

Xu, W., Portanova, J., Chander, A., Ben-Zeev, D., & Cohen, T. (2020b). *The Centroid Cannot Hold: Comparing Sequential and Global Estimates of Coherence as Indicators of Formal Thought Disorder*. PsyArXiv. https://doi.org/10.31234/osf.io/sfkqc

Xu, W., Wang, W., Portanova, J., Chander, A., Campbell, A., Pakhomov, S., Ben-Zeev, D., & Cohen, T. (2022). Fully automated detection of formal thought disorder with Time-series Augmented Representations for Detection of Incoherent Speech (TARDIS). *Journal of Biomedical Informatics*, *126*. https://doi.org/10.1016/j.jbi.2022.103998

Zipitria, I., Arruarte, A., & Elorriaga, J. A. (2006). Observing Lemmatization Effect in LSA Coherence and Comprehension Grading of Learner Summaries. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4053 LNCS*, 595–603. https://doi.org/10.1007/11774303_59